Check for updates

Exploring chemical compound space with quantum-based machine learning

O. Anatole von Lilienfeld, Klaus-Robert Müller and Alexandre Tkatchenko

Abstract | Rational design of compounds with specific properties requires understanding and fast evaluation of molecular properties throughout chemical compound space — the huge set of all potentially stable molecules. Recent advances in combining quantum-mechanical calculations with machine learning provide powerful tools for exploring wide swathes of chemical compound space. We present our perspective on this exciting and quickly developing field by discussing key advances in the development and applications of quantum-mechanics-based machine-learning methods to diverse compounds and properties, and outlining the challenges ahead. We argue that significant progress in the exploration and understanding of chemical compound space can be made through a systematic combination of rigorous physical theories, comprehensive synthetic data sets of microscopic and macroscopic properties, and modern machine-learning methods that account for physical and chemical knowledge.

Due to an unfathomably large number of possible molecules and materials[1,2], and the combinatorially many ways for them to undergo chemical transformations, our understanding of chemistry requires a first-principles approach with proper roots in quantum mechanics (QM) and statistical mechanics (SM). QM gives us the ability to calculate accurate microscopic properties (energies, atomic forces, electronic-energy levels, electrostatic multipoles and polarizabilities) for fixed molecular geometries, and SM allows us to sample QM energy surfaces in a given statistical ensemble and calculate macroscopic properties. Accurate QM and SM simulations are computationally demanding, even for a single molecule or simple material; hence, more efficient computational approaches are urgently needed to address the molecular processes over multiple length scales and timescales with sufficient accuracy in order to obtain insights into the evolution of properties throughout chemical compound space (CCS). Such efficient methods may eventually enable the long-held dream of in silico chemical and materials design.

CCS is large but finite and accounts for the set of all feasible metastable atomic configurations resulting from solving Schrödinger's equation. Non-equilibrium molecular configurations provide smooth interpolations between points in this high-dimensional chemical space. While accounting for all possible conformations of all stereoisomers of all possible constitutions of all possible compositions, CCS also encodes repeating patterns, abundant signatures and low-dimensional building blocks. Nowadays, most quantum machine learning (QML) approaches decompose predicted properties into atomic contributions. For example, computing quantum properties, such as the atomization energy, as a series expansion in select fragments called 'AM-ons' has, by now, enabled learning properties for chemically diverse molecules. AM-on stands for 'atom in a molecule' and the suffix 'on' indicates that each AM-on can be considered as a building-block dictionary entry that is selected whenever relevant in the ensemble of fragments that constitute any larger query molecule[3]. The pervasiveness of the underpinning constituting patterns when charting CCS suggests that there is an analogy to constellations of stars in the universe. Constellations, like molecules, have names and, more importantly, have been useful for orientation and navigation. Similarly, property patterns throughout chemical space can be combined in 'constellations', from which properties of new molecules of interest can be calculated using linear or non-linear combination of properties of known molecules or molecular fragments. Although relationships for stars and planets are rather well understood, a rigorous understanding of CCS in terms of molecular components has not yet been achieved but would be of utmost usefulness for rational compound design. In this Perspective, we argue that the recently developed machine learning (ML) approaches will significantly aid in achieving a deeper understanding of CCS. We give several examples that illustrate the substantial progress achieved in this field and outline the many remaining challenges yet to be addressed.

The targeted exploration of CCS aiming to obtain compounds with desired properties is a longstanding endeavour. Many efforts in cheminformatics or materials informatics have relied on statistics and ML to search CCS for relevant pharmaceutical properties, such as receptor binding, toxicity[4,5] or materials stability[6–8]. Despite often being useful and computationally efficient, the main drawback of these approaches is that they are not transferable to molecules and properties outside of their domain of applicability, which results from a lack of underlying principles of physics, as also pointed out in REF.[9].

QM describes the electronic structure of any material compound, thereby, determining the behaviour of matter at large and dictating all the mutual relationships between observable microscopic properties[10]. In light of such a large domain of applicability of QM, it is not a surprise that fundamental contributions to density functional theory (DFT) — one of the most efficient formulations of QM — are highly cited[11]. The unbiased study of chemical space for the purpose of exploration as well as exploitation (computational compound design) strictly requires sampling algorithms with maximal efficiency.

Although QM-based design of materials has already been successfully applied to some specific materials-design challenges[12–14], it imposes a prohibitive computational cost. Consequently, the improvement in efficiency and robustness of electronic-structure calculations play an increasingly important role in current and future materials-design efforts[15–18].

ML has already enabled many applications in many fields[19,20], including medical diagnostics[21–23], particle physics[24], bioinformatics[25], brain–computer interfaces[26], social-media analysis[27], robotics[28] and team, social or board games[29–31]. Here, we focus on recent fundamental ML developments aimed at a quantum-based understanding of CCS (see BOX 1 for an overview of the key concepts in the field of QML). The key idea is that any observable property for any system can be obtained from solving the relevant quantum-mechanical equations. A more comprehensive QM-based understanding of CCS is now foreseeable because of maturity, efficiency and reproducibility of electronic-structure methods, such as DFT and post-Hartree–Fock wave-function methods, and codes[32]. Furthermore, fast-paced developments in high-performance computing hardware have also helped to improve our QM-based understanding of CCS. The scientific codes in the electronic-structure community have matured to such an extent that a considerable fraction of the world's top high-performance computing centres busy themselves with QM calculations. Finally, conceptual adaptations of statistical mechanics and continuous advances in statistical learning have, nowadays, enabled the performance of intelligent data analysis on both small and large data sets, and extraction of valuable quantitative insights in a systematic manner. Many recent publications, including special journal issues and reviews of quantum-based ML approaches[33–36], have highlighted the fact that combining quantum calculations with ML can lead to considerable leaps in exploring and understanding chemical and materials spaces. In addition to learning quantum-mechanical observables (by integrating over electronic degrees of freedom), evidence has been presented that it is equally possible to build ML models of SM ensemble averages (by integrating over relevant atomistic configurational degrees of freedom), such as free energy, entropy or kinetic pathways[37,38].

To distinguish the emerging field of physics-based ML from preceding efforts in cheminformatics, bioinformatics and materials informatics, we refer to the combinations of QM and SM approaches with ML as QML models. QML refers to the idea of applying modern statistical learning theory to predict electronic and atomistic properties and processes in molecules and materials. We also remark that the goals and reaches of QML models should not be confused with quantum ML

Box 1 | **Explanation of quantum machine learning terms**

Here, we give a compact explanation for various keywords discussed in this Perspective.

- Machine learning (ML): methods based on statistical learning theory for obtaining numerical models from data samples that generalize well on unseen data. Generally, ML models improve with the availability of more data; hence, the models are said to 'learn from data'. ML models are inductive, meaning that they are typically not based on any underlying physical model, but can, in principle, reconstruct physical models from the provided data. In the context of exploring chemical compound space, available data are less abundant than in typical ML applications such as computer vision. It becomes, therefore, important to most efficiently make use of available data by combining prior knowledge about physics and chemistry with powerful ML models, as we will argue throughout this Perspective.
- Representation: the model that encodes the structure of and relations between atoms. It is crucial for quantifying geometric and chemical similarities of molecules. The representation needs to be unique and invariant to atom indexing, as well as to molecular translations and rotations in space (BOX 2).
- Supervised, unsupervised and semi-supervised learning: ML with labels is called supervised learning. Examples of supervised learning are classification or regression, in which the class label or regression value for every sample is given in the training data. On the contrary, in unsupervised learning, label information is not included in the training data set. Clustering and dimensionality reduction are typical unsupervised learning problems. Semi-supervised learning assumes that most samples have no label and only for very few samples are labels provided for training.
- Parametric and non-parametric models: parametric models assume a finite set of model parameters that need to be estimated (for example, by using a mean of a Gaussian), whereas non-parametric models do not rely on this assumption. Popular non-parametric models, such as Gaussian processes, can be viewed as having infinitely many parameters.
- Regression: in regression, the relationship between an input representation and continuous output variables is estimated. The most common simple-regression analysis is linear regression, in which a linear function is fitted to the data according to some loss function, such as mean squared error. A widely used classical model for non-linear regression is kernel–ridge regression that generalizes well to unseen data with limited scalability in larger data sets.
- Deep neural networks: widely used and flexible non-linear regression models based on neural networks. Deep neural networks refer to structured architectures that have a large number of hidden layers, offering large flexibility and rich, multiscale representations. Due to their scalability, they are ideally suited to extract complex, non-linear relations from large data sets.
- Cross validation: a common ML procedure used for ensuring generalization to unseen data and avoiding overfitting.
- Learning curves: measure the performance of ML models upon increasing the number of data samples used for training the model.
- Density functional theory (DFT): the workhorse method for electronic-structure calculations on molecules and materials. While DFT is, in principle, an exact theory, in practice, approximations are made for electronic quantum (exchange and correlation) effects. DFT implementations often provide a good compromise between accuracy and efficiency. Most current quantum machine learning data sets for molecules and materials are based on DFT calculations.
- Hartree–Fock (HF): fundamental electronic-structure method used as a starting point for essentially all practical calculations of quantum correlation energy in molecular systems. HF provides an exact treatment of electronic-exchange effects due to Pauli repulsion.
- Coupled cluster (CC) methods: a set of methods to obtain and systematically improve the calculated estimate of electronic-correlation energy based on the HF wave function. This is achieved by increasing the level of modelled electronic excitations: single excitations, double excitations and so on. However, higher-level treatment requires orders of magnitude more computational resources. In particular, the coupled cluster single double (triple) (CCSD(T)) method is the so-called 'gold standard' of computational quantum chemistry, including single, double and perturbative triple excitations (to the fourth order in perturbation theory). Within converged basis sets, CCSD(T) typically yields so-called 'chemical accuracy' of ~1 kcal mol^{-1} in atomization energies of molecules with single reference character.

algorithms executed on quantum computers. As such, QML models aim to provide a feedback mechanism between QM and/or SM, and (statistical) ML. Given sufficient reference data obtained from QM and SM simulations, queries of properly trained ML models can yield accurate properties within milliseconds[39] — as opposed to the many CPU hours or days necessary to solve the corresponding quantum and statistical mechanics problems for representative compounds. Because of the rigorous interpolation of QML in complex non-linear spaces and their consequently controlled predictive accuracy[40], the door has now opened for an extensive analysis and study of these interpolated spaces, which was previously impossible due to the prohibitive computational cost of direct QM and SM simulations.

