npj | Computational Materials

www.nature.com/npjcompumats

REVIEW ARTICLE OPEN

Machine learning in materials informatics: recent applications and prospects

Rampi Ramprasad[1], Rohit Batra[1], Ghanshyam Pilania[2,3], Arun Mannodi-Kanakkithodi[1,4] and Chiho Kim[1]

Propelled partly by the Materials Genome Initiative, and partly by the algorithmic developments and the resounding successes of data-driven efforts in other domains, informatics strategies are beginning to take shape within materials science. These approaches lead to surrogate machine learning models that enable rapid predictions based purely on past data rather than by direct experimentation or by computations/simulations in which fundamental equations are explicitly solved. Data-centric informatics methods are becoming useful to determine material properties that are hard to measure or compute using traditional methods—due to the cost, time or effort involved—but for which reliable data either already exists or can be generated for at least a subset of the critical cases. Predictions are typically interpolative, involving fingerprinting a material numerically first, and then following a mapping (established via a learning algorithm) between the fingerprint and the property of interest. Fingerprints, also referred to as "descriptors", may be of many types and scales, as dictated by the application domain and needs. Predictions may also be extrapolative—extending into new materials spaces—provided prediction uncertainties are properly taken into account. This article attempts to provide an overview of some of the recent successful data-driven "materials informatics" strategies undertaken in the last decade, with particular emphasis on the fingerprint or descriptor choices. The review also identifies some challenges the community is facing and those that should be overcome in the near future.

npj Computational Materials (2017)3:54 ; doi:10.1038/s41524-017-0056-5

OVERARCHING PERSPECTIVES

When a new situation is encountered, cognitive systems (including humans) have a natural tendency to make decisions based on past similar encounters. When the new situation is distinctly different from those encountered in the past, errors in judgment may occur and lessons may be learned. The sum total of such past scenarios, decisions made and the lessons learned may be viewed collectively as "experience", "intuition" or even as "common sense". Ideally, depending on the intrinsic capability of the cognitive system, its ability to make decisions should progressively improve as the richness of scenarios encountered increases.

In recent decades, the artificial intelligence (AI) and statistics communities have made these seemingly vague notions quantitative and mathematically precise.[1,2] These efforts have resulted in practical machines that learn from past experiences (or "examples"). Classic exemplars of such machine learning approaches include facial, fingerprint or object recognition systems, machines that can play sophisticated games such as chess, Go or poker, and automation systems such as in robotics or self-driving cars. In each of these cases, a large data set of past examples is required, e.g., images and their identities, configuration of pieces in a board game and the best moves, and scenarios encountered while driving and the best actions.

On the surface, it may appear as though the "data-driven" approach for determining the best decision or answer when a new situation or problem is encountered is radically different from approaches based on fundamental science in which predictions are made by solving equations that govern the pertinent phenomena. But viewed differently, is not the scientific process itself—which begins with observations, followed by intuition, then construction of a quantitative theory that explains the observations, and subsequently, refinement of the theory based on new observations—the ultimate culmination of such data-driven inquiries?

For instance, consider how the ancient people from India and Sri Lanka figured out, through persistent tinkering, the alloying elements to add to iron to impede its tendency to rust, using only their experience and creativity[3,4] (and little "steel science", which arose from this empiricism much later)—an early example of the reality and power of "chemical intuition." Or, more recently, over the last century, consider the enormously practical Hume–Rothery rules to determine the solubility tendency of one metal in another,[5] the Hall–Petch studies that have led to empirical relationships between grain sizes and mechanical strength (not just for metals but for ceramics as well),[6,7] and the group contribution approach to predict complex properties of organic and polymeric materials based just on the identity of the chemical structure,[8] all of which arose from data-driven pursuits (although they were not called as such), and later rationalized using physical principles. It would thus be fair to say that data—either directly or indirectly—drives the creation of both complex fundamental and simple empirical scientific theories. Figure 1 charts the timeline for some classic historical and diverse examples of data-driven efforts.

In more modern times, in the last decade or so, thanks to the implicit or explicit acceptance of the above notions, the "data-

[1]Department of Materials Science & Engineering and Institute of Materials Science, University of Connecticut, 97 North Eagleville Rd., Unit 3136, Storrs, CT 06269-3136, USA; [2]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany; [3]Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA and [4]Center for Nanoscale Materials, Lamont National Laboratory, 9700 S. Cass Ave., Lemont, IL 60439, USA
Correspondence: Rampi Ramprasad (rampi.ramprasad@uconn.edu)

Received: 19 July 2017 Revised: 13 November 2017 Accepted: 17 November 2017
Published online: 13 December 2017

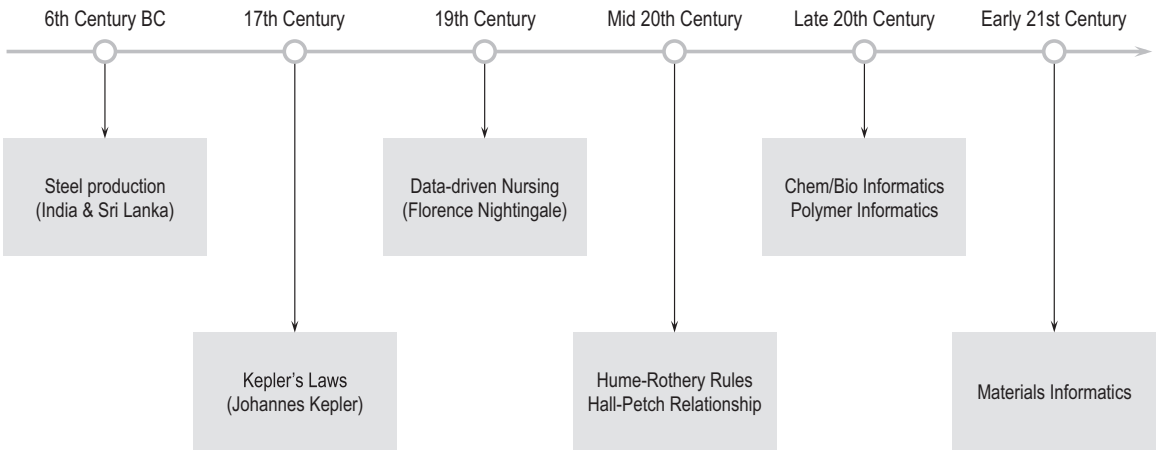

Fig. 1 Some classic historical examples of data-driven science and engineering efforts

driven", "machine learning", or "materials informatics" paradigms (with these terms used interchangeably by the community) are rapidly becoming an essential part of the materials research portfolio.[9–12] The availability of robust and trustworthy in silico simulation methods and systematic synthesis and characterization capabilities, although time-consuming and sometimes expensive, provides a pathway to generate at least a subset of the required critical data in a targeted and organized manner (e.g., via "high-throughput" experiments or computations). Indeed, such efforts are already underway, which have lead to the burgeoning of a number of enormously useful repositories such as NOMAD (http://nomad-coe.eu), Materials Project (http://materialsproject.org), Aflowlib (http://www.aflowlib.org), and OQMD (http://oqmd.org). Mining or learning from these resources or other reliable extant data can lead to the recognition of previously unknown correlations between properties, and the discovery of qualitative and quantitative rules—also referred to as surrogate models—that can be used to predict material properties orders of magnitude faster and cheaper, and with reduced human effort than required by the benchmark simulation or experimental methods utilized to create the data in the first place.

With excitement and opportunities come challenges. Questions constantly arise as to what sort of materials science problems are most appropriate for, or can benefit most from, a data-driven approach. A satisfactory understanding of this aspect is essential before one makes a decision on using machine learning methods for their problem of interest. Perhaps the most dangerous aspect of data-driven approaches is the unwitting application of machine learning models to cases that fall outside the domain of prior data. A rich and largely uncharted area of inquiry is to recognize when such a scenario ensues, and to be able to quantify the uncertainties of the machine learning predictions especially when models veer out-of-domain. Solutions for handling these perilous situations may open up pathways for adaptive learning models that can progressively improve in quality through systematic infusion of new data—an aspect critical to the further burgeoning of machine learning within the hard sciences.

