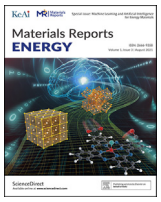
Materials Reports: Energy 1 (2021) 100047

Contents lists available at ScienceDirect

Materials Reports: Energy

journal homepage: www.keaipublishing.com/en/journals/materials-reports-energy



Review

# Computational discovery of energy materials in the era of big data and machine learning: A critical review

Ziheng Lu

*Department of Materials Science & Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge, CB3 0FS, UK*

ARTICLE INFO

*Keywords:*
Machine learning
Material discovery
Crystal structure prediction
Deep learning
Generative model
Inverse material design
High throughput screening
Density functional theory

ABSTRACT

The discovery of novel materials with desired properties is essential to the advancements of energy-related technologies. Despite the rapid development of computational infrastructures and theoretical approaches, progress so far has been limited by the empirical and serial nature of experimental work. Fortunately, the situation is changing thanks to the maturation of theoretical tools such as density functional theory, high-throughput screening, crystal structure prediction, and emerging approaches based on machine learning. Together these recent innovations in computational chemistry, data informatics, and machine learning have acted as catalysts for revolutionizing material design and hopefully will lead to faster kinetics in the development of energy-related industries. In this report, recent advances in material discovery methods are reviewed for energy devices. Three paradigms based on empiricism-driven experiments, database-driven high-throughput screening, and data informatics-driven machine learning are discussed critically. Key methodological advancements involved are reviewed including high-throughput screening, crystal structure prediction, and generative models for target material design. Their applications in energy-related devices such as batteries, catalysts, and photovoltaics are selectively showcased.

## 1. Introduction

Energy is undoubtedly one of the grand challenges to mankind. A survey on the energy section by the International Energy Agency (IEA) forecasts a ~15% increase in global energy demand by 2030.[1,2] Achieving so in a sustainable way is a difficult task but is crucial to the future prosperity and economic development of a modern world. It requires dedicated effort to shift to renewable energy sources and a near-term surge of investment in clean energy technologies.

From a technological perspective, the physiochemical properties of the key materials shape the bottleneck of the advancements in energy-related fields as illustrated in Fig. 1. For example, in the field of photovoltaics, the band structure and the defect tolerance of the absorber material critically define the upper bound of the final efficiency of the cell. When it comes to batteries, the redox potential and the specific capacity of medium ions such as $Li^+$ of electrodes determine the upper bound of the energy density and the competitiveness of such with internal combustion engines. In the field of emerging technologies, the situation is more pressing. For example, solid-state batteries are promising alternatives to current lithium-ion batteries (LIBs) for electrochemical energy storage in terms of safety, energy density, and manufacturing costs.[3,4] However, it is still not to a competitive commercial stand with LIB due to the bottleneck of the electrolyte material.[5] Currently, the ionic mobility in these solid replacements is still too low and the stability at the interfaces also poses significant challenges.[6] Similarly, fuel cells are promising alternatives to internal combustion engines and hold the promise of providing zero carbon emission. However, the slow catalytic kinetics, the hindered mass transport, and limited durability of the electrolyte practically block its use.[7] All of these devices concerning energy conversion and storage are to some extent bottlenecked by one or several of the key materials. Discovery of novel materials with desired properties is thus essential to the performance enhancement of existing technologies and the enabling of emerging ones.

Conventionally, the discovery of novel materials is solely based on

*E-mail addresses:* zl462@cam.ac.uk, zluag@connect.ust.hk.

Production and Hosting by Elsevier on behalf of KeAi

https://doi.org/10.1016/j.matre.2021.100047
Received 8 May 2021; Accepted 24 May 2021
Available online 28 June 2021

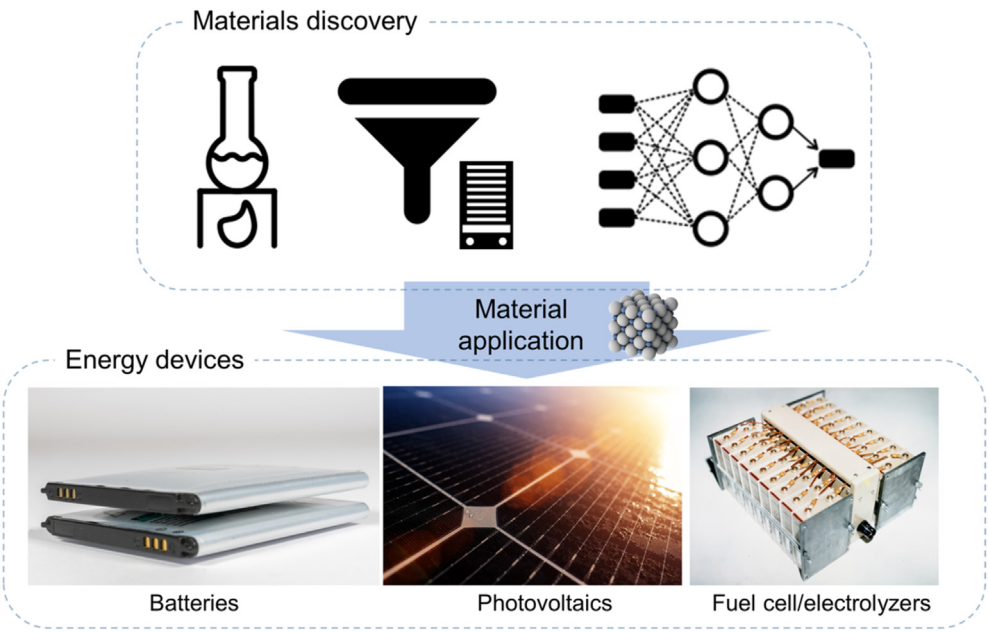



**Fig. 1.** Schematics showing advancement of energy research driven by materials discovery.

experimental trial-and-error. While such empiricism-driven processes have led to important discoveries, they are also characterized by slow kinetics. The situation has fundamentally changed from the late 60s thanks to the development of density functional theory (DFT) which laid the foundation of calculating electronic properties of practical materials from first principles.[8–10] In recent decades, the fast development of computational infrastructure further facilitates advancements in this field.[11] A number of codes have been made available to the community and the supercomputers can provide computing power on the scale of hundreds of petaFLOPS.[12–14] Computational chemists and theoretical physicists can now routinely compute the properties of a material containing up to hundreds of atoms in its unit cell with quantum mechanical-level of details. Achieving such fast computation of material properties has driven the material design based on high-throughput screening.[15] By assembling the structures from known databases and calculating the properties entry by entry using DFT, desired materials can be found.[16] However, such a process is fundamentally constrained because the possible outcomes are bounded by the initial decision concerning its chemical and structural space. A typical example is the hybrid organic-inorganic halide perovskites for photovoltaics.[17,18] Before its discovery by experimentalists, it is completely off the search range by computational chemists. Therefore, it is crucial to extend the chemical and structural spaces for new material discovery. Crystal structure prediction methods stand out in this context.[19,20] By arranging the atoms in different manners, one can explore the potential energy surface (PES) of a specific composition. By locating the local energy minima, one can in principle find the most stable structure and the meta-stable ones. These methods can significantly extend the database approach for materials screening. Nevertheless, the encounter of materials with desired property using such a screening approach is still dependent on a matter of luck. In this context, the recent invasion of data informatics and machine learning is starting to revolutionize the field of materials discovery by providing novel tools to inversely design materials with specific properties. New algorisms are emerging rapidly and the data-informatics infrastructures catering to these new methods are being developed. Hugh amount of data, or big data, including atomic structures, formation energies, and electronic band structures is generated on a day-to-day basis. The validation and correct usage of these data are being intensively studied.[21,22]

In view of the rapid development of the field of materials discovery of energy materials, we assemble this review. In this contribution, we review the recent methodological advancements of materials discovery. The above development of research approaches which constitutes the three critical paradigms of materials discovery will be discussed in Section 2. The high-throughput screening approach, the relevant database development, and the addition of crystal structure prediction methods to extend such an approach will be reviewed in Section 3. In Section 4, materials design approaches based on machine learning and data informatics will be discussed. Non-comprehensive examples of the successful application of the above methods will be provided in Section 5 in the area of batteries, catalysts, and photovoltaics. Finally, critical remarks will be given to the outstanding issues in the field and an outlook to future developments will be provided.

## 2. Paradigms of materials discovery

Materials discovery can be dated back to the early days. Apart from noble metals such as gold, other metals with stronger reactivity usually need to be reduced from their compounds, especially when large quantities are needed. One example is copper. It can be extracted from its sulfide ores at high temperatures under a reducing atmosphere. Despite its crude nature, one can still find common characteristics of this with today's materials discovery. By experimental trial-and-error and post-mortem characterization, new materials are identified. For example, in terms of inorganic materials, the most typical method of synthesis is still high-temperature sintering. To look for potential new materials, precursors are put into a crucible and are subject to sintering at elevated temperatures. Such a process represents the first paradigm of materials discovery, i.e., empiricism-driven experiments. It has laid the foundation for the development of chemistry and materials science, see Fig. 2. One should note that despite the emergence of more advanced methods to be discussed shortly, this conventional approach is of great technological importance and should not be overlooked. In fact, before the popularization of modern computers, almost all critical materials are discovered in this manner, with input from chemistry and metallurgy. For example, in a LIB, the cathode materials were discovered through ad hoc design.[23,24] By studying the intercalation reaction between guest ions with solid hosts in $TiS_2$, Whittingham used such a layered material as a cathode, and demonstrated the first rechargeable lithium battery.[25] Further development was achieved by Goodenough et al. Based on the knowledge that the $S^{2-}$ 3p band lies at higher energy as compared with

that of $O^{2-}$ 2p, they used S to replace O to enable the use of the $Co^{3+/4+}$ redox couple.[23] Such substitution leads to significantly higher energy density and to the discovery of one of the most successful cathode materials for LIBs to date, i.e., $LiCoO_2$.[26] However, the drawback of experimental trial-and-error is also trivial to see. The efficiency of such an empiricism-driven approach is low and is becoming less and less favorable due to the emergence of the ever-complicating materials requirements. In this field of energy storage and conversion, almost each component of a device is bottlenecked by a functional material with highly specified materials property. Therefore, a shift towards a new materials discovery paradigm is necessary.

The second paradigm is the high-throughput screening of candidate materials. Although such screening can be done both computationally and experimentally, to limit the current report within its scope, we will focus on the advancements on the computational part despite the rapid development of high-throughput synthesis. Work in this area can be found in a number of recent works.[27–30] Computational screening of candidate materials is achieved by computing properties of the material from their structures. By screening a large pool of candidate materials, one can find targets to synthesize and characterize.[31–37] Such a method heavily relies on the theoretical advancements of quantum mechanical approaches and the development of computational infrastructures. In 1964 and 1965, Hohenberg, Kohn, and Sham published the two seminal papers on electronic structure theory. By reformulating the ground-state solution of the Schrödinger equation as a problem of minimizing energy as a functional of the charge density, the computational costs are significantly reduced compared to traditional methods such as quantum Monte Carlo and post-Hartree-Fock theory.[9,10] Despite that the exact energy functional has yet to be determined, DFT is able to tackle the property prediction task for practical materials from the fundamental laws of quantum mechanics. Later development of plane-wave electronic structure codes such as CASTEP,[13] VASP,[14] Quantum Espresso,[38] and local basis codes such as Gaussian[39] and ORCA[40] further facilitates the widespread usage of DFT beyond quantum chemists. To date, with minimal training, a non-specialist in electronic structure theory can carry out simple DFT tasks such as formation energy calculation, geometry optimizations, and band structure computations. More importantly, over the past decades, computational power has been increasing exponentially, thanks to the fast development of computational infrastructures. This allowed the entry-by-entry computation of existing databases such as the Inorganic Crystal Structure Database (ICSD).[41] Assembling these data entries into computational databases required careful handling of the caveats of current formalisms of DFT. In this context, a number of corrections have been proposed which laid the foundation of a various data informatics projects.[42,43] A number of different databases with different standards have been developed which will be discussed in the next subsection. Using these databases, fast screening can be carried out for materials with desired properties. For example, the electronic band structure is contained in almost all of these databases, which could serve the purpose of screening photovoltaic materials. For properties that are not included, it is almost trivial to write scripts to do high-throughput screen once the computation of such property can be streamlined. While such screening can ideally facilitate the materials design process, the materials discovery process along this pipeline is fundamentally constrained by its search space. One way to enlarge the search space is to do element exchange. While this has led to a significant increase in the number of material entries in these databases and many successful cases, the structural space is still limited. In this context, crystal structure prediction methods provide a good complementary despite that it constitutes an independent research area on its own. By exploring the potential energy surfaces (PESs) of specific number and types of atoms, the local energy minima are located and the corresponding structures may serve as candidate if their properties fulfill the requirements and their energy is low enough. The computational screening approach constitutes the second paradigm of materials discovery and will be discussed in detail in Section 3.

The third paradigm of materials discovery is based on data informatics and machine learning. From the discussions above, the success of high-throughput screening is inherently determined by the choice of search space. Even if the structural space is exhausted using crystal structure prediction methods (which is usually not the case), it is possible that the specific material with desired properties simply does not exist within the predetermined composition range. In other words, the success of the screening approach is still somewhat dependent on a matter of luck. One may or may not find new materials that are synthesizable and the materials found might not cater to the initially targeted application. In this context, methods to find the bounds of properties within a specific composition range and to carry out target-specific material design are critically needed.[44–46] Such a target has been partially met by ad hoc design using knowledge from chemistry and physics while the other half might be solved by the development of machine learning and data informatics. By learning from the large quantities of data generated from high-throughput computations, it is possible to teach computers chemistry and even train them to be better chemists than us, which is a lesson we learned from AlphaGo.[47] This field only started in the recent decades and is advancing most rapidly within the past decade. However, many exciting achievements have been made including the development of predictive models of structures, feature engineering techniques,[48,49] machine-learning force-fields,[50] and the recent generative models for target-specific materials design.[50,51] This approach is a natural

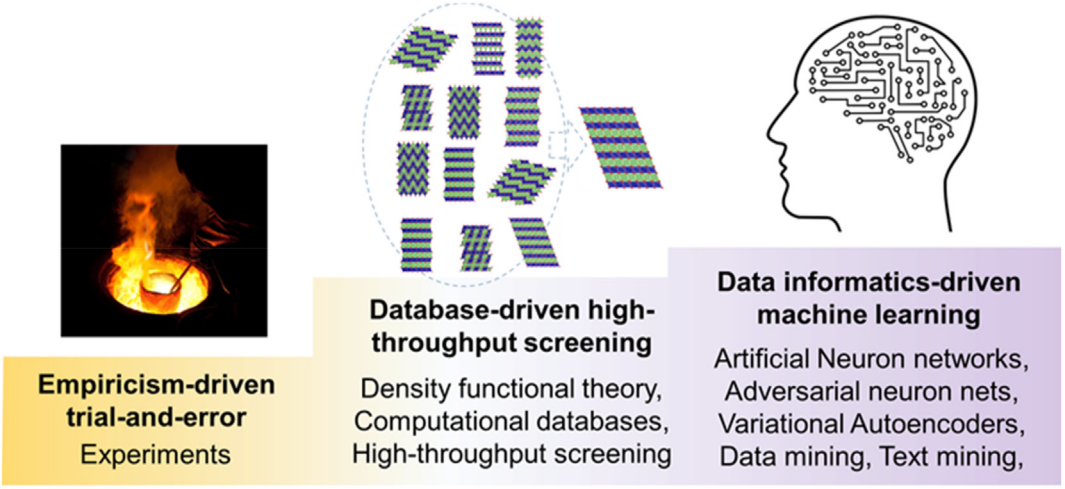



**Fig. 2.** Schematics of the three paradigms of materials discovery.

descendent of the high-throughput materials screening and data science. Despite that this field is still in its infancy, it holds the promise of significantly speeding up materials discovery and increase the chance of finding novel materials with desired properties. The details will be discussed in Section 4 with details.