Given the substantial progress in QML discussed in this Perspective, we argue that meaningful progress in the exploration and understanding of CCS can be made through systematic combination of rigorous physical theories, comprehensive data sets of QM and SM properties, and sophisticated ML methods that incorporate physical and chemical knowledge. The authors have witnessed the quick development of QML from the perspective of electronic-structure calculations and, hence, the focus in this Perspective is on combining QM and ML with the goal of enhanced exploration of CCS. Efforts to use ML to capture SM properties in analogous ways is the subject of active current research[41,42].

Goals and advances of QML

The overarching goal of QML is to develop reliable models with the accuracy of high-level electronic-structure calculations. Depending on the application, the reference data can be obtained from high-level quantum chemistry, such as coupled cluster single double (triple) (CCSD(T)), or from DFT calculations. Although much work remains to be done to reach the 'dream' of exact QML models, many key advances have been recently achieved that we discuss in this section and connect to important remaining challenges for which we deem that urgent progress is needed.

All QML advances hinge on the availability of trustworthy QM data. These data need to cover a certain important domain, for example, the CCS of organic drug-like compounds, as explored by Reymond and colleagues through their generated database (GDB) list of simplified molecular-input line-entry system (SMILES) strings[43–46]. QM calculations on these molecular graphs led to the publication of data sets that collect equilibrium structures and properties of many thousands of small molecules (QM7 and QM9)[47,48], their molecular-dynamics trajectories (MD17)[49] and non-equilibrium molecular structures (ANI-1)[50]. One can also calculate equilibrium structures and properties of solids[51–53], or generate equilibrium and non-equilibrium molecular dynamics (MD) data for a single element (for example, silicon)[54]. The ultimate goal of QML is to develop a universal and efficient model for the whole CCS that enables the accurate description of molecules and materials on equal footing and possibly leads to new insights on CCS underlying regularity and chemical relationships. Reorganizing the periodic table (in the sense of revisiting and generalizing Pettifor's concept of Mendeleev number)[55,56] represents a first and important step in this direction[53,57]. Initially, various models have been developed focusing either on molecular or materials data, but versatile models have been more recently proposed that can be applied to both molecules and solids[58–60].

CCS is commonly explored using cheminformatics-based approaches. In contrast, QML rigorously adheres to its roots in fundamental physics, such that it is consistent with the laws of QM and SM. One of the first QML applications in which ML techniques were used for non-linear interpolation of QM data aimed to construct reliable system-specific interatomic potentials or potential-energy surfaces, going beyond conventional force fields in terms of universality (atom-type specificity no longer required) and accuracy[61–66]. Further developments aimed at transferable QML models that are trained and applicable throughout CCS for the description of QM properties, as shown for the QM7 set of organic molecules[39], highlighting the potential of QML for efficient and accurate exploration of CCS. This idea was rapidly demonstrated to be applicable to many electronic properties using neural networks as well as kernel–ridge regression[47,67,68], or to search for polymers with useful properties[69], explore chemical properties of crystalline solids[53,70–73] and design materials for a variety of technological applications[74,75].

A crucial aspect that determines the reliability and applicability of any QML model is its generalization accuracy that is assessed on the calculated QM properties of a sufficiently large out-of-sample (hold-out) test data set. It is remarkable how quickly the generalization accuracy and data efficiency of QML models has improved during the past few years. As shown in FIG. 1 on the example of the QM9 data set, the QML prediction errors have decreased by 40-fold — from 8 kcal mol^{-1} (0.340 eV) to 0.2 kcal mol^{-1} (0.008 eV) in 2018 (REF.[58]), using exactly the same training. This noticeable increase in accuracy mainly stems from incorporation of physical prior knowledge into the QML models, such as proper description of permutational symmetries of atoms in a molecule[49,58,60,76], as well as explicit inclusion of physically

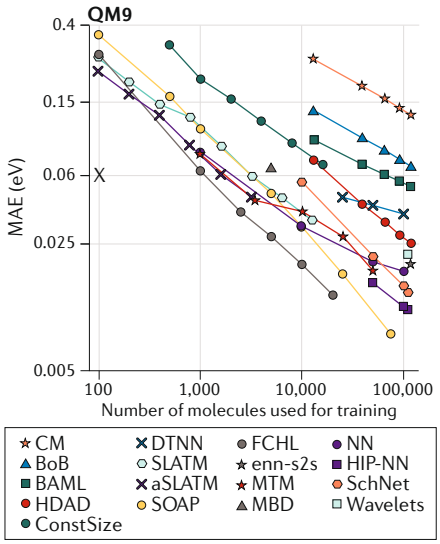

Fig. 1 | **Learning curves illustrate the progress of QML models of atomization energies of molecules over the past few years.** This plot shows the mean absolute error (MAE) in eV on atomization energies of small molecules in the quantum-mechanics-based data set for organic molecules with up to nine non-hydrogen atoms (QM9)[48]. The compared quantum machine learning (QML) models differ solely by representation and model architecture, and correspond to Coulomb matrix (CM)[39], bag of bonds (BoB)[83], bonds, angles, machine learning (BAML)[77], histogram of distances, angles, dihedrals (HDAD)[137], constant-size descriptors (ConstSize)[103], deep tensor neural network (DTNN)[86], spectrum of London and Axilrod–Teller–Muto (SLATM)[3], atomic SLATM (aSLATM), smooth overlap of atomic positions (SOAP)[60], Faber, Christensen, Huang, Lilienfeld (FCHL)[58], message passing node and edge-based neural network with set-to-set readout function (enn-s2s)[162], moment tensor model (MTM)[135], many-body-based (MBD) kernel–ridge regression[78], reactive neural network (NN)[87], Hierarchically Interacting Particle Neural Network (HIP-NN)[163], SchNet[59] and wavelets[164]. The black X on the left indicates the target value in the 'QM9 challenge', in which QML models should be developed to reach 1 kcal mol^{-1} (0.043 eV) accuracy on the QM9 data set using only information of 100 molecules for training. To date, this challenge has not been met. Adapted from REF.[165], Springer Nature Limited.

motivated pairwise and many-body terms for interatomic interactions into QML descriptors[77,78]. A discussion of several widely used representations of molecules in the context of kernel-based ML is shown in BOX 2. A noticeable conclusion from FIG. 1 is that the most advanced existing QML models are extremely data-efficient, achieving chemical accuracy of 1 kcal mol^{-1} for the QM9 data set of 134,000 organic molecules[48] using only 1,000 molecules (0.7%) for training. This result hints on the potential sparsity of CCS, implying a low complexity and dimensionality[79] of the property-prediction problem throughout CCS. In other words, this evidence suggests that (unknown) properties of query molecules can be predicted as non-linear combinations of (known) properties of only a few other molecules, which are not necessarily chemically similar. To search for better QML models, the authors, together with other researchers, agreed to award US$1,500 as of May 2020 to the scientist(s) who devise(s) a QML model that meets the Institute for Pure and Applied Mathematics (IPAM) QM9 challenge; that is, a QML model that reaches predictive accuracy of ~1 kcal mol^{-1} after training on only 100 QM9 molecules.

A critical component of every QML model lies in the representation (sometimes also referred to as descriptor) of an atomic system composed of nuclei and electrons. Uniqueness of the representation is a necessary condition for QML models to converge down to arbitrary accuracy[80], and an increase in similarity between representation and target function typically lowers the offset in learning curves[77]. In order to afford efficient learning, representations should also capture well-known translational, rotational and permutational invariances of atomistic systems, but can also be significantly improved by explicitly including other physical priors, such as temporal and spatial symmetries of interatomic interactions[49] or differential relationships of response properties[81]. An additional requirement is that atomistic representations should be as computationally efficient as possible in order to benefit from QML's computational efficiency over the numerical solvers used in conventional QM calculations. Although many and various representations have been proposed[39,58,70,78,82–85], there is still an ongoing debate on the advantages and limitations of different representations for different application domains. While kernel-based ML models require explicit formulation for the representation[39,67,82], using one scale per kernel, deep neural networks, such as deep tensor neural network (DTNN)-based approaches[59,86,87], can result in an implicit multiscale representation from a scalable learning process.

