Machine learning for molecular and materials science

Keith T. Butler, Daniel W. Davies, Hugh Cartwright,
Olexandr Isayev & Aron Walsh

Published version information

Citation: KT Butler et al. "Machine learning for molecular and materials science." Nature, vol. 559, no. 7715 (2018): 547-555.

DOI: 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]

1. ISIS Facility, Rutherford Appleton Laboratory, Harwell Campus OX11 0QX, UK
2. Department of Chemistry, University of Bath, Bath BA2 7AY, UK
3. Department of Chemistry, Oxford University, Oxford OX1 3QZ, UK
4. UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
5. Department of Materials Science and Engineering, Yonsei University, Seoul 03722, Korea
6. Department of Materials, Imperial College London, London SW7 2AZ, UK

In this Perspective, we summarise recent progress in machine learning for the chemical sciences. Machine learning techniques suitable for addressing research questions in this domain are outlined, as well as future directions. We envisage a future where the design, synthesis, characterisation, and application of molecules and materials is accelerated by artificial intelligence.

A powerful structure-property relationship for molecules and materials is provided by the Schrödinger equation. 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, realising the importance of rapidly developing computer technologies, created a program called *Gaussian 70* that could do what scientists call *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 be described today 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 21st Century, with activity in applications ranging from developing catalysts for greenhouse gas conversion, discovering materials for energy harvesting and storage, to computer-assisted drug design[7]. The modern chemical simulation toolkit allows the properties of a compound to be anticipated (with reasonable accuracy) even 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] is now a mature technique for calculating the structure and behaviour of solids,[10] which has enabled the development of extensive databases covering 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 (AI) methods has the potential to significantly alter, and enhance, the role of computers in science and engineering. The combination of big data and AI 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 AI that has evolved rapidly in recent years is machine learning (ML). At the heart of ML applications lie statistical algorithms whose performance, much like that of a novice chemical researcher, improves with experience. There is a growing infrastructure of machine learning tools for generating, testing, and refining scientific models. Such techniques are suitable for addressing complex problems involving massive combinatorial spaces or nonlinear processes, which conventional procedures either cannot solve or can only tackle at great computational cost.

As the machinery for AI and ML matures, significant advances are being made not only by those in mainstream AI research, but also by experts in other fields (domain experts) who have the vision and the drive to adopt these approaches for their purposes. As we detail in the *Learning to learn* box, the resources and tools that facilitate the application of ML techniques by non-computer scientists mean that the barrier to entry is lower than ever.

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

1. Nuts and Bolts of Machine Learning

Given enough data, could a computer determine all known physical laws (and potentially also those that are currently unknown) without human input? Yes, given a rule-discovery algorithm. In traditional computational approaches, the computer is little more than a calculator, employing an algorithm provided by a human expert. By contrast, ML approaches learn the rules that underlie a dataset through assessment of a portion of that data. We consider in turn the basic steps involved in the construction of a model, as illustrated in Figure 1; this constitutes a blueprint of the generic workflow required for successful application of ML in a materials discovery process.

Data collection

Machine learning comprises models that learn from existing (training) data. Data may require initial pre-processing, during which missing or spurious elements are identified and handled. For example, the inorganic crystal structure database (ICSD) currently contains 188,000 entries, which have been checked for technical mistakes, but are still subject to human and measurement errors. Identifying and removing such errors is essential if ML algorithms are not to be misled by their presence. There is a growing public concern about the lack of reproducibility and error propagation of experimental data published in peer-reviewed scientific literature. In certain fields like cheminformatics, best practices and guidelines are established to address these problems.[16]

The training of an ML model may be *supervised, semi-supervised* or *unsupervised*, depending upon 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 data set 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 values.

Supervised learning is the most mature and powerful of these approaches, and is used in most ML studies in the physical sciences, for example, 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 unrecognised 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 time-domain but for interpretation it is converted to a frequency-domain with the Fourier transform. Just like scientists, a ML algorithm may learn more effectively using one format rather than the other. The process of converting raw data into something more suitable for an algorithm is called *featurisation* 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 may require insight into both the underlying scientific problem and the operation of the learning algorithm, since 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. For example, the Coulomb matrix[19] contains information on atomic nuclear repulsion, as well as the potential energy of free atoms; the matrix is invariant to molecular translations and rotations. Molecular systems also lend themselves to description as graphs.[20] In the solid-state, the conventional description of crystal structures by translation vectors and fractional coordinates of the atoms is not appropriate for ML, since 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] and property-labelled materials fragments[23] are amongst the new ways in which this problem is being tackled.