This article attempts to provide an overview of some of the recent successful data-driven materials research strategies undertaken in the last decade, and identifies challenges that the community is facing and those that should be overcome in the near future.

ELEMENTS OF MACHINE LEARNING (WITHIN MATERIALS SCIENCE)

Regardless of the specific problem under study, a prerequisite for machine learning is the existence of past data. Thus, either clean, curated and reliable data corresponding to the problem under study should already be available, or an effort has to be put in place upfront for the creation of such data. An example data set may be an enumeration of a variety of materials that fall within a well-defined chemical class of interest and a relevant measured or computed property of those materials (see Fig. 2a). Within the machine learning parlance, the former, i.e., the material, is referred to as "input", and the latter, i.e., the property of interest, is referred to as the "target" or "output." A learning problem (Fig. 2b) is then defined as follows: Given a {materials → property} data set, what is the best estimate of the property for a new material not in the original data set? Provided that there are sufficient examples, i.e., that the data set is sufficiently large, and provided that the new material falls within the same chemo-structural class as the materials in the original data set, we expect that it should be possible to make such an estimate. Ideally, uncertainties in the prediction should also be reported, which can give a sense of whether the new case is within or outside the domain of the original data set.

All data-driven strategies that attempt to address the problem posed above are composed of two distinct steps, both aimed at satisfying the need for quantitative predictions. The first step is to represent numerically the various input cases (or materials) in the data set. At the end of this step, each input case would have been reduced to a string of numbers (or "fingerprints"; see Fig. 2c). This is such an enormously important step, requiring significant expertise and knowledge of the materials class and the application, i.e., "domain expertise", that we devote a separate Section to its discussion below.

The second step establishes a mapping between the fingerprinted input and the target property, and is entirely numerical in nature, largely devoid of the need for domain knowledge. Both the fingerprinting and mapping/learning steps are schematically illustrated in Fig. 2. Several algorithms, ranging from elementary (e.g., linear regression) to highly sophisticated (kernel ridge regression, decision trees, deep neural networks), are available to establish this mapping and the creation of surrogate prediction models.[13–15] While some algorithms provide actual functional forms that relate input to output (e.g., regression based schemes), others do not (e.g., decision trees). Moreover, the amount of available data may also dictate the choice of learning algorithms. For instance, tens to thousands of data points may be adequately handled using regression algorithms such as kernel ridge regression or gaussian process regression, but the availability of much larger data sets (e.g., hundreds of thousands or millions) may warrant deep neural networks, simply due to considerations of favorable scalability of the prediction models with data set size. In the above discussion, it was implicitly assumed that the target

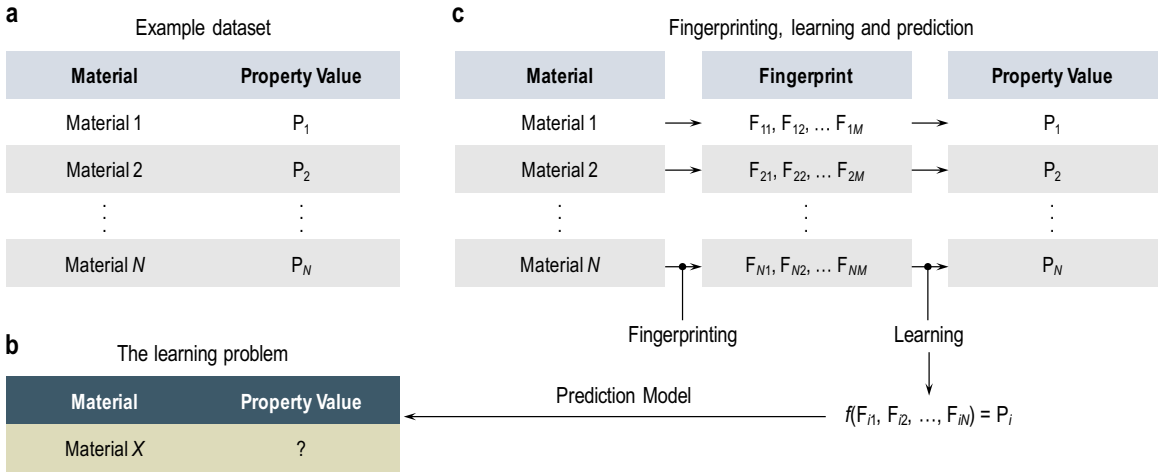

Fig. 2 The key elements of machine learning in materials science. **a** Schematic view of an example data set, **b** statement of the learning problem, and **c** creation of a surrogate prediction model via the fingerprinting and learning steps. *N* and *M* are, respectively, the number of training examples and the number of fingerprint (or descriptor or feature) components

property is a continuous quantity (e.g., bulk modulus, band gap, melting temperature, etc.). Problems can also involve discrete targets (e.g., crystal structure, specific structural motifs, etc.), which are referred to as classification problems. At this point, it is worth mentioning that the learning problem as described above for the most part involving a mapping between the fingerprints and target properties is referred to as "supervised learning"; "unsupervised learning", on the other hand, involves using just the fingerprints to recognize patterns in the data (e.g., for classification purposes or for reduction of the dimensionality of the fingerprint vector).[9,15]

Throughout the learning process, it is typical (and essential) to adhere to rigorous statistical practices. Central to this are the notions of cross-validation and testing on unseen data, which attempt to ensure that a learning model developed based on the original data set can truly handle a new case without falling prey to the perils of "overfitting".[9,15] Indeed, it should be noted here that some of the original and most successful applications of machine learning, including statistical treatments and practices such as regularization and cross-validation, were first introduced into materials research in the field of alloy theory, cluster expansions and lattice models.[16–24] These ideas, along with machine learning techniques such as compressive sensing, are further taking shape within the last decade.[25,26]

Machine learning should be viewed as the sum total of the organized creation of the initial data set, the fingerprinting and learning steps, and a necessary subsequent step (discussed at the end of this article) of progressive and targeted new data infusion, ultimately leading to an expert recommendation system that can continuously and adaptively improve.

HIERARCHY OF FINGERPRINTS OR DESCRIPTORS

We now elaborate on what is perhaps the most important component of the machine learning paradigm, the one that deals with the numerical representation of the input cases or materials. A numerical representation is essential to make the prediction scheme quantitative (i.e., moving it away from the "vague" notions alluded to in the first paragraph of this article). The choice of the numerical representation can be effectively accomplished only with adequate knowledge of the problem and goals (i.e., domain expertise or experience), and typically proceeds in an iterative manner by duly considering aspects of the material that the target property may be correlated with. Given that the numerical representation serves as the proxy for the real material, it is also referred to as the fingerprint of the material or its descriptors (in machine learning parlance, it is also referred to as the feature vector).

Depending on the problem under study and the accuracy requirements of the predictions, the fingerprint can be defined at varying levels of granularity. For instance, if the goal is to obtain a high-level understanding of the factors underlying a complex phenomenon—such as the mechanical or electrical strength of materials, catalytic activity, etc.—and prediction accuracy is less critical, then the fingerprint may be defined at a gross level, e.g., in terms of the general attributes of the atoms the material is made up of, other potentially relevant properties (e.g., the band gap) or higher-level structural features (e.g., typical grain size). On the other hand, if the goal is to predict specific properties at a reasonable level of accuracy across a wide materials chemical space—such as the dielectric constant of an insulator or the glass transition temperature of a polymer—the fingerprint may have to include information pertaining to key atomic-level structural fragments that may control these properties. If extreme (chemical) accuracy in predictions is demanded—such as total energies and atomic forces, precise identification of structural features, space groups or phases—the fingerprint has to be fine enough so that it is able to encode details of atomic-level structural information with sub-Angstrom-scale resolution. Several examples of learning based on this hierarchy of fingerprints or descriptors are provided in subsequent Sections.