In QM calculations, different properties of an atomistic system (such as electronic energy, atomic forces, multipole moments, polarizability and electronic-energy levels) can be evaluated as QM operators acting on the electronic wave function. The situation is more intricate in the case of QML models. Initially, separate QML models were used to describe different properties, for example, one model for total energy and a different one for the polarizability. Therefore, one of the first goals of neural networks was to develop transferable neural networks that can simultaneously predict multiple electronic properties as the one applied to the QM7 data set[47]. Kernel methods and deep neural networks have also been extended to reflect multi-property prediction[68,88]. However, beyond that, it would be very advantageous to view QML models as a coarse-grained surrogate for electronic interactions in an atomistic system, akin to a downfolded version of the wave function. In such a model, electronic properties in QML could be calculated similarly to explicit QM calculations, namely, as operators on manifolds. This idea of using response operators for ML in chemical space has been recently introduced by Christensen and colleagues[81] and various QML models of electron densities

Box 2 | **Summary of various molecular representations**

A key aspect of quantum machine learning approaches is the definition of a molecular representation that enables one to measure similarity between molecules. There is no 'universal' representation that satisfies all desirable properties at once, being general, accurate, efficient and transferable at the same time. This has led to a flurry of developments of different molecular representations, each of which satisfies only part of the general requirements. Although all the existing models are fundamentally based on the same information — atomic positions and nuclear charges — they provide a different mapping between discrete atomic information and continuous spatial degrees of freedom. In principle, both kernel methods and neural networks start with a molecular representation, although neural networks can also generate the representation as part of the learning process[59,86]. Here, we provide a brief summary of the different existing models for representing molecules.

- Atom-centred symmetry functions[166]: products of radial and angular symmetry functions are used to represent molecular environments. The advantage is a compact and physically inspired representation. The limitation is that the complexity grows quickly with the increase in the number of atom types.
- Coulomb matrix[39]: inverse distance matrix that represents internuclear Coulomb repulsion between atoms. A very efficient, global and elegant representation. It does not satisfy permutation symmetry for equivalent atoms and can lead to discontinuities.
- Bag of bonds[83]: in this case, the Coulomb matrix representation is vectorized, improving its similarity measure and efficiency. Permutation symmetry and discontinuities still pose a problem.
- Many-body-based representation[78]: localized extension of bag of bonds with explicit treatment of different distance scales, including three-body and higher-body terms. It solves the permutation symmetry. However, it can become inefficient when using higher than three-body terms.
- Smooth overlap of atomic positions[60]: in this case, molecules are represented as a superposition of radial Gaussian functions and angular momentum terms. Rigorously treats rotation and permutation symmetries and avoids discontinuities. Can become very expensive for accurate predictions and for several atom types.
- Sine matrix[84]: generalization of Coulomb matrix to periodic systems.
- Many-body tensor representation[85]: a tensor representation for molecules and solids. Generally applicable to many different systems. Can become expensive to evaluate for large systems.
- Partial radial distribution functions[70]: a representation based on atomic radial distribution functions for solids. Only tested for predicting electronic properties of simple solids.
- Wavelet scattering transform[164]: a multiscale representation for molecular properties based on wavelets. Naturally captures multiscale nature of molecular properties without imposing localization to those properties. Expertise is needed in constructing an appropriate wavelet basis for new systems.
- Moment tensor potentials[97]: atomic potential representation based on atomic moments. General and efficient and mainly tested on elemental solids. Requires wide testing especially on multicomponent systems.
- Faber, Christensen, Huang, Lilienfeld[58]: encodes both the chemical composition and structural degrees of freedom in the representation for enabling alchemical design. Can be applied to both molecules and solids. Can become expensive when accurate predictions are desired.

and wave functions have, by now, also been proposed[89–95].

In general, the chemical space of molecules and materials can be explored by looking at compositional (that is, elements forming bonds) and configurational (that refers to the same kind of bonds that form different structures) degrees of freedom. Up to now, we have mainly discussed models that explore compositional degrees of freedom in CCS. However, configurational degrees of freedom are crucial to understand dynamics of molecules under given external conditions. As indicated above, the construction of reliable ML force fields from QM data was one of the first examples of successful applications of QML[61,64,65] and, by now, the use of QML models has become routine in order to enable MD calculations that go beyond the realm of classical, system-dependent force fields[49,59,60,82,96,97]. Recently, the focus has been shifting towards emphasizing data efficiency of QML force fields. For example, by sampling only a few hundred molecular conformations from an MD trajectory, the symmetrized gradient-domain machine learning (sGDML) model[76] can give global force fields for small molecules (≤25 atoms) that reach accuracy comparable to those achievable at coupled cluster level of theory — the 'gold standard' in quantum chemistry. In this way, sGDML reaches accuracy of 0.2 kcal mol^{-1} in energies and 1 kcal (mol Å)$^{-1}$ in atomic forces, relative to coupled cluster calculations. This level of accuracy is crucial when modelling conformational transitions and vibrational spectroscopy of even small molecules such as ethanol and aspirin[98]. Hence, QML has already enabled us to obtain essentially exact dynamics for small molecules using computationally efficient MD simulations that correspond to a full QM treatment of both electrons and nuclei. Likewise, MD simulations of materials have largely benefited from the application of QML approaches. For example, QML-enabled simulations have been used to study the growth mechanism of tetrahedral amorphous carbon[99,100], demonstrating the possibility to develop transferable DFT-level force fields for the investigation of complex processes in elemental solids.

In light of the theoretical advances offered by QML, it is reasonable to wonder how QML-based predictions compare to experiments, if they could facilitate the analysis of the experimental measurements and possibly even guide the design of new experiments. It is encouraging to note that QML-based MD simulations can already reach the accuracy and efficiency necessary to predict experimental outcomes. For example, path-integral MD using the sGDML QML force field fitted with CCSD(T) atomic forces has been used to demonstrate that low-frequency excitations of ethanol arise from highly anharmonic combination of vibrational normal modes[101], thus, resolving a long-standing experimental controversy (see REF.[101] for a detailed discussion) (FIG. 2).

Scalability of QML models has also been demonstrated by the applicability of models to large systems after being trained on small systems. Atom-by-atom-based training exploiting the locality of the chemical environment around each atom by on-the-fly training-set selection of only the most representative small fragments (such as AM-ons) was shown to yield promising results for energies, forces, NMR shifts and other QM properties for diverse systems of varying sizes reaching up to hundreds of atoms (not counting hydrogens)[3]. Albeit without the query-tailored selection of training fragments within the AM-on approach, other extensive atom-by-atom-based kernel–ridge regression and neural-network-based models have also demonstrated scalability (REFS[86,87,102–104]).

Another way to reduce complexity and increase accuracy of QML models consists of combining various levels of theory for training and testing[37,105–107]. For example, the Δ-ML approach corresponds to generating QML models of corrections to a lower level efficient electronic-structure method in order to reach the accuracy of a much higher and more expensive level of theory[37].

Creating universal QML-based force fields trained and applicable across CCS still remains an open challenge due to the diversity of systems and non-local quantum interactions encountered when navigating simultaneously configurational and compositional space[59,81,86,87,102,108–110].

Finally, we note that robust software codes are crucial for enabling widespread usage of ML techniques for chemical applications. Here, some pointers are provided to software packages of the discussed ML methods for quantum chemistry. General code for running QML is found in REF.[111]. Deep learning methods such as DTNN and SchNet can be readily implemented using the SchNetPack software[112]. The alternative accurate neural-network engine for molecular energies (ANI) neural-network package is also available[102]. The sGDML package, in which prior information has been included in the ML method (spatial and temporal conservation laws), is readily and easily usable in REF.[76]. The iNNvestigate toolbox that can explain non-linear learning methods such as deep learning is described in REF.[113].

QML-based insights into CCS

In the previous section, we discussed how QML approaches can aid and extend the reach of QM calculations in several important directions. However, possibly the major appeal of QML is to provide new insights into QM properties of molecules and materials and, ultimately, enable efficient exploration of CCS and rational design of molecules and materials with tailored properties. Most commonly, ML methods are employed in fundamental and applied sciences to categorize and structure data, and to develop predictive models. Only recently, ML models have been used to learn and gain insights about the unknown underlying regularities hidden in data[86]. Given that the ML models achieve this by using rigorous statistical theory[114], they can facilitate the formulation of new insights into chemical properties[86] or actionable hypotheses[115] that can be further validated and tested with high-level theory or experiments. In other words, ML modelling has become a powerful and indispensable part of the scientific discovery process itself.

FIGURE 2 illustrates six examples selected from our work and the literature of how such insights were gained from QML. The development of an accurate and data-efficient sGDML force-field model[49,101] (FIG. 2a) enabled quantum MD simulations with essentially exact atomic forces computed at CCSD(T) level of theory. For the first time, it was possible to compute accurate thermodynamic and spectroscopic properties for molecules as large as aspirin without compromising the accuracy of atomic forces with the timescale accessible in MD simulations. Future work to improve the scalability of frameworks such as sGDML will enable us to perform fully predictive MD simulations in which both electrons and nuclei are treated rigorously with exact QM equations without unnecessary compromises between accuracy and efficiency of molecular simulations.