Choice of learner

When the data set has been collected and represented appropriately, it is time to choose a model to represent 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 (e.g. the categorisation of a material as a metal or an insulator) or a continuous set (e.g. polarisability). Building a model for the former requires *classification*, while the latter requires *regression*. A range of different learning algorithms can be applied (see Figure 2), 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, ("bagging" or "stacking") to create a more robust overall model.

Common algorithms (learners) include:

Naïve Bayes[25] is a collection of classification algorithms based on Bayes' theorem that identify the most probable hypothesis, given the data as our prior knowledge about the problem. Bayes' theorem provides a formal way to calculate 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 one can select the hypothesis (or model) with the highest probability of correctly representing the data.

In *nearest neighbour (k-NN)*[26] methods the distances between samples and training data in a descriptor hyperspace are calculated. *k*-NN methods are so-called because the output value for a prediction relies on the values of the *k* nearest neighbours, where *k* is an integer. *k*-NN 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, while in a regressor the value is the average of the *k* nearest points.

Decision trees[27] are flowchart-like diagrams used to determine a course of action or outcomes. 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 the tree. Both root and leaf nodes contain questions or criteria to be answered. Branches are arrows connecting nodes, showing the flow from question to answer. Decision trees are often used in ensemble methods (meta-algorithms) that combine multiple trees into one predictive model in order to improve performance.

Kernel methods are a class of algorithms; whose best known members are the *support vector machine* (SVM) and *kernel ridge regression* (KRR).[28] The name "kernel" comes from use of the kernel function, a "trick" that transforms input data into a high-dimensional representation, where the problem is easier to solve. In a sense, a kernel is a similarity function provided by the domain expert. It takes two inputs and, from them, creates an output that quantifies how similar they are.

Artificial neural networks (ANNs) and deep neural networks (DNNs)[29] 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 network's stored knowledge. 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 may substantially improve or impair learning, and the selection of optimal values is often problematic. Consequently, the development of automatic optimisation algorithms is an area of active investigation, as is their incorporation into accessible packages for non-expert users (see Table 1).

Model optimisation

When the learner (or set of learners) has been chosen and predictions are being made, a trial model must be evaluated to allow for optimisation 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.

Total Error = **Bias** + **Variance** + **Irreducible Errors**

Bias is the error from incorrect assumptions in the algorithm and can result in the model missing underlying relationships. Variance on the other hand is sensitivity to small fluctuations in the training set. Even well-trained ML models may contain errors arising from 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. 3.

High bias (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 (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, whilst 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 to determine the quality of a model is to withhold a randomly-selected portion of data during training. This withheld data set, known as a *test set*, is shown to the model once training is complete (Figure 3). 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 to evaluate the transferability and applicability of a model are required in such cases.

2. 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 ML is helping to progress, and reduce the barriers between, the areas of chemical/materials design, synthesis, characterisation and modelling. We finally describe some of the important developments in the field of AI for data-mining existing literature.

Guiding chemical synthesis

Organic chemists were amongst the first scientists to recognise the potential of computational methods in laboratory practice. E.J. Corey's OCSS program,[33] developed

more than 50 years ago, was an attempt to automate retrosynthetic analysis. In a synthetic chemistry route, the number of possible transformations per step can range from around 80 to several thousand,[34] which compares to the order of tens of potential moves at each game position in chess.[35] In chemical synthesis, human experts are required to specify conditional and contextual rules, which exclude large sets of potential reagents at a given step, thus limiting the number of choices available to the algorithm. The contextual rules (typically many thousands of them) are of the 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]

The combination of extremely complex systems and huge numbers of potential solutions, arising from competing objective functions (cost, purity, time, toxicity etc.) 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 most commonly rely on many-layered ANNs or a combination of ANNs with 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 candidates,[36] or rank the likelihood of a predicted product by applying the rules.[37] One ANN that learned from chemical literature examples was able to achieve a level of sophistication such that trained chemists could not distinguish between computer and human expert designed routes.[34] 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, based on so-called 'sequence-to-sequence' approaches, rooted in the relationships between organic chemistry and linguistics. By casting molecules as text strings, these relationships have been applied in several chemical design studies.[38,39] In sequence-to-sequence approaches a model is fed an input of products and then outputs reactants as a SMILES string.[40] A similar approach has also been applied to retrosynthesis.[41] Future developments in areas such as one-shot learning (as recently applied to drug discovery)[42] could lead to wider application in fields like natural product synthesis, where training data are scarce.