The general rule of thumb is that finer the fingerprint, greater is the expected accuracy, and more laborious, more data-intensive and less conceptual is the learning framework. A corollary to the last point is that rapid coarse-level initial screening of materials should generally be targeted using coarser fingerprints.

Regardless of the specific choice of representation, the fingerprints should also be invariant to certain transformations. Consider the facial recognition scenario. The numerical representation of a face should not depend on the actual placement location of the face in an image, nor should it matter whether the face has been rotated or enlarged with respect to the examples the machine has seen before. Likewise, the representation of a material should be invariant to the rigid translation or rotation of the material. If the representation is fine enough that it includes atomic position information, permutation of like atoms should not alter the fingerprint. These invariance properties are easy to incorporate in coarser fingerprint definitions but non-trivial in fine-level descriptors. Furthermore, ensuring that a fingerprint contains all the relevant components (and only the relevant components)

for a given problem requires careful analysis, for example, using unsupervised learning algorithms.[9,15] For these reasons, construction of a fingerprint for a problem at hand is not always straightforward or obvious.

EXAMPLES OF LEARNING BASED ON GROSS-LEVEL PROPERTY-BASED DESCRIPTORS

Two historic efforts in which gross-level descriptors were utilized to create surrogate models (although they were not couched under those terms) have lead to the Hume–Rothery rules[5] and Hall–Petch relationships[6,7] (Fig. 1). The former effort may be viewed as a classification exercise in which the target is to determine whether a mixture of two metals will form a solid solution; the gross-level descriptors considered were the atomic sizes, crystal structures, electronegativities, and oxidation states of the two metal elements involved. In the latter example, the strength of a polycrystalline material is the target property, which was successfully related to the average grain size; specifically a linear relationship was found between the strength and the reciprocal of the square root of the average grain size. While a careful manual analysis of data gathered from experimentation was key to developing such rules in the past, modern machine learning and data mining approaches provide powerful pathways for such knowledge discovery, especially when the dependencies are multivariate and highly nonlinear.

To identify potential nonlinear multivariate relationships efficiently, one may start from a moderate number of potentially relevant primary descriptors (e.g., electronegativity, E, ionic radius, R, etc.), and create millions or even billions of compound descriptors by forming algebraic combinations of the primary descriptors (e.g., E/R^2, R log(E), etc.); see Fig. 3a, b. This large space of nonlinear mathematical functions needs to be "searched" for a subset that is highly correlated with the target property. Dedicated methodological approaches to accomplish such a task have emerged from recent work in genetic programing,[27] compressed sensing,[28,29] and information science.[30]

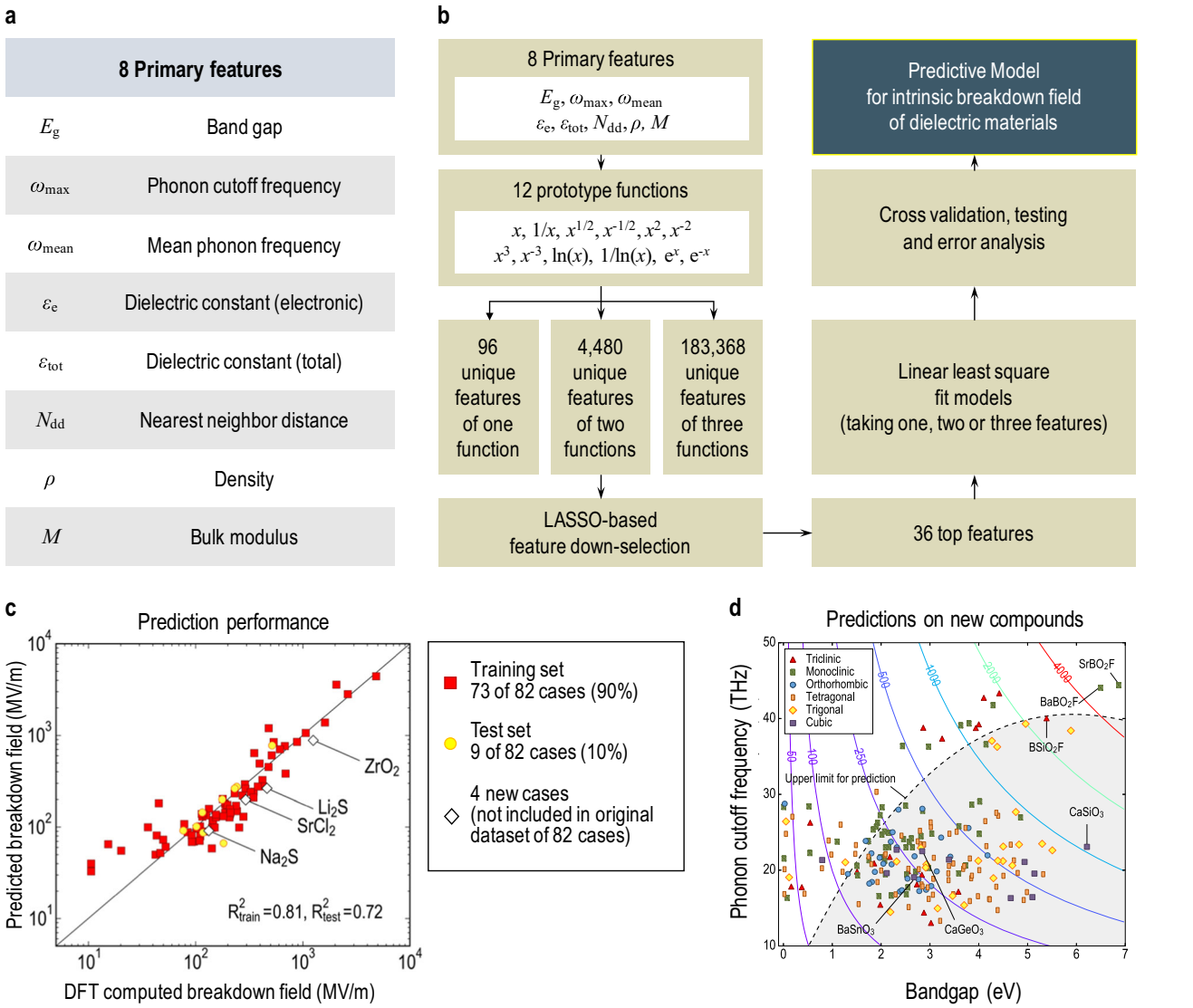

Fig. 3 Building phenomenological models for the prediction of the intrinsic electrical breakdown field of insulators. **a** Primary features expected to correlate to the intrinsic breakdown field; **b** Creation of compound features, down-selection of a subset of critical compound features using LASSO and predictive model building; **c** Final phenomenological model performance versus DFT computations for the binary octet data set (adapted with permission from ref. [31] Copyright (2017) American Chemical Society); and **d** Application of the model for the identification of new breakdown resistant perovskite type materials (contours represent predicted breakdown field in MV/m and the model's prediction domain is depicted in gray color) (adapted with permission from ref. [32] Copyright (2017) American Chemical Society)

One such approach—based on the least absolute shrinkage and selection operator (LASSO)—was recently demonstrated to be highly effective for determining key physical factors that control a complex phenomenon through identification of simple empirical relationships.[28,29] An example of such complex behavior is the tendency of insulators to fail when subjected to extreme electric fields.[31,32] The critical field at which this failure occurs in a defect-free material—referred to as the intrinsic electrical breakdown field—is related to the balance between energy gained by charge carriers from the electric field to the energy lost due to collisions with phonons. The intrinsic breakdown field may be computed from first principles by treatment of electron-phonon interactions, but this computation process is enormously laborious. Recently, the breakdown field was computed from first principles using density functional theory (DFT) for a benchmark set of 82 binary octet insulators.[31] This data set included alkali metal halides, transition metal halides, alkaline earth metal chalcogenides, transition metal oxides, and group III, II–VI, I–VII semiconductors. After validating the theoretical results by comparing against available experimental data, this data set was used to build simple predictive phenomenological surrogate models of dielectric breakdown using LASSO as well as other advanced machine learning schemes. The general flow of the LASSO-based procedure, starting from the primary descriptors considered (Fig. 3a), is charted in Fig. 3b. The trained and validated surrogate models were able to reveal key correlations and analytical relationships between the breakdown field and other easily accessible material properties such as the band gap and the phonon cutoff frequency. Figure 3c shows the agreement between such a discovered analytical relationship and the DFT results (spanning three orders of magnitude) for the benchmark data set of 82 insulators, as well as for four new ones that were not included in the original training data set.

