REVIEW

https://doi.org/10.1038/s41586-018-0337-2

Machine learning for molecular and materials science

Keith T. Butler[1], Daniel W. Davies[2], Hugh Cartwright[3], Olexandr Isayev[4]* & Aron Walsh[5,6]*

Here we summarize recent progress in machine learning for the chemical sciences. We outline machine-learning techniques that are suitable for addressing research questions in this domain, as well as future directions for the field. We envisage a future in which the design, synthesis, characterization and application of molecules and materials is accelerated by artificial intelligence.

The Schrödinger equation provides a powerful structure–property relationship for molecules and materials. For a given spatial arrangement of chemical elements, the distribution of electrons and a wide range of physical responses can be described. The development of quantum mechanics provided a rigorous theoretical foundation for the chemical bond. In 1929, Paul Dirac famously proclaimed that the underlying physical laws for the whole of chemistry are "completely known"[1]. John Pople, realizing the importance of rapidly developing computer technologies, created a program—Gaussian 70—that could perform ab initio calculations: predicting the behaviour, for molecules of modest size, purely from the fundamental laws of physics[2]. In the 1960s, the Quantum Chemistry Program Exchange brought quantum chemistry to the masses in the form of useful practical tools[3]. Suddenly, experimentalists with little or no theoretical training could perform quantum calculations too. Using modern algorithms and supercomputers, systems containing thousands of interacting ions and electrons can now be described using approximations to the physical laws that govern the world on the atomic scale[4–6].

The field of computational chemistry has become increasingly predictive in the twenty-first century, with activity in applications as wide ranging as catalyst development for greenhouse gas conversion, materials discovery for energy harvesting and storage, and computer-assisted drug design[7]. The modern chemical-simulation toolkit allows the properties of a compound to be anticipated (with reasonable accuracy) before it has been made in the laboratory. High-throughput computational screening has become routine, giving scientists the ability to calculate the properties of thousands of compounds as part of a single study. In particular, density functional theory (DFT)[8,9], now a mature technique for calculating the structure and behaviour of solids[10], has enabled the development of extensive databases that cover the calculated properties of known and hypothetical systems, including organic and inorganic crystals, single molecules and metal alloys[11–13].

The emergence of contemporary artificial-intelligence methods has the potential to substantially alter and enhance the role of computers in science and engineering. The combination of big data and artificial intelligence has been referred to as both the "fourth paradigm of science"[14] and the "fourth industrial revolution"[15], and the number of applications in the chemical domain is growing at an astounding rate. A subfield of artificial intelligence that has evolved rapidly in recent years is machine learning. At the heart of machine-learning applications lie statistical algorithms whose performance, much like that of a researcher, improves with training. There is a growing infrastructure of machine-learning tools for generating, testing and refining scientific models. Such techniques are suitable for addressing complex problems that involve massive combinatorial spaces or nonlinear processes, which conventional procedures either cannot solve or can tackle only at great computational cost.

As the machinery for artificial intelligence and machine learning matures, important advances are being made not only by those in mainstream artificial-intelligence research, but also by experts in other fields (domain experts) who adopt these approaches for their own purposes. As we detail in Box 1, the resources and tools that facilitate the application of machine-learning techniques mean that the barrier to entry is lower than ever.

In the rest of this Review, we discuss progress in the application of machine learning to address challenges in molecular and materials research. We review the basics of machine-learning approaches, identify areas in which existing methods have the potential to accelerate research and consider the developments that are required to enable more wide-ranging impacts.

Nuts and bolts of machine learning

With machine learning, given enough data and a rule-discovery algorithm, a computer has the ability to determine all known physical laws (and potentially those that are currently unknown) without human input. In traditional computational approaches, the computer is little more than a calculator, employing a hard-coded algorithm provided by a human expert. By contrast, machine-learning approaches learn the rules that underlie a dataset by assessing a portion of that data and building a model to make predictions. We consider the basic steps involved in the construction of a model, as illustrated in Fig. 1; this constitutes a blueprint of the generic workflow that is required for the successful application of machine learning in a materials-discovery process.

Data collection

Machine learning comprises models that learn from existing (training) data. Data may require initial preprocessing, during which missing or spurious elements are identified and handled. For example, the Inorganic Crystal Structure Database (ICSD) currently contains more than 190,000 entries, which have been checked for technical mistakes but are still subject to human and measurement errors. Identifying and removing such errors is essential to avoid machine-learning algorithms being misled. There is a growing public concern about the lack of reproducibility and error propagation of experimental data

[1]ISIS Facility, Rutherford Appleton Laboratory, Harwell Campus, Harwell, UK. [2]Department of Chemistry, University of Bath, Bath, UK. [3]Department of Chemistry, Oxford University, Oxford, UK. [4]Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [5]Department of Materials Science and Engineering, Yonsei University, Seoul, South Korea. [6]Department of Materials, Imperial College London, London, UK. *e-mail: olexandr@olexandrisayev.com; a.walsh@imperial.ac.uk

Box 1

Learning to learn

One of the most exciting aspects of machine-learning techniques is their potential to democratize molecular and materials modelling by reducing the computer power and prior knowledge required for entry. Just as Pople's Gaussian software made quantum chemistry more accessible to a generation of experimental chemists, machine-learning approaches, if developed and implemented correctly, can broaden the routine application of computer models by non-specialists. The accessibility of machine-learning technology relies on three factors: open data, open software and open education. There is an increasing drive for open data within the physical sciences, with an ideal best practice outlined recently[98,99]. Some of the open software being developed is listed in Table 2. There are also many excellent open education resources available, such as massive open online courses (MOOCs).

fast.ai (http://www.fast.ai) is a course that is "making neural nets uncool again" by making them accessible to a wider community of researchers. One of the advantages of this course is that users start to build working machine-learning models almost immediately. However, it is not for absolute beginners, requiring a working knowledge of computer programming and high-school-level mathematics.

DataCamp (https://www.datacamp.com) offers an excellent introduction to coding for data-driven science and covers many practical analysis tools relevant to chemical datasets. This course features interactive environments for developing and testing code and is suitable for non-coders because it teaches Python at the same time as machine learning.

Academic MOOCs are useful courses for those wishing to get more involved with the theory and principles of artificial intelligence and machine learning, as well as the practice. The Stanford MOOC (https://www.coursera.org/learn/machine-learning) is popular, with excellent alternatives available from sources such as https://www.edx.org (see, for example, 'Learning from data (introductory machine learning)') and https://www.udemy.com (search for 'Machine learning A–Z'). The underlying mathematics is the topic of a course from Imperial College London (https://www.coursera.org/specializations/mathematics-machine-learning).