Beyond the synthesis of a target molecule, ML models can been applied to assess the likelihood that a product will crystallise. By applying feature selection techniques, Wicker and Cooper developed a two-parameter model, capable of predicting the propensity of a given molecule to crystallise with an accuracy of ~ 80%.[43] 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.[44] Another study trained a model to predict the reaction conditions for new organically templated inorganic product formation with a success rate of 89%.[45]

A less explored avenue of ML is how to best sample the set of possible experimental set-ups. *Active learning* predicts the optimal future experiments required to better understand a given problem. It was recently applied to understand the conditions for the synthesis and crystallisation of complex polyoxometalate clusters.[46] Starting from initial data on failed and successful experiments, the ML approach then directed future experiments and was shown

to be capable of covering six times as much crystallisation 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 AI-based chemical planners, with great advances in robotic synthesis[46] promises a rich new frontier in the production of new compounds.

Assisting multi-dimensional characterisation

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.[47] A solution would be to incorporate real-time data into the modelling with results that are then returned to the experiment, forming a feedback loop.[48] ML offers the promise of a unifying framework allowing synergy of synthesis, imaging, theory and simulations.

The power of ML methods for enhancing the link between modelling and experiment has been demonstrated in the field of surface science. Combining *ab initio* simulations with multi-stage pattern recognition systems that use convolutional neural networks Ziatdinov and co-workers were able to characterize complex surface reconstructions.[49] ML methods have also shown recent promise in areas such as microstructural characterisation[50] and the identification of interesting regions in large complex neutron scattering 3D volumetric datasets.[51] A different example of ML opening new avenues in an area of complicated characterisation is phase transitions of highly-correlated systems; neural networks have been trained to encode phases of matter and thus identify transitions.[52]

Enhancing theoretical chemistry

Modelling is now commonly considered as an equally important component to synthesis and characterisation 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 rapidly grow in complexity as the size of the system increases, so considerable effort is devoted to finding short-cuts and approximations that might allow one to calculate properties to an acceptable degree of fidelity, without the need for unreasonable amounts of computer time.

Approaches based on DFT have been successful in predicting properties of many classes of compounds, offering generally high accuracy at reasonable cost. However, the Achilles heel of DFT remains the exchange-correlation functional that describes non-classical interactions between electrons. There are notable limitations of current approximations for weak chemical interactions (e.g. layered materials), highly correlated (d and f electron) systems, and the latest generation of quantum materials (e.g. iron pnictide superconductors), which often require a more expensive many-body Hamiltonian. Drawing from the growing number of structure-property databases (Table 2), accurate universal density functionals can be learned from data.[53,54] Early examples include the Bayesian error estimation functional

(BEEF)[55] as well as combinatorially-optimised DFT functionals.[56] Going beyond the standard approach to DFT, the need to solve the Kohn-Sham equations is by-passed by learning density-to-energy and density-to-potential maps directly from training systems.[57]

Equally challenging is the description of chemical processes across length and time scales, for example, the ubiquitous corrosion of metals in the presence of oxygen and water. The description of realistic chemical interactions (bond forming and breaking) including solvents, interfaces, and disorder is still limited by the computational cost of quantum mechanical approaches. The task of developing transferrable analytic forcefields is a well-defined problem for machine learning.[58,59] 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.[60,61] Whilst the combination of methods with varying levels of approximation is promising, much work is needed in the quantification and minimisation of error propagation across methods. In this context, initiatives for error estimation such as the DAKOTA package[62] are critically important.

Targeting discovery of new compounds

Until now we have considered how ML can be used to enhance and integrate the areas of synthesis, characterisation and modelling. However, ML can be used to reveal new ways to discover compounds. Models that relate system descriptors to desirable properties are already used to reveal structure-property relationships.[63,64] So far, the fields of molecular (primarily pharmaceutical/medicinal) and materials chemistry have experienced different degrees of uptake of ML approaches to the design of new compounds, in part due to the challenges of representing the crystal structure and morphology of extended solids.

Crystalline solids

The application of ML to the discovery of functional materials is an emerging field. An early report in 1998 applied ML to the prediction of new magnetic and optoelectronic materials,[65] but the number of studies has only risen significantly since 2010.[66–68] The complexity of games like "Go" is reminiscent of certain problems in materials science,[69,70] for example the description of on-lattice interactions that govern chemical disorder, magnetism, and ferroelectricity. Even for small unit cell representations, 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 realise specific electronic structure features.[71] Similar inverse design approaches have also been applied in molecular chemistry to tailor ground and excited state properties.[72]