The phenomenological model was later employed to systematically screen and identify perovskite compounds with high breakdown strength. The purely machine learning based screening revealed that boron-containing compounds are of particular interest, some of which were predicted to exhibit remarkable intrinsic breakdown strength of ~1 GV/m (see Fig. 3d). These predictions were subsequently confirmed using first principles computations.[32]

The LASSO-based and related schemes have also been shown to be enormously effective at predicting the preferred crystal structures of materials. In a pioneering study that utilized the LASSO-based approach, Ghiringelli and co-workers were able to classify binary octet insulators into tendencies for the formation of rock salt versus zinc blende structures.[28,29,33] More recently, Bialon and co-workers[34] aimed to classify 64 different prototypical crystal structures formed by A_xB_y type compounds, where A and B are *sp*-block and transition metal elements, respectively. After searching over a set of 1.7×10^5 non-linear descriptors formed by physically meaningful functions of primary coarse-level descriptors such as band-filling, atomic volume, and different electronegativity scales of the *sp* and *d* elements, the authors were able to find a set of three optimal descriptors. A three-dimensional structure-map—built on the identified descriptor set—was used to classify 2105 experimentally known training examples available from the Pearson's Crystal Database[35] with an 86% probability of predicting the correct crystal structure. Likewise, Oliynyk and co-workers recently used a set of elemental descriptors to train a machine-learning model, built on a random forest algorithm,[36] with an aim to accelerate the search for Heusler compounds. After training the model on available crystallographic data from Pearson's Crystal Database[35] and the ASM Alloy Phase Diagram Database[37] the model was used to evaluate the probabilities at which compounds with the formula AB_2C will adopt Heusler structures. This approach was exceptionally successful in distinguishing between Heusler and non-Heusler compounds (with a true positive rate of 94%), including the prediction of unknown compounds and flagging erroneously assigned entries in the literature and in crystallographic databases. As a proof of concept, 12 novel predicted candidates (Gallides with formulae MRu_2Ga and RuM_2Ga, where M = Ti, V, Cr, Mn, Fe, and Co) were synthesized and confirmed to be Heusler compounds. One point to be cautious about when creating an enormous number of compound descriptors (starting from a small initial set of primary descriptors) is model interpretability. Efforts must be taken to ensure that the final set of shortlisted descriptors (e.g., the output of the LASSO process) is stable, i.e., the same or similar set of compound descriptors is obtained during internal cross-validation steps, lest the process becomes a victim of the "curse of dimensionality."

Yet another application of the gross-level descriptors relate to the prediction of the band gap of insulators.[38–42] Rajan and co-workers[38] have used experimentally available band gaps of ABC_2 chalcopyrite compounds to train regression models with electronegativity, atomic number, melting point, pseudopotential radii, and the valence for each of the A, B, and C elements as features. Just using the gross-level elemental features, the developed machine learning models were able to predict the experimental band gaps with moderate accuracy. In a different study, Pilania and co-workers[41] used a database consisting of computed band gaps of ~1300 $AA'BB'O_6$ type double perovskites to train a kernel ridge regression (KRR) machine learning model, a scheme that allows for nonlinear relationships based on measures of (dis) similarity between fingerprints, for efficient predictions of the band gaps. A set of descriptors with increasing complexity was identified by searching across a large portion of the feature space using LASSO, with $\geq$1.2 million compound descriptors created from primary elemental features such as electronegativities, ionization potentials, electronic energy levels, and valence orbital radii of the constituent atomic species. One of the most important chemical insights that emerged from this effort was that the band gap in the double perovskites is primarily controlled (and therefore effectively learned) by the lowest occupied energy levels of the A-site elements and electronegativities of the B-site elements.

Other successful attempts of using gross-level descriptors include the creation of surrogate models for the estimation of formation enthalpies,[43–45] free energies,[46] defect energetics,[47] melting temperatures,[48,49] mechanical properties,[50–52] thermal conductivity,[53] catalytic activity,[54,55] and radiation damage resistance.[56] Efforts are also underway for the identification of novel shape memory alloys,[57] improved piezoelectrics,[58] MAX phases,[59] novel perovskite[60] and double perovskite halides,[43,60] CO_2 capture materials,[61] and potential candidates for water splitting.[62]

Emerging materials informatics tools also offer tremendous potential and new avenues for mining for structure-property-processing linkages from aggregated and curated materials data sets.[63] While a large fraction of such efforts in the current literature has considered relatively simple definitions of the material that included mainly the overall chemical composition of the material, Kalidindi and co-workers[64–67] have recently proposed a new materials data science framework known as Materials Knowledge Systems[68,69] that explicitly accounts for the complex hierarchical material structure in terms of *n*-point spatial correlations (also frequently referred to as *n*-point statistics). Further adopting the *n*-point statistics as measures to quantify materials microstructure, a flexible computational framework has been developed to customize toolsets to understand structure-property-processing linkages in materials science.[70]

EXAMPLES OF LEARNING BASED ON MOLECULAR FRAGMENT-LEVEL DESCRIPTORS

The next in the hierarchy of descriptor types are those that encode finer details than those captured by the gross-level properties. Within this class, materials are described in terms of the basic building blocks they are made of. The origins of "block-level" or "molecular fragment" based descriptors can be traced back to cheminformatics, which is a field of theoretical chemistry that deals with correlating properties such as biological activity, physio-chemical properties and reactivity with molecular structure and fragments,[71–73] leading up to what is today referred to as quantitative structure activity/property relationships (QSAR/QSPR).

Within materials science, specifically, within polymer science, the notions underlying QSAR/QSPR ultimately led to the successful group contribution methods.[8] Van Krevelen and co-workers studied the properties of polymers and discovered that they were strongly correlated to the chemical structure (i.e., nature of the polymer repeat unit, end groups, etc.) and the molecular weight distribution. They observed that polymer properties such as glass transition temperature, solubility parameter and bulk modulus (which were, and still are, difficult to compute using traditional computational methods) were correlated with the presence of chemical groups and combinations of different groups in the repeat unit. Based on a purely data-driven approach, they developed an "atomic group contribution method" to express various properties as a linear weighted sum of the contribution (called atomic group parameter) from every atomic group that constituted the repeat unit. These groups could be units like CH_2, C_6H_4, CH_2-CO, etc., that make up the polymer. It was also noticed that factors such as the presence of aromatic rings, long side chains and cis/trans conformations influence the properties, prompting their introduction into the group additivity scheme. For instance, a CH_2 group attached to an aromatic ring would have a different atomic group parameter than a CH_2 group attached to an aliphatic group. In this fashion, nearly all the important contributing factors were taken into account, and linear empirical relationships were devised for thermal, elastic and other polymer properties. However, widespread usage of these surrogate models is still restricted because (1) the definition of atomic groups is somewhat *ad hoc*, and (2) the target properties are assumed to be linearly related to the group parameters.