Many machine-learning professionals run informative blogs and podcasts that deal with specific aspects of machine-learning practice. These are useful resources for general interest as well as for broadening and deepening knowledge. There are too many to provide an exhaustive list here, but we recommend https://machinelearningmastery.com and http://lineardigressions.com as a starting point.

published in peer-reviewed scientific literature. In certain fields, such as cheminformatics, best practices and guidelines have been established to address these problems[16].

The training of a machine-learning model may be supervised, semi-supervised or unsupervised, depending on the type and amount of available data. In supervised learning, the training data consist of sets of input and associated output values. The goal of the algorithm is to derive a function that, given a specific set of input values, predicts the output values to an acceptable degree of fidelity. If the available dataset consists of only input values, unsupervised learning can be used in an attempt to identify trends, patterns or clustering in the data. Semi-supervised learning may be of value if there is a large amount of input data, but only a limited amount of corresponding output data.

Supervised learning is the most mature and powerful of these approaches, and is used in the majority of machine-learning studies in the physical sciences, such as in the mapping of chemical composition to a property of interest. Unsupervised learning is less common, but can be used for more general analysis and classification of data or to identify previously unrecognized patterns in large datasets[17].

Data representation

Even though raw scientific data are usually numerical, the form in which data are presented often affects learning. In many types of spectroscopy, the signal is acquired in the time domain, but for interpretation it is converted to the frequency domain using the Fourier transform. Like scientists, a machine-learning algorithm might learn more effectively using one format than the other. The process of converting raw data into something more suitable for an algorithm is called featurization or feature engineering.

The more suitable the representation of the input data, the more accurately can an algorithm map it to the output data. Selecting how best to represent the data could require insight into both the underlying scientific problem and the operation of the learning algorithm, because it is not always obvious which choice of representation will give the best performance; this is an active topic of research for chemical systems[18].

Many representations are available to encode structures and properties. One example is the Coulomb matrix[19], which contains information on atomic nuclear repulsion and the potential energy of free atoms, and is invariant to molecular translations and rotation. Molecular systems also lend themselves to descriptions as graphs[20]. In the solid state, the conventional description of crystal structures that uses translation vectors and fractional coordinates of the atoms is not appropriate for machine learning because a lattice can be represented in an infinite number of ways by choosing a different coordinate system. Representations based on radial distribution functions[21], Voronoi tessellations[22] or property-labelled materials fragments[23] are among the new ways in which this problem is being tackled.

Choice of learner

When the dataset has been collected and represented appropriately, it is time to choose a model to learn from it. A wide range of model types (or learners) exists for model building and prediction. Supervised-learning models may predict output values within a discrete set (such as categorizing a material as a metal or an insulator) or a continuous set (such as polarizability). Building a model for the former requires classification, whereas the latter requires regression. A range of learning algorithms can be applied (see Table 1), depending on the type of data and the question posed. It may be helpful to use an ensemble of different algorithms, or of similar algorithms with different values for their internal parameters (known as 'bagging' or 'stacking'), to create a more robust overall model. We outline some of the common algorithms (learners) in the following.

Naive Bayes classifiers[24] are a collection of classification algorithms based on Bayes' theorem that identify the most probable hypothesis, given the data as prior knowledge about the problem. Bayes' theorem provides a formal way of calculating the probability that a hypothesis is correct, given a set of existing data. New hypotheses can then be tested and the prior knowledge updated. In this way, the hypothesis (or model) with the highest probability of correctly representing the data can be selected.

In *k*-nearest-neighbour[25] methods, the distances between samples and training data in a descriptor hyperspace are calculated. They are so called because the output value for a prediction relies on the values of the *k* 'nearest neighbours' in the data, where *k* is an integer. Nearest-neighbour models can be used in both classification and regression models: in classification, the prediction is determined by the class of the majority of the *k* nearest points; in regression, it is determined by the average of the *k* nearest points.

Decision trees[26] are flowchart-like diagrams used to determine a course of action or outcome. Each branch of the tree represents a possible decision, occurrence or reaction. The tree is structured to show how and why one choice may lead to the next, with branches indicating that each option is mutually exclusive. Decision trees comprise a root node, leaf nodes and branches. The root node is the starting point of

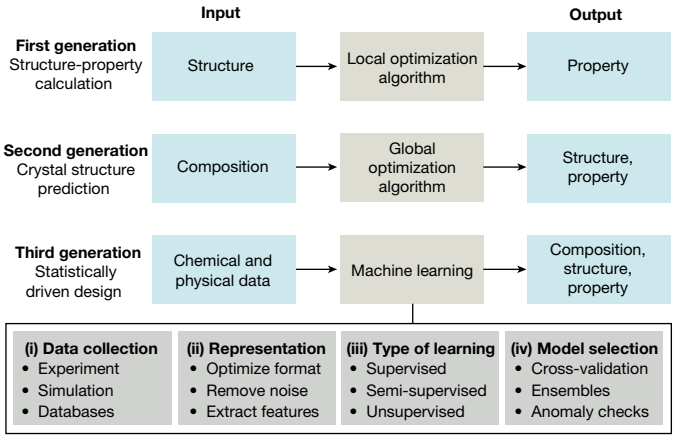

Fig. 1 | Evolution of the research workflow in computational chemistry. The standard paradigm in the first-generation approach is to calculate the physical properties of an input structure, which is often performed via an approximation to the Schrödinger equation combined with local optimization of the atomic forces. In the second-generation approach, by using global optimization (for example, an evolutionary algorithm) an input of chemical composition is mapped to an output that contains predictions of the structure or ensemble of structures that the combination of elements are likely to adopt. The emerging third-generation approach is to use machine-learning techniques with the ability to predict composition, structure and properties provided that sufficient data are available and an appropriate model is trained. Four stages of training a machine-learning model with some of the common choices are listed in the bottom panel.

the tree. Both root and leaf nodes contain questions or criteria to be addressed. Branches are arrows connecting nodes, showing the flow from question to answer. Decision trees are often used in ensemble methods (meta-algorithms), which combine multiple trees into one predictive model to improve performance.

Kernel methods are a class of algorithms, the best known members of which are support vector machine and kernel ridge regression[27]. The name 'kernel' comes from the use of a kernel function—a function that transforms input data into a higher-dimensional representation that makes the problem easier to solve. In a sense, a kernel is a similarity function provided by the domain expert: it takes two inputs and creates an output that quantifies how similar they are.