Prediction of the likelihood of a composition to adopt a given crystal structure is a good example of a supervised classification problem in ML. Some recent examples involve the prediction of how likely a given composition is to adopt the so-called Heusler and half-Heusler crystal structures. One method predicts the likelihood a given composition will adopt the Heusler structure and is trained on experimental data.[73] This approach was applied to screen hypothetical compositions and successfully identified 12 new gallide compounds, which were subsequently experimentally verified. Similarly, a random forest model was trained on experimental data to learn the probability that a given ABC stoichiometry would adopt the half-Heusler structure.[74]

As an alternative to learning from experimental data, calculated properties can be used as a training set for ML. Moot and co-workers showed how assessing the degree of similarity between electronic band structures could yield improved photocathodes for dye-sensitised solar cells.[75] A ML model, trained to reproduce energies for the elpasolite crystal structure (ABC_2D_6), was applied to screen all 2×10^6 possible combinations of elements that satisfy the formula, revealing chemical trends and identifying 128 new materials.[76] Such models are expected to become a central feature in the next generation of high-throughput virtual screening procedures.

It is notable that the majority of crystal solid ML studies to date have concentrated on a particular crystal structure type. This is because of the difficulty of representing crystalline solids in a format which can easily be fed 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 critical areas in ML for crystalline solids (see section 2 subsection "Data representation"). As we will see below, the use of ML in molecular chemistry is more advanced than in the solid state, to a large extent this is due to greater ease with which molecules can be described in a manner amenable to algorithmic interpretation.

Molecular science

The QSAR (Quantitative Structure-Activity Relationship) approach is now a firmly established tool for drug discovery and molecular design. With the development of massive databases of assayed and virtual molecules,[77,78] methods for rapid, reliable virtual screening of these molecules for pharmacological (or other) activity are required to unlock their potential. QSARs 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 (most commonly linear) mathematical transformation and the D_i are calculated or measured structural properties.[79] ML has a long history in the development of QSARs, stretching back over half a century.[80]

Molecular science is benefitting from cutting edge algorithmic developments in ML such as generative adversarial networks (GANs)[81] and reinforcement learning for the computational design of novel putative biologically active compounds. In a GAN, two models are trained simultaneously: a generative model G captures the distribution of data, and a discriminative model D estimates the probability that a sample came from the training set rather than G. The training procedure for G is to maximize the probability of D making an error (Figure 4). The ORGAN (Objective-Reinforced Generative Adversarial Networks)[82] model is capable of generating novel organic molecules from scratch. Such a model 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, one could bias newly generated chemical structures towards those with desired physical and biological properties (*de novo* design).

Reclaiming the literature

A final area for which we consider the recent progress of ML (across all disciplines) is tapping into the vast wealth of knowledge that already exists. While the scientific literature provides a wealth of information to researchers, it is increasingly difficult to navigate due to the proliferation of journals, articles, and databases. Text mining has become a popular approach to identify and extract information from unstructured text sources. This approach can be used to extract facts and relationships in a structured form to create specialised databases, to transfer knowledge between domains, and more generally to support research decision-making.[83] Text mining is applied to answer many different research questions, ranging from the discovery of novel drug–protein target associations, or analysis of high throughput experiments, to developing systematic materials databases.[84] Due 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 specialised field where text processing and machine learning techniques are combined.

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

3. Frontiers in Machine Learning

Many opportunities exist for further breakthroughs in ML to provide even greater advances in the automated design and discovery of molecules and materials. Here we highlight some frontiers in the field.

1. **More knowledge from smaller data sets.** ML approaches typically require large amounts of data for learning to be effective. While this is rarely an issue in fields such as image recognition, in which millions of input data sets are available, in chemistry or materials science. We are often limited to hundreds or thousands, if not fewer, high-quality data points. We researchers need to become better at making the data associated with our publications accessible in computer readable form. Another promising solution to the problem of limited datasets is *meta-learning,* where knowledge is learned *within* and *across* problems.[85] New developments such as neural Turing machines[86] or imitation learning[87] are enabling the realisation of this process. A Bayesian framework has recently been reported to achieve human-level performance on one-shot learning problems with limited data[88], which has consequences for molecular and materials science where data is sparse and generally expensive and slow to obtain.
2. **Efficient chemical representations.** The standard description of chemical reactions, in terms of composition, structure and properties has been optimised 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.[89] In the context of materials, useful descriptors[90] and new approaches for adapting simple existing

heuristics for machine learning have been outlined;[91] however, much work remains to develop powerful new descriptions. In the field of molecular reactions exciting advances, such as the use of neural networks to create fingerprints for molecules in reactions are leading to advances in synthesis prediction.[92] 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. Chemists have a lot to learn from a field of *representation learning* i.e., learning representations of the data that make it easier to extract new information and knowledge.