Beyond energies and forces, a general QML model needs to predict accurate electronic response properties (akin to evaluating properties as expectation values of the QM wave functions). For example, the application of response operator theory to QML models is illustrated in FIG. 2b. Here, the QML model of the binding

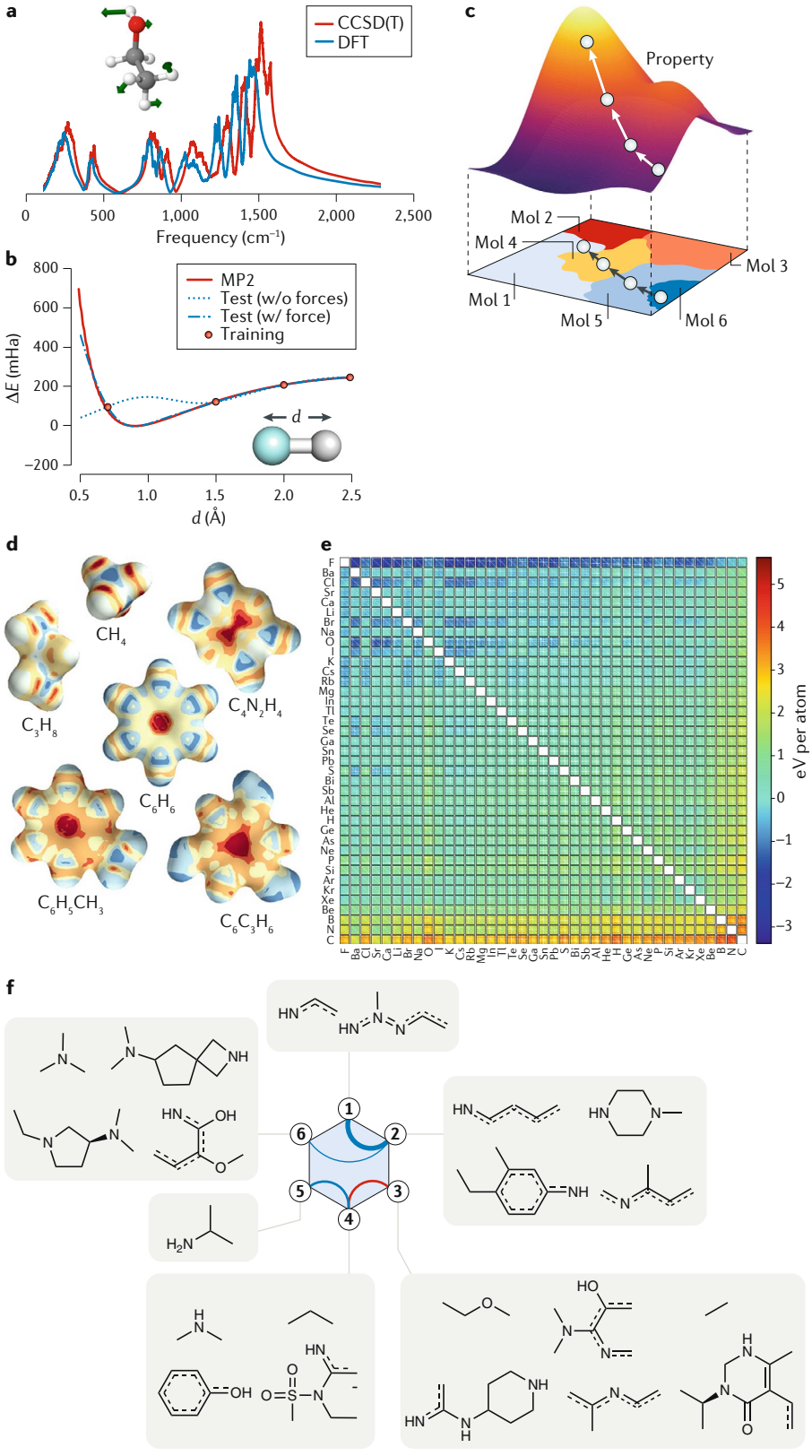

Fig. 2 | **Insights from QML models. a** | Physico-chemical insights can be gained from quantum machine learning (QML) models such as the dynamical modelling of vibrational anharmonicities in ethanol at the quantum-chemical coupled cluster single double (triple) (CCSD(T)) level obtaining physical insights from essentially exact dynamics and explaining experimental low-frequency vibrational modes (data replotted from REF.[101]). **b** | Binding potential of hydrogen fluoride demonstrating that inclusion of forces, in addition to the energies from the MP2 potential, in the training set results in improved kernel–ridge-regression-based QML models[81]. **c** | Smooth property optimization in a continuous latent chemical space of reduced dimensionality[116]. The discrete chemical space (Mol 1–Mol 6) is mapped to a continuous space of reduced dimensionality (white dots), enabling gradient-based optimization of molecular properties. **d** | Spatially resolved chemical potentials inferred for several molecules with the deep tensor neural network model demonstrating the ability of this model to provide quantum-mechanical insights that were not part of its training procedure[86]. **e** | QML-based formation energy estimates of ~2 million elpasolite (ABC_2D_6) crystals, the components of which are main-group elements, have enabled the identification of nearly 90 new crystal candidates[53]. **f** | Top six most frequent chemical motifs occurring in agonist candidate to bind the human muscarinic acetylcholine receptor M1 according to random-matrix theory[126]. Connecting lines indicate cooperativity of these motifs in terms of enhancement (red) or weakening (blue) of binding (when both motifs are simultaneously present). The width of the lines corresponds to the magnitude of the effect. DFT, density functional theory. Part **a** is adapted from REF.[101], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Part **b** is adapted with permission from REF.[81], AIP. Part **c** is adapted with permission from REF.[116], ACS (https://pubs.acs.org/doi/10.1021/acscentsci.7b00572); further permissions related to the material excerpted should be directed to the ACS. Part **d** is adapted from REF.[86], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Part **e** is adapted from REF.[53], CC BY 3.0 (https://creativecommons.org/licenses/by/3.0/). Part **f** is adapted with permission from REF.[126], PNAS.

curve of hydrogen fluoride improves qualitatively through the explicit inclusion of its derivative with respect to interatomic distance. The same formalism upon inclusion of the corresponding electric-field derivatives also improves dramatically the prediction of the electrostatic dipole moment[81]. The ultimate goal of QML should be creating a unified surrogate quantum model that can simultaneously learn many QM properties in a data-efficient and accurate manner.

In other work by Gómez-Bombarelli and co-workers[116], the optimization of new compounds was enabled through the use of generative models in latent space, a compressed variant of CCS with reduced dimensionality (FIG. 2c), resulting in the prediction of promising light-harvesting materials candidates. Conceptually, this way of tackling the 'inverse design problem' (that is, the material is designed in order to exhibit a desired property)[117] through an intermediate step of effectively coarse-graining chemical space is intriguing. This inverse design

of compounds constitutes an alternative approach to iteratively solving the 'forward design problem', whereby compounds with target properties are designed, using gradient-based[118–120], Monte Carlo[12] or genetic algorithms[13,117], or combinations thereof[121]. Furthermore, the problem of generating meaningful compounds is solved elegantly in an implicit fashion by directly training on valid molecular graphs (SMILES strings), automatically ensuring that the 'grammar' of newly generated molecules is not being violated.

Another way to obtain new insights is by analysing what the QML models have learned from the data, in the spirit of explainable artificial intelligence, in which ML models are dissected to analyse their inner mechanisms that lead to their respective predictions[114,122–124]. An example of the application of this concept is provided by the analysis of the molecular representation learned by DTNN[86]. DTNN models and other flexible, non-parametric ML models are trained on QM molecular energies and, in the limit of infinite data, can learn the exact map of the solution of the Schrödinger equation for different molecular structures. Because the exact solution can only be formally achieved in the limit of an exact representation of the wave function, there is mathematically no other choice for DTNN than to attain an exact representation of the wave function. In practice, the representation is trained on a finite number of molecules, hence, DTNN learns the 'Schrödinger mapping' on a finite set of molecules and is, therefore, not necessarily in a one-to-one relation with the wave function. One can query the learned representation by adding a probe atom to a given molecule[86]. By visualizing the energy isosurface of the probe atom, one can immediately see that the obtained representation exhibits features that closely resemble electron densities or electrostatic potentials. This indicates that the model is able to infer QM features in the representation directly from a restricted set of QM energies (FIG. 2d). Hence, the DTNN approach is attempting to solve an inverse design problem[12] by constructing a coarse-grained QM representation from a finite set of molecular energies or other QM properties (FIG. 2d). Despite being trained only on total energies of molecules, the DTNN approach grasps fundamental chemical concepts such as bond saturation and different degrees of aromaticity. For example, the DTNN model predicts the $C_6O_3H_6$ molecule to be 'more aromatic' than benzene or toluene[86]. $C_6O_3H_6$ does have higher ring stability than both benzene and toluene, and DTNN predicts $C_6O_3H_6$ to be the molecule with the most stable aromatic carbon ring among all molecules in the QM9 database[86]. Interestingly, the mathematical construction of the DTNN model and other flexible non-parametric models[3,87], based on atomic contributions, provide statistically rigorous partitioning of extensive molecular properties into atomic contributions — an interesting alternative to QM-based partitioning schemes[125].

We note that, although DTNN in principle can generate a 'Schrödinger map' between molecular Hamiltonians and molecular quantum properties, detailed analysis of the underlying representation learned by DTNN amounts to a complex inverse problem. To address this problem, a generalized SchNOrb architecture has been developed that learns the DFT wave function directly[92]. It would be desirable to unify both ML architectures (DTNN or SchNet and SchNOrb) to combine the direct 'Schrödinger map' (SchNOrb) with the inverse 'Schrödinger map' (DTNN or SchNet) and gain further understanding into the QM of molecules.

QML is evidently also applicable to solids. For example, it has been used to calculate the formation energies of ~2 million elpasolite crystals (of ABC_2D_6 sum formula with the components being main-group elements)[53] (FIG. 2e). All the crystal candidates were ranked according to their estimated thermodynamic stability on the convex hull, resulting in the identification of nearly 100 potentially stable new crystals that were then added to the Materials Project database[51]. Furthermore, detailed analysis of oxidation states resulted in the discovery of an exotic crystal, $NFAl_2Ca_6$, in which Al carries an unusual negative oxidation state. This surprising finding was possible only through the systematic combination of QM calculations and ML.