Artificial neural networks and deep neural networks[28] loosely mimic the operation of the brain, with artificial neurons (the processing unit) arranged in input, output and hidden layers. In the hidden layers, each neuron receives input signals from other neurons, integrates those signals and then uses the result in a straightforward computation. Connections between neurons have weights, the values of which represent the stored knowledge of the network. Learning is the process of adjusting the weights so that the training data are reproduced as accurately as possible.

Whatever the model, most learners are not fully autonomous, requiring at least some guidance. The values of internal variables (hyperparameters) are estimated beforehand using systematic and random searches, or heuristics. Even modest changes in the values of hyperparameters can improve or impair learning considerably, and the selection of optimal values is often problematic. Consequently, the development of automatic optimization algorithms is an area of active investigation, as is their incorporation into accessible packages for non-expert users (see Table 2).

Model optimization

When the learner (or set of learners) has been chosen and predictions are being made, a trial model must be evaluated to allow for optimization and ultimate selection of the best model. Three principal sources of error arise and must be taken into account: model bias, model variance and irreducible errors, with the total error being the sum of these. Bias is the error from incorrect assumptions in the algorithm and can result in the model missing underlying relationships. Variance is sensitivity to small fluctuations in the training set. Even well-trained machine-learning models may contain errors due to noise in the training data, measurement limitations, calculation uncertainties, or simply outliers or missing data. Poor model performance usually indicates a high bias or a high variance, as illustrated in Fig. 2.

High bias (also known as underfitting) occurs when the model is not flexible enough to adequately describe the relationship between inputs and predicted outputs, or when the data are insufficiently detailed to allow the discovery of suitable rules. High variance (or overfitting) occurs when a model becomes too complex; typically, this occurs as the number of parameters is increased. The diagnostic test for overfitting is that the accuracy of a model in representing training data continues to improve, while the performance in estimating test data plateaus or declines.

The key test for the accuracy of a machine-learning model is its successful application to unseen data. A widely used method for determining the quality of a model involves withholding a randomly selected portion of data during training. This withheld dataset, known as a test set, is shown to the model once training is complete (Fig. 2). The extent to which the output data in the validation set is accurately predicted then provides a measure of the effectiveness of training. Cross-validation is reliable only when the samples used for training and validation are representative of the whole population, which may present problems if the sample size is small or if the model is applied to data from compounds that are very different to those in the original dataset. A careful selection of methods for evaluating the transferability and applicability of a model is required in such cases.

Accelerating the scientific method

Whether through the enumeration and analysis of experimental data or the codification of chemical intuition, the application of informatics to guide laboratory chemists is advancing rapidly. In this section, we explore how machine learning is helping to progress and to reduce the barriers between chemical and materials design, synthesis, characterization and modelling. We also describe some of the important developments in the field of artificial intelligence for data-mining existing literature.

Guiding chemical synthesis

Organic chemists were among the first scientists to recognize the potential of computational methods in laboratory practice. E. J. Corey's Organic Chemical Simulation of Synthesis (OCSS) program[29], developed 50 years ago, was an attempt to automate retrosynthetic analysis. In a synthetic chemistry route, the number of possible transformations

Table 1 | Classes of machine-learning techniques and some chemical questions they could answer

Class	Bayesian	Evolutionary[a]	Symbolist	Connectionist	Analogist
Method	Probabilistic inference	Evolving structures	Logical inference	Pattern recognition	Constrained optimization
Algorithms include	Naive Bayes Bayesian networks	Genetic algorithm Particle swarm	Rules Decision trees	Artificial neural networks Back propagation	Nearest neighbour Support vectors
Chemical query	Is my new theory valid?	What molecule gives this property?	How do I make this material?	What compound did I synthesize?	Find a structure–property relation

The classes shown were chosen following ref. [97].
[a]Although evolutionary algorithms are often integrated into machine-learning procedures, they form part of a wider class of stochastic search algorithms.

Table 2 | Publicly accessible learning resources and tools related to machine learning

Name	Description	URL
General-purpose machine-learning frameworks		
Caret	Package for machine learning in R	https://topepo.github.io/caret
Deeplearning4j	Distributed deep learning for Java	https://deeplearning4j.org
H2O.ai	Machine-learning platform written in Java that can be imported as a Python or R library	https://h2o.ai
Keras	High-level neural-network API written in Python	https://keras.io
Mlpack	Scalable machine-learning library written in C++	https://mlpack.org
Scikit-learn	Machine-learning and data-mining member of the scikit family of toolboxes built around the SciPy Python library	http://scikit-learn.org
Weka	Collection of machine-learning algorithms and tasks written in Java	https://cs.waikato.ac.nz/ml/weka
Machine-learning tools for molecules and materials		
Amp	Package to facilitate machine learning for atomistic calculations	https://bitbucket.org/andrewpeterson/amp
ANI	Neural-network potentials for organic molecules with Python interface	https://github.com/isayev/ASE_ANI
COMBO	Python library with emphasis on scalability and efficiency	https://github.com/tsudalab/combo
DeepChem	Python library for deep learning of chemical systems	https://deepchem.io
GAP	Gaussian approximation potentials	http://libatoms.org/Home/Software
MatMiner	Python library for assisting machine learning in materials science	https://hackingmaterials.github.io/matminer
NOMAD	Collection of tools to explore correlations in materials datasets	https://analytics-toolkit.nomad-coe.eu
PROPhet	Code to integrate machine-learning techniques with quantum-chemistry approaches	https://github.com/biklooost/PROPhet
TensorMol	Neural-network chemistry package	https://github.com/jparkhill/TensorMol

per step can range from around 80 to several thousand[30]; for comparison, there are only tens of potential moves at each game position in chess[31]. In chemical synthesis, human experts are required to specify conditional and contextual rules, which exclude large sets of potential reagents at a given step, limiting the number of choices available to the algorithm. The contextual rules (typically many thousands of them) are of utmost importance if a machine relying on a traditional algorithm is to compete with an expert. Recent breakthroughs in the Chematica program have shown that computers can be more efficient than humans in these tasks[32].