## 3. Big data and high-throughput screening

### 3.1. The high-throughput screening approach

High-throughput screening represents the second generation materials discovery approach. By computing the properties of a large number of structures, those that meet the requirements and have low energies are chosen to be synthesized. As shown in Fig. 3, a typical process of computational high-throughput screening constitutes four steps, i.e., identification of target properties, defining screening spaces, property prediction, and selection of candidate materials. Among these steps, identifying target properties is one of the most critical and difficult steps. Usually, materials scientists can easily name the desired macroscopic properties of functional materials in an energy conversion/storage device. For example, for electrocatalysts, the material needs to give low overpotential at relatively small current densities.[52,53] This is sometimes also reflected by its Tafel slope and the turnover frequency. These quantities measure the catalytic activity of the material. While requirements on these quantities can be determined by back-tracing the final device-level performance, how they relate microscopic quantities, especially those computable from DFT is non-trivial. For some applications such as photovoltaics, the light absorption properties, and charge carrier transport can be calculated from the band structure, with a basic understanding of solid-state physics. However, for other applications such as the previously mentioned electrocatalyst, one may need to go through rigorous derivation to obtain a relatively simple quantity to characterize the property, which is also called a descriptor.[52] One example is the hydrogen adsorption free energy ($\Delta G_H$) for the hydrogen evolution reaction.[54–57] It requires the computation of the adsorption energy of H on the catalyst surface.[37] Unfortunately, such a descriptor is non-trivial to calculate as it requires the consideration of different surfaces of the catalyst and the H coverage. A further simplification can be done by correlating such energy with the electronic band structure of the catalyst, e.g., the d band center of metals.[58,59] For other applications, such descriptor may beyond the limit of DFT calculations and therefore requires other theoretical tools. For example, the mechanical strength of polycrystalline alloys and intermetallic materials is dependent on the distribution of grains within the material.[60,61] To capture such a distribution, a model easily reaches tens of thousands of atoms. While it can be computed using the most up-to-date DFT framework, the computational cost is extremely high and such calculations cannot be carried out in a high-throughput manner. Therefore, one of the major requirements for a potential high-throughput screening discovery of novel material is that a clear relationship can be built between the macroscopic measurables and the microscopic computables. Also, calculating such microscopic quantities should not be too computational demanding. In fact, there is a tradeoff between the computational cost and the quality of the screening process. For instance, in searching for good hydrogen evolution catalysts, in principle, one should consider different coverage of H on the surface. However, this requires the enumeration of the possible arrangement of adsorbents attached to potential active sites, which is computationally extremely demanding. Therefore, one may simplify such computation to the dilute limit, i.e., hydrogen atoms are adsorbed on the surface that is large enough to ignore adsorbents-adsorbent interactions.[61]

After identification of the target properties and correlation it to a set of microscopic computables, the next step is to choose the composition and the structural space to carry out the screening. Currently, the screening space is usually constrained to a specific set of structures, e.g., perovskites.[62–64] By changing substituting the elements in the structure, one can obtain a large number of independent candidates to compute. Such a method has a relatively high rate of success because the structure template chosen usually has a relatively high chance of being genetically favorable. However, it also suffers from the drawback of constrained structural space. For example, a binary compound with a one-to-one elemental ratio may form a number of different structures like NaCl-type rocksalt, α-HgS-type trigonal cinnabars, and β-HgS-type metacinnabars. Simple elemental substitution on predefined structural templates may lead to a loss of potentially stable structures. A way to get around with this is to carry out crystal structure predictions. Using algorithms such as random sampling and particle swarm optimization, the potential energy surfaces (PES) of a given composition are explored and the low energy structures can be obtained.[65,66] The details of crystal structure prediction by this significant method will be discussed in Section 3.3. By incorporating such methods, one can partially resolve the constraint on the structural space. However, making predictions on stable structures is usually computationally expensive. Also, the composition space is still constrained, and a wise choice of such needs yet other theoretical considerations. Finally, in occasional cases, large databases may be used for initial screening purposes. These databases usually contain structure entries exceeding $10^6$ and can provide wide coverage of the structural and the compositional space.[67] Searching for materials in such a huge space without care will lead to a low encounter rate and is usually not carried out. It is worthwhile to note that these databases may not only serve as the pool for actual screening but also as the base for


**Fig. 3.** Illustration of essential steps of high-throughput screening for materials discovery.

determining the thermodynamic stability of a material. Considering its importance, these databases will be introduced separately in Section 3.2.

After choosing the screening space, the actual property prediction is carried out. Since the descriptor is already determined, one only needs to streamline the workflow of such computations using DFT or other theoretical tools. For DFT calculations, structural relaxation is likely to be necessary. A number of bulk properties only require solving the ground-state electronic structure of the relaxed candidate material. These properties include but are not limited to the band gaps and the formation energies. Other properties such as elastic modulus and spectroscopic features usually need the consideration of lattice vibrations, i.e., phonons.[68] Beyond these properties featuring bulk characteristics, many others require the computation of modified structures. For example, the adsorption energy of a specific molecule to the surface of a material and the work function is critically dependent on the creation of a proper surface model. It is worthwhile to notice that some caveats of the current DFT formalism may affect the accuracy of the predictions and should be addressed properly.[69–73] For example, the electronic band structure of strongly correlated systems such as transition metal oxides cannot be precisely captured using functionals under the semi-local general gradient approximations.[70] Hubbard U corrections and hybrid functionals are usually used to mitigate this issue.[73]

The final step is to conduct the screening. A set of selection criteria needs to be determined for the screening process. Usually, the first step is to eliminate those structures that are not likely to be synthesized. This is usually done by calculating the energy above hull of the specific material within its compositional range and requires the calculation of the energetics of relevant phases within such range, leading to high computational cost. Fortunately, a number of databases are available where the common stable phases have been pre-calculated.[16] As long as the computational setting is in line with those used for preparing the database, the energy above hull of the candidate structure can be calculated using these existing energy values. For solid materials, the energy above hull is a good estimate of the phase stability. A value of zero means the material shapes the hull and will be unlikely to decompose into other phases.[69] Sometimes relatively small values, e.g., 10 meV atom$^{-1}$ are also acceptable as the entropy term arising from temperature may stabilize the material. These less stable phases are sometimes also called meta-stable. It is worthwhile to note that this term holds a different meaning in the field of chemical physics. Rigorously, a meta-stable material refers to whose free energy above hull is zero and therefore is stabilized only due to kinetic reasons. Following the determination of phase stability, the set of selection criteria can be applied to sieve potential candidates from the computed material entries.

After obtaining a number of potential candidate materials, the composition, the structural, and the predicted property information are transferred to a synthesis chemist. A series of materials characterizations such as X-ray diffraction will be carried out to confirm the structural correctness. The performance of the material will be assessed by incorporating it into energy devices. If the performance is not ideal, the four steps need to be reviewed critically to verify the validity of each assumption. Another round of screening may be necessary, closing the design loop.

### 3.2. Computational databases

The development of modern computational databases laid the foundation for fast materials screening. A number of general-purpose databases to date have been developed including the Materials Project, Materials Cloud,[74] Open Quantum Materials Database,[75] NOMAD,[75] AFLOW,[76] JARVIS,[77] NRELMatDB.[78] These databases are mostly built upon existing experimental structural data such as ICSD. Therefore, they store a large number of structural data and the corresponding properties calculated using electronic structure theory such as DFT. Due to the wide coverage of stable phases in these databases, it is possible to compute the phase stability of new materials without recalculating the relevant phases within the composition range. This requires corrections to the DFT entries based on the widely used Perdew-Burke-Ernzerhof (PBE) functional under the GGA approximation.[79] For example, the $O_2$ molecule is known to overbind using GGA-PBE.[80] To correctly reflect the formation energies of oxide materials, the energy of the $O_2$ molecule is shifted by comparing the formation enthalpy from experiments and the computations.[81] Correlated systems such as transition metal oxides are another family of materials that is difficult to be precisely described using GGA-PBE. Hubbard U corrections have been added to the d channels of transition metals to overcome this issue. Further hybridizing schemes enabling the mix of materials entries with pure transition metals and transition metal oxides have also been developed.[81] With these developments, the energy above hull of a newly predicted material can be trivially obtained using basic DFT calculations and a few lines of scripts.

Apart from these general-purpose databases, many dedicated counterparts have also been created. For example, for two-dimensional materials, hypothetical structures have been generated via computational exfoliation, elemental substitution, and machine learning methods. Databases such as 2DMatPedia have been constructed based on these data entries and could serve as the foundation for the screening of photovoltaic materials, electronic materials, and beyond.[82–84] For other dedicated applications such as topological materials,[85,86] organic crystals,[87] and even simple inorganic perovskites,[88] high-throughput screening has been carried out and databases have been generated.

While these databases have enabled both theorists and experimentalists quick searches of candidate materials and the corresponding material properties, they have inherent drawbacks. First, the number of data entries for a material with a dedicated application is still limited, despite a large number of structures of the entire database. Second, the composition coverage is still not ideal. Especially for multi-element composition, the data entries are still scarce. This leads to an overestimation of phase stability of newly computed materials. It is especially true when it comes to compositions that are less explored and are less technologically relevant. To circumvent such issues, new structure entries should be generated via crystal structure prediction methods, which will be discussed in the next subsection. Third, the distribution of data entries is significantly biased. With the rise of machine learning and other materials informatics methods, the data from these databases have served as the training set for the models. The behavior of such models is significantly influenced by their training set. However, the distribution of data entries usually cannot make a uniform coverage of the entire spectrum of materials space. For example, the number of structures with fewer elements is significantly higher than that with a larger number of elements, which is not the case in nature. Additionally, compositions with potential technological relevance are covered better than those less investigated. In this context, it is necessary to develop novel methods to redistribute the effort of adding new entries into these databases with modern methods from data informatics.

### 3.3. Crystal structure prediction

Crystal structure prediction methods have been developed to find stable and meta-stable structures that are potentially synthesizable.[20,89] While they represent an independent area of active research, these methods have been widely applied in many technological-relevant fields such as high-pressure physics, superconductivity, semiconductors, and electrochemical energy storage.[90–92] Formally, the task of crystal prediction can be defined as the determination of low-lying energy minima on the Born-Oppenheimer energy surfaces, as illustrated in Fig. 4(a). The inputs of a crystal structure prediction computation could be the numbers and types of atoms in the system. By manipulating the arrangements of these atoms, the configurational space of this set of atoms is explored. Usually, a large number of basins exist and the probability of finding a low-energy structure is dependent on the size of the hyper-volume of a specific basin.[93–95] Unfortunately, the exact shape of the potential energy surface is unknown and requires calculation in a point-by-point manner.

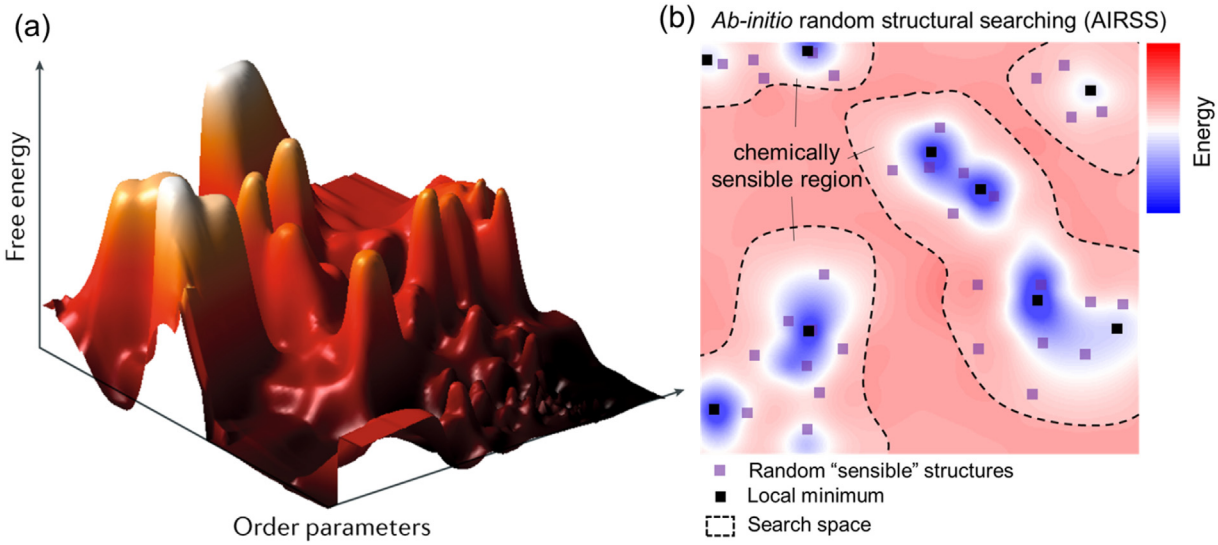



**Fig. 4.** (a) Energy landscape of crystalline $Au_8Pd_4$. Reproduced with permission from Ref. 89. (b) Illustration of the distribution of local energy minima in a typical PES and the random sampling approach. Reproduced with permission from Ref. 92.

The energy and its gradient can be evaluated using either electronic structure methods such as DFT or using empirical models such as forcefields. Currently, DFT is the most widely used energy engine for such a task. It gives a good balance of accuracy and computational speed for systems that are not too large (up to a hundred atoms). For large systems, empirical forcefields need to be fitted but are usually less accurate. Modern approaches using machine learning offer a way to achieve DFT-level accuracy and low computational cost that is on the level of empirical forcefields.[96,97] The output of a crystal structure prediction is a set of crystal structures. As mentioned previously, these structures could serve as the base for an independent screening process for task-specific materials using Pareto optimization.[98] They can also be added to existing databases for later use which will partially resolve the constraints on the configurational space when carrying out high-throughput screening using the database and the elemental substitution approaches.