The fundamental nature of QML is not restricted to precalculated data sets. Within seminal work, Lee and co-workers have applied ML to experimental data in order to understand and control ligand–protein binding[126] (FIG. 2f). In particular, random-matrix theory was used to identify the chemical groups and features that strongly affect binding and those that do not. Such analysis can provide invaluable information on how to exploit local chemical properties to steer a complex mechanism, such as drug–target binding.

All of these examples demonstrate the great potential of QML for extracting statistical insights and new knowledge of quantum properties throughout CCS that cannot be directly obtained from conventional quantum calculations.

The data-driven nature of QML approaches that are based on exploring increasingly larger swathes of CCS also offers the possibility of rationally designing molecules with multiple desired properties. For example, in a hypothetical drug-design scenario, one could be interested in finding a particularly stable molecule with a large polarizability α (that would stabilize the drug–protein van der Waals interaction) and a large electronic HOMO–LUMO gap E_{gap} (that will afford stability with respect to external electrostatic fields). These three requirements would normally be considered contradictory to each other. Firstly, stability is typically inversely correlated with polarizability — stable molecules are normally thought to have small polarizability[127–129]. Secondly, the HOMO–LUMO gap is the leading-order contribution in the denominator of the polarizability formula, hence, it is often assumed that polarizable molecules should have small HOMO–LUMO gaps. One is then faced with a difficult question of whether the formulated design problem of low E, high α and high E_{gap} is achievable. This question can be partially answered by analysing the pairwise correlation between different molecular properties for a large but finite set of drug-like molecules. These correlations are shown in FIG. 3 for roughly 131,000 molecules in the QM9 data set[48] using the same analysis initially performed by Montavon and co-workers[47]. The first observation is that the correlation between most electronic properties is rather weak, if at all present. Most strikingly, above the lower bound of polarizability and atomization energy (both must be bounded from below), we observe no visible correlation. The same observation is made for polarizability versus the HOMO–LUMO gap. This pairwise comparison of three different properties suggests that one can find many drug-like molecules that satisfy the seemingly contrasting requirements of high stability, high polarizability and a large HOMO–LUMO gap. Similar freedom of design is observed for HOMO versus LUMO eigenvalues, as well as for HOMO–LUMO gap versus heat capacity. This data-driven analysis, spurred by the QML approach, illustrates a novel way to look at rational design in CCS, breaking conventional descriptor-property rules, as well as notions of restricted chemical diversity[47].

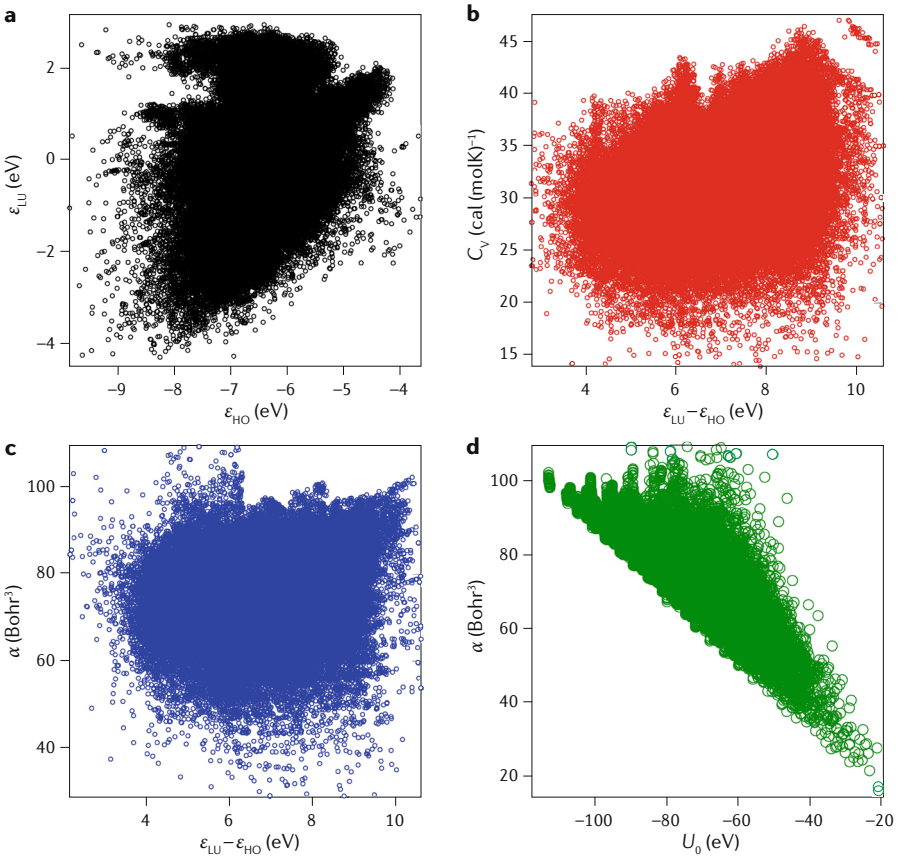

Fig. 3 | **Lack of visible correlation between pairs of molecular properties.** Weak dependency of molecular property pairs in chemical compound space illustrated by eigenvalues of highest occupied (HO) and lowest unoccupied (LU) molecular orbitals (part **a**), heat capacity and electronic gap (part **b**), polarizability and electronic gap (part **c**), and polarizability and atomization energy (part **d**). These results have been obtained using the analysis proposed in REF.[47] based on data from the QM9 data set[48].

Developments in QML approaches can potentially change the way we perform and use atomistic simulations. This is due to the use of rigorous QM and SM priors and data instead of heuristic cheminformatics, a holistic view of CCS rather than the traditional encyclopedic view in which systems are studied one at a time, and because insights provided by new tools are relevant in different scientific fields (electronic-structure prediction, materials science, organic chemistry, molecular dynamics and drug discovery). FIGURE 4 illustrates how this new approach could leverage conventional computational compound design applications and contribute substantially to ongoing experimental efforts, as also recently reviewed in the context of catalyst design[36]. As such, QML is clearly already taking the first steps in the direction of generally tackling the inverse design question in CCS[12,75,117].

We are confident that QML is not just limited to direct and inverse problems but can offer a fresh view on more fundamental aspects of molecules and materials. For example, QML represents a unique opportunity to rigorously test and assess known rules in chemistry, derived from human intuition and empiricism, up to an unprecedented degree of statistical confidence if sufficient data are made available. Furthermore, QML models could also help us discover and extract new concepts, which have, hitherto, escaped the notion of chemists. These developments may provide a further boost to attempts to gain a holistic understanding of CCS, its structure and what it holds in store for us in terms of interesting materials, properties or processes.

Challenges and outlook

Although QML has seen tremendous progress over the past few years, many more challenges remain to be addressed. In the following, we proceed by charting some of the challenges we consider as the most interesting and pressing.

Towards big data in CCS

An important limitation for the QML field is the lack of large, comprehensive data sets. Although data sets like QM7 (REFS[39,47]), QM9 (REF.[48]), Materials Project[51], OQMD[52], elpasolites[53], MD17 (REF.[49]), ANI-1 data set[50], silicon structures[54] and others have served well for the development and testing of new ML techniques, there is an inherent danger to overfit to benchmarks — an issue that has driven the field of computer vision, for example, to exploring increasingly large and complex data sets[130,131]. It is, therefore, important to establish large, high-quality data resources that can enable the development of a new model to explore both composition and configuration of an increasingly wide portion of CCS[86,102].

Despite the advantage of having comprehensive big data in CCS, it remains crucial to develop efficient models that only rely on small amounts of data because of the combinatorial scaling in CCS. Clearly, learning with abundant data is straightforward, but it becomes more challenging to reliably learn from small data sets. In this case, it becomes crucial to include prior physics-based knowledge and invariance information to achieve data efficiency without compromising the robustness and accuracy of the QML model[49,76,101].

Furthermore, the new generation of models should ideally quantify the uncertainty of its own prediction[89,132,133], possibly in combination with active learning strategies that may lead to improved sampling of CCS and effectively lead to smaller model uncertainties[134,135]. It is important to note that active learning actually induces non-stationarity in learning, since active learning by construction does not simply randomly sample from the underlying data distribution but actively biases the choice using the active learning score function[136]. In addition, models need to explain their prediction[114,122], such that they can also be a source of insight[59,86]. In other words, future developments need to consider data generation, model building, explanation, insight extraction and sampling in a single, comprehensive framework. The recently introduced AM-on approach, which selects molecular fragments and trains QML models on the fly, represents a first step in this direction[3].

Learning complex electronic properties

Current limitations and shortcomings of QML models are related to the prediction of intensive properties such as the eigenvalues of molecular orbitals[137] or excited-state

properties such as excitation energies[138,139]. Transferable yet accurate QML models of electron densities[140] and molecular orbitals in molecules and band structures in solids also remain a challenge. Another issue, lurking behind rigorous and robust statistical-learning procedures such as *k*-fold cross-validation and converged learning curves, is the selection bias encoded in many of the training sets used in the field. Stability or property distributions are typically unknown in CCS and, therefore, hamper the rigorous assessment of the degree to which any given data set is truly representative of broader chemical spaces. Similar problems of representability were also encountered in other fields of ML, for example, when trying to measure to what extent different search engines reflect all accessible website content on the internet. Practically, only a few random exemplary websites can be analysed while making a strong assumption that these selected websites are indeed representative of all the web page content on the internet[141,142].