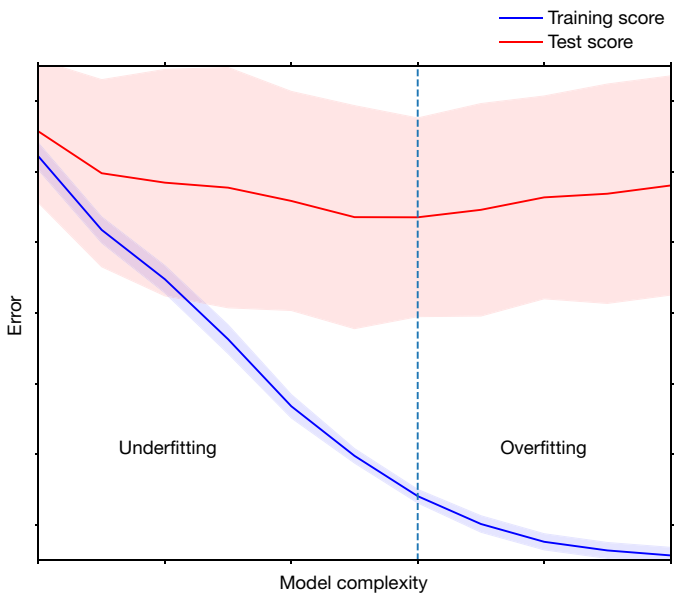

Fig. 2 | Errors that arise in machine-learning approaches. Errors can arise during both the training of a new model (blue line) and the application of a built model (red line). A simple model may suffer from high bias (underfitting), whereas a complex model may suffer from high variance (overfitting), which leads to a bias–variance trade-off. In the underfitting region the model performance can improve with further parameterization, whereas in the overfitting region the model performance will decrease. The optimal point for a model is just before the performance on the testing set starts to deteriorate with increased parameterization, which is indicated by the dashed vertical line. The model shown here is built on an example from https://kaggle.com, available at https://keeeto.github.io/blog/bias_variance. The shaded areas show the standard deviations of the fits for model training (blue) and testing (red).

The combination of extremely complex systems and huge numbers of potential solutions, which arise from competing objective functions (such as cost, purity, time and toxicity), make synthetic chemistry ill-suited to the application of traditional algorithmic approaches. However, because of this complexity, synthesis is one area of research that can benefit most from the application of artificial intelligence.

Deep-learning approaches, which typically rely on many-layered artificial neural networks or a combination of artificial neural networks and other learning techniques such as Boltzmann machines, are showing particular promise for predicting chemical-synthesis routes by combining rules-based expert systems with neural networks that rank the candidate synthetic pathways[33] or the likelihood of a predicted product by applying the rules[34]. One artificial neural network that learned from examples taken from the chemical literature was able to achieve a level of sophistication such that trained chemists could not distinguish between computer- and human-expert-designed routes[30]. However, a severe drawback of rules-based systems is that they have difficulty operating outside their knowledge base.

Alternatives to rules-based synthesis prediction have also been proposed, for example, so-called sequence-to-sequence approaches, which are based on the relationships between organic chemistry and linguistics. By casting molecules as text strings, these relationships have been applied in several chemical-design studies[35,36]. In sequence-to-sequence approaches, a model is fed an input of products and then outputs reactants as a SMILES string[37]. A similar approach has also been applied to retrosynthesis[38]. Future developments in areas such as one-shot learning (as recently applied to drug discovery)[39] could lead to wider application of non-rules-based methods in fields such as natural product synthesis, for which training data are scarce.

Beyond the synthesis of a target molecule, machine-learning models can be applied to assess the likelihood that a product will crystallize. By applying feature-selection techniques, a two-parameter model capable of predicting the propensity of a given molecule to crystallize

with an accuracy of around 80% has been demonstrated[40]. Crucially, this model had access to a training set of more than 20,000 crystalline and non-crystalline compounds. The availability of such open-access databases is pivotal for the further development of similar predictive models[41]. In another study, a model was trained to predict the reaction conditions for new organically templated inorganic-product formation with a success rate of 89%[42].

A less explored avenue of machine learning is how to best sample the set of possible experimental set-ups. Active learning predicts the optimal future experiments that are required to better understand a given problem. It was recently applied to help to understand the conditions for the synthesis and crystallization of complex polyoxometalate clusters[43]. Starting from initial data on failed and successful experiments, the machine-learning approach directed future experiments and was shown to be capable of covering six times as much crystallization space as a human researcher in the same number of experiments.

Computational assistance for the planning and direction of chemical synthesis has come a long way since the early days of hand-coded expert systems. Much of this progress has been achieved in the past five years. Incorporation of artificial-intelligence-based chemical planners, with advances in robotic synthesis[43], promises a rich new frontier in the production of novel compounds.

Assisting multi-dimensional characterization

The structure of molecules and materials is typically deduced by a combination of experimental methods, such as X-ray and neutron diffraction, magnetic and spin resonance, and vibrational spectroscopy. Each approach has a certain sensitivity and length scale, and information from each method is complementary. Unfortunately, it is rare that data are fully assimilated into a coherent description of atomic structure. Analyses of individual streams often result in conflicting descriptions of the same compound[44]. A solution could be to incorporate real-time data into the modelling, with results then returned to the experiment, forming a feedback loop[45]. Machine learning represents a unifying framework that could enable the synergy of synthesis, imaging, theory and simulations.

The power of machine-learning methods for enhancing the link between modelling and experiment has been demonstrated in the field of surface science. Complex surface reconstructions have been characterized by combining ab initio simulations with multi-stage pattern-recognition systems that use convolutional neural networks[46]. Machine-learning methods have also recently shown promise in areas such as microstructural characterization[47] and the identification of interesting regions in large, complex, neutron-scattering, volumetric (three-dimensional) datasets[48]. Another example of machine learning opening new avenues in an area of complicated characterization is phase transitions of highly correlated systems; neural networks have been trained to encode topological phases of matter and thus identify transitions between them[49].

Enhancing theoretical chemistry

Modelling is now commonly considered as being equally important as synthesis and characterization for successful programmes of research. Using atomistic simulations, the properties of a molecule or material can, in principle, be calculated for any chemical composition and atomic structure. In practice, the computations grow rapidly in complexity as the size of the system increases, so considerable effort is devoted to finding short-cuts and approximations that enable the properties of the material to be calculated to an acceptable degree of fidelity, without the need for unreasonable amounts of computer time.

Approaches based on DFT have been successful in predicting the properties of many classes of compounds, offering generally high accuracy at reasonable cost. However, DFT and related electronic-structure techniques are limited by the exchange-correlation functional that describes non-classical interactions between electrons. There are notable limitations of the current approximations for weak chemical interactions (such as in layered materials), for highly correlated (*d* and *f* electron) systems and for the latest generation of quantum materials (such as iron pnictide superconductors), which often require a more sophisticated many-body Hamiltonian. Drawing from the growing number of structure–property databases (Table 3), accurate universal density functionals can be learned from data[50,51]. Early examples include the Bayesian error-estimation functional[52] and combinatorially optimized DFT functionals[53]. Going beyond the standard approach to DFT, the need to solve the Kohn–Sham equations can be by-passed by learning density-to-energy and density-to-potential maps directly from training systems[54].