Random sampling is the baseline method among a large number of crystal structure prediction methods. It samples the local energy minima in the configurational hyperspace by randomly distributing the atoms followed by local minimization along the path of steepest descent on the potential energy surface.[65] Unfortunately, the number of local energy minima increases exponentially with the number of atoms in the system.[99] Therefore, it will be nearly impossible to sample the entire configurational space for the global minima. In fact, a no-free lunch theorem has been proved by Wolpert and Macready.[100] It is highly possible that we will not be able to find an algorithm that works well in all circumstances for global optimization. Fortunately, the potential energy surface of physical systems has several unique features that can be used for reducing the computational demand. As shown in Fig. 4(b), the portion of the potential energy surface we are interested in is in fact a very small subset of the entire space.[101] First, atoms are never too close to each other in physical systems. When they are too close, strong repulsive forces exert on each of the atoms. Therefore, there are no local energy minima within such a portion of the potential energy surface. Second, those energy minima with low energy values usually correspond to highly symmetrical structures.[102] Third, certain local arrangement of chemical species leads to low energies due to unique chemical interactions. For example, in systems containing P and O, the structure containing phosphate groups is highly likely to have low energies.[92] Utilizing these features, one can constrain the search space and reduce the cost of random sampling. Such spirit has been reflected by the ab initio random structure searching method. By biasing the search using chemically-informed pairwise minimum separation between different chemical species and a number of other constraints, this method is one of the most efficient despite its simplicity. Such a method also gives good control of the distribution of configurational space being accessed and can be used for gather unbiased data for construction of database, for training machine learning models, and for carrying out materials informatics analysis.

Beyond random sampling, many global optimization techniques have been applied to crystal structure prediction. In fact, global optimization itself is a very active and large field in applied mathematics.[103] Some of the most important and widely-used methods include simulated annealing,[104,105] basin hopping,[106] metadynamics,[107] minima hopping,[108] and evolution algorithms. The last consist of two of the most successful methods that have been applied in the field crystal structure prediction,[66,109–116] i.e., the Oganov-Glass evolution algorithm[109] and the Wang's version of particle swarm optimization.[66] Accordingly, a number of codes have been developed along with the development and application of these global optimization methods to predict new structures. Some of the mostly used and publicly available codes include AIRSS (ab initio random structure searching), CALYPSO (crystal structure analysis by particle swarm optimization), CrySPY, DMACRYS (energy minimization package to simulate rigid molecules with multipoles), GASP, Gator (first-principles genetic algorithm for molecular crystal structure prediction), GRACE (generation ranking and characterization engineer), MAISE (module for ab initio structure evolution), Molpak (molecular packing), UPack (Utrecht Crystal Packer), USPEX (Universal Structure Predictor: Evolutionary Xtallography or uspekh in Russian), and Xtalopt (evolutionary crystal structure prediction).[20]

While crystal structure prediction has provided a method to predict new materials that have never been discovered or synthesized before and have been successfully applied in a number of fields, some outstanding issues need to be addressed in this area. First, the prediction of materials with large unit cells and many atoms within the unit cell has always been a tough task. In this context, better algorithms need to be developed despite that it might be difficult to create a universal one. Also, the computation for energies and forces needs to be accelerated since this is the most time-consuming step of any crystal structure prediction method at the moment. Second, the accuracy of current DFT methods in describing the potential energy surfaces needs to be increased. Currently, the widely used PBE functional has a number of caveats in correctly reflecting the features of systems like molecular crystals with weak interaction and strongly correlated systems such as transition metal oxide.[69–73] Third, the crystal structure prediction method is inherently

incapable to determine the composition for the search. In fact, the current crystal structure prediction is still materials discovery method instead of target-specific materials design. The search space is constrained by the initial choice of composition. Higher-level methods using machine learning and data informatics need to be developed to help determine the composition and promote the success rate of materials design using crystal structure prediction.

## 4. Machine learning approaches

### 4.1. Basic principles

Machine learning and related data informatics approaches represent the third generation materials design approaches. In principle, they can resolve the outstanding issues of the high-throughput screening approach and achieve the target-specific material design.[117–119] Machine learning itself is an independent research area branching off artificial intelligence. Its major task is to develop algorithms and models that can learn patterns and perform tasks like human beings. It is especially useful to deal with tasks featuring combinatorial or exponential complexity, which cannot easily be streamlined using conventional methods. In the context of materials discovery, the ability of machine learning to generalize from known data to explore the unknown is well suited for identifying new materials using existing databases. Such a task is difficult to carry out using the first and second generation approaches, i.e., experimental trial-and-error and high-throughput screening, which are bottlenecked by their efficiency. In fact, to date, there are $\sim 10^6$ crystal structures being reported which constitutes only a very limited portion of the possible materials universe.[67,120–123] Using machine learning approaches, it is possible to make predictions on new materials and their properties. In return, these data could be refined using high-level tools such as experiments or DFT calculations and fed back to the machine learning model for better accuracy. In general, machine learning models for materials discovery can be classified into two categories, depending on how they are used to generate potential candidate materials. The first type of method refers to those that can accelerate conventional screening approach by bypassing the time-consuming DFT computations. The second type is generative models. Instead of feeding the model with structures to get predicted properties, properties are used as input and new structures with desired properties are generated. The details on this will be discussed in the coming part on the construction of a machine learning model in Sections 4.2 and 4.3.

Building a machine learning model for materials discovery usually constitutes five essential steps, i.e., target determination, data collection, featurization, and model selection, and training.[118] Target determination is the first step and arguably the most critical step because it determines if the model will be generalizable and if it will have an enormously large error. Target determination is to identify the goals and the prediction target of the machine learning model. A prediction target can be a single property or a set of features of a material. For example, in the context of solid electrolytes for a solid-state battery, the most critical task is to find solid materials with high ionic conductivity.[124,125] In addition, the materials need to have a wide electrochemical window and should be synthesizable, i.e., the energy above hull needs to be zero or close to zero.[126] It is worthwhile to note that, the task of a machine learning model can often be classification or regression.[127,128] A regression process builds the mapping between the input structure and the target properties such as band gap, heat capacity, formation energy, and Young's modulus. Classification can be viewed as a highly specialized regression process. Instead of building a mapping between the structure and the continuous numerical properties, it cauterized the structure into different categories. Being metallic vs. non-metallic, being superconductive vs. non-superconductive, and being thermodynamically stable vs. unstable are typical examples.[127,128]

The second step is data collection. A machine learning model generalizes the knowledge learned from existing data to make predictions. The quality and quantity of data significantly affect the outcome of the model. The quality of the training data is essential to the accuracy of the model. The "garbage in, garbage out" situation should be avoided. Usually, there are several levels of accuracy of data. The experimentally measured structures and properties are usually the most reliable despite there might be some experimental error. However, caution needs to be taken for properties that are less easy to measure such as catalytic activities. In electrocatalysis, the difference in the overpotential can be as high as several orders of magnitude for the same composition depending on the measurement setup and the reliability of the report.[129–138] Beyond that, computational data such as the formation energies of materials using DFT calculations is also relatively reliable. Especially those from large computational databases which have been systematically constructed to avoid potential error from different input sets. The accuracy also depends on the level of theory used in the computation. For example, the band gaps calculated using the PBE functional have systemic error and its accuracy is inferior to those obtained using hybrid functionals.[139] The experimentally measured band gap is further considered to be more accurate and the gold standard.[140] The quantity of the data is also of great importance. The current machine learning models are generally more suited for interpolation instead of extrapolation. Therefore, a large number of data entries are of great significance to avoiding undersampling of the materials space of interest. The databases mentioned in Section 3.2 may serve as a good starting point for data collection. In fact, many codes are available to retrieve, convert, and manipulate the material entries in these databases. Examples include pymatgen,[141] AiiDa,[142] atomate,[143] and ASE.[144] The number of data entries required to train a model depends on the type and complexity of the model and there is no universal value for this. However, a rule of thumb might be 50. For a machine learning model, there should be no fewer data entries at least. Whether the data is enough can be verified by testing the model by making predictions on known data outside the training set. Beyond the quantity of the data, the distribution is also important. In the hyperspace of input vectors of the model, the distribution of the data should not clump together but should be ideally distributed uniformly within the range of interest. For example, if one wants to learn the band gap of a binary compound with the composition of $AB_x$, the data being fed into the model should be ideally cover all possible $x$ values instead of all concentrated on several specific values. To resolve the distribution issue of the data, one can use newly developed tools from the field of material informatics such as SOAP (smooth overlap of atomic positions),[145] ASAP (automatic selection and prediction tools),[146] and SHEAP (stochastic hyperspace embedding and projection).[92] As illustrated in Fig. 5, these methods can be used to quantify the distances between different structures and to do the following analysis on the distribution. For example, the SOAP descriptor uses a series of spherical harmonics as the basis set. By placing Gaussian density distributions at each atom, the local coordination of the atom is expanded in the spherical power spectrum, corresponding to the neighbor density. Due to its unique formulation, it forms compact support which is essential for the calculation of local energies. Using such a descriptor, not only the structural similarity between two structures can be evaluated, but also machine learning forcefields can be built.

The third step is featurization. It refers to the process of converting the training data into readable numerical forms by machines. These values are also called descriptors, features, or fingerprints. Featurization is a critical data pre-processing process. They can significantly influence the accuracy of a model. In fact, it determines its upper bound. Structures are the most important data to featurize as a major bulk of the machine learning models to date use the structural information as an input and make predictions based on it. A feature for such a purpose is also known as a representation. A good representation has three key criteria, i.e., uniqueness, universality, and efficiency.[147,148] First, the representation needs to be invariant to the symmetries of the system. For example, after translation, rotation, and atomic permutation, the numbers or vectors to represent a molecule should not change.[145,149] In this context, the

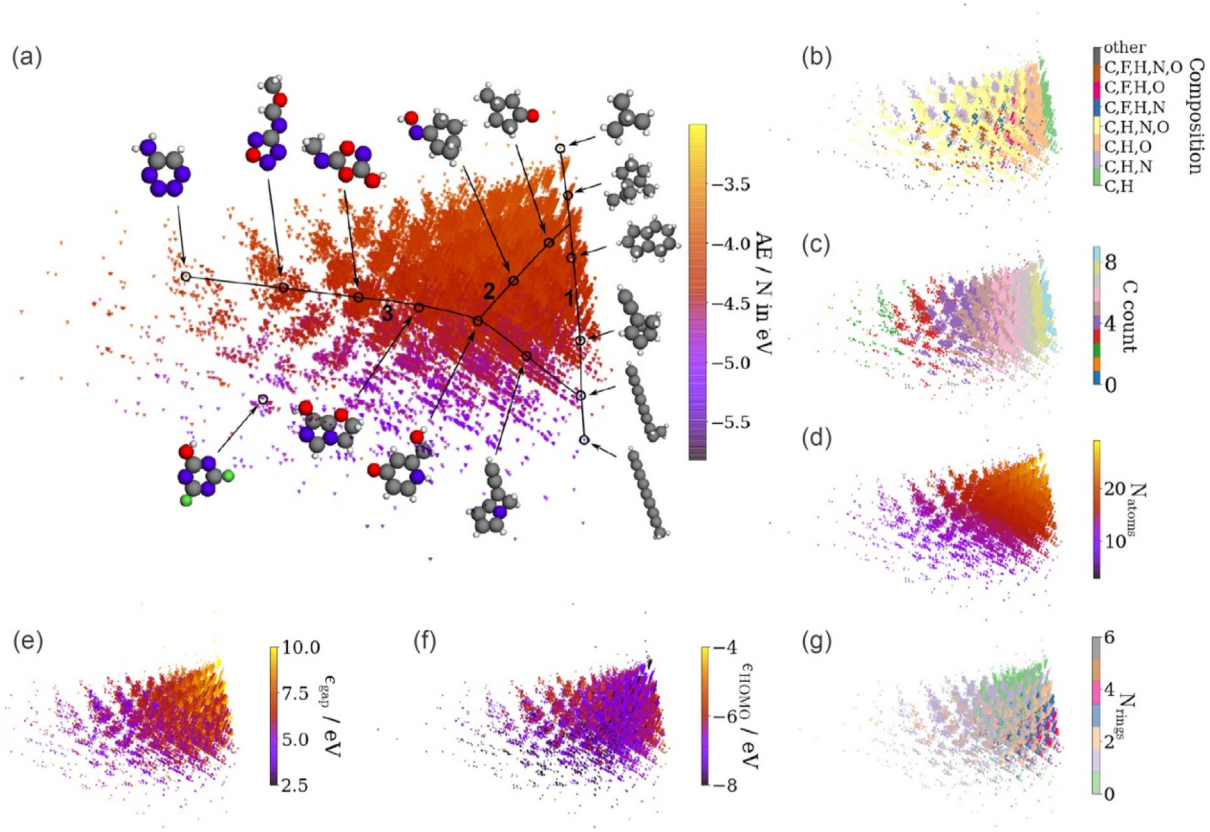



**Fig. 5.** Kernel PCA (KPCA) maps of the QM9 database using a global SOAP kernel. The frames are color-coded according to structural descriptors (b, c, d, g) and quantum mechanical properties (a, e, f). Reproduced with permission from Ref. 146.