Other challenges include the determination of the irreducible set of variables (that is, formal scaling of CCS is combinatorial in number of atoms and elemental species is only an upper bound but what is its effective dimensionality?)[79]. Assuring constant prediction errors throughout CCS is another challenge, as is reaching a quantitative understanding of the relation between the QML models' learning efficiency (as manifested in learning curve) and the dimensionalities of CCS as encoded in the training data.

Multiscale QML models

A very promising research direction is the integration of QML models across different levels of theory. By exploiting decades of research on the validity and applicability of the various approximations made when solving Schrödinger's equation, ample data obtained with computationally less demanding approximations can be combined with fewer but more accurate data points. As a result, the QML models must only learn the differences between the various levels of theory, which is substantially less demanding in terms of data needs. As such, these Δ-learning approaches allow us to invest the model complexity on the truly difficult aspects[37,105,106,143–145].

Many studies so far have successfully explored structure–property relationships in restricted chemical spaces. However, the final goal is to enable global and universal exploration of CCS exploiting the appropriate framework offered by the combination of QM, SM and ML. Here, we have connected some of the ongoing efforts in this endeavour and have provided pointers into the broad activities of the scientific field that has emerged. We would like to stress that the tools available now have reached a level of maturity that should become helpful to a wider community of researchers and practitioners.

Towards molecular design with QML

In the following, we would like to outline three specific and outstanding open challenges that require further development of QML tools to find broad application.

More observables. The use of QML to estimate statistical mechanics observables, for example, for the prediction of free-energy profiles of rare events, remains an outstanding challenge. This will enable direct comparison to experimental gas-phase rate constants from the literature, as well as to vibrational or linear free-energy-perturbation-based free-energy estimates. The comparison of the predictions to experimental results obtained in solution also requires the inclusion of solvent effects (that can be calculated using continuum solvent models, through addition of shells of solvent or through periodic boundary conditions). If necessary, one should also include nuclear quantum effects by performing path-integral simulations. These developments will serve the goal of establishing once and for all the validity of QML approaches by direct comparison to experiment, rather than to precalculated quantum results.

Experimental design. QML forms the natural basis for software that can run assisted experiments (automatized or robot) or help scientists with experimental-design decisions. For example, QML approaches along these lines have been used to guide the design of new materials[53,116,146,147]. Latent-space applications and computational alchemy can be combined with state-of-the-art optimizers, in essence, paving the way towards the experimental realization of aforementioned multi-property design tasks relevant to the identification of promising drug, photovoltaic, battery or catalyst candidates.

Reaction design using QML. Computer-based reaction planning and discovery has a long-standing history in chemistry, dating back to the 1960s, including contributions by Corey and co-workers in 1972 (REF.[148]) and Herges and Hoock in 1992 (REF.[149]). A comprehensive review of the field[150] also discusses the Chematica software that identifies unexplored synthetic routes based on new combinations of already established reactions reported in the literature. Laino and colleagues[151] and Segler et al.[152] introduced literature-based ML models for chemical reactions. However, these approaches can be problematic when it comes to combinations or reaction conditions for which previously published reactions are not representative. Moreover, the approach is inherently biased to established chemistry knowledge, limiting the possibility to discover entirely new reaction mechanisms and synthesis pathways. The latter point, however, is critical, for example, in the context of developing new catalysts and, more fundamentally, to fill any existing gaps in our understanding of potentially useful reactions. In order to computationally predict new reaction profiles, we must rely on universal first-principles numerical simulation of the relevant quantum and

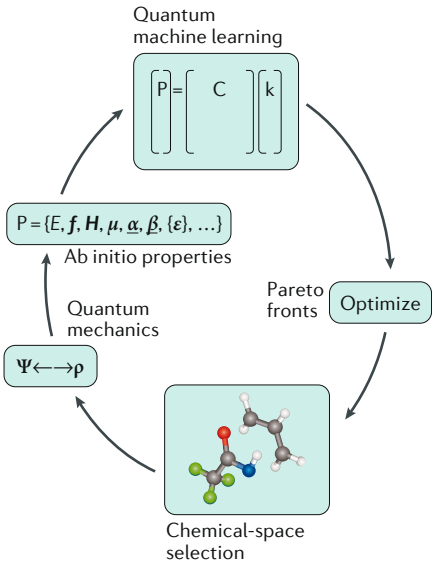

Fig. 4 | **Application concept of QML.** Multiple physico-chemical properties can be iteratively optimized in chemical space, spanned by structural and compositional degrees of freedom. After selection of some initial representative target compounds, quantum mechanics is invoked to calculate relevant ab initio properties (such as energies, forces, Hessian, dipole moment, polarizability, hyperpolarizability and electronic eigenvalues) and to subsequently generate and update a quantum machine learning (QML) model, in this case, exemplified by a kernel regression model with P, C and k corresponding to the property vector, regression coefficient matrix and kernel vector. Multi-objective property-optimization algorithms can subsequently create a new list of candidate structures until convergence is reached.

statistical mechanics, accounting for the electronic and atomic rearrangements that occur in a reaction pathway[153–155]. QM and SM account for the appropriate physics framework necessary to describe the electronic rearrangements occurring during a reaction and enable the user to dial in atomic configurations and chemical composition at will. When optimizing reactions through screens of prospective combinations of reactants, products and external conditions, the role of solvents and catalysts is crucial, as they can alter the ranking and even make the reactions possible. Therefore, one has to expand training sets to include libraries of solvents and simple catalysts, and apply extended QML models not only of energies and forces but also of statistical mechanical averages. Once trained, these QML models could be used to optimize reaction conditions (such as solvents, ions, temperature and pressure) in chemical space using gradient-free optimizers such as Monte Carlo, genetic or simplex algorithms. First steps in this direction were already taken in 2012 by Pozun and colleagues[156].

Conclusions

Over recent years, overwhelming evidence has been gathered by the community suggesting that QML models can be truly generalized throughout CCS. As such, our approach has changed, moving from globally fitting parameters in fixed functional forms, inspired by physics-informed models (such as universal force fields[157] or semi-empirical methods[158–160]), to locally optimizing regression weights in generic basis-set expansions that can be converged in size. Resulting QML models enable rapid predictions of relevant quantum properties for new out-of-sample systems (after training) and achieve converged predictive power through sufficiently large training sets (as evinced by convergence properties of learning curves). Thanks to the tremendous reduction in computational cost of query tasks, QML models can shift the focus from individual instances in CCS towards entire ensembles of compounds. Recovery (or rejection) of known and discovery and elucidation of unknown structure–property relationships have, therefore, become feasible, realistic and valuable goals that were previously not accessible.

Conceptual challenges include the definition of locality in CCS, that is, when the QML models are interpolating or, rather, extrapolating. Despite ensemble methods and Gaussian processes providing a first direction of uncertainty quantification in a limited domain of applicability, mathematically and physically more well-founded methods are still waiting to be discovered. Rigorous definitions of diversity[161] properly rooted in QM and SM might also be necessary to tackle the selection-bias problem and to maximize data efficiency. First-principles based diversity measures would have to properly account for all sorts of systems, including metal-organic frameworks, nanomaterials, organic materials, functional materials, inorganic crystals, metastable solids, liquid mixtures and biosystems.

An educational challenge corresponds to the establishment of the academic curriculum for this interdisciplinary young field, in which research programmes in chemistry, physics and computer science need to be tightly interwoven. Conventional curriculae in traditional departments of chemistry, materials science, physics, computer science or biology do not cover the coursework necessary for students to appropriately reach a level by which they can meaningfully contribute to this line of research.

Finally, we would like to stress that the progress made and described herein is dwarfed by the scope of the problem: gaining virtual control of CCS through physics-based understanding has remained elusive for all of humanity's past scientific efforts. One of the many rewards of reaching this goal would be the routine discovery and design of interesting molecules and materials with desired properties. As such, the community has, so far, just been scratching the surface of what is to come. To further push the frontier of this field of science, sustained and increasing investments are necessary in terms of computer power, interdisciplinary education and training, funding agencies and, most importantly, human interdisciplinary creativity.

O. Anatole von Lilienfeld[1]✉, *Klaus-Robert Müller*[2,3,4]✉ *and Alexandre Tkatchenko*[5]✉

[1]*Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, Basel, Switzerland.*

[2]*Machine Learning Group, Technische Universität Berlin, Berlin, Germany.*

[3]*Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea.*

[4]*Max Planck Institute for Informatics, Saarbrücken, Germany.*

[5]*Department of Physics and Materials Science, University of Luxembourg, Luxembourg City, Luxembourg.*