3. **Quantum learning.** While 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.[93] This parallelisation leads to an exponential speedup in computational efficiency as the number of (qu)bits used increases.[94] Quantum chemistry is a strong candidate to benefit, because solving Schrödinger's equation on a quantum computer has a natural fit.[95] One of the challenges for quantum computing is knowing how to detect and correct errors that may occur in the data. Despite significant efforts in industry and academia, no error-corrected qubits has been built so far. Quantum machine learning explores the application of ML approaches to quantum problems, and *vice versa*, the application of quantum computing to ML problems. The possibility of exponential speedups in optimisation problems means that quantum machine learning has enormous potential. In problems such as optimising synthetic routes[96] or improving a given metric (e.g. 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.[97,98]
4. **Establishing new principles.** Automatic discovery of scientific laws and principles[99-100] by inspection of the weights of trained ML 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 ML model represents knowledge rarely maps directly onto forms that scientists are familiar with. Given suitable data, an ANN might discover the Ideal Gas Law, pV=nRT, but the translation of connection weights to a formula, typically through *statistical learning*, is not trivial, even for a law this simple. A more subtle issue exists: the laws that underlie the behaviour of a material might depend upon knowledge that scientists do not yet possess, e.g. a many-body interaction giving rise to a new type of superconductivity. If an advanced ML 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 which at heart are so complex that, were they to be discovered by a ML system, would be too challenging for even a knowledgeable scientist to understand. A ML 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 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 revolutionise the molecular and materials discovery process.

Acknowledgments This work has been supported by the EPSRC (grant no. 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 Eshelman Institute for Innovation award.

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

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to O.I. (olexandr@olexandrisayev.com) or A.W. (a.walsh@imperial.ac.uk).

NAME	DESCRIPTION	URL
GENERAL PURPOSE MACHINE LEARNING FRAMEWORKS		
CARET	Package for machine learning in R	topepo.github.io/caret
DEEPLEARNING4J	Distributed deep learning for Java	deeplearning4j.org
H2O.AI	Machine learning platform written in Java that can be imported as a Python or R library	h2o.ai
KERAS	High-level neural networks API written in Python	keras.io
MLPACK	Scalable machine learning library written in C++	mlpack.org
SCIKIT-LEARN	Machine learning and data mining member of the 'scikit' family of toolboxes built around the SciPy Python library	scikit-learn.org
STATISTICS AND MACHINE LEARNING TOOLBOX	Machine learning library for MATLAB	mathworks.com/machinelearning
WEKA	Collection of machine learning algorithms and tasks that can be applied directly or from Java code	cs.waikato.ac.nz/ml/weka
MACHINE LEARNING TOOLS FOR MOLECULES AND MATERIALS		
AMP	Package to facilitate machine-learning for atomistic calculations	bitbucket.org/andrewpeterson/amp
ANI	Neural network potentials for organic molecules with python interface	github.com/isayev/ASE_ANI
COMBO	Python library with emphasis on scalability and efficiency	github.com/tsudalab/combo
DEEPCHEM	Python library for deep learning of chemical systems	deepchem.io
GAP	Gaussian Approximation Potentials	libatoms.org/Home/Software
MATMINER	Python library for assisting machine learning in materials science	hackingmaterials.github.io/matminer
NOMAD	Collection of tools to explore correlations in materials datasets	analytics-toolkit.nomad-coe.eu
PROPHET	Code to integrate machine learning techniques with quantum chemistry approaches	github.com/biklooost/PROPhet
TENSORMOL	Neural network chemistry package	github.com/jparkhill/TensorMol

Table 1. A collection of publically-accessible learning resources and tools relating to machine learning.