Equally challenging is the description of chemical processes across length scales and timescales, such as the corrosion of metals in the presence of oxygen and water. A realistic description of chemical interactions (bond forming and breaking) including solvents, interfaces and disorder is still limited by the computational cost of available quantum-mechanical approaches. The task of developing transferrable analytic force fields is a well-defined problem for machine learning[55,56]. It has been demonstrated that, in simple materials, approximate potential-energy surfaces learned from quantum-mechanical data can save orders of magnitude in processing cost[57,58]. Although the combination of methods with varying levels of approximation is promising, much work is needed in the quantification and minimization of error propagation across methods. In this context, initiatives for error estimation such as the DAKOTA package (https://dakota.sandia.gov) are critically important.

Targeting discovery of new compounds

We have considered how machine learning can be used to enhance and integrate synthesis, characterization and modelling. However, machine learning can also reveal new ways of discovering compounds. Models that relate system descriptors to desirable properties are already used to reveal previously unknown structure–property relationships[59,60]. So far, the fields of molecular (primarily pharmaceutical and medicinal) and materials chemistry have experienced different degrees of uptake of machine-learning approaches to the design of new compounds, in part owing to the challenges of representing the crystal structure and morphology of extended solids.

Crystalline solids. The application of machine learning to the discovery of functional materials is an emerging field. An early report in 1998 applied machine learning to the prediction of magnetic and optoelectronic materials[61], but the number of studies has risen substantially only since 2010[62–64]. The complexity of games like Go is reminiscent of certain problems in materials science[65,66], such as the description of on-lattice interactions that govern chemical disorder, magnetism and ferroelectricity. Even for representations of small unit cells, the number of configurations of a disordered crystal can quickly exceed the limitations of conventional approaches. An inverse-design procedure illustrated how such a combinatorial space for an alloy could be harnessed to realize specific electronic structure features[67]. Similar inverse-design approaches have also been applied in molecular chemistry to tailor ground- and excited-state properties[68].

Predicting the likelihood of a composition to adopt a given crystal structure is a good example of a supervised classification problem in machine learning. Two recent examples involve predicting how likely a given composition is to adopt the so-called Heusler and half-Heusler crystal structures. The first predicts the likelihood that a given composition will adopt the Heusler structure and is trained on experimental data[69]. This approach was applied to screen hypothetical compositions and successfully identified 12 new gallide compounds, which were subsequently verified experimentally. In the second, a random forest model was trained on experimental data to learn the probability that a given ABC stoichiometry would adopt the half-Heusler structure[70].

As an alternative to learning from experimental data, calculated properties can be used as a training set for machine learning. Assessing the degree of similarity between electronic band structures has been shown to yield improved photocathodes for dye-sensitized solar cells[71]. A machine-learning model, trained to reproduce energies for the

Table 3 | Publicly accessible structure and property databases for molecules and solids

Name	Description	URL
Computed structures and properties		
AFLOWLIB	Structure and property repository from high-throughput ab initio calculations of inorganic materials	http://aflowlib.org
Computational Materials Repository	Infrastructure to enable collection, storage, retrieval and analysis of data from electronic-structure codes	https://cmr.fysik.dtu.dk
GDB	Databases of hypothetical small organic molecules	http://gdb.unibe.ch/downloads
Harvard Clean Energy Project	Computed properties of candidate organic solar absorber materials	https://cepdb.molecularspace.org
Materials Project	Computed properties of known and hypothetical materials carried out using a standard calculation scheme	https://materialsproject.org
NOMAD	Input and output files from calculations using a wide variety of electronic-structure codes	https://nomad-repository.eu
Open Quantum Materials Database	Computed properties of mostly hypothetical structures carried out using a standard calculation scheme	http://oqmd.org
NREL Materials Database	Computed properties of materials for renewable-energy applications	https://materials.nrel.gov
TEDesignLab	Experimental and computed properties to aid the design of new thermo-electric materials	http://tedesignlab.org
ZINC	Commercially available organic molecules in 2D and 3D formats	https://zinc15.docking.org
Experimental structures and properties		
ChEMBL	Bioactive molecules with drug-like properties	https://www.ebi.ac.uk/chembl
ChemSpider	Royal Society of Chemistry's structure database, featuring calculated and experimental properties from a range of sources	https://chemspider.com
Citrination	Computed and experimental properties of materials	https://citrination.com
Crystallography Open Database	Structures of organic, inorganic, metal–organic compounds and minerals	http://crystallography.net
CSD	Repository for small-molecule organic and metal–organic crystal structures	https://www.ccdc.cam.ac.uk
ICSD	Inorganic Crystal Structure Database	https://icsd.fiz-karlsruhe.de
MatNavi	Multiple databases targeting properties such as superconductivity and thermal conductance	http://mits.nims.go.jp
MatWeb	Datasheets for various engineering materials, including thermoplastics, semi-conductors and fibres	http://matweb.com
NIST Chemistry WebBook	High-accuracy gas-phase thermochemistry and spectroscopic data	https://webbook.nist.gov/chemistry
NIST Materials Data Repository	Repository to upload materials data associated with specific publications	https://materialsdata.nist.gov
PubChem	Biological activities of small molecules	https://pubchem.ncbi.nlm.nih.gov

elpasolite crystal structure (ABC_2D_6), was applied to screen all two million possible combinations of elements that satisfy the formula, revealing chemical trends and identifying 128 new materials[72]. Such models are expected to become a central feature in the next generation of high-throughput virtual screening procedures.

The majority of crystal-solid machine-learning studies so far have concentrated on a particular type of crystal structure. This is because of the difficulty of representing crystalline solids in a format that can be fed easily to a statistical learning procedure. By concentrating on a single structure type, the representation is inherently built into the model. Developing flexible, transferrable representations is one of the important areas of research in machine learning for crystalline solids (see subsection 'Data representation'). As we will see below, the use of machine learning in molecular chemistry is more advanced than in the solid state, to a large extent owing to the greater ease with which molecules can be described in a manner amenable to algorithmic interpretation.

Molecular science. The quantitative structure–activity relationship is now a firmly established tool for drug discovery and molecular design. With the development of massive databases of assayed and virtual molecules[73,74], methods for rapid, reliable, virtual screening of these molecules for pharmacological (or other) activity are required to unlock the potential of such molecules. Models based on quantitative structure–activity relationships can be described as the application of statistical methods to the problem of finding empirical relationships of the type $P_i = k'(D_1, D_2, \ldots, D_n)$, where P_i is the property of interest, k' is a (typically linear) mathematical transformation and D_i are calculated or measured structural properties[75]. Machine learning has a long history in the development of quantitative structure–activity relationships, stretching back over half a century[76].