Cartesian coordinate is not a good representation. A one-to-one correspondence is highly desired despite the difficulty in obtaining so. This is especially important for generative models as the output of the model should be able to be converted to a unique structure. Beyond these requirements on uniqueness, the representation should also be ideally capable of representing different systems, e.g., molecules and crystals. A vector containing three atomic numbers ($Z$) of elements is suitable to represent an $ABX_3$ perovskite but incapable of representing other structures. Efficiency is also critical to make the machine learning process as fast as possible. A number of representation have been developed for molecular and extended systems, including Coulomb matrix,[150] SMILES,[151,152] bag of bonds,[153] radial distribution functions,[149] BAML,[154] molecular graphs,[155] extended-connectivity fingerprints,[156] translation vectors, fractional coordinates of the atoms, Voronoi tessellations,[157] property-labeled materials fragments,[158] and SOAP.[145] They are capable of different tasks and the detailed descriptions should be found in the original publication. Besides the structure, materials properties also can/need to be featurized, e.g., band gaps, electronic density of states, highest occupied orbitals, and lowest unoccupied orbitals. Sometimes these features need to be selected and preprocessed, techniques such as least absolute shrinkage and selection operator (LASSO) regularization,[159–161] principal component analysis (PCA), and t-distributed stochastic neighbor embedding ($t$-SNE) can be used.[162] Detailed review on such aspects can be found in the work by Chen et al.[118]

The fourth step is model selection. Many machine learning models are available and this area itself is developing rapidly. The choice of model should be determined based on the target of the task as mentioned previously. In general, there are three types of machine learning models, i.e., supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is used to build the mapping between the input features and the output labels or values. In materials discovery, the inputs are usually the structures while the output is usually properties such as formation energies and band gaps. Using such models, one can bypass the time-consuming DFT or experiment steps to directly predict the properties of new materials.[163,164] Unsupervised learning is used to find patterns from the data. As disclosed in its name, there are not input labels for these data. For example, clustering is a typical task used to classify data into different categories without knowing an explicit numerical threshold.[165] Generative models such as generative adversarial nets can find patterns in existing data and "fake" new data that is structurally similar to those existing ones, as illustrated in Fig. 6[166–168] In the area of materials discovery, these algorithms are extremely important and are under intensive development and will be introduced in Section 4.3. Reinforcement learning mimics how human learns from interactions with the environment and is used to improve the ability performing specific tasks via giving rewards or punishments after decision is made. Such methods are also under intensive development in the area of materials optimization and discovery.[169] For example, the composition optimization task can be accelerated using such methods.[170]

Finally, the model is being trained. The process of training is highly dependent on the model. Usually, a loss function needs to be defined. This is especially true for regression models. Such loss function is defined to reflect the difference between the predicted properties and the labels. Mean absolute error on a number of weighted properties is used in this context. It is worthwhile to note that the loss function needs to be differentiable. During the training process, overfitting is an often encountered problem. It is a modeling error that occurs when the fitted function is aligned too closely to the training data so that the predictivity of the model is lost or the model cannot be generalized. This usually happens when the model complexity is too high and the quantity of training data is relatively small. To avoid so, methods such as early stop, dropout, regularization, and removal of anomalies and redundant features. Fortunately, many codes are available to set up, train, and validate

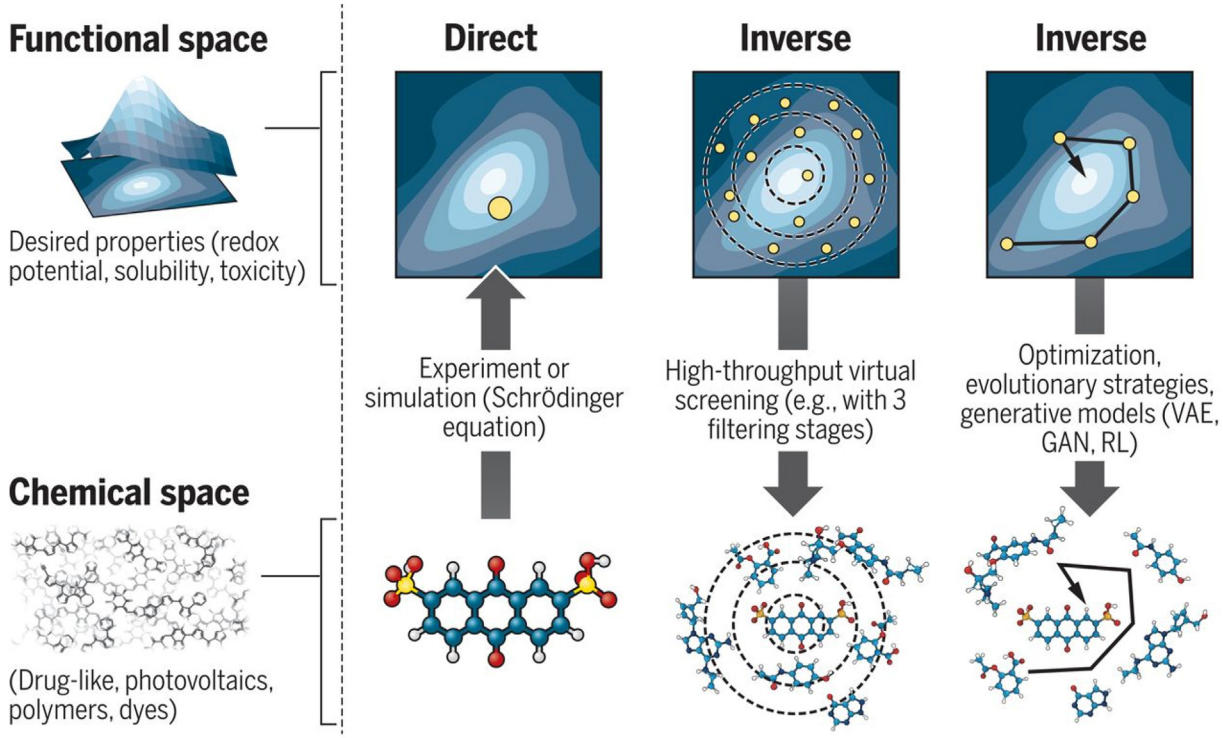



**Fig. 6.** Schematic of the different approaches toward molecular design. Reproduced with permission from Ref. 176.

a machine learning model. General-purpose packages include scikit-learn,[171] tensorflow,[172] and Pytorch.[173] For materials modeling and discovery, specialized tools such as AutoMatminer[174] and PROPhet[175] have been developed.

### 4.2. Accelerating high-throughput computations

Despite the advancements in computational power, the prediction of materials properties using DFT calculations and other means has become the absolute bottleneck for high-throughput screening. Many machine learning algorithms/models have been developed to learn the mapping between the input features such as the structure and the properties. Using these methods, the materials discovery process can be significantly accelerated. Here we provide a non-comprehensive list of machine learning models for screening acceleration.

Linear and generalized linear models are a family of machine learning models that are developed based on linear models. Due to their simplicity, they sometimes are not regarded as machine learning models. In a linear model, the property (or label) is linearly related to each element of the feature. These models have a mathematical form as follow:

$$y = X\beta \tag{1}$$

where $X$ is a matrix describing the features, y is a vector of the target property, and $\beta$ is the coefficient vector. Examples of such models are the correlation between the electronic d band center or p band center with the catalytic performance of electrocatalyst.[56,177] Sometimes, the features or target properties are bounded, and generalized linear models have to be incorporated. In these models, linking functions need to be added to covert the bounded target.

While the linear models are elegant and can usually be interpreted physically, most properties of a material do not linearly dependent on the features. In this context, a number of models that are able to capture non-linearity have been developed, and the most well-known ones are kernel-based models, tree-based methods, and deep learning.

In the kernel-based models, a similarity measure, i.e., the kernel, is introduced to allow us to construct algorithms in dot product space. For example, for a non-linear model expanded using a polynomial basis, the mathematical form takes:

$$y = \varphi(x)^T\beta \tag{2}$$

where $\varphi(x) = [1, x, x^2, \ldots, x^{m-1}]^T$ and $m$ is the dimension of the feature. The coefficient vector can thus be determined by a set of combination coefficients $\lambda$:

$$\beta = \sum_{i=1}^{n} \lambda_i \varphi(x^{(i)}) \tag{3}$$

$$y^* = \varphi(x^*)^T\beta = \sum_{i=1}^{n} \lambda_i \langle \varphi(x^*), \varphi(x^{(i)}) \rangle \tag{4}$$

where $n$ is the size of the data, $\lambda_i$ is the combination coefficients, and $\langle , \rangle$ is the inner product. The inner product computes the similarity between the two inputs and can be generalized to other kernels, with the most well-known being the Gaussian kernel. Gaussian process regression, kernel ridge regression, and support vector machine are typical examples of kernel-based methods.

Three-based models are usually used to refer to those based on decision trees which classify the label by answering a series of yes and no questions on the input features. Random forest is a model based on decision trees. By using an ensemble of decision trees and taking the average, it resolves the overfitting issue and is robust in predicating properties from features.

Deep learning is a subset of the broader family of machine learning methods. These methods are based on artificial neural networks which loosely mimics its biological counterpart. In these models, a collection of connected units (neurons or nodes) are constructed. These nodes receive a set of weighted inputs and evaluate whether they will give output to other nodes. A non-linear activation function determines whether the signal is sent out and its value. During the training of an artificial neuron network, the weights are adjusted so that the predicted value matches the

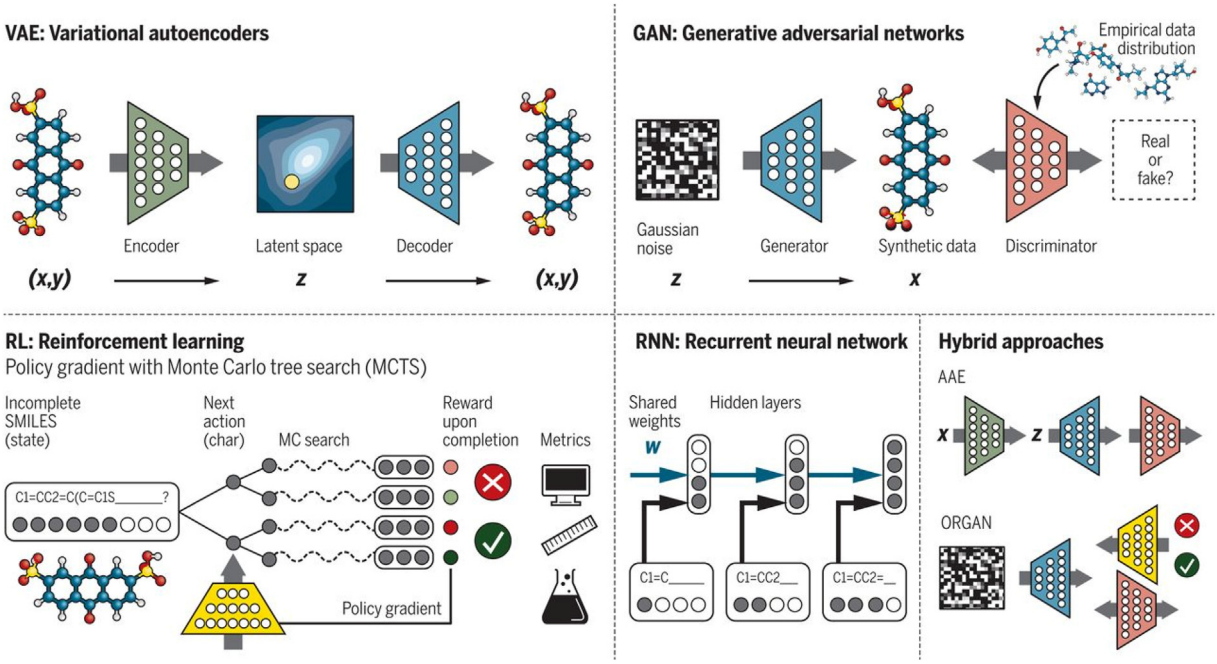



**Fig. 7.** Schematic representation of several architectures found in generative models. Reproduced with permission from Ref. 176.

actual label. The "deep" feature of deep learning comes from the multi-layered characteristic of neurons. A unique feature of deep learning differed from conventional methods is that it allows the learning of representation of the features and allows less or no feature engineering. Many deep learning models have been proposed and the number is growing rapidly. A detailed discussion needs dedicated review and can be found elsewhere.[178,179]

It is worthwhile to note that a critical prediction problem in materials discovery is synthesizability. The energy of a predicted structure is highly relevant to whether the material can be synthesized or not. In this context, finding the mapping between the structure and the energy using machining is highly beneficial because it reduces the time of using DFT to evaluate such a property. In fact, learning the relation between the atomic arrangements and the energy, i.e., fitting the PES, is an independent field. Conventionally, empirical forcefields have been built but suffer from a number of inherent drawbacks. Considering the highly non-linear nature of PES concerning the atomic coordinate, the non-linear methods above have been utilized. A number of machine learning potentials have been developed to date and are used in the field of predicting phase transition,[180–182] phase diagrams,[182] or even crystal structure prediction.[182] Excellent reviews have been assembled in this area.[96]

All the models above are used to bypass the time-consuming DFT computations by fitting the mapping between the input features and the output labels for property prediction. Therefore, material discovery via this route is also known as the direct method. Unfortunately, the success of such is dependent on the predefined search space. To resolve this issue, a new type of model has been developed to generate structures from required properties. These generative models are also known as inverse materials design.

### 4.3. Generative models

As mentioned previously, there are two mapping directions for materials discovery, i.e., the forward mapping and the inverse mapping. In the prior, the features such as structures are fed into the model and the properties are predicted. On the contrary, inverse mapping directly gives structures or compositions which are essential for synthesis from the required properties. Therefore, the latter is more suited to materials design.

To date, the most widely used three types of generative models are the recurrent neural networks (RNN),[183,184] the variational autoencoder (VAE),[185] and the generative adversarial nets (GAN), see Fig. 7[186] RNN is a method to generate sequences and has been used as a starting point for generating new materials in the molecular realm. This is because organic molecules can easily be represented in the form of SMILES strings. A number of techniques such as long short-term memory cells, attention mechanism, and memory effects can be incorporated into the model to consider the time-dependency on the patterns.[184,187,188] The RNN provides a good starting point to generate new molecules. However, in real materials design tasks, the process needs to be biased. For example, one may need molecules with low energies so that they can be synthesized. A proper set of HOMO and LUMO values is critical for its functionality in real devices. In this context, VAE and GAN can be used. To discuss the features of a VAE, the autoencoder needs to be introduced first. An AE consists of an encoding and a decoding network. The encoding network maps the structure of a molecule to low dimensional space, i.e., the latent space. The decoding network inversely maps the vector in the latent space to its original representation, e.g., a SMILES string. Such structure of an AE gives it the capability of capturing some of the features of the data. By further sampling new vectors from the latent space, one can generate new structures. In a VAE, the encoding network is constrained so that the latent vectors are generated to follow a Gaussian distribution. A molecule is thus represented no long by a fix point but a probability distribution in the latent space, giving it better generalizability. An interesting method to incorporate supervision is to train the VAE to reproduce the structure and the properties at the same time. This is achievable because the latent space itself is continuous and differentiable. Therefore, molecules with similar properties will be close to each other in the latent space and biased generation of new materials can be achieved.[189]

GAN provides another method to generate new materials.[51] GAN is initially proposed by Goodfellow and co-workers and is used to generate fake (but indistinguishable) images from known ones.[186] A GAN consists of two subnetworks, similar to that of a VAE. A generative network (generator) generates candidates while the discriminative network (discriminator) evaluates them. In a training process, the generator tries to generate synthetic data by mapping points from the latent space. On the other hand, the discriminator will try to evaluate whether the data is

fake or real. The model is trained in an alternative manner and the goal of training the two networks is different. For a generator, the goal is to synthesize new data that are part of the true data distribution, or in other words, to increase the error rate of a discriminator. For the discriminator, the goal is to distinguish the synthetic data from the real data with better accuracy. In the realm of materials discovery, a GAN is able to generate new materials that are realistic (have the physical feature of a real material). More importantly, by adding features to the structures, the latent space could be sampled in a biased way to generate materials with specific properties, similar to that of a VAE. However, it is worthwhile to note that the training of a GAN is non-trivial. Convergence is difficult to achieve even for image generation where the features have been well-engineered. Currently, improving the convergence of training a GAN is an important topic in the field of machine learning.[190]

The abovementioned models have been developed to fit the generation of the molecule because the representation of a molecule is relatively simple using SMILES. For extended systems such as crystals or solids, this task is non-trivial. The representation needs to have several properties as mentioned in Section 4.1. To date, the one-to-one correspondence and the symmetry invariance haven't been achieved at the same time with high efficiency to the best of our knowledge. A way around is to use image-based method. Noh and co-workers transformed three-dimensional crystals by sampling the real space with Gaussians.[191] The atoms in the cells are replaced by Gaussian functions with different intensities. By doing so, the representation becomes invertible and image-based learning methods could be used. In their work, they used a VAE to achieve the generation of new structures. It is worthwhile to note that, data augmentation is still necessary to help the machine to understand the symmetry invariant property.[192,193]

## 5. Applications

The above-mentioned computational methods have been applied to a wide range of energy-related fields and achieved a partial success. In this section, we will compile the examples of applying screening-based and machine learning-based methods for materials discovery in the area of energy research, with an emphasis on the latter. It is worthwhile to note that the current compilation is not aimed to be comprehensive but to showcase how these methods can be used.