NAME	DESCRIPTION	URL
	COMPUTED STRUCTURES AND PROPERTIES	
AFLOWLIB	Distributed properties repository from high-throughput *ab initio* calculations of inorganic materials	aflowlib.org
COMPUTATIONAL MATERIALS REPOSITORY	Infrastructure to enable collection, storage, retrieval and analysis of data from electronic structure codes	cmr.fysik.dtu.dk
GDB	Databases of hypothetical small organic molecules	gdb.unibe.ch
HARVARD CLEAN ENERGY PROJECT	Computed properties of candidate organic solar absorber materials	cepdb.molecularspace.org
MATERIALS PROJECT	Computed properties of known and hypothetical materials carried out using a standard calculation scheme	materialsproject.org
NOMAD	Input and output files from calculations using a wide variety of electronic structure codes	nomad-repository.eu
OPEN QUANTUM MATERIALS DATABASE	Computed properties of mostly hypothetical structures carried out using a standard calculation scheme	oqmd.org
NREL MATERIALS DATABASE	Computed properties of materials for renewable energy applications	materials.nrel.gov
TEDESIGNLAB	Experimental and computed properties to aid the design of new thermoelectric materials	tedesignlab.org
ZINC	Commercially available organic molecules in 2D and 3D formats	zinc15.docking.org
	EXPERIMENTAL STRUCTURES AND PROPERTIES	
CHEMBL	Bioactive molecules with drug-like properties	ebi.ac.uk/chembl
CHEMSPIDER	Royal Society of Chemistry's structure database featuring calculated and experimental properties from a range of sources	chemspider.com
CITRINATION	Computed and experimental properties of materials	citrination.com
CRYSTALLOGRAPHY OPEN DATABASE	Structures of organic, inorganic, metal-organic compounds and minerals	crystallography.net
CSD	Repository for small-molecule organic and metal-organic crystal structures	www.ccdc.cam.ac.uk
ICSD	Inorganic Crystal Structure Database	icsd.fiz-karlsruhe.de
MATNAVI	Multiple databases targeting properties such as superconductivity and thermal conductance	mits.nims.go.jp
MATWEB	Datasheets for various engineering materials including thermoplastics, semiconductors and fibres	matweb.com
NIST CHEMISTRY WEBBOOK	High accuracy gas-phase, thermochemistry and spectroscopic data	webbook.nist.gov/chemistry/
NIST MATERIALS DATA REPOSITORY	Repository to upload materials data associated with specific publications	materialsdata.nist.gov
PUBCHEM	Biological activities of small molecules	pubchem.ncbi.nlm.nih.gov

Table 2. A representative collection of publically-accessible structure and property databases for molecules and solids that can be used to feed machine learning approaches.

Figure legends

Figure 1 Illustration of a machine learning workflow applied to interpret real world observations. It consists of four basic steps: (i) *data collection* – acquisition of data from experiment, simulations or other sources; (ii) *data representation* – processing of data to ensure its correctness, integrity and transformation into a form suitable for ML; (iii) *choice of learner* – selection of the types of ML model used to represent the problem; (iv) *model optimisation* – rigorous testing of the resultant model(s) to minimise error and choose the optimal representation.

Figure 2 Classes of machine learning techniques (following Ref. 24) and examples of problems that can be posed to them by a curious scientist. Whilst evolutionary algorithms are often integrated into machine learning procedures, they form part of a wider class of stochastic search algorithms.

Figure 3 Errors that arise in machine learning approaches, both during 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), while a complex model may suffer from high variance (overfitting) leading to a bias-variance trade-off. The model shown here is built on an example from kaggle.com, available at https://keeeto.github.io/blog/bias_variance/.

Figure 4 The Generative Adversarial Networks (GAN)[81] approach to molecular discovery. Two models G (generator) and D (discriminator) play a continuous "game", where the generator is learning to produce more and more realistic samples, which can vary in structure and composition, and the discriminator is learning to get better and better at distinguishing fake data from real data.

Box 1

Learning to Learn

One of the most exciting aspects of machine learning techniques is their promise to democratise 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, ML approaches, if developed and implemented correctly, can broaden routine application of computer models by non-specialists. The accessibility of ML technology relies critically on three factors: *open data, open software* and *open education.* There is an increasing drive to open data within the physical sciences and the best practice has been outlined in recent articles.[30,31] Some of the open software being developed is listed in Table 1. There are also many excellent open education resources, such as massive open online courses (MOOCs) available.

http://www.fast.ai is a course that aims to "make neural nets uncool again"! One of the great advantages of fast.ai is that the novice user starts to build working machine learning models almost immediately. The course, however, is not for absolute beginners, and requires a working knowledge of computer programming and high-school level mathematics.

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 extremely useful interactive environments to develop and test code and is suitable for non-coders, as it teaches the student Python at the same time as ML.

Academic MOOCs are the best locations for those who wish to get more involved with the theory and principles of AI and ML, 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 (Learning from Data) and https://www.udemy.com (Machine Learning A-Z). The underlying mathematics is the topic of a course from Imperial College (https://www.coursera.org/specializations/mathematics-machine-learning).

Data blogs and podcasts. Many ML professionals run informative blogs and podcasts dealing with specific aspects of ML practice. These are useful resources for general interest as well as broadening and deepening knowledge. There are too many to provide an exhaustive list here, but we do recommend https://machinelearningmastery.com and http://lineardigressions.com to get started.