Modern data-driven methods have significantly improved on these earlier ideas with regards to both issues mentioned above. Recently, in order to enable the accelerated discovery of polymer dielectrics,[74–79] hundreds of polymers built from a chemically allowed combination of seven possible basic units, namely, CH_2, CO, CS, O, NH, C_6H_4, and C_4H_2S, were considered, inclusive of van der Waals interactions,[80] and a set of properties relevant for dielectric applications, namely, the dielectric constant and band gap, were computed using DFT.[74,81] These polymers were then fingerprinted by keeping track of the occurrence of a fixed set of molecular fragments in the polymers in terms of their number fractions.[81,82] A particular molecular fragment could be a triplet of contiguous blocks such as –NH–CO–CH_2– (or, at a finer level, a triplet of contiguous atoms, such as C_4–O_2–C_3 or C_3–N_3–H_1, where X_n represents an *n*-fold coordinated X atom).[83,84] All possible triplets were considered (some examples are shown in Fig. 4a), and the corresponding number fractions in a specific order formed the fingerprint of a particular polymer (see Fig. 4b). This

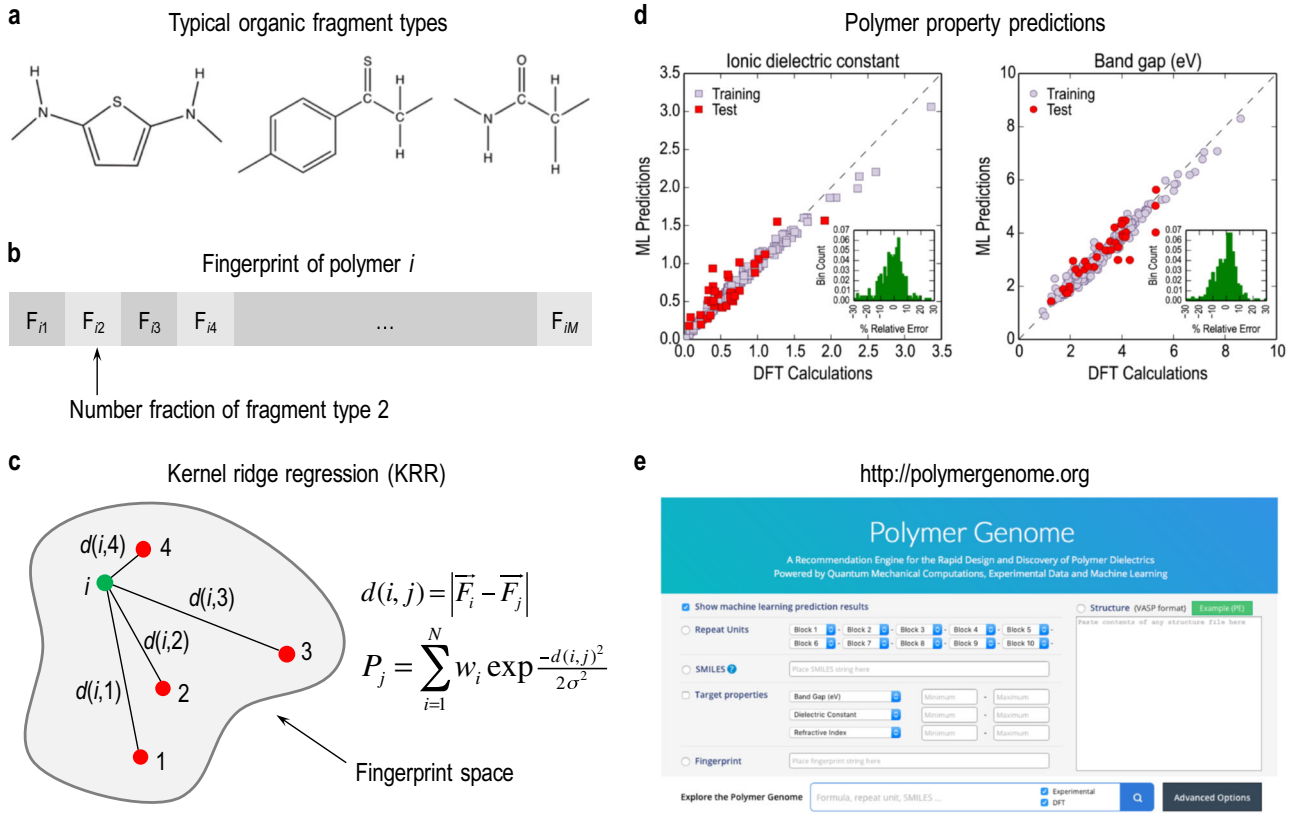

Fig. 4 Learning polymer properties using fragment-level fingerprints. **a** Typical fragments that can be used for the case of organic molecules, crystals or polymers; **b** Schematic of organic polymer fingerprint construction; **c** Schematic of the kernel ridge regression (KRR) scheme showing the example cases in fingerprint (*F*) space. The distance, $d(i, j)$, between the point (in fingerprint space) corresponding to a new case, *j*, and each of the training example cases, *i*, is used to predict the property, P_j, of case *j*; **d** Surrogate machine learning (ML) model predictions versus DFT results for key dielectric polymer properties;[81] **e** Snapshot of the Polymer Genome online application for polymer property prediction

procedure provides a uniform and seamless pathway to represent all polymers within this class, and the procedure can be indefinitely generalized by considering higher order fragments (i.e., quadruples, quintuples, etc., of atom types). Furthermore, relationships between the fingerprint and properties have been established using the KRR learning algorithm; a schematic of how this algorithm works is shown in Fig. 4c. The capability of this scheme for dielectric constant and band gap predictions is portrayed in Fig. 4d. These predictive tools are available online (Fig. 4e) and are constantly being updated.[85] The power of such modern data-driven molecular fragment-based learning approaches (like its group contribution predecessor) lies in the realization that any type of property related to the molecular structure—whether computable using DFT (e.g., band gap, dielectric constant) or measurable experimentally (e.g., glass transition temperature, dielectric loss)—can be learned and predicted.

The molecular fragment-based representation is not restricted to polymeric materials. Novel compositions of $A_xB_yO_z$ ternary oxides and their most probable crystal structures have been predicted using a probabilistic model built on an experimental crystal structure database.[86] The descriptors used in this study are a combination of the type of crystal structure (spinel, olivine, etc.) and the composition information, i.e., the elements that constitute the compound. Likewise, surrogate machine learning models have been developed for predicting the formation energies of $A_xB_yO_z$ ternary compounds using only compositional information as descriptors, trained on a data set of 15,000 compounds from the Inorganic Crystal Structure Database.[44] Using this approach, 4500 new stable materials have been discovered. Finally, surrogate models have been developed for predicting the formation energies of elpasolite crystals with the general formula A_2BCD_6, based mainly on compositional information. The descriptors used take into account the periodic table row and column of elements A, B, C, and D that constitute the compound (although this fingerprint could have been classified as a gross-level one, we choose to place this example in the present Section as the prototypical structure of the elpasolite was implicitly assumed in this work and fingerprint). Important correlations and trends were revealed between atom types and the energies; for example, it was found that the preferred element for the D site is F, and that for the A and B sites are late group II elements.[43]

EXAMPLES OF LEARNING BASED ON SUB-ANGSTROM-LEVEL DESCRIPTORS

We now turn to representing materials at the finest possible scale, such that the fingerprint captures precise details of atomic configurations with high fidelity. Such a representation is useful in many scenarios. For instance, one may attempt to connect this fine-scale fingerprint directly with the corresponding total potential energy with chemical accuracy, or with structural phases/motifs (e.g., crystal structure or the presence/absence of a stacking fault). The former capability can lead to purely data-driven accelerated atomistic computational methods, and the latter to refined and efficient on-the-fly characterization schemes.