Molecular science is benefitting from cutting-edge algorithmic developments in machine learning such as generative adversarial networks[77] and reinforcement learning for the computational design of biologically active compounds. In a generative adversarial network, two models are trained simultaneously: a generative model (or generator) captures the distribution of data while a discriminative model (or discriminator) estimates the probability that a sample came from the training set rather than the generator. The training procedure for the generator is to maximize the probability of the discriminator making an error (Fig. 3). Models based on objective-reinforced generative adversarial networks[78] are capable of generating new organic molecules from scratch. Such models can be trained to produce diverse molecules that contain specific chemical features and physical responses, through a reward mechanism that resembles classical conditioning in psychology. Using reinforcement learning, newly generated chemical structures can be biased towards those with the desired physical and biological properties (de novo design).

Reclaiming the literature

A final area for which we consider the recent progress of machine learning (across all disciplines) is tapping into the vast amount of knowledge that already exists. Although the scientific literature provides a wealth of information to researchers, it is increasingly difficult to navigate owing to the proliferation of journals, articles and databases. Text mining has become a popular approach to identifying and extracting

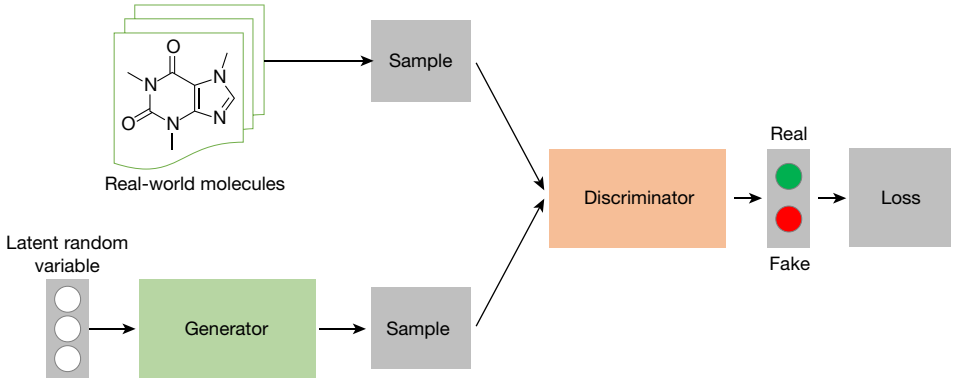

Fig. 3 | The generative adversarial network (GAN) approach to molecular discovery. Two models—a generator and a discriminator—play a continuous 'game'. The generator learns to map from a latent random variable (noise) to a particular distribution of molecules. The discriminator learns to get better and better at distinguishing fake data from real data. The two artificial neural networks are optimizing a different and opposing objective function, or loss function, in a zero-sum game. The general mathematical formulation for the GAN approach to unsupervised machine learning is outlined in ref. [77].

information from unstructured text sources. This approach can be used to extract facts and relationships in a structured form to create specialized databases, to transfer knowledge between domains and, more generally, to support research decision-making[79]. Text mining is applied to answer many different research questions, including in the discovery of drug–protein target associations, the analysis of high-throughput experiments and the development of systematic materials databases[80]. Owing to the heterogeneous nature of written resources, the automated extraction of relevant information is far from trivial. To address this, text mining has evolved into a sophisticated and specialized field where text-processing and machine-learning techniques are combined.

In cases where supplementary data are provided with a publication, it is made available in various formats and databases, often without validated or standardized metadata. The issue of data and metadata interoperability is key. Some examples of forward-looking initiatives that are pushing accessible, reusable data in scientific research are The Molecular Sciences Software Institute (http://molssi.org) and the Open Science Monitor (https://ec.europa.eu/research/openscience).

Frontiers in machine learning

There are many opportunities for further breakthroughs in machine learning to provide even greater advances in the automated design and discovery of molecules and materials. Here we highlight some frontiers in the field.

More knowledge from smaller datasets

Machine-learning approaches typically require large amounts of data for learning to be effective. Although this is rarely an issue in fields such as image recognition, in which millions of input datasets are available, in chemistry or materials science we are often limited to hundreds or thousands, if not fewer, high-quality data points. Researchers need to become better at making the data associated with our publications accessible in a computer-readable form. Another promising solution to the problem of limited datasets is meta-learning, whereby knowledge is learned within and across problems[81]. New developments such as neural Turing machines[82] and imitation learning[83] are enabling the realization of this process. A Bayesian framework has recently been reported to achieve human-level performance on one-shot learning problems with limited data[84], which has consequences for molecular and materials science where data are sparse and generally expensive and slow to obtain.

Efficient chemical representations

The standard description of chemical reactions, in terms of composition, structure and properties, has been optimized for human learning. Most machine-learning approaches for chemical reactions or properties use molecular or atomic descriptors to build models, the success of which is determined by the validity and relevance of these descriptors. A good descriptor must be simpler to obtain than the target property and of as low dimensionality as possible[85]. In the context of materials, useful descriptors[86] and approaches for adapting simple existing heuristics for machine learning have been outlined[87]; however, much work remains to develop powerful new descriptions. In the field of molecular reactions, advances such as the use of neural networks to create fingerprints for molecules in reactions are leading to improvements in synthesis prediction[88]. As has been demonstrated by the successful adoption of the concept of molecular fragments[23], the field of crystalline-materials design can learn much from advances in molecular nomenclature and representation.

Quantum learning

Whereas classical computing processes bits that are either 1 or 0, quantum computers use the quantum superposition of states to process qubits that are both 1 and 0 at the same time[89]. This parallelization leads to an exponential speedup in computational efficiency as the number of (qu)bits used increases[90]. Computational chemistry is a strong candidate to benefit from quantum computing because solving Schrödinger's equation on a quantum computer that consists of interacting electrons has a natural fit[91]. One of the challenges for quantum computing is knowing how to detect and correct errors that may occur in the data. Despite substantial efforts in industry and academia, no error-corrected qubits have been built so far. Quantum machine learning explores the application of machine-learning approaches to quantum problems, and vice versa—the application of quantum computing to machine-learning problems. The possibility of exponential speedups in optimization problems means that quantum machine learning has enormous potential. In problems such as optimizing synthetic routes[92] or improving a given metric (for example, optical absorption for solar energy materials) where multiple acceptable solutions exist, loss of qubit fidelity is less serious than when certainty is required. The physical sciences could prove a particularly rich field for quantum-learning applications[93,94].