References

1. Dirac, P. A. M. Quantum mechanics of many-electron systems. *Proc. R. Soc. London A Math. Phys. Eng. Sci.* **123,** 714 (1929).
2. Pople, J. A. Quantum chemical models (Nobel Lecture). *Angew. Chemie 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. Walsh, A., Sokol, A. A. & Catlow, C. R. A. *Computational approaches to energy materials*. (Wiley-Blackwell, 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,** 3000 (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,** 11002 (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,** 53208 (2016).
15. Schwab, K. The fourth industrial revolution. *Foreign Affairs* (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,** 58301 (2012).
20. Bonchev, D. & Rouvray, D. H. *Chemical graph theory: introduction and fundamentals*. *Erkenntnis* **68,** (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 ML 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,** 24104 (2017).
23. Isayev, O. *et al.* Universal fragment descriptors for predicting electronic properties of inorganic crystals. *Nat. Commun.* **8,** 15679 (2017).
24. Domingos, P. *The Master Algorithm*. (Basic Books, 2015).
25. Hand, D. J. & Yu, K. Idiot's Bayes: not so stupid after all? *Int. Stat. Rev. / Rev. Int. Stat.* **69,** 385 (2001).
26. Shakhnarovich, G., Darrell, T. & Indyk, P. *Nearest-neighbor methods in learning and vision: theory and practice*. (The MIT Press).
27. Rokach, L. & Maimon, O. Decision trees. in *Data Mining and Knowledge Discovery Handbook* 165–192 (2010).
28. Shawe-Taylor, J. & Cristianini, N. *Kernel methods for pattern analysis*. *Elements* (2004).
29. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Networks* **61,** 85–117 (2015).

30. Coudert, F.-X. Reproducible research in computational chemistry of materials. *Chem. Mater.* **29,** 2615–2617 (2017).
31. 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).
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. Pensak, D. A. & Corey, E. J. Lhasa—logic and heuristics applied to synthetic analysis. in *Computer-Assisted Organic Synthesis* 1–32 (1977).
34. 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.
35. Szymkuć, S. *et al.* Computer-assisted synthetic planning: the end of the beginning. *Angew. Chemie Int. Ed.* **55,** 5904–5937 (2016).
36. Segler, M. H. S. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. - A Eur. J.* **23,** 5966–5971 (2017).
37. Cole, J. C. *et al.* Generation of crystal structures using known crystal structures as analogues. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72,** 530–541 (2016).
38. 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 ML to guide all stages of a materials discovery workflow from quantum chemical calculations to materials synthesis.
39. Jastrzębski, S., Leśniak, D. & Czarnecki, W. M. Learning to smile(s). *ArXiv* 1602.06289 (2016).
40. Nam, J. & Kim, J. Linking the neural machine translation and the prediction of organic chemistry reactions. *ArXiv* 1612.09529 (2016).
41. Liu, B. *et al.* Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3,** 1103–1113 (2017).
42. 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).
43. Wicker, J. G. P. *et al.* Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm* **17,** 1927–1934 (2015).
A crystal engineering application of ML to assess the probability of a given molecule forming a high-quality crystal.
44. Pillong, M. *et al.* A publicly available crystallisation data set and its application in machine learning. *CrystEngComm* (2017).
45. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533,** 73–76 (2016).
The study trains an ML model to predict the success of a chemical reaction, incorporating the results of uncessfull attempts rather than siimply known (sucessful) reactions.
46. 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).
47. Billinge, S. J. L. & Levin, I. The problem with determining atomic structure at the nanoscale. *Science* **316,** 561–565 (2007).
48. 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).
49. Ziatdinov, M., Maksov, A. & Kalinin, S. V. Learning surface molecular structures via machine vision. *npj Comput. Mater.* **3,** 31 (2017).
50. 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).
51. Hui, Y. & Liu, Y. Volumetric data exploration with machine learning-aided visualization in neutron science. *ArXiv* 1710.05994 (2017).
52. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13,** 431–434 (2017).
53. Christensen, R., Hansen, H. A. & Vegge, T. Identifying systematic dft errors in catalytic reactions. *Catal. Sci. Technol.* **5,** 4946–4949 (2015).

54. Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R. & Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **108,** 253002 (2012).
55. Wellendorff, J. *et al.* Density functionals for surface science: exchange-correlation model development with bayesian error estimation. *Phys. Rev. B* **85,** 235149 (2012).
56. 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,** (2016).
57. 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.
58. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chemie Int. Ed.* **56,** 12828–12840 (2017).
59. 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).
60. 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, ML is used to fit interatomic potentials that reproduce the total energy and derivates from quantum mechanical calculations and enable accurate low-cost simulations.
61. Handley, C. M. & Popelier, P. L. A. Potential energy surfaces fitted by artificial neural networks. *J. Phys. Chem. A* **114,** 3371–3383 (2010).
62. Dakota | explore and predict with confidence. Available at: https://dakota.sandia.gov/. (Accessed: 5th April 2018).
63. Pulido, A. *et al.* Functional materials discovery using energy–structure–function maps. *Nature* **543,** 657–664 (2017).
64. Hill, J. *et al.* Materials science with large-scale data and informatics: unlocking new opportunities. *MRS Bull.* **41,** 399–409 (2016).
65. Kiselyova, N. N., Gladun, V. P. & Vashchenko, N. D. Computational materials design using artificial intelligence methods. *J. Alloys Compd.* **279,** 8–13 (1998).
66. 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).
67. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3,** 2810 (2013).
68. Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding natures missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22,** 3762–3767 (2010).
An early example of harnessing materials databases. Information on known compounds is used to construct an ML model to predict the viability of previously unreported chemisties.
69. Walsh, A. The quest for new functionality. *Nat. Chem.* **7,** 274–275 (2015).
70. Davies, D. W. *et al.* Computational screening of all stoichiometric inorganic materials. *Chem* **1,** 617–627 (2016).
71. Franceschetti, A. & Zunger, A. The inverse band-structure problem of finding an atomic configurationwith given electronic properties. *Nature* **402,** 60–63 (1999).
72. Kuhn, C. & Beratan, D.N. Inverse strategies for molecular design. J. Phys. Chem. **100**, 10595-10599 (1996).
73. Oliynyk, A. O. *et al.* High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chem. Mater.* **28,** 7324–7331 (2016).
74. 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).
75. Moot, T. *et al.* Material informatics driven design and experimental validation of lead titanate as an aqueous solar photocathode. *Mater. Discov.* **6,** 9–16 (2017).
76. 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).
77. Oprea, T. I. & Tropsha, A. Target, chemical and bioactivity databases – integration is key. *Drug Discov. Today Technol.* **3,** 357–365 (2006).

78. Sterling, T. & Irwin, J. J. Zinc 15 – ligand discovery for everyone. *J. Chem. Inf. Model.* **55,** 2324–2337 (2015).
79. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* **29,** 476–488 (2010).
80. 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).
81. Goodfellow, I. J. *et al.* Generative adversarial networks. in *Advances in Neural Information Processing Systems 27* (2014).
82. 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. *arXiv* 1705.10843 (2017).
83. Fleuren, W. W. M. & Alkema, W. Application of text mining in the biomedical domain. *Methods* **74,** 97–106 (2015).
84. Kim, E. *et al.* Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29,** 9436–9444 (2017).
85. Jankowski, N., Duch, W. & Grąbczewski, K. *Meta-learning in computational intelligence*. (Springer, 2011).
86. Graves, A., Wayne, G. & Danihelka, I. Neural turing machines. *arXiv* 1410.5401 (2014).
87. Duan, Y. *et al.* One-shot imitation learning. *ArXiv* 1703.07326 (2017).
88. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350,** (2015).
89. 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).
90. Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nat. Mater.* **12,** 191–201 (2013).
91. Seko, A., Togo, A. & Tanaka, I. Descriptors for machine learning of materials data. in *Nanoinformatics* 3–23 (2018).
92. Duvenaud, D. *et al.* Convolutional networks on graphs for learning molecular fingerprints. *Arxiv* 1509.09292 (2015).
93. Steane, A., Quantum computing. *Rep. Prog. Phys.* **61**, 117 (1998)
94. Harrow, A. W., Hassidim, A. & Lloyd, S. Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.* **103,** 150502 (2009).
95. Aspuru-Guzik, A., Dutoi, A. D., Love, P. J., Head-Gordon, M. Simulated quantum computation of molecular energies. *Science* **309,** 1704–1707 (2005).
An early application of quantum computing to molecular problems. A quantum algorithm that scales linearly in the number of basis functions is demonstrated for calculating properties of chemical interest.
96. Reiher, M., Wiebe, N., Svore, K. M., Wecker, D. & Troyer, M. Elucidating reaction mechanisms on quantum computers. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 7555–7560 (2017).
97. Dunjko, V., Taylor, J. M. & Briegel, H. J. Quantum-enhanced machine learning. *Phys. Rev. Lett.* **117,** 130501 (2016).
98. Biamonte, J. *et al.* Quantum machine learning. *Nature* **549,** 195–202 (2017).
99. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324,** 81 LP-85 (2009).
100. Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Sci. Adv.* **3,** e1602614 (2017).