"Chemical accuracy" specifically refers to potential energy and reaction enthalpy predictions with errors of < 1 kcal/mol, and atomic force predictions (the input quantity for molecular dynamics, or MD, simulations) with errors of < 0.05 eV/Å. Chemical accuracy is key to enable reliable MD simulations (or for precise identification of the appropriate structural phases or motifs), and is only possible with fine-level fingerprints that offer sufficiently high configurational resolution, more than those in the examples encountered thus far.

The last decade has seen spectacular activity and successes in the general area of data-driven atomistic computations. All modern atomistic computations use either some form of quantum mechanical scheme (e.g., DFT) or a suitably parameterized semi-empirical method to predict the properties of materials, given just the atomic configuration. Quantum mechanical methods are versatile, i.e., they can be used to study any material, in principle. However, they are computationally demanding, as complex differential equations governing the behavior of electrons are solved for every given atomic configuration. Systems involving at most about 1000 atoms can be simulated routinely in a practical setting today. In contrast, semi-empirical methods use prior knowledge about interatomic interactions under known conditions and utilize parameterized analytical equations to determine properties such as the total potential energies, atomic forces, etc. These semi-empirical force fields are several orders of magnitude faster than quantum mechanical methods, and are the choice today for routinely simulating systems containing millions to billions of atoms, as well as the dynamical evolution of systems at nonzero temperatures (using the MD method) at timescales of nanoseconds to milliseconds. However, a major drawback of traditional semi-empirical force fields is that they lack versatility, i.e., they are not transferable to situations or materials for which the original functional forms and parameterizations do not apply.

Machine learning is rapidly bridging the chasm between the two extremes of quantum mechanical and semi-empirical methods, and has offered surrogate models that combine the best of both worlds. Rather than resort to specific functional forms and parameterizations adopted in semi-empirical methods (the aspects that restrict their versatility), machine learning methods use an {atomic configuration → property} data set, carefully prepared, e.g., using DFT, to make interpolative predictions of the property of a new configuration at speeds several orders of magnitude faster than DFT. Any material for which adequate reference DFT computations may be performed ahead of time can be handled using such a machine learning scheme. Thus, the lack of versatility issue of traditional semi-empirical approach and the time-intensive nature of quantum mechanical calculations are simultaneously addressed, while also preserving quantum mechanical and chemical accuracy.

The primary challenge though has been the creation of suitable fine-level fingerprinting schemes for materials, as these fingerprints are required to be strictly invariant with respect to arbitrary translations, rotations, and exchange of like atoms, in addition to being continuous and differentiable (i.e., "smooth") with respect to small variations in atomic positions. Several candidates, including those based on symmetry functions,[87–89] bispectra of neighborhood atomic densities,[90] Coulomb matrices (and its variants),[91,92] smooth overlap of atomic positions (SOAP),[93–96] and others,[97,98] have been proposed. Most fingerprinting approaches use sophisticated versions of distribution functions (the simplest one being the radial distribution function) to represent the distribution of atoms around a reference atom, as qualitatively captured in Fig. 5a. The Coulomb matrix is an exception, which elegantly represents a molecule, with the dimensionality of the matrix being equal to the total number of atoms in the molecule. Although questions have arisen with respect to smoothness considerations and whether the representation is under/over-determined (depending on whether the eigenspectrum or the entire matrix is used as the fingerprint),[93] this approach has been shown to be able to predict various molecular properties accurately.[92]

Figure 5b also shows a general schema typically used in the construction of machine learning force fields, to be used in MD simulations. Numerous learning algorithms—ranging from neural networks, KRR, Gaussian process regression (GPR), etc.—have been utilized to accurately map the fingerprints to various materials properties of interest. A variety of fingerprinting schemes, as well as learning schemes that lead up to force fields have been recently reviewed.[9,93,99] One of the most successful and widespread machine learning force field schemes to date is the one by Behler and co-workers,[87] which uses symmetry function

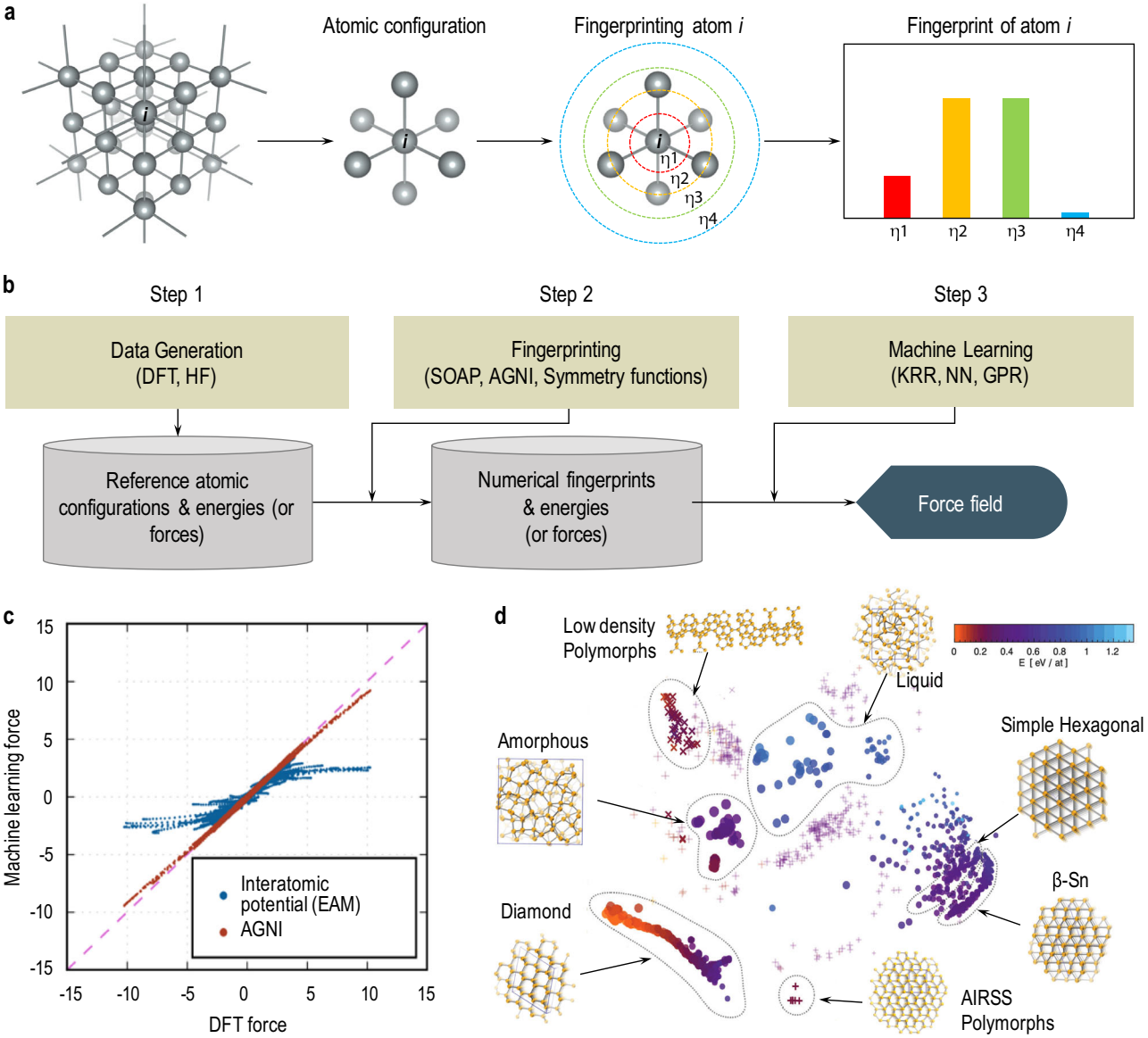

Fig. 5 Learning from fine-level fingerprints. **a** A schematic portrayal of the sub-Angstrom-level atomic environment fingerprinting scheme adopted by Behler and co-workers. η_js denote the widths of Gaussians, indexed by *j*, placed at the reference atom *i* whose environment needs to be fingerprinted. The histograms in the right represent the integrated number of atoms within each Gaussian sphere; **b** Schematic of a typical workflow for the construction of machine learning force fields; **c** Prediction of atomic forces in the neighborhood of an edge dislocation in bulk Al using the atomic force-learning scheme AGNI and the embedded atom method (EAM), and comparison with the corresponding DFT results (adapted with permission from ref. [105] Copyright (2017) American Chemical Society); **d** Classifying atomic environments in Si using the SOAP fingerprinting scheme and the Sketch Map program for dimensionality reduction (adapted with permission from ref. [117] Copyright (2017) Royal Society of Chemistry)