### 5.1. Batteries

A battery is an electrochemical storage device that stores energy in terms of redox pairs and releases it through electrochemical reactions. Specifically, rechargeable alkali-ion batteries use alkali ions as a media, and the shuttling of these ions between the cathode and the anode leads to redox reactions. The reaction is reversible so that the battery can be charged and discharged multiple times.

A number of material properties are crucial to the performance of a battery. One of the most critical is the diffusion properties. The diffusion of alkali ions in the bulk of electrodes and in the electrolyte determines the rate capability of a battery. Such a property has been extensively studied using machine learning approaches because standard DFT-based approaches are too expensive for screening purposes. Jalem and co-workers carried out a high-throughput screening computation of the Li diffusion barrier in the $LiMXO_4$ polyanion cathodes and their $LiMTO_4F$ relatives.[194,195] Using these data, they extracted 42 descriptors that are potentially relevant to the diffusion barrier. A least-squares model was built to map the correlation between the descriptor and the barriers. Interestingly, by comparing the coefficient, it is found that the quadratic elongation and the bond angle variance of the M octahedron as well as the angle of the Li–O–M edge-sharing affect the diffusion barrier most significantly. These results were further supported by the neural network model trained on similar systems. Recently, they took a step further and developed a Bayesian-driven approach to screen 318 favorite structures.[196] Using a Gaussian process model, they used only 5 structures as a starting point and inferred which composition should be computed next. In this manner, they were able to minimize the cost during the exploration of the compositional space for property calculations. Similar to the initial model built by Jalem et al. Sendek et al. developed a logistic regression model to predict whether material is superionic or not base on existing ionic conductivity data.[197] They also used a large set of potential features including the bond iconicity, the anion coordination number, and the Li–Li distances. Despite the small number of data points used for training, the results have been cross-verified. Using the model, they screened 12,831 materials and found 21 potential superionic conductors. Fujimura and co-workers used a support vector machine to fit the non-linear mapping between the structural features and the ionic conductivity of ionic conductors.[198] Specifically, they performed molecular dynamics simulations using DFT-based force engines on LISICON-type structures to construct the training set instead of using experimental data. It is important to note that the above models only use a small number of data points to train the model. Therefore, generalizability is a potential issue, especially for artificial neural network models. Low-grade data from the empirical method have been generated to saturate the training set. For example, the bond valence approach has been used to generate Li migration barriers in 400 compounds. These data were used to train regression models.[199] While the methodological development is plausible, the practical usage of such a model is questionable because the bond valence model itself is fast enough for screening purposes.

In terms of electrode materials, apart from the diffusion properties, the redox potential and the specific capacity are also critical parameters to look at. They determine the upper bound of the energy density of a battery. In this context, high-throughput screening via both the database approach and the crystal structure prediction approaches has been carried out. Chen and co-workers performed high-throughput screening using structure templates from a family of carbonophosphate.[200] By substituting elements and carry out energy and property calculations, they found a novel carbonophosphate compound. Through experimental synthesis and electrochemical tests, these materials are confirmed. Despite the relatively low energy density, this study shows the capability of computational discovery in the area of batteries. Despite the success, the screening approach based on the existing database and element substitution is inherently bottlenecked by its predefined structural space. Crystal structure prediction methods are useful in this context. Lu et al. recently showed that via a random sampling approach using AIRSS, a number of important cathode materials can be re-identified.[92] Beyond the application aspect, they also showed how to bias the search to enhance the search efficiency. It is found that the distance between atoms, the symmetry of the generated random structures, and the information on structural units need to be incorporated to boost the speed of search. They also made predictions using the AIRSS-based cathode searching framework, as shown in Fig. 8. In particular, they found a number of polymorphs of the $LiTM(C_2O_4)_2$ oxalates which have the potential to serve as high-rate, energy-dense, and cheap cathode materials. Their study remarks a significant step forward to saturating the structural space for materials discovery in the field of energy research, which is important for the later development of training sets for machine learning.

### 5.2. Catalysts

Catalysts, especially electrocatalysts are crucial to the development of a number of energy-related technologies including fuel cells, electrolyzers for hydrogen generation, and fuel converter for $CO_2$ reduction. These catalysts can modify the kinetics of the electrochemical reaction and facilitate not only the efficiency of these devices but also increase the selectivity when multiple reactions are involved.

The electronic structure of a bulk catalyst is known to be correlated to its catalytic performance. The most widely used electronic descriptor in this sense is the center of the d electronic band. For transition metals, such a value is correlated to the bonding between the surface and the

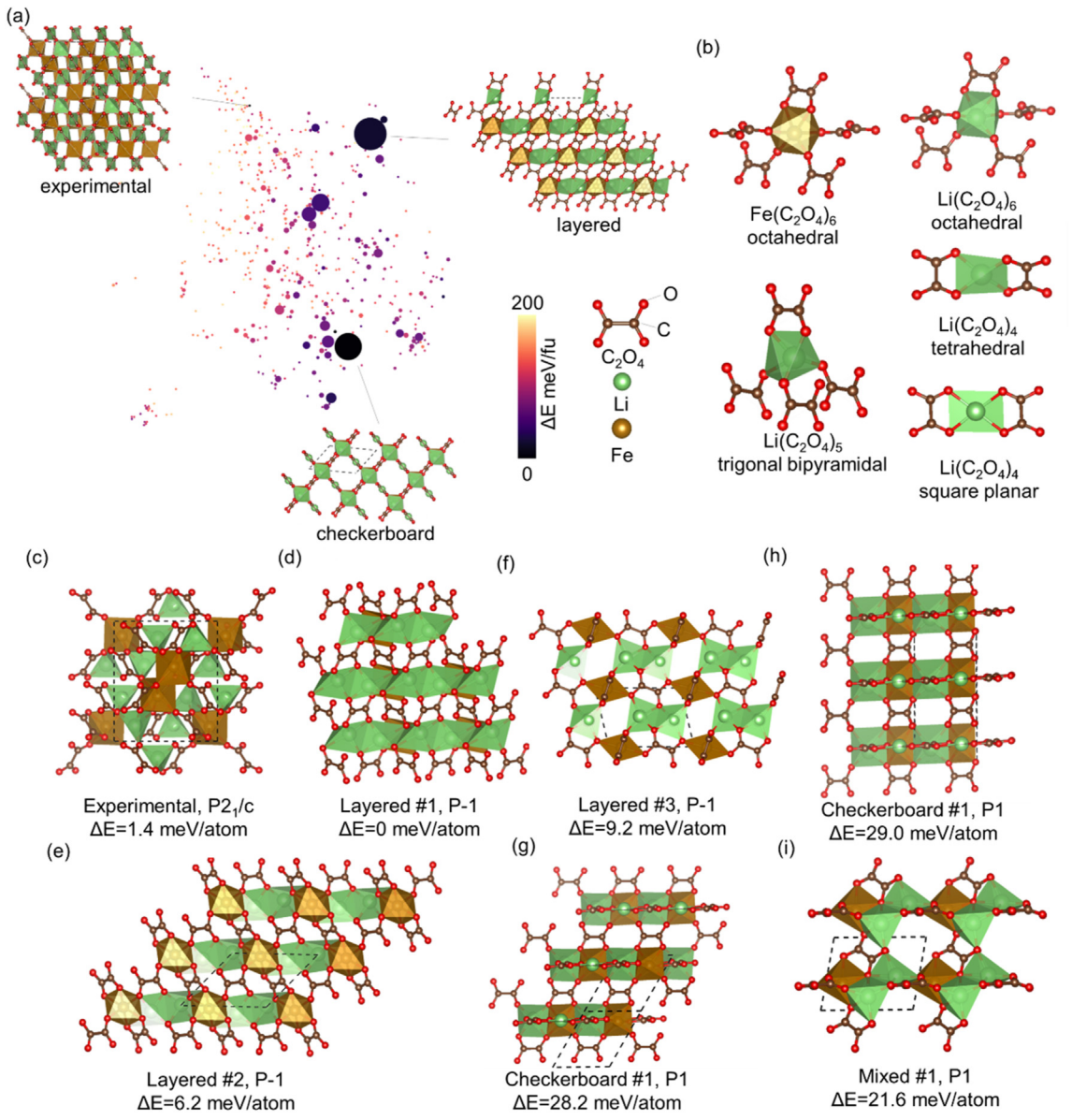



**Fig. 8.** Computational discovery of novel oxalate-based cathode materials for LIBs. (a) SHEAP map of AIRSS search results for $Li_2Fe(C_2O_4)_2$. (b) Ternary slice of the Li–Fe–C–O phase diagram containing the decomposition products of $Li_2Fe(C_2O_4)_2$. (c) Structural density of states of the AIRSS search results for $Li_2TM(C_2O_4)_2$ where TM = Fe, Co, Ni, V, and Mn. (d)–(i) Structures of the low energy polymorphs of $Li_2Fe(C_2O_4)_2$. Reproduced with permission from Ref. 92.

adsorbents. Unfortunately, the computation of the d band usually involves time-consuming DFT modeling. Direct mapping between the structures with the d band center to bypass such rate-limiting step is therefore favorable.[201,202] Takigawa and co-workers used a regression model to capture the non-linear relation using the elemental types as the feature.[203] They were able to achieve a relatively small root mean squared error of 0.5 eV in the predicted d-center values. Similarly, Niu et al. fitted a non-linear relation between the adsorption energy of *OH on defective g-$C_3N_4$ with the structure. The adsorption energy was chosen as a descriptor as the catalytic activity. Their model is highly efficient and achieves a better description of the structural-property relation compared with the linear model.[204]

Apart from the bulk properties, the catalytic process actually happens at the surface of the catalyst. Therefore, for a physical model, the adsorption energy of adsorbents needs to be calculated. Unfortunately, the potential surface termination and the active sites lead to a large number of combinations. Modeling such becomes a time-consuming step using DFT. In this context, a direct mapping between the electronic structure and the adsorption energy is possible using a machine learning model, with a typical example shown in Fig. 9. Using artificial neuron networks, Ma and co-workers fitted the correlation between the atom-projected electronic structure calculated using DFT and the adsorption energy of CO molecule on metal alloys.[205] They were able to achieve a better predictivity compared with the d-band center model. Similarly, the adsorption energy of other molecular species such as OH on a number of surfaces has been modeled.[206–208] Beyond these, a further step to give a more detailed description of the catalyst system but with the manageable computational cost is to develop machine learning-based forcefields. Neural network potentials have been developed in this context. These potentials mimic the PES of the system using artificial neural networks which can capture highly non-linear and high-dimensional functions. Surfaces of Pd, Ga–Ni alloys, Au, Au–Fe alloys have been modeled for CO reduction and a number of other applications.[209–211]

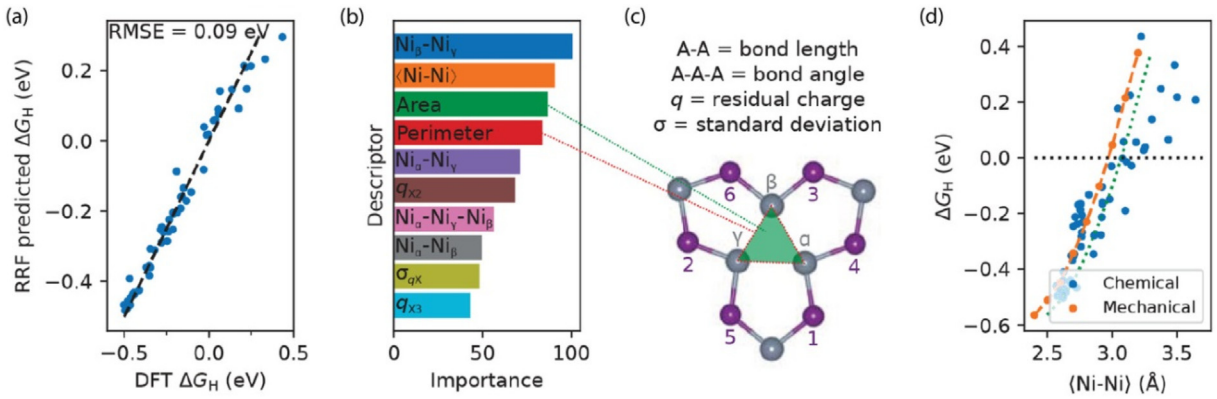



**Fig. 9.** (a) Parity plot of H binding free energy by regularized random forests and DFT. (b) Top 10 feature importance of descriptors obtained from RRF models. (c) The geometry of $Ni_3$-hollow site and the descriptor visualization. (d) The correlation between average Ni–Ni bond distance and the H binding free energy induced by chemical and mechanical pressure. Reproduced with permission from Ref. 201.