Establishing new principles

Automatic discovery of scientific laws and principles[95,96] by inspection of the weights of trained machine-learning systems is a potentially transformational development in science. Although models developed from machine learning are predictive, they are not necessarily (or even usually) interpretable; there are several reasons for this. First, the way in which a machine-learning model represents knowledge rarely maps directly to forms that scientists are familiar with. Given suitable data, an artificial neural network might discover the ideal gas law ($pV = nRT$), but the translation of connection weights into a formula, typically through statistical learning, is non-trivial, even for a simple law such as this. The second reason is more subtle: the laws that underlie the

behaviour of a compound might depend on knowledge that scientists do not yet possess, such as a many-body interaction giving rise to a new type of superconductivity. If an advanced machine-learning system was able to learn key aspects of quantum mechanics, it is hard to envisage how its connection weights could be turned into a comprehensible theory if the scientist lacked understanding of a fundamental component of it. Finally, there may be scientific laws that are so complex that, were they to be discovered by a machine-learning system, they would be too challenging for even a knowledgeable scientist to understand. A machine-learning system that could discern and use such laws would truly be a computational black box.

As scientists embrace the inclusion of machine learning with statistically driven design in their research programmes, the number of reported applications is growing at an extraordinary rate. This new generation of computational science, supported by a platform of open-source tools and data sharing, has the potential to revolutionize molecular and materials discovery.

Received: 20 October 2017; Accepted: 9 May 2018;
Published online 25 July 2018.

1. Dirac, P. A. M. Quantum mechanics of many-electron systems. *Proc. R. Soc. Lond. A* **123**, 714–733 (1929).
2. Pople, J. A. Quantum chemical models (Nobel lecture). *Angew. Chem. Int. Ed.* **38**, 1894–1902 (1999).
3. Boyd, D. B. Quantum chemistry program exchange, facilitator of theoretical and computational chemistry in pre-internet history. *ACS Symp. Ser.* **1122**, 221–273 (2013).
4. Arita, M., Bowler, D. R. & Miyazaki, T. Stable and efficient linear scaling first-principles molecular dynamics for 10000+ atoms. *J. Chem. Theory Comput.* **10**, 5419–5425 (2014).
5. Wilkinson, K. A., Hine, N. D. M. & Skylaris, C.-K. Hybrid mpi-openmp parallelism in the Onetep linear-scaling electronic structure code: application to the delamination of cellulose nanofibrils. *J. Chem. Theory Comput.* **10**, 4782–4794 (2014).
6. Havu, V., Blum, V., Havu, P. & Scheffler, M. Efficient *O*(*N*) integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **228**, 8367–8379 (2009).
7. Catlow, C. R. A., Sokol, A. A. & Walsh, A. *Computational Approaches to Energy Materials* (Wiley-Blackwell, New York, 2013).
8. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
9. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
10. Lejaeghere, K. et al. Reproducibility in density functional theory calculations of solids. *Science* **351**, aad3000 (2016).
11. 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).
12. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
13. Calderon, C. E. et al. The AFLOW standard for high-throughput materials science calculations. *Comput. Mater. Sci.* **108**, 233–238 (2015).
14. Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: realization of the 'fourth paradigm' of science in materials science. *APL Mater.* **4**, 053208 (2016).
15. Schwab, K. The fourth industrial revolution. *Foreign Affairs* https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution (2015).
16. Fourches, D., Muratov, E. & Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **50**, 1189–1204 (2010).
17. Kireeva, N. et al. Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison. *Mol. Inform.* **31**, 301–312 (2012).
18. 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).
19. 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).
20. Bonchev, D. & Rouvray, D. H. *Chemical Graph Theory: Introduction and Fundamentals* (Abacus Press, New York, 1991).
21. Schütt, K. T. et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014). **A radial-distribution-function description of periodic solids is adapted for machine-learning models and applied to predict the electronic density of states for a range of materials**.
22. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
23. Isayev, O. et al. Universal fragment descriptors for predicting electronic properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
24. Hand, D. J. & Yu, K. Idiot's Bayes—not so stupid after all? *Int. Stat. Rev.* **69**, 385–398 (2001).
25. Shakhnarovich, G., Darrell, T. & Indyk, P. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice* (MIT Press, Boston, 2005).
26. Rokach, L. & Maimon, O. in *Data Mining and Knowledge Discovery Handbook* (eds Maimon, O. & Rokach, L.) 149–174 (Springer, New York, 2010).
27. Shawe-Taylor, J. & Cristianini, N. *Kernel Methods for Pattern Analysis* (Cambridge Univ. Press, Cambridge, 2004).
28. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
29. Corey, E. J. & Wipke, W. T. Computer-assisted design of complex organic synthesis. *Science* **166**, 178–192 (1969).
30. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018). **A computer-driven retrosynthesis tool was trained on most published reactions in organic chemistry**.
31. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
32. Klucznik, T. et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522–532 (2018).
33. Segler, M. H. S. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. Eur. J.* **23**, 5966–5971 (2017).
34. Cole, J. C. et al. Generation of crystal structures using known crystal structures as analogues. *Acta Crystallogr. B* **72**, 530–541 (2016).
35. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016). **This study uses machine learning to guide all stages of a materials discovery workflow from quantum-chemical calculations to materials synthesis**.
36. Jastrzębski, S., Leśniak, D. & Czarnecki, W. M. Learning to SMILE(S). Preprint at https://arxiv.org/abs/1602.06289 (2016).
37. Nam, J. & Kim, J. Linking the neural machine translation and the prediction of organic chemistry reactions. Preprint at https://arxiv.org/abs/1612.09529 (2016).
38. Liu, B. et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
39. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* **3**, 283–293 (2017).
40. Wicker, J. G. P. & Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm* **17**, 1927–1934 (2015). **This paper presents a crystal engineering application of machine learning to assess the probability of a given molecule forming a high-quality crystal**.
41. Pillong, M. et al. A publicly available crystallisation data set and its application in machine learning. *CrystEngComm* **19**, 3737–3745 (2017).
42. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016). **The study trains a machine-learning model to predict the success of a chemical reaction, incorporating the results of unsuccessful attempts as well as known (successful) reactions**.
43. Dragone, V., Sans, V., Henson, A. B., Granda, J. M. & Cronin, L. An autonomous organic reaction search engine for chemical reactivity. *Nat. Commun.* **8**, 15733 (2017).
44. Billinge, S. J. L. & Levin, I. The problem with determining atomic structure at the nanoscale. *Science* **316**, 561–565 (2007).
45. Kalinin, S. V., Sumpter, B. G. & Archibald, R. K. Big–deep–smart data in imaging for guiding materials design. *Nat. Mater.* **14**, 973–980 (2015).
46. Ziatdinov, M., Maksov, A. & Kalinin, S. V. Learning surface molecular structures via machine vision. *npj Comput. Mater.* **3**, 31 (2017).
47. de Albuquerque, V. H. C., Cortez, P. C., de Alexandria, A. R. & Tavares, J. M. R. S. A new solution for automatic microstructures analysis from images based on a backpropagation artificial neural network. *Nondestruct. Test. Eval.* **23**, 273–283 (2008).
48. Hui, Y. & Liu, Y. Volumetric data exploration with machine learning-aided visualization in neutron science. Preprint at https://arxiv.org/abs/1710.05994 (2017).
49. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431–434 (2017).
50. Christensen, R., Hansen, H. A. & Vegge, T. Identifying systematic DFT errors in catalytic reactions. *Catal. Sci. Technol.* **5**, 4946–4949 (2015).
51. Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R. & Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).
52. Wellendorff, J. et al. Density functionals for surface science: exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **85**, 235149 (2012).
53. Mardirossian, N. & Head-Gordon, M. ωB97M-V a combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **144**, 214110 (2016).
54. Brockherde, F. et al. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **8**, 872 (2017). **This study transcends the standard approach to DFT by providing a direct mapping from density to energy, paving the way for higher-accuracy approaches**.
55. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed.* **56**, 12828–12840 (2017).