fingerprints mapped to the total potential energy using a neural network. Several applications have been studied, including surface diffusion, liquids, phase equilibria in bulk materials, etc. This approach is also quite versatile in that multiple elements can be considered. Bispectra based fingerprints combined with GPR learning schemes have lead to Gaussian approximation potentials,[87,90] which have also been demonstrated to provide chemical accuracy, versatility and efficiency.

A new development within the area of machine learning force fields is to learn and predict the atomic forces directly;[100–105] the total potential energy is determined through appropriate integration of the forces along a reaction coordinate or MD trajectory.[105] These approaches are inspired by Feynman's original idea that it should be possible to predict atomic forces given just the atomic configuration, without going through the agency of the total potential energy.[106] An added attraction of this perspective is that the atomic force can be uniquely assigned to an individual atom, while the potential energy is a global property of the entire system (partitioning the potential energy to atomic contributions does not have a formal basis). Mapping atomic fingerprints to purely atomic properties can thus lead to powerful and accurate prescriptions. Figure 5c, for instance, compares the atomic forces at the core of an edge dislocation in Al, predicted using a machine learning force prediction recipe called AGNI, with the DFT forces for the same atomic configuration. Also shown are forces predicted using the embedded atom method (EAM), a popular

classical force field, for the same configuration. EAM tends to severely under-predict large forces while the machine learning scheme predicts forces with high fidelity (neither EAM nor the machine learning force field were explicitly trained on dislocation data). This general behavior is consistent with recent detailed comparisons of EAM with machine learning force fields.[107] It is worth noting that although this outlook of using atomic forces data during force field development is reminiscent of the "force-matching" approach of Ercolessi and Adams,[108] this new development is distinct from that approach in that it attempts to predict the atomic force given just the atomic configuration.

Another notable application of fine-level fingerprints has been in the use of the electronic charge density itself as the representation to learn various properties[82] or density functionals,[109–111] thus going to the very heart of DFT. While these efforts are in a state of infancy—as they have dealt with mainly toy problems and learning the kinetic energy functional—such efforts have great promise as they attempt to integrate machine learning methods within DFT (all other DFT-related informatics efforts so far have utilized machine learning external to DFT).

Fine-level fingerprints have also been used to characterize structure in various settings. Within a general crystallographic structure refinement problem, one has to estimate the structural parameters of a system, i.e., the unit cell parameters (a, b, c, α, β, and γ) that best fit measured X-ray diffraction (XRD) data. Using a Bayesian learning approach and a Markov chain Monte Carlo algorithm to sample multiple combinations of possible structural parameters for the case of Si, Fancher and co-workers[112] not only accurately determined the estimates of the structural parameters, but also quantified the associated uncertainty (thus going beyond the conventional Rietveld refinement method).

Unsupervised learning using fine-level fingerprints (and clustering based on these fingerprints) has led to the classification of materials based on their phases or structural characteristics.[11,12] Using the XRD spectrum itself as the fingerprint, high-throughput XRD measurements for various compositional spreads[11,12,113–116] have been used to automate the creation of phase diagrams. Essentially, features of the XRD spectra are used to distinguish between phases of a material as a function of composition. Likewise, on the computational side, the SOAP fingerprints have been effectively used to distinguish between different allotropes of materials, as well as different motifs that emerge during the course of a MD simulation (see Fig. 5d for an example).[117]

CRITICAL STEPS GOING FORWARD

Quantifying the uncertainties of predictions

Given that machine learning predictions are inherently statistical in nature, uncertainties must be expected in the predictions. Moreover, predictions are typically and ideally interpolative between data points corresponding to previously seen data. To what extent a new case for which a prediction needs to be made falls in or out of the domain of the original data set (i.e., to what extent the predictions are interpolative or extrapolative) may be quantified using the predicted uncertainty. While strategies are available to prescribe prediction uncertainties, these ideas have been explored only to a limited extent within materials science.[57,118] Bayesian methods (e.g., Gaussian process regression)[15] provide a natural pathway for estimating the uncertainty of the prediction in addition to the prediction itself. This approach assumes that a Gaussian distribution of models fit the available data, and thus a distribution of predictions may be made. The mean and variance of these predictions—the natural outcomes of Bayesian approaches—are the most likely predicted value and the uncertainty of the prediction, respectively, within the spectrum of models and the fingerprint considered. Other methods may also be utilized to estimate uncertainties, but at significant added cost. A straightforward and versatile scheme is bootstrapping,[119] in which different (but small) subsets of the data are randomly excluded, and several prediction models are developed based on these closely related but modified data sets. The mean and variance of the predictions from these bootstrapped models provide the property value and expected uncertainty. Essentially, this approach attempts to probe how sensitive the model is with respect to slight "perturbations" to the data set. Another related methodology is to explicitly consider a variety of closely related models, e.g., neural networks or decision trees with slightly different architectures, and to use the distribution of predictions to estimate uncertainty.[89]

Adaptive learning and design

Uncertainty quantification has a second important benefit. It can be used to continuously and progressively improve a prediction model, i.e., render it a truly learning model. Ideally, the learning model should adaptively and iteratively improve by asking questions such as *"what should be the next new material system to consider or include in the training set that would lead to an improvement of the model or the material?"* This may be accomplished by balancing the tradeoffs between exploration and exploitation.[118,120] That is, at any given stage of an iterative learning process, a number of new candidates may be predicted to have certain properties with uncertainties. The tradeoff is between exploiting the results by choosing to perform the next computation (or experiment) on the material predicted to have the optimal target property or further improving the model through exploration by performing the calculation (or experiment) on a material where the predictions have the largest uncertainties. This can be done rigorously by adopting well-established information theoretic selector frameworks such as the knowledge gradient.[121,122] In the initial stages of the iterative process, it is desired to "explore and learn" the property landscape. As the machine learning predictions improve and the associated uncertainties shrink, the adaptive design scheme allows one to gradually move away from exploration towards exploitation. Such an approach, schematically portrayed in Fig. 6a, enables one to systematically expand the training data towards a target chemical space, where materials with desired functionality are expected to reside.

Some of the first examples of using adaptive design for targeted materials discovery include identification of shape memory alloys with low thermal hysteresis[57] and accelerated search for $BaTiO_3$-based piezoelectrics with optimized morphotropic phase boundary.[58] In the first example, Xue and co-workers[57] employed the aforementioned adaptive design framework to find NiTi-based shape memory alloys that may display low thermal hysteresis. Starting from a limited number of 22 training examples and going through the iterative process 9 times, 36 predicted compositions were synthesized and tested from a potential space of ~800,000 compound possibilities. It was shown that 14 out of these 36 new compounds were better (i.e., had a smaller thermal hysteresis) than any of the 22 compounds in the original data set. The second successful demonstration of the adaptive design approach combined informatics and Landau–Devonshire theory to guide experiments in the design of lead-free piezoelectrics.[58] Guided by predictions from the machine learning model, an optimized solid solution, $(Ba_{0.5}Ca_{0.5})TiO_3$–$Ba(Ti_{0.7}Zr_{0.3})O_3$, with piezoelectric properties was synthesized and characterized to show better temperature reliability than other $BaTiO_3$-based piezoelectrics in the initial training data.