### 5.3. Photovoltaics

Solar cells are photoelectrical devices that harvest the energy of photons and convert it to electricity. Solar energy is a clean energy source and therefore is critical to the reduction of $CO_2$ emission. In a solar cell, the most critical component is arguably the semiconducting light absorber layer. In fact, the three generations of photovoltaics are defined according to the absorber material. Si-based cells are the first generation. These cells are now cheap to produce but the efficiency is limited. The second generation is based on $CuInGaS_2$ (CIGS) and related compounds. These cells can be made through fully integrated thin-film methods and therefore can be flexible and used in multiple scenarios.[212] Also, the efficiency of such cells can be made higher when coupled with other types of solar cells as the band structure of CIGS can be tuned by modifying its composition.[213] The deficiency of such technology is the cost - not only the Indium sources are limited but also the thin-film sputtering process is time-consuming and equipment-relying. The third-generation solar cells are based on perovskite materials. In particular, solar cells in single-junction architectures based on organic-inorganic perovskites have risen quickly from 3.8% in 2009 to 25.5% in 2021.[17,214] These cells are extremely promising if the stability issue can be resolved.

For single-junction cells, the Shockley-Queisser limit state with the maximum efficiency of 33% can be reached when the band gap of the absorber layer is 1.34 eV.[215,216] Since the band gap is a direct output of plane-wave DFT, one can in principle use high-throughput screening to search for potential materials. Despite the relatively fast speed of calculation of band gaps using semi-local functionals such as PBE, people have tried to further accelerate such a process by mapping these quantities using machine learning models. Structures with predefined templates such as the Ruddlesden–Popper perovskites and hybrid perovskites have been modeled using models including neural networks and gradient boosting regression.[217,218] For example, Lu and co-workers developed a data driven-method to discover stable Pb-free hybrid perovskites by training a gradient boosting regression model on 212 reported band gap values (see Fig. 10).[218] They used the model to predict the band gap of 5158 new perovskites. By further screening their stability, 6 stable promising absorbers are selected as candidates. Similar approaches have been taken to explore other chemical/structural systems.[219,220] In these models, due to the predefined structure templates, the input features can be simplified into simple vectors. However, this also means the model can only be applied in a limited space. In this context, general descriptions such as SOAP and graph-based descriptors have been developed.[145,221] Using these descriptors, large databases such as the Materials Projects have been used as training sets and the models obtained are applicable to wide structural and composition space.

A major problem of these studies is the low quality of the band gaps in the training set. It is well-known that the PBE-based gaps are systematically smaller than the actual value. Hybrid functionals and GW calculation can improve the accuracy of the band gap prediction.[222–224] However, these methods are significantly more expensive and almost impossible to be used for large-scale screening with the computational power to date. Nevertheless, the application of such on small subsets of materials is possible. For example, Agiorgousis and co-workers computed 220 double perovskites using hybrids and trained the results using random forests and support vector machines.[225] Recently, several groups have developed multi-fidelity models to resolve the conflict between the

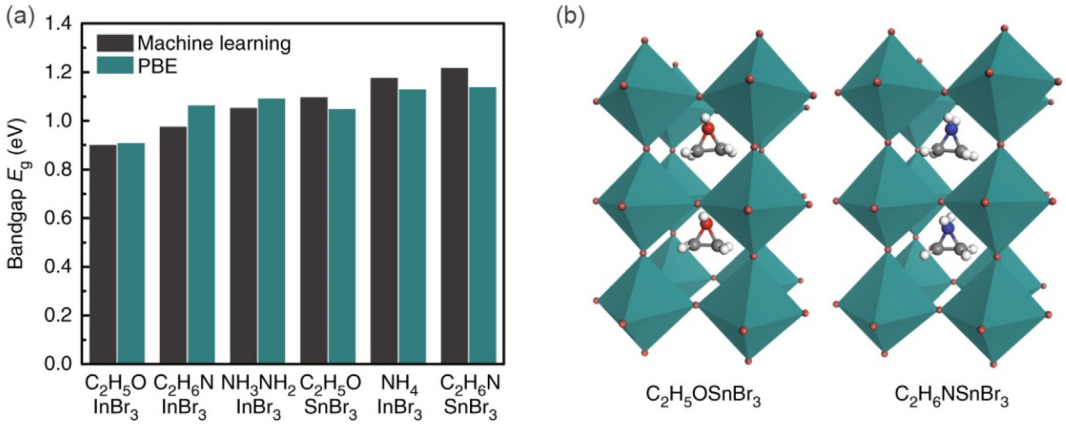



**Fig. 10.** Comparison of the machine learning predictions with DFT calculations. (a) A comparison between the predicted and the DFT-calculated results of 6 selected perovskites. (b) Optimized structures of typical perovskites. Reproduced with permission from Ref. 218.

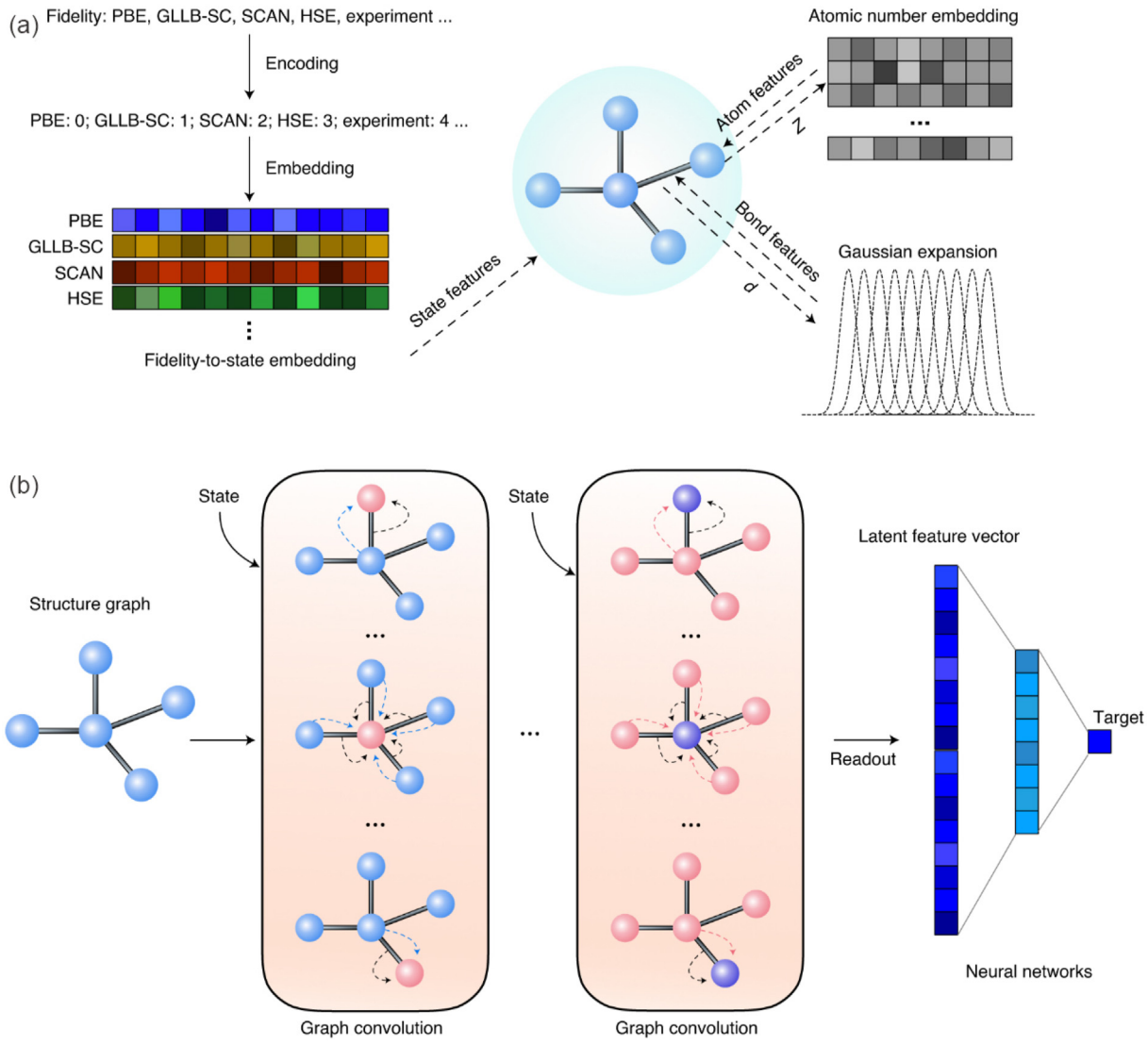



**Fig. 11.** Illustration of the multi-fidelity graph network model. (a) Representation of material in a graph network model with the fidelity of each data is encoded as an integer. (b) Construction of a materials graph network model. Reproduced with permission from Ref. 140.

quality of the data and the computational speed.[140,226] Chen et al. developed a multi-fidelity graph network model as a universal tool to make accurate predictions of materials properties with relatively small data size, see Fig. 11[140] Especially, they used the low-quality but large-quantity PBE band gaps to enhance the resolution of the latent structural features and increased the accuracy of prediction on experimental band gap errors. A decrease in the mean absolute error of up to 45% is achieved.

## 6. Conclusions and perspectives

Computational discovery of novel materials has now become an indispensable part of the research in the area of energy devices. Many examples of success have proved that these techniques are invaluable tools for accelerating the kinetics of future design of energy devices. Despite the success, efforts still need to be made. Especially, we are now in the process of a paradigm shift from the 2nd-gen high-throughput screening to the more robust 3rd-gen machine learning-based methods. This shift needs us to better integrate the new developments in the artificial intelligence field into the materials science and compositional chemistry community. Several basic issues pressingly need to be resolved. A universal, computationally efficient, one-to-one corresponding and symmetry invariant representation needs to be invented to enable computers to understand our materials world. Databases with better record quality, larger quantity, wider coverage, and evener distribution across the entire material space need to be constructed for better training sets. New algorithms for the inverse material design displayed better robustness and better trainability need to be developed.

Now, we already stepped our first step into the new era of materials discovery and we are witnessing an avalanche of development of new ideas and novel methods. Hopefully, with a hand-in-hand collaboration between the computer scientists, the computational chemists, the physicists, the materials scientists, and the energy researchers, we will be able to give a firm answer to the outstanding question: computational materials discovery - it has become a reality.[227]

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The author thanks Dr. Chi Chen and Dr. Zhaofu Zhang for the helpful discussions.