✉*e-mail: anatole.vonlilienfeld@unibas.ch; klaus-robert.mueller@tu-berlin.de; alexandre.tkatchenko@uni.lu*

https://doi.org/10.1038/s41570-020-0189-9
Published online: 12 June 2020

1. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823 (2004).
2. Mullard, A. The drug-maker's guide to the galaxy. *Nat. News* **549**, 445 (2017).
3. Huang, B. & von Lilienfeld, O. A. Efficient accurate scalable and transferable quantum machine learning with am-ons. Preprint at *arXiv* https://arxiv.org/abs/1707.04146 (2017).
4. Oprea T. I. et al. in *Molecular Interaction Fields* (Wiley-VCH, 2006).
5. Butina, D., Segall, M. D. & Frankcombe, K. Predicting ADME properties in silico: methods and models. *Drug Discov. Today* **7**, S83–S88 (2002).
6. Rajan, K. Materials informatics. *Mater. Today* **8**, 38–45 (2005).
7. 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.* **22**, 3762–3767 (2010).
8. Ward, L. & Wolverton, C. Atomistic calculations and materials informatics: a review. *Curr. Opin. Solid State Mater. Sci.* **21**, 167–176 (2017).
9. Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* **9**, 273–276 (2010).
10. von Lilienfeld, O. A. First principles view on chemical compound space: gaining rigorous atomistic control of molecular properties. *Int. J. Quantum Chem.* **113**, 1676–1689 (2013).
11. Van Noorden, R., Maher, B. & Nuzzo, R. The top 100 papers. *Nat. News* **514**, 550–553 (2014).
12. Franceschetti, A. & Zunger, A. The inverse band-structure problem of finding an atomic configuration with given electronic properties. *Nature* **402**, 60–63 (1999).
13. Jóhannesson, G. H. et al. Combined electronic structure and evolutionary search approach to materials design. *Phys. Rev. Lett.* **88**, 255506 (2002).
14. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
15. Hafner, J., Wolverton, C. & Ceder, G. Toward computational materials design: the impact of density functional theory on materials research. *MRS Bull.* **31**, 659–668 (2006).
16. Hachmann, J. et al. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2**, 2241–2251 (2011).
17. Marzari, N. Materials modelling: the frontiers and the challenges. *Nat. Mater.* **15**, 381–382 (2016).
18. Alberi, K. et al. The 2019 materials by design roadmap. *J. Phys. D Appl. Phys.* **52**, 013001 (2018).
19. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
20. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
21. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
22. Klauschen, F. et al. Scoring of tumor-infiltrating lymphocytes: from visual estimation to machine learning. *Semin. Cancer Biol.* **52**, 151–157 (2018).
23. Jurmeister, P. et al. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci. Transl Med.* **11**, eaaw8513 (2019).
24. Baldi, P., Sadowski, P. & Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun.* **5**, 4308 (2014).
25. Lengauer, T., Sander, O., Sierra, S., Thielen, A. & Kaiser, R. Bioinformatics prediction of HIV coreceptor usage. *Nat. Biotechnol.* **25**, 1407–1410 (2007).
26. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M. & Muller, K.-R. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal. Process. Mag.* **25**, 41–56 (2008).
27. Perozzi, B., Al-Rfou, R. & Skiena, S. in *Proc. ACM SIGKDD Int. Conf. Knowledge Discov. Data Mining*, 701–710 (ACM, 2014).
28. Thrun, S. Burgard, W. & Fox, D. *Probabilistic Robotics* (MIT Press, 2005).
29. Lewis, M. M. *Moneyball: The Art of Winning an Unfair Game* (Norton, W. W., 2003).
30. Ferrucci, D., Levas, A., Bagchi, S., Gondek, D. & Mueller, E. T. Watson: beyond jeopardy! *Artif. Intell.* **199**, 93–105 (2013).
31. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).

32. Lejaeghere, K. et al. Reproducibility in density functional theory calculations of solids. *Science* **351**, aad3000 (2016).
33. Rupp, M., von Lilienfeld, O. A. & Burke, K. Guest editorial: special topic on data-enabled theoretical chemistry. *J. Chem. Phys.* **148**, 241401 (2018).
34. Schneider, W. F. & Guo, H. Machine learning. *J. Phys. Chem. A* **122**, 879–879 (2018).
35. von Lilienfeld, O. A. Quantum machine learning in chemical compound space. *Angew. Chem. Int. Ed.* **57**, 4164–4169 (2018).
36. Freeze, J. G., Kelly, H. R. & Batista, V. S. Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. *Chem. Rev.* **119**, 6595–6612 (2019).
37. Ramakrishnan, R. et al. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
38. Mardt, A., Pasquali, L., Wu, H. & Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **9**, 5 (2018).
39. 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.* **108**, 058301 (2012).
40. Cortes, C., Jackel, L. D., Solla, S. A., Vapnik, V. & Denker, J. S. in *Advances in Neural Information Processing Systems*. 327–334 (1994).
41. Noé, F. Machine learning for molecular dynamics on long timescales. Preprint at *arXiv* https://arxiv.org/abs/1812.07669 (2018).
42. Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).
43. Fink, T., Bruggesser, H. & Reymond, J.-L. Virtual exploration of the small-molecule chemical universe below 160 daltons. *Angew. Chem. Int. Ed.* **44**, 1504–1508 (2005).
44. 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.* **47**, 342–353 (2007).
45. 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.* **131**, 8732–8733 (2009).
46. Ruddigkeit, L., van Deursen, R., Blum, L. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52**, 2684–2875 (2012).
47. Montavon, G. et al. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003 (2013).
48. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
49. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
50. 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* **4**, 170193 (2017).
51. Ong, S. et al. The materials project. *Materials Project* http://materialsproject.org/ (2011).
52. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
53. 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.* **117**, 135502 (2016).
54. Bartók, A., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X.* **8**, 041048 (2018).
55. Pettifor, D. G. The structures of binary compounds. I. Phenomenological structure maps. *J. Phys. C. Solid State Phys.* **19**, 285–313 (1986).
56. Pettifor, D. G. Structure maps for pseudobinary and ternary phases. *Mater. Sci. Technol.* **4**, 675–691 (1988).
57. Willatt, M. J., Musil, F. & Ceriotti, M. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Phys. Chem. Chem. Phys.* **20**, 29661–29668 (2018).
58. 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.* **148**, 241717 (2018).
59. 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.* **148**, 241722 (2018).
60. Bartók, A. et al. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
61. Sumpter, B. G. & Noid, D. W. Potential energy surfaces for macromolecules. A neural network technique. *Chem. Phys. Lett.* **192**, 455–462 (1992).
62. Ho, T. S. & Rabitz, H. A general method for constructing multidimensional molecular potential energy surfaces from *ab initio* calculations. *J. Chem. Phys.* **104**, 2584–2597 (1996).
63. Lorenz, S., Gross, A. & Scheffler, M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chem. Phys. Lett.* **395**, 210–215 (2004).
64. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
65. Bartók, A., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
66. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
67. Hansen, K. et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
68. Ramakrishnan, R. & von Lilienfeld, O. A. Many molecular properties from one kernel in chemical space. *CHIMIA* **69**, 182–186 (2015).
69. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
70. Schütt, K. et al. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
71. Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
72. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
73. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
74. Pyzer-Knapp, E. O., Li, K. & Aspuru-Guzik, A. Learning from the Harvard clean energy project: The use of neural networks to accelerate materials discovery. *Adv. Funct. Mater.* **25**, 6495–6502 (2015).
75. Jørgensen, M. S., Larsen, U. F., Jacobsen, K. W. & Hammer, B. Exploration versus exploitation in global atomistic structure optimization. *J. Phys. Chem. A* **122**, 1504–1509 (2018).
76. 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.* **240**, 38–45 (2019).
77. Huang, B. & von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **145**, 161102 (2016).
78. 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.* **14**, 2991–3003 (2018).
79. Braun, M. L., Buhmann, J. M. & Müller, K. R. On relevant dimensions in kernel feature spaces. *J. Mach. Learn. Res.* **9**, 1875–1906 (2008).
80. 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.* **115**, 1084–1093 (2015).
81. Christensen, A. S., Faber, F. A. & von Lilienfeld, O. A. Operators in quantum machine learning: response properties in chemical space. *J. Chem. Phys.* **150**, 064105 (2019).
82. Bartók, A., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
83. Hansen, K., Biegler, F., von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
84. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).
85. Huo, H. & Rupp, M. Unified representation for machine learning of molecules and crystals. Preprint at *arXiv* https://arxiv.org/abs/1704.06439 (2017).
86. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).
87. 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.* **148**, 241708 (2018).
88. Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **5**, eaav6490 (2019).
89. Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R. & Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).
90. Carleo, G. & Troyer, M. Solving the quantum many-body problem with artificial neural networks. *Science* **355**, 602–606 (2017).
91. Brockherde, F., Li, L., Tuckerman, M. E., Burke, K. & Müller, K.-R. Bypassing the Kohn–Sham equations with machine learning. *Nat. Commun.* **8**, 872 (2017).
92. Schütt, K., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10**, 5024 (2019).
93. Fabrizio, A., Grisafi, A., Meyer, B., Ceriotti, M. & Corminboeuf, C. Electron density learning of non-covalent systems. *Chem. Sci.* **10**, 9424–9432 (2019).
94. Hermann, J., Schätzle, Z. & Noé, F. Deep neural network solution of the electronic Schrödinger equation. Preprint at *arXiv* https://arxiv.org/abs/1909.08423 (2019).
95. Pfau, D., Spencer, J. S. de A., Matthews, G. G. & Foulkes, W. M. C. Ab-initio solution of the many-electron Schrödinger equation with deep neural networks. Preprint at *arXiv* https://arxiv.org/abs/1909.02487 (2019).
96. Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **115**, 1032–1050 (2015).
97. Shapeev, A. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2016).
98. Sauceda, H. E., Chmiela, S., Poltavsky, I., Müller, K.-R. & Tkatchenko, A. Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *J. Chem. Phys.* **150**, 114102 (2019).
99. Deringer, V. L. et al. Computational surface chemistry of tetrahedral amorphous carbon by combining machine learning and density functional theory. *Chem. Mater.* **30**, 7438–7445 (2018).
100. 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.* **30**, 7446–7455 (2018).
101. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).
102. 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.* **8**, 3192–3203 (2017).
103. 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.* **148**, 241718 (2018).
104. Chen, X., Jørgensen, M. S., Li, J. & Hammer, B. Atomic energies from a convolutional neural network. *J. Chem. Theory Comput.* **14**, 3933–3942 (2018).

105. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).
106. Zaspel, B., Huang, H., Harbrecht & von Lilienfeld, O. A. Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited. *J. Chem. Theory Comput.* **15**, 1546–1559 (2018).
107. Batra, R., Pilania, G., Uberuaga, B. & Ramprasad, R. Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia. *ACS Appl. Mater. Interfaces* **11**, 24906–24918 (2019).
108. Rupp, M., Ramakrishnan, R. & von Lilienfeld, O. A. Machine learning for quantum mechanical properties of atoms in molecules. *J. Phys. Chem. Lett.* **6**, 3309–3313 (2015).
109. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate *ab initio* molecular dynamics. *Int. J. Quantum Chem.* **115**, 1074–1083 (2015).
110. Jacobsen, T. L., Jørgensen, M. S. & Hammer, B. On-the-fly machine learning of atomic potential in density functional theory structure optimization. *Phys. Rev. Lett.* **120**, 026102 (2018).
111. Christensen, A. S. et al. QML: a Python toolkit for quantum machine learning. *GitHub* https://github.com/qmlcode/qml (2017).
112. Schütt, K. et al. SchNetPack: a deep learning toolbox for atomistic systems. *J. Chem. Theory Comput.* **15**, 448–455 (2018).
113. Alber, M. et al. iNNvestigate neural networks! *J. Mach. Learn. Res.* **20**, 1–8 (2019).
114. Lapuschkin, S. et al. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1096 (2019).
115. Binder, A. et al. Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles. Preprint at *arXiv* https://arxiv.org/abs/1805.11178 (2018).
116. Gõmez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
117. Zunger, A. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* **2**, 0121 (2018).
118. Kuhn, C. & Beratan, D. N. Inverse strategies for molecular design. *J. Phys. Chem.* **100**, 10595–10599 (1996).
119. von Lilienfeld, O. A., Lins, R. & Rothlisberger, U. Variational particle number approach for rational compound design. *Phys. Rev. Lett.* **95**, 153002 (2005).
120. Wang, M., Hu, X., Beratan, D. N. & Yang, W. Designing molecules by optimizing potentials. *J. Am. Chem. Soc.* **128**, 3228–3232 (2006).
121. d'Avezac, M. & Zunger, A. Identifying the minimum-energy atomic configuration on a lattice: Lamarckian twist on Darwinian evolution. *Phys. Rev. B* **78**, 064102 (2008).
122. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* **10**, e0130140 (2015).
123. Ribeiro, M. T., Singh, S. & Guestrin, C. in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discov. Data Mining* 1135–1144 (ACM, 2016).
124. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal. Process.* **73**, 1–15 (2018).
125. Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta.* **44**, 129–138 (1977).
126. Lee, A. A. et al. Ligand biological activity predicted by cleaning positive and negative chemical correlations. *Proc. Natl Acad. Sci. USA* **116**, 3373–3378 (2019).
127. Hohm, U. Dipole polarizability and bond dissociation energy. *J. Chem. Phys.* **101**, 6362–6364 (1994).
128. Hohm, U. Is there a minimum polarizability principle in chemical reactions? *J. Phys. Chem. A.* **104**, 8418–8423 (2000).
129. Geerlings, P., De Proft, F. & Langenaeker, W. Conceptual density functional theory. *Chem. Rev.* **103**, 1793–1874 (2003).
130. Deng, J. et al. in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.* 248–255 (IEEE, 2009).
131. Rohrbach, M., Amin, S., Andriluka, M. & Schiele, B. in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.* 1194–1201 (IEEE, 2012).
132. Schwaighofer, A., Schroeter, T., Mika, S. & Blanchard, G. How wrong can we get? A review of machine learning approaches and error bars. *Comb. Chem. High Throughput Screen.* **12**, 453–468 (2009).
133. Smith, R. C. *Uncertainty Quantification: Theory, Implementation, and Applications* (SIAM, 2013).
134. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
135. Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. Machine learning of molecular properties: Locality and active learning. *J. Chem. Phys.* **148**, 241727 (2018).
136. Sugiyama, M. & Kawanabe, M. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation* (MIT Press, 2012).
137. Faber, F. A. et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).
138. Ramakrishnan, R., Hartmann, M., Tapavicza, E. & von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *J. Chem. Phys.* **143**, 084111 (2015).
139. Pronobis, W., Schütt, K. T., Tkatchenko, A. & Müller, K.-R. Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning. *Eur. Phys. J. B* **91**, 178 (2018).
140. Grisafi, A. et al. Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **5**, 57–64 (2019).
141. Lawrence, S. & Giles, C. L. Accessibility of information on the web. *Nature* **400**, 107 (1999).
142. Lawrence, S. & Giles, C. L. Searching the world wide web. *Science* **280**, 98–100 (1998).
143. Ginzburg I. & Horn, D. in *Advances in Neural Information Processing Systems* (eds Jordan, M. I., LeCun, Y. & Solla, S. A.) 224–231 (MIT Press, 1994).
144. Bogojeski, M., Vogt-Maranto, L., Tuckerman, M. E., Mueller, K.-R. & Burke, K. Density functionals with quantum chemical accuracy: from machine learning to molecular dynamics. Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv.8079917.v1 (2019).
145. Smith, J. S. et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10**, 2903 (2019).
146. 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.* **19**, 3931–3935 (2016).
147. 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.* **9**, 7069–7077 (2018).
148. Corey, E. J., Wipke, W. T., Cramer, R. D. & Howe, W. J. Computer-assisted synthetic analysis. facile man-machine communication of chemical structure by interactive computer graphics. *J. Am. Chem. Soc.* **94**, 421–430 (1972).
149. Herges, R. & Hoock, C. Reaction planning: Computer-aided discovery of a novel elimination reaction. *Science* **255**, 711–713 (1992).
150. Szymkuć, S. et al. Computer-assisted synthetic planning: The end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
151. Schwaller, T., Gaudin, D., Lanyi, C., Bekas & Laino, T. "Found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).
152. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
153. Leach, A. R. *Molecular Modelling: Principles and Applications* (Addison-Wesley Longman, 1998).
154. Helgaker, T., Jørgensen, P. & Olsen, J. *Molecular Electronic-Structure Theory* (Wiley, 2000).
155. Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation* (Oxford Univ. Press, 2010).
156. Pozun, Z. D. et al. Optimizing transition states via kernel-based machine learning. *J. Chem. Phys.* **136**, 174101–174109 (2012).
157. Rappé, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. III & Skid, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
158. Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **13**, 1173–1213 (2007).
159. Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **19**, 1–32 (2013).
160. Aradi, B., Hourahine, B. & Frauenheim, T. DFTB+, a sparse matrix-based implementation of the DFTB method. *J. Phys. Chem. A* **111**, 5678–5684 (2007).
161. Marienwald, H., Pronobis, W., Müller, K.-R. & Nakajima, S. Tight bound of incremental cover trees for dynamic diversification. Preprint at *arXiv* https://arxiv.org/abs/1806.06126 (2018).
162. Gilmer, J., Schoenholz, S. S., Riley, F., Vinyals, O. & Dahl, G. E. in *Proc. Int. Conf. Mach. Learn.* 1263–1272 (2017).
163. Nebgen, B. et al. Transferable dynamic molecular charge assignment using deep neural networks. *J. Chem. Theory Comput.* **14**, 4687–4698 (2018).
164. Eickenberg, M., Exarchakis, G., Hirn, M., Mallat, S. & Thiry, L. Solid harmonic wavelet scattering for predictions of molecule properties. *J. Chem. Phys.* **148**, 241732 (2018).
165. Faber, F. A., Christensen, A. S. & von Lilienfeld O. A. in *Machine Learning meets Quantum Physics, Lecture Notes in Physics* (eds Schütt, K. T. et al.) (Springer, 2020).
166. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural networks potentials. *J. Chem. Phys.* **134**, 074106 (2011).

Acknowledgements
All authors thank F. A. Faber and J. Wagner for preparing the graphics in Fig. 1 and the cover image related to this article, respectively. O.A.v.L. acknowledges funding from the Swiss National Science foundation (nos. PP00P2_138932 and 407540_167186 NFP 75 Big Data) and from the European Research Council (ERC-CoG grant QML). This work was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. A.T. acknowledges financial support from the European Research Council (ERC-CoG grant BeStMo). K.-R.M. acknowledges partial financial support by the German Federal Ministry of Education and Research (BMBF) under grants 01IS14013A-E, 01GQ1115 and 01GQ0850; Deutsche Forschungsgesellschaft (DFG) under grant Math+, EXC 2046/1, project ID 390685689 and by the Institute for Information & Communication Technology Promotion (IITP) grant funded by the Korea government (nos. 2017-0-00451 and 2017-0-01779). Correspondence to O.A.v.L., K.-R.M. and A.T.

Author contributions
All authors contributed equally to the preparation of this manuscript.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Chemistry thanks F. Noé, G. Csanyi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

RELATED LINKS
Accurate neural-network engine for molecular energies (ANI) neural-network package: https://github.com/isayev/ASE_ANI
QM9 challenge: https://tinyurl.com/y2e589wj
Repository of data sets for quantum machine learning: http://quantum-machine.org
SchNetPack: https://github.com/atomistic-machine-learning/schnetpack
Symmetrized gradient-domain machine learning (sGDML): http://quantum-machine.org/gdml/#code

© Springer Nature Limited 2020