56. 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).
57. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
In this study, machine learning is used to fit interatomic potentials that reproduce the total energy and energy derivatives from quantum-mechanical calculations and enable accurate low-cost simulations.
58. Handley, C. M. & Popelier, P. L. A. Potential energy surfaces fitted by artificial neural networks. *J. Phys. Chem. A* **114**, 3371–3383 (2010).
59. Pulido, A. et al. Functional materials discovery using energy–structure–function maps. *Nature* **543**, 657–664 (2017).
60. Hill, J. et al. Materials science with large-scale data and informatics: unlocking new opportunities. *MRS Bull.* **41**, 399–409 (2016).
61. Kiselyova, N. N., Gladun, V. P. & Vashchenko, N. D. Computational materials design using artificial intelligence methods. *J. Alloys Compd.* **279**, 8–13 (1998).
62. 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).
63. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
64. 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).
In an early example of harnessing materials databases, information on known compounds is used to construct a machine-learning model to predict the viability of previously unreported chemistries.
65. Walsh, A. The quest for new functionality. *Nat. Chem.* **7**, 274–275 (2015).
66. Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. *Chem* **1**, 617–627 (2016).
67. Franceschetti, A. & Zunger, A. The inverse band-structure problem of finding an atomic configuration with given electronic properties. *Nature* **402**, 60–63 (1999).
68. Kuhn, C. & Beratan, D. N. Inverse strategies for molecular design. *J. Phys. Chem.* **100**, 10595–10599 (1996).
69. Oliynyk, A. O. et al. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
70. Legrain, F., Carrete, J., van Roekeghem, A., Madsen, G. K. H. & Mingo, N. Materials screening for the discovery of new half-heuslers: machine learning versus ab initio methods. *J. Phys. Chem. B* **122**, 625–632 (2018).
71. Moot, T. et al. Material informatics driven design and experimental validation of lead titanate as an aqueous solar photocathode. *Mater. Discov.* **6**, 9–16 (2016).
72. 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).
73. Oprea, T. I. & Tropsha, A. Target, chemical and bioactivity databases – integration is key. *Drug Discov. Today. Technol.* **3**, 357–365 (2006).
74. Sterling, T. & Irwin, J. J. ZINC 15 – ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
75. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* **29**, 476–488 (2010).
76. Hansch, C. & Fujita, T. p-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **86**, 1616–1626 (1964).
77. Goodfellow, I. J. et al. Generative adversarial networks. Preprint at https://arxiv.org/abs/1406.2661 (2014).
78. Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C. & Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. Preprint at https://arxiv.org/abs/1705.10843 (2017).
79. Fleuren, W. W. M. & Alkema, W. Application of text mining in the biomedical domain. *Methods* **74**, 97–106 (2015).
80. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
81. Jankowski, N., Duch, W. & Grąbczewski, K. (eds) *Meta-Learning in Computational Intelligence* (Springer, Berlin, 2011).
82. Graves, A., Wayne, G. & Danihelka, I. Neural Turing machines. Preprint at https://arxiv.org/abs/1410.5401 (2014).
83. Duan, Y. et al. One-shot imitation learning. Preprint at https://arxiv.org/abs/1703.07326 (2017).
84. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
85. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
86. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
87. Seko, A., Togo, A. & Tanaka, I. in *Nanoinformatics* (ed. Tanaka, I.) 3–23 (Springer, Singapore, 2018).
88. Duvenaud, D. et al. Convolutional networks on graphs for learning molecular fingerprints. Preprint at https://arxiv.org/abs/1509.09292 (2015).
89. Steane, A. Quantum computing. *Rep. Prog. Phys.* **61**, 117 (1998).
90. Harrow, A. W., Hassidim, A. & Lloyd, S. Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.* **103**, 150502 (2009).
91. Aspuru-Guzik, A., Dutoi, A. D., Love, P. J. & Head-Gordon, M. Simulated quantum computation of molecular energies. *Science* **309**, 1704–1707 (2005).
In an early application of quantum computing to molecular problems, a quantum algorithm that scales linearly with the number of basis functions is demonstrated for calculating properties of chemical interest.
92. Reiher, M., Wiebe, N., Svore, K. M., Wecker, D. & Troyer, M. Elucidating reaction mechanisms on quantum computers. *Proc. Natl Acad. Sci. USA* **114**, 7555–7560 (2017).
93. Dunjko, V., Taylor, J. M. & Briegel, H. J. Quantum-enhanced machine learning. *Phys. Rev. Lett.* **117**, 130501 (2016).
94. Biamonte, J. et al. Quantum machine learning. *Nature* **549**, 195–202 (2017).
95. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
96. Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614 (2017).
97. Domingos, P. *The Master Algorithm* (Basic Books, New York, 2015).
98. Coudert, F.-X. Reproducible research in computational chemistry of materials. *Chem. Mater.* **29**, 2615–2617 (2017).
99. Tetko, I. V., Maran, U. & Tropsha, A. Public (Q)SAR services, integrated modeling environments, and model repositories on the web: state of the art and perspectives for future development. *Mol. Inform.* **36**, 1600082 (2017).

Acknowledgements This work was supported by the EPSRC (grant numbers EP/M009580/1, EP/K016288/1 and EP/L016354/1), the Royal Society and the Leverhulme Trust. O.I. acknowledges support from DOD-ONR (N00014-16-1-2311) and an Eshelman Institute for Innovation award.

Reviewer information *Nature* thanks F.-X. Coudert, M. Waller and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors contributed equally to the design, writing and editing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to O.I. or A.W.