## References

1. Semieniuk G, Taylor L, Rezai A, et al. Plausible energy demand patterns in a growing global economy with climate policy. *Nat Clim Change*. 2021:1–6.
2. Newell RG, Raimi D. *Global Energy Outlook Comparison Methods: 2020 Update*. Resources for the Future; 2020.
3. Famprikis T, Canepa P, Dawson JA, et al. Fundamentals of inorganic solid-state electrolytes for batteries. *Nat Mater*. 2019;18:1278–1291.
4. Manthiram A, Yu X, Wang S. Lithium battery chemistries enabled by solid-state electrolytes. *Nat. Rev. Mater.* 2017;2:1–16.
5. Huang KJ, Ceder G, Olivetti EA. *Manufacturing Scalability Implications of Materials Choice in Inorganic Solid-State Batteries*. 2021. Joule.
6. Xiao Y, Wang Y, Bo S-H, et al. Understanding interface stability in solid-state batteries. *Nat. Rev. Mater.* 2020;5:105–126.
7. Kodama K, Nagai T, Kuwaki A, et al. Challenges in applying highly active Pt-based nanostructured catalysts for oxygen reduction reactions to fuel cell vehicles. *Nat Nanotechnol*. 2021:1–8.
8. Geerlings P, De Proft F, Langenaeker W. Conceptual density functional theory. *Chem Rev*. 2003;103:1793–1874.
9. Hohenberg P, Kohn W. Inhomogeneous electron gas. *Phys Rev*. 1964;136:B864.
10. Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. *Phys Rev*. 1965;140:A1133.
11. Lundstrom M. Moore's law forever? *Science*. 2003;299:210–211.
12. Gonze X, Amadon B, Anglade P-M, et al. ABINIT: first-principles approach to material and nanosystem properties. *Comput Phys Commun*. 2009;180:2582–2615.
13. Segall M, Lindan PJ, Probert Ma, et al. First-principles simulation: ideas, illustrations and the CASTEP code. *J Phys Condens Matter*. 2002;14:2717.
14. Hafner J. Ab-initio simulations of materials using VASP: density-functional theory and beyond. *J Comput Chem*. 2008;29:2044–2078.
15. Pollice R, dos Passos Gomes G, Aldeghi M, et al. Data-driven strategies for accelerated materials design. *Accounts Chem Res*. 2021;54:849–860.
16. Horton M, Dwaraknath S, Persson K. Promises and perils of computational materials databases. *Nature Computational Science*. 2021;1:3–5.
17. Kojima A, Teshima K, Shirai Y, et al. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *J Am Chem Soc*. 2009;131:6050–6051.
18. Yin W-J, Yang J-H, Kang J, et al. Halide perovskite materials for solar cells: a theoretical review. *J Mater Chem*. 2015;3:8926–8942.
19. Woodley SM, Catlow R. Crystal structure prediction from first principles. *Nat Mater*. 2008;7:937–946.
20. Oganov AR, Pickard CJ, Zhu Q, et al. Structure prediction drives materials discovery. *Nat. Rev. Mater.* 2019;4:331–348.
21. Liu Y, Zhao T, Ju W, et al. Materials discovery and design using machine learning. *Journal of Materiomics*. 2017;3:159–177.
22. Lu W, Xiao R, Yang J, et al. Data mining-aided materials discovery and optimization. *Journal of materiomics*. 2017;3:191–201.
23. Manthiram A. A reflection on lithium-ion battery cathode chemistry. *Nat Commun*. 2020;11:1–9.
24. Xie J, Lu Y-C. A retrospective on lithium-ion batteries. *Nat Commun*. 2020;11:1–4.
25. Whittingham MS. Electrical energy storage and intercalation chemistry. *Science*. 1976;192:1126–1127.
26. Goodenough JB. Metallic oxides. *Prog Solid State Chem*. 1971;5:145–399.
27. He J, Dettelbach KE, Salvatore DA, et al. High-throughput synthesis of mixed-metal electrocatalysts for $CO_2$ reduction. *Angew Chem Int Ed*. 2017;56:6068–6072.
28. Nursam NM, Wang X, Caruso RA. High-throughput synthesis and screening of titania-based photocatalysts. *ACS Comb Sci*. 2015;17:548–569.
29. Pyzer-Knapp EO, Suh C, Gómez-Bombarelli R, et al. What is high-throughput virtual screening? A perspective from organic materials discovery. *Annu Rev Mater Res*. 2015;45:195–216.
30. Hautier G. Finding the needle in the haystack: materials discovery and design through computational ab initio high-throughput screening. *Comput Mater Sci*. 2019;163:108–116.
31. Lang PT, Kuntz ID, Maggiora GM, et al. Evaluating the high-throughput screening computations. *J Biomol Screen*. 2005;10:649–652.
32. Curtarolo S, Hart GL, Nardelli MB, et al. The high-throughput highway to computational materials design. *Nat Mater*. 2013;12:191–201.
33. Luo S, Li T, Wang X, et al. High-throughput computational materials screening and discovery of optoelectronic semiconductors. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2021;11:e1489.
34. Bleicher KH, Böhm H-J, Müller K, et al. Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov*. 2003;2:369–378.
35. Cheng L, Assary RS, Qu X, et al. Accelerating electrolyte discovery for energy storage with high-throughput screening. *J Phys Chem Lett*. 2015;6:283–291.
36. Dudek AZ, Arodz T, Gálvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb Chem High Throughput Screen*. 2006;9:213–228.
37. Greeley J, Jaramillo TF, Bonde J, et al. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat Mater*. 2006;5:909–913.
38. Giannozzi P, Baroni S, Bonini N, et al. Quantum ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J Phys Condens Matter*. 2009;21:395502.
39. Frisch MJ, Trucks GW, Schlegel HB, et al. *Gaussian 16*. Wallingford, CT: Rev. C.01; 2016.
40. Neese F. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2012;2:73–78.
41. Belsky A, Hellenbrandt M, Karen VL, et al. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr Sect B Struct Sci*. 2002;58:364–369.
42. Jain A, Hautier G, Moore CJ, et al. A high-throughput infrastructure for density functional theory calculations. *Comput Mater Sci*. 2011;50:2295–2310.
43. Jain A, Hautier G, Ong SP, et al. Formation enthalpies by mixing GGA and GGA+ U calculations. *Phys Rev B*. 2011;84, 045115.
44. Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. *Apl Mater*. 2016;4, 053208.
45. Rajan K. Materials informatics: the materials "gene" and big data. *Annu Rev Mater Res*. 2015;45:153–169.
46. Takahashi K, Tanaka Y. Materials informatics: a journey towards material design and synthesis. *Dalton Trans*. 2016;45:10497–10499.
47. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529:484–489.
48. Li Z, Ma X, Xin H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catal Today*. 2017;280:232–238.
49. Kalidindi SR. Feature engineering of material structure for AI-based materials knowledge systems. *J Appl Phys*. 2020;128, 041103.
50. Unke OT, Chmiela S, Sauceda HE, et al. *Machine Learning Force Fields*. 2020. arXiv preprint arXiv:2010.07067.
51. Kim S, Noh J, Gu GH, et al. Generative adversarial networks for crystal structure prediction. *ACS Cent Sci*. 2020;6:1412–1420.
52. Debe MK. Electrocatalyst approaches and challenges for automotive fuel cells. *Nature*. 2012;486:43–51.
53. Huang ZF, Wang J, Peng Y, et al. Design of efficient bifunctional oxygen reduction/evolution electrocatalyst: recent advances and perspectives. *Adv. Energy Mater.* 2017;7:1700544.
54. Nørskov JK, Rossmeisl J, Logadottir A, et al. Origin of the overpotential for oxygen reduction at a fuel-cell cathode. *J Phys Chem B*. 2004;108:17886–17892.
55. Seh ZW, Kibsgaard J, Dickens CF, et al. Combining theory and experiment in electrocatalysis: insights into materials design. *Science*. 2017;355.
56. Hammer B, Nørskov JK. Theoretical surface science and catalysis—calculations and concepts. *Adv Catal*. 2000;45:71–129.
57. Hinnemann B, Moses PG, Bonde J, et al. Biomimetic hydrogen evolution: $MoS_2$ nanoparticles as catalyst for hydrogen evolution. *J Am Chem Soc*. 2005;127:5308–5309.
58. Nørskov J. Electronic factors in catalysis. *Prog Surf Sci*. 1991;38:103–144.
59. Hammer B, Norskov JK. Why gold is the noblest of all the metals. *Nature*. 1995;376:238–240.
60. Chen G, Peng Y, Zheng G, et al. Polysynthetic twinned TiAl single crystals for high-temperature applications. *Nat Mater*. 2016;15:876–881.
61. Fan L, Yang T, Zhao Y, et al. Ultrahigh strength and ductility in newly developed materials with coherent nanolamellar architectures. *Nat Commun*. 2020;11:1–8.
62. Körbel S, Marques MA, Botti S. Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations. *J Mater Chem C*. 2016;4:3157–3167.
63. Huo Z, Wei S-H, Yin W-J. High-throughput screening of chalcogenide single perovskites by first-principles calculations for photovoltaics. *J Phys Appl Phys*. 2018;51:474003.
64. Chen S, Hou Y, Chen H, et al. Exploring the stability of novel wide bandgap perovskites by a robot based high throughput approach. *Adv. Energy Mater.* 2018;8:1701543.
65. Pickard CJ, Needs R. Ab initio random structure searching. *J Phys Condens Matter*. 2011;23, 053201.
66. Wang Y, Lv J, Zhu L, et al. Crystal structure prediction via particle-swarm optimization. *Phys Rev B*. 2010;82, 094116.
67. Jain A, Ong SP, Hautier G, et al. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. *Apl Mater*. 2013;1, 011002.
68. Togo A, Tanaka I. First principles phonon calculations in materials science. *Scripta Mater*. 2015;108:1–5.
69. Grimme S, Hansen A, Brandenburg JG, et al. Dispersion-corrected mean-field electronic structure methods. *Chem Rev*. 2016;116:5105–5154.
70. Svane A, Gunnarsson O. Transition-metal oxides in the self-interaction–corrected density-functional formalism. *Phys Rev Lett*. 1990;65:1148.
71. Anisimov VI, Zaanen J, Andersen OK. Band theory and mott insulators: hubbard U instead of stoner I. *Phys Rev B*. 1991;44:943.
72. Schreiner PR. Relative energy computations with approximate density functional theory—a caveat!. *Angew Chem Int Ed*. 2007;46:4217–4219.
73. Liechtenstein A, Anisimov VI, Zaanen J. Density-functional theory and strong interactions: orbital ordering in Mott-Hubbard insulators. *Phys Rev B*. 1995;52:R5467.
74. Talirz L, Kumbhar S, Passaro E, et al. Materials Cloud, a platform for open computational science. *Scientific data*. 2020;7:1–12.
75. Saal JE, Kirklin S, Aykol M, et al. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM (J Occup Med)*. 2013;65:1501–1509.

76. Curtarolo S, Setyawan W, Hart GL, et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comput Mater Sci*. 2012;58:218–226.
77. Choudhary K, Garrity KF, Reid AC, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*. 2020;6:1–13.
78. Stevanović V, Lany S, Zhang X, et al. Correcting density functional theory for accurate predictions of compound enthalpies of formation: fitted elemental-phase reference energies. *Phys Rev B*. 2012;85:115104.
79. Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. *Phys Rev Lett*. 1996;77:3865.
80. Ernzerhof M, Scuseria GE. Assessment of the perdew–burke–ernzerhof exchange-correlation functional. *J Chem Phys*. 1999;110:5029–5036.
81. Wang L, Maxisch T, Ceder G. Oxidation energies of transition metal oxides within the GGA+ U framework. *Phys Rev B*. 2006;73:195107.
82. Zhou J, Shen L, Costa MD, et al. 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Scientific data*. 2019;6:1–10.
83. Choudhary K, Kalish I, Beams R, et al. High-throughput identification and characterization of two-dimensional materials using density functional theory. *Sci Rep*. 2017;7:1–16.
84. Mounet N, Gibertini M, Schwaller P, et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat Nanotechnol*. 2018;13:246–252.
85. Li X, Zhang Z, Yao Y, et al. High throughput screening for two-dimensional topological insulators. *2D Mater*. 2018;5, 045023.
86. Jin L, Zhang X, Dai X, et al. Screening topological materials with a CsCl-type structure in crystallographic databases. *IUCrJ*. 2019;6:688–694.
87. Borysov SS, Geilhufe RM, Balatsky AV. Organic materials database: an open-access online database for data mining. *PloS One*. 2017;12, e0171501.
88. Mao X, Sun L, Wu T, et al. First-principles screening of all-inorganic lead-free ABX3 perovskites. *J Phys Chem C*. 2018;122:7670–7675.
89. Oganov AR, Lyakhov AO, Valle M. How evolutionary crystal structure prediction worksand why. *Accounts Chem Res*. 2011;44:227–237.
90. Wang Y, Ma Y. Perspective: crystal structure prediction at high pressures. *J Chem Phys*. 2014;140, 040901.
91. Lv J, Wang Y, Zhu L, et al. Predicted novel high-pressure phases of lithium. *Phys Rev Lett*. 2011;106, 015503.
92. Lu Z, Zhu B, Shires BW, et al. *Ab Initio Random Structure Searching for Battery Cathode Materials*. 2021. arXiv preprint arXiv:2104.00441.
93. Doye JP, Wales DJ, Miller MA. Thermodynamics and the global optimization of Lennard-Jones clusters. *J Chem Phys*. 1998;109:8143–8153.
94. Massen CP, Doye JP. Identifying communities within energy landscapes. *Phys Rev*. 2005;71, 046101.
95. Massen CP, Doye JP. Power-law distributions for the areas of the basins of attraction on a potential energy landscape. *Phys Rev*. 2007;75, 037101.
96. Behler J. Perspective: machine learning potentials for atomistic simulations. *J Chem Phys*. 2016;145:170901.
97. Keith JA, Vassilev-Galindo V, Cheng B, et al. *Combining Machine Learning and Computational Chemistry for Predictive Insights into Chemical Systems*. 2021. arXiv preprint arXiv:2102.06321.
98. Kvashnin AG, Oganov AR, Samtsevich AI, et al. Computational search for novel hard chromium-based materials. *J Phys Chem Lett*. 2017;8:755–764.
99. Stillinger FH. Exponential multiplicity of inherent structures. *Phys Rev*. 1999;59:48.
100. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput*. 1997;1:67–82.
101. Oganov AR, Valle M. How to quantify energy landscapes of solids. *J Chem Phys*. 2009;130:104504.
102. Wales DJ. Symmetry, near-symmetry and energetics. *Chem Phys Lett*. 1998;285:330–336.
103. Floudas CA, Gounaris CE. A review of recent advances in global optimization. *J Global Optim*. 2009;45:3.
104. Pannetier J, Bassas-Alsina J, Rodriguez-Carvajal J, et al. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature*. 1990;346:343–345.
105. Schön JC, Jansen M. First step towards planning of syntheses in solid-state chemistry: determination of promising structure candidates by global optimization. *Angew Chem Int Ed Engl*. 1996;35:1286–1304.
106. Wales DJ, Doye JP. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J Phys Chem*. 1997;101:5111–5116.
107. Martoňák R, Laio A, Parrinello M. Predicting crystal structures: the Parrinello-Rahman method revisited. *Phys Rev Lett*. 2003;90, 075503.
108. Goedecker S. Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J Chem Phys*. 2004;120:9911–9917.
109. Oganov AR, Glass CW. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J Chem Phys*. 2006;124:244704.
110. Deaven DM, Ho K-M. Molecular geometry optimization with a genetic algorithm. *Phys Rev Lett*. 1995;75:288.
111. Call ST, Zubarev DY, Boldyrev AI. Global minimum structure searches via particle swarm optimization. *J Comput Chem*. 2007;28:1177–1186.
112. Lonie DC, Zurek E. XtalOpt: an open-source evolutionary algorithm for crystal structure prediction. *Comput Phys Commun*. 2011;182:372–387.
113. Tipton WW, Hennig RG. A grand canonical genetic algorithm for the prediction of multi-component phase diagrams and testing of empirical potentials. *J Phys Condens Matter*. 2013;25:495401.
114. Judson RS, Jaeger EP, Treasurywala AM, et al. Conformational searching methods for small molecules. II. Genetic algorithm approach. *J Comput Chem*. 1993;14:1407–1414.
115. Bush T, Catlow CRA, Battle P. Evolutionary programming techniques for predicting inorganic crystal structures. *J Mater Chem*. 1995;5:1269–1272.
116. Curtis F, Li X, Rose T, et al. GAtor: a first-principles genetic algorithm for molecular crystal structure prediction. *J Chem Theor Comput*. 2018;14:2246–2264.
117. Gubernatis J, Lookman T. Machine learning in materials design and discovery: examples from the present and suggestions for the future. *Physical Review Materials*. 2018;2:120301.
118. Chen C, Zuo Y, Ye W, et al. A critical review of machine learning of energy materials. *Adv. Energy Mater*. 2020;10, 1903242.
119. Schleder GR, Padilha AC, Acosta CM, et al. From DFT to machine learning: recent approaches to materials science–a review. *J Phys: Materials*. 2019;2, 032001.
120. Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr Sect B Struct Sci*. 2002;58:380–388.
121. Villars P, Onodera N, Iwata S. The Linus Pauling file (LPF) and its application to materials design. *J Alloys Compd*. 1998;279:1–7.
122. Reymond J-L. The chemical space project. *Accounts Chem Res*. 2015;48:722–730.
123. Curtarolo S, Setyawan W, Wang S, et al. AFLOWLIB. ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comput Mater Sci*. 2012;58:227–235.
124. Kamaya N, Homma K, Yamakawa Y, et al. A lithium superionic conductor. *Nat Mater*. 2011;10:682–686.
125. Zou Z, Li Y, Lu Z, et al. Mobile ions in composite solids. *Chem Rev*. 2020;120:4169–4221.
126. Zhao Q, Stalin S, Zhao C-Z, et al. Designing solid-state electrolytes for safe, energy-dense batteries. *Nat. Rev. Mater*. 2020;5:229–252.
127. Isayev O, Fourches D, Muratov EN, et al. Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chem Mater*. 2015;27:735–743.
128. Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials*. 2018;4:1–8.
129. Niu S, Li S, Du Y, et al. How to reliably report the overpotential of an electrocatalyst. *ACS Energy Lett*. 2020;5:1083–1087.
130. Zheng X, Ji Y, Tang J, et al. Theory-guided Sn/Cu alloying for efficient CO 2 electroreduction at low overpotentials. *Nature Catalysis*. 2019;2:55–61.
131. Zhu Y, Zhou W, Sunarso J, et al. Phosphorus-doped perovskite oxide as highly efficient water oxidation electrocatalyst in alkaline solution. *Adv Funct Mater*. 2016;26:5862–5872.
132. Zhu Y, Zhou W, Chen Y, et al. A high-performance electrocatalyst for oxygen evolution reaction: LiCo0. 8Fe0. 2O2. *Adv Mater*. 2015;27:7150–7155.
133. Zhu Y, Zhou W, Yu J, et al. Enhancing electrocatalytic activity of perovskite oxides by tuning cation deficiency for oxygen reduction and evolution reactions. *Chem Mater*. 2016;28:1691–1697.
134. Zhu Y, Chen G, Xu X, et al. Enhancing electrocatalytic activity for hydrogen evolution by strongly coupled molybdenum nitride@ nitrogen-doped carbon porous nano-octahedrons. *ACS Catal*. 2017;7:3540–3547.
135. Zhu Y, Zhou W, Zhong Y, et al. A perovskite nanorod as bifunctional electrocatalyst for overall water splitting. *Adv. Energy Mater*. 2017;7, 1602122.
136. Zhu Y, Zhou W, Chen ZG, et al. SrNb0. 1Co0. 7Fe0. 2O3− δ perovskite as a next-generation electrocatalyst for oxygen evolution in alkaline solution. *Angew Chem*. 2015;127:3969–3973.
137. Zhu Y, Zhou W, Ran R, et al. Promotion of oxygen reduction by exsolved silver nanoparticles on a perovskite scaffold for low-temperature solid oxide fuel cells. *Nano Lett*. 2016;16:512–518.
138. Zhu Y, Zhou W, Shao Z. Perovskite/carbon composites: applications in oxygen electrocatalysis. *Small*. 2017;13, 1603793.
139. Jain M, Chelikowsky JR, Louie SG. Reliability of hybrid functionals in predicting band gaps. *Phys Rev Lett*. 2011;107:216806.
140. Chen C, Zuo Y, Ye W, et al. Learning properties of ordered and disordered materials from multi-fidelity data. *Nature Computational Science*. 2021;1:46–53.
141. Ong SP, Richards WD, Jain A, et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput Mater Sci*. 2013;68:314–319.
142. Pizzi G, Cepellotti A, Sabatini R, et al. AiiDA: automated interactive infrastructure and database for computational science. *Comput Mater Sci*. 2016;111:218–230.
143. Mathew K, Montoya JH, Faghaninia A, et al. Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows. *Comput Mater Sci*. 2017;139:140–152.
144. Larsen AH, Mortensen JJ, Blomqvist J, et al. The atomic simulation environment—a Python library for working with atoms. *J Phys Condens Matter*. 2017;29:273002.
145. Bartók AP, Kondor R, Csányi G. On representing chemical environments. *Phys Rev B*. 2013;87:184115.
146. Cheng B, Griffiths R-R, Wengert S, et al. Mapping materials and molecules. *Accounts Chem Res*. 2020;53:1981–1991.
147. Haghighatlari M, Li J, Heidar-Zadeh F, et al. *Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods*. 2020. Chem.
148. Himanen L, Jäger MO, Morooka EV, et al. DScribe: library of descriptors for machine learning in materials science. *Comput Phys Commun*. 2020;247:106949.
149. Von Lilienfeld OA, Ramakrishnan R, Rupp M, et al. Fourier series of atomic radial distribution functions: a molecular fingerprint for machine learning models of quantum chemical properties. *Int J Quant Chem*. 2015;115:1084–1093.
150. Neese F. An improvement of the resolution of the identity approximation for the formation of the Coulomb matrix. *J Comput Chem*. 2003;24:1740–1747.

151. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*. 1988;28:31–36.
152. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci*. 1989;29:97–101.
153. Hansen K, Biegler F, Ramakrishnan R, et al. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J Phys Chem Lett*. 2015;6:2326–2331.
154. Huang B, Von Lilienfeld OA. *Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity*. AIP Publishing LLC; 2016.
155. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*. 2018;9:513–530.
156. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50:742–754.
157. Jalem R, Nakayama M, Noda Y, et al. A general representation scheme for crystalline solids based on Voronoi-tessellation real feature values and atomic property data. *Sci Technol Adv Mater*. 2018;19:231–242.
158. Tawfik SA, Isayev O, Spencer MJ, et al. Predicting thermal properties of crystals using machine learning. *Advanced Theory and Simulations*. 2020;3:1900208.
159. Ghiringhelli LM, Vybiral J, Levchenko SV, et al. Big data of materials science: critical role of the descriptor. *Phys Rev Lett*. 2015;114:105503.
160. Kim C, Pilania G, Ramprasad R. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. *Chem Mater*. 2016;28:1304–1311.
161. Ouyang R, Curtarolo S, Ahmetcik E, et al. SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*. 2018;2, 083802.
162. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9.
163. Tabor DP, Roch LM, Saikin SK, et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* 2018;3:5–20.
164. Butler KT, Davies DW, Cartwright H, et al. Machine learning for molecular and materials science. *Nature*. 2018;559:547–555.
165. Chen C, Baiyee ZM, Ciucci F. Unraveling the effect of La A-site substitution on oxygen ion diffusion and oxygen catalysis in perovskite BaFeO 3 by data-mining molecular dynamics and density functional theory. *Phys Chem Chem Phys*. 2015;17:24011–24019.
166. Yao Z, Sánchez-Lengeling B, Bobbitt NS, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*. 2021;3:76–86.
167. D. Schwalbe-Koda, R. Gómez-Bombarelli. Generative models for automatic chemical design, Machine Learning Meets Quantum Physics, Springer2020, pp. 445-467.
168. Chen CT, Gu GX. Generative deep neural networks for inverse materials design using backpropagation and active learning. *Advanced Science*. 2020;7:1902607.
169. Kireeva N, Pervov VS. Materials space of solid-state electrolytes: unraveling chemical composition–structure–ionic conductivity relationships in garnet-type metal oxides using cheminformatics virtual screening approaches. *Phys Chem Chem Phys*. 2017;19:20904–20918.
170. Xue D, Balachandran PV, Hogden J, et al. Accelerated search for materials with targeted properties by adaptive design. *Nat Commun*. 2016;7:1–9.
171. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
172. Abadi M, Barham P, Chen J, et al. *Tensorflow: A System for Large-Scale Machine Learning, 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016:265–283.
173. Paszke A, Gross S, Massa F, et al. *Pytorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. arXiv preprint arXiv:1912.01703.
174. Dunn A, Wang Q, Ganose A, et al. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*. 2020;6:1–10.
175. Kolb B, Lentz LC, Kolpak AM. Discovering charge density functionals and structure-property relationships with PROPhet: a general framework for coupling machine learning and first-principles methods. *Sci Rep*. 2017;7:1–9.
176. Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: generative models for matter engineering. *Science*. 2018;361:360–365.
177. Lee Y-L, Kleis J, Rossmeisl J, et al. Prediction of solid oxide fuel cell cathode activity with first-principles descriptors. *Energy Environ Sci*. 2011;4:3966–3970.
178. Guo Y, Liu Y, Oerlemans A, et al. Deep learning for visual understanding: a review. *Neurocomputing*. 2016;187:27–48.
179. Agrawal A, Choudhary A. Deep materials informatics: applications of deep learning in materials science. *MRS Communications*. 2019;9:779–792.
180. Deringer VL, Bernstein N, Csányi G, et al. Origins of structural and electronic transitions in disordered silicon. *Nature*. 2021;589:59–64.
181. Monserrat B, Brandenburg JG, Engel EA, et al. Liquid water contains the building blocks of diverse ice phases. *Nat Commun*. 2020;11:1–8.
182. Seko A. Machine learning potentials for multicomponent systems: the Ti-Al binary system. *Phys Rev B*. 2020;102:174104.
183. Bowman SR, Vilnis L, Vinyals O, et al. *Generating Sentences from a Continuous Space*. 2015. arXiv preprint arXiv:1511.06349.
184. Graves A. *Generating Sequences with Recurrent Neural Networks*. 2013. arXiv preprint arXiv:1308.0850.
185. Kingma DP, Welling M. *Auto-encoding Variational Bayes*. 2013. arXiv preprint arXiv:1312.6114.
186. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. *Generative Adversarial Networks*. 2014. arXiv preprint arXiv:1406.2661.
187. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9:1735–1780.
188. Olah C, Carter S. Attention and augmented recurrent neural networks. *Distill*. 2016;1:e1.
189. Dai H, Tian Y, Dai B, et al. *Syntax-directed Variational Autoencoder for Structured Data*. 2018. arXiv preprint arXiv:1802.08786.
190. Arjovsky M, Chintala S, Bottou L. *Wasserstein Generative Adversarial Networks, International Conference on Machine Learning*. PMLR; 2017:214–223.
191. Noh J, Kim J, Stein HS, et al. Inverse design of solid-state materials via a continuous representation. *Matter*. 2019;1:1370–1384.
192. Noh J, Gu GH, Kim S, et al. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chem Sci*. 2020;11:4871–4881.
193. Ren Z, Noh J, Tian S, et al. *Inverse Design of Crystals Using Generalized Invertible Crystallographic Representation*. 2020. arXiv preprint arXiv:2005.07609.
194. Jalem R, Aoyama T, Nakayama M, et al. Multivariate method-assisted Ab initio study of olivine-type LiMXO4 (Main Group M2+–X5+ and M3+–X4+) compositions as potential solid electrolytes. *Chem Mater*. 2012;24:1357–1364.
195. Jalem R, Kimura M, Nakayama M, et al. Informatics-aided density functional theory study on the Li ion transport of Tavorite-type LiMTO4F (M3+–T5+, M2+–T6+). *J Chem Inf Model*. 2015;55:1158–1168.
196. Jalem R, Kanamori K, Takeuchi I, et al. Bayesian-driven first-principles calculations for accelerating exploration of fast ion conductors for rechargeable battery application. *Sci Rep*. 2018;8:1–10.
197. Sendek AD, Yang Q, Cubuk ED, et al. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. *Energy Environ Sci*. 2017;10:306–320.
198. Fujimura K, Seko A, Koyama Y, et al. Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms. *Adv. Energy Mater*. 2013;3:980–985.
199. Nakayama M, Kanamori K, Nakano K, et al. Data-driven materials exploration for Li-ion conductive ceramics by exhaustive and informatics-aided computations. *Chem Rec*. 2019;19:771–778.
200. Chen H, Hautier G, Jain A, et al. Carbonophosphates: a new family of cathode materials for Li-ion batteries identified computationally. *Chem Mater*. 2012;24:2009–2016.
201. Wexler RB, Martirez JMP, Rappe AM. Chemical pressure-driven enhancement of the hydrogen evolving activity of Ni2P from nonmetal surface doping interpreted via machine learning. *J Am Chem Soc*. 2018;140:4678–4683.
202. Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for $CO_2$ reduction and $H_2$ evolution. *Nature Catalysis*. 2018;1:696–703.
203. Takigawa I, Shimizu K-i, Tsuda K, et al. Machine-learning prediction of the d-band center for metals and bimetals. *RSC Adv*. 2016;6:52587–52595.
204. Niu H, Wan X, Wang X, et al. Single-atom rhodium on defective g-C3N4: a promising bifunctional oxygen electrocatalyst. *ACS Sustainable Chem Eng*. 2021;9:3590–3599.
205. Ma X, Li Z, Achenie LE, et al. Machine-learning-augmented chemisorption model for $CO_2$ electroreduction catalyst screening. *J Phys Chem Lett*. 2015;6:3528–3533.
206. Li Z, Wang S, Chin WS, et al. High-throughput screening of bimetallic catalysts enabled by machine learning. *J Mater Chem*. 2017;5:24131–24138.
207. Gasper R, Shi H, Ramasubramaniam A. Adsorption of CO on low-energy, low-symmetry Pt nanoparticles: energy decomposition analysis and prediction via machine-learning models. *J Phys Chem C*. 2017;121:5612–5619.
208. Noh J, Back S, Kim J, et al. Active learning with non-ab initio input features toward efficient $CO_2$ reduction catalysts. *Chem Sci*. 2018;9:5152–5159.
209. Boes JR, Kitchin JR. Neural network predictions of oxygen interactions on a dynamic Pd surface. *Mol Simulat*. 2017;43:346–354.
210. Ulissi ZW, Tang MT, Xiao J, et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for $CO_2$ reduction. *ACS Catal*. 2017;7:6600–6608.
211. Chen Y, Huang Y, Cheng T, et al. Identifying active sites for $CO_2$ reduction on dealloyed gold surfaces by combining machine learning with multiscale simulations. *J Am Chem Soc*. 2019;141:11651–11657.
212. Shi S, Yao L, Ma P, et al. Recent progress in high temperature resistance PI substrate with low CTE for CIGS thin film solar cells, Mater. *Today Energy*. 2021:100640.
213. Han Q, Hsieh Y-T, Meng L, et al. High-performance perovskite/Cu (In, Ga) $Se_2$ monolithic tandem solar cells. *Science*. 2018;361:904–908.
214. Jeong J, Kim M, Seo J, et al. Pseudo-halide anion engineering for α-$FAPbI_3$ perovskite solar cells. *Nature*. 2021:1–5.
215. Shockley W, Queisser HJ. Detailed balance limit of efficiency of p-n junction solar cells. *J Appl Phys*. 1961;32:510–519.
216. Tang G, Ghosez P, Hong J. Band-edge orbital engineering of perovskite semiconductors for optoelectronic applications. *J Phys Chem Lett*. 2021;12:4227–4239.
217. Allam O, Holmes C, Greenberg Z, et al. Density functional theory–machine learning approach to analyze the bandgap of elemental halide perovskites and ruddlesden-popper phases. *ChemPhysChem*. 2018;19:2559–2565.
218. Lu S, Zhou Q, Ouyang Y, et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat Commun*. 2018;9:1–8.
219. Zhang T, Cai Z, Chen S. Chemical trends in the thermodynamic stability and band gaps of 980 halide double perovskites: a high-throughput first-principles study. *ACS Appl Mater Interfaces*. 2020;12:20680–20690.
220. Fu P, Hu S, Tang J, et al. Material exploration via designing spatial arrangement of octahedral units: a case study of lead halide perovskites. *Front Optoelectron*. 2021:1–8.

221. Chen C, Ye W, Zuo Y, et al. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater*. 2019;31:3564–3572.
222. Heyd J, Scuseria GE, Ernzerhof M. Hybrid functionals based on a screened Coulomb potential. *J Chem Phys*. 2003;118:8207–8215.
223. Paier J, Marsman M, Hummer K, et al. Screened hybrid density functionals applied to solids. *J Chem Phys*. 2006;124:154709.
224. Aryasetiawan F, Gunnarsson O. The GW method. *Rep Prog Phys*. 1998;61:237.
225. Agiorgousis ML, Sun YY, Choe DH, et al. Machine learning augmented discovery of chalcogenide double perovskites for photovoltaics. *Advanced Theory and Simulations*. 2019;2, 1800173.
226. Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput Mater Sci*. 2017;129:156–163.
227. Oganov AR, Kvashnin AG, Saleh G. *Computational Materials Discovery: Dream or Reality?*. 2018.


**Ziheng Lu** is currently a research associate at the University of Cambridge. He received his Ph.D. in 2018 from the Hong Kong University of Science and Technology and then joined the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences as a junior research scientist. In 2020, he moved to Cambridge under the support of the Faraday Institution as a Faraday Institution Research Fellow. His current research focuses on computational-aided design of materials and interfaces as well as laser-assisted synthesis of novel compounds for energy storage and conversion.