Other algorithms

The materials science community is just beginning to explore and utilize the plethora of available information theoretic algorithms to mine and learn from data. The usage of an algorithm is driven

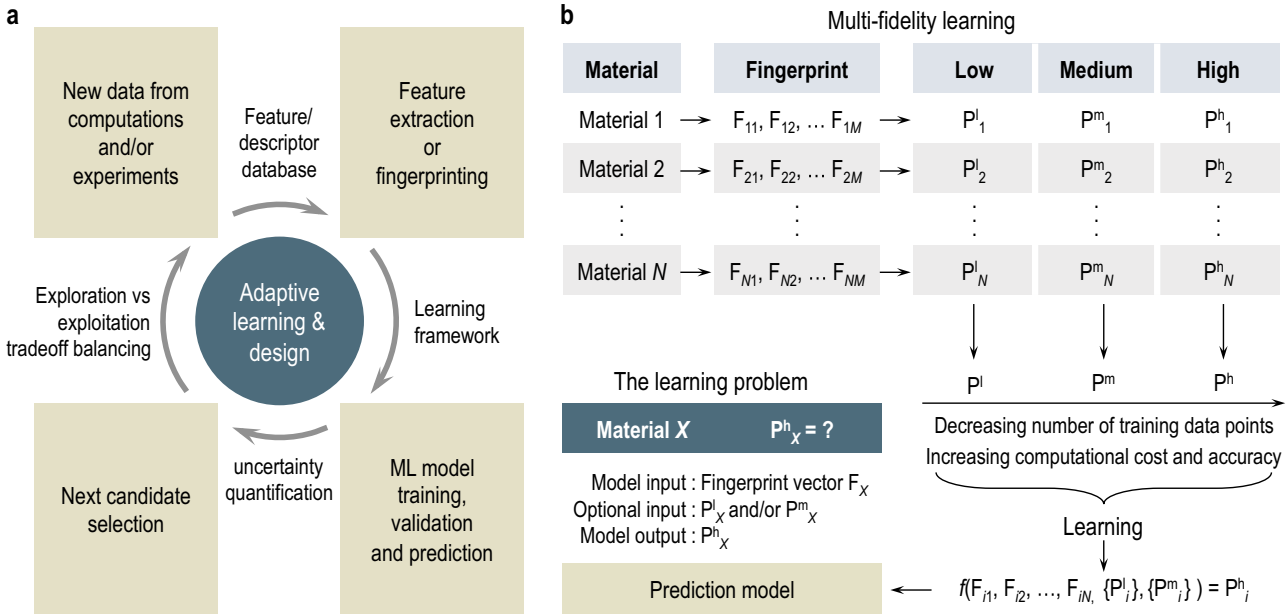

Fig. 6 **a** Schematic illustration of adaptive design via balanced exploration and exploitation enabled by uncertainty quantification. **b** An example data set used in a multi-fidelity learning setting involving target properties obtained at various levels of fidelity and expense, and the statement of the multi-fidelity learning problem

largely by need, as it should. One such need is to be able to learn and predict vectorial quantities. Examples include functions, such as the electronic or vibrational density of states (which are functions of energy or frequency). Although, the target property in these cases may be viewed as a set of scalar quantities at each energy or frequency (for a given structure) to be learned and predicted independently, it is desirable to learn and predict the entire function simultaneously. This is because the value of the function at a particular energy or frequency is correlated to the function values at other energy or frequency values. Properly learning the function of interest requires machine learning algorithms that can handle vectorial outputs. Such algorithms are indeed available,[123,124] and if exploited can lead to prediction schemes of the electronic structure for new configurations of atoms. Another class of examples where vector learning is appropriate includes cases where the target property is truly a vector (e.g., atomic force) or a tensor (e.g., stress). In these cases, the vector or tensor transforms in a particular way as the material itself is transformed, e.g., if it is rotated (in the examples of functions discussed above, the vectors, i.e., the functions, are invariant to any unitary transformation of the material). These truly vectorial or tensorial target property cases will thus have to be handled with care, as has been done recently using vector learning and covariant kernels.[102]

Another algorithm that is beginning to show value within material science falls under multi-fidelity learning.[125] This learning method can be used when a property of interest can be computed at several levels of fidelities, exhibiting a natural hierarchy in both computational cost and accuracy. A good materials science example is the band gap of insulators computed at an inexpensive lower level of theory, e.g., using a semilocal electronic exchange-correlation functional (the low-fidelity value), and the band gap computed using an more accurate, but expensive, approach, e.g., using a hybrid exchange-correlation functional (the high-fidelity value). A naive approach in such a scenario can be to use a low-fidelity property value as a feature in a machine learning model to predict the corresponding higher fidelity value. However, using low-fidelity estimates as features strictly requires the low-fidelity data for all materials for which predictions are to be made using the trained model. This can be particularly challenging and extremely computationally demanding when faced with a combinatorial problem that targets exploring vast chemical and configurational spaces. A multi-fidelity co-kriging framework, on the other hand, can seamlessly combine inputs from two or more levels of fidelities to make accurate predictions of the target property for the highest fidelity. Such an approach, schematically represented in Fig. 6b, requires high-fidelity training data only on a subset of compounds for which low-fidelity training data is available. More importantly, the trained model can make efficient highest-fidelity predictions even in the absence of the low-fidelity data for the prediction set compounds. While multi-fidelity learning is routinely used in several fields to address computationally challenging engineering design problems,[125,126] it is only beginning to find applications in materials informatics.[42]

Finally, machine learning algorithms may also lead to strategies for making the so-called "inverse design" of materials possible. Inverse design refers to the paradigm whereby one seeks to identify materials that satisfy a target set of desired properties (in this parlance, the "forward" process refers to predicting the properties of a given material).[127] Within the machine learning context, although the backward process of going from a desired set of properties to the appropriate fingerprints is straightforward, the process of inverting the fingerprint to actual physically and chemically meaningful materials continues to be a major hurdle. Two strategies that are adopted to achieve inverse design within the context of machine learning involves either inverting the desired properties to only fingerprints that correspond to physically realizable materials (through imposition of constraints that fingerprint components are required to satisfy),[83,127] or adopting schemes such as the genetic algorithm or simulated annealing to determine iteratively a population of materials that meet the given target property requirements.[81,83] Despite these developments, true inverse design continues to remain a challenge (although materials design through adaptive learning discussed above appears to have somewhat mitigated this challenge).

DECISIONS ON WHEN TO USE MACHINE LEARNING

Perhaps the most important question that plagues new researchers eager to use data-driven methods is whether their problem lends itself to such methods. Needless to say, the existence of past reliable data, or efforts devoted to its generation for at least a subset of the critical cases in a uniform and controlled manner, is a prerequisite for the adoption of machine learning. Even so, the question is the appropriateness of machine learning for the problem at hand. Ideally, data-driven methods should be aimed at (1) properties very difficult or expensive to compute or measure using traditional methods, (2) phenomena that are complex enough (or nondeterministic) that there is no hope for a direct solution based on solving fundamental equations, or (3) phenomena whose governing equations are not (yet) known, providing a rationale for the creation of surrogate models. Such scenarios are replete in the social, cognitive and biological sciences, explaining the pervasive applications of data-driven methods in such domains. Materials science examples ideal for studies using machine learning methods include properties such as the glass transition temperature of polymers, dielectric loss of polycrystalline materials over a wide frequency and temperature range, mechanical strength of composites, failure time of engineering materials (e.g., due to electrical, mechanical or thermal stresses), friction coefficient of materials, etc., all of which involve the inherent complexity of materials, i.e., their polycrystalline or amorphous nature, multi-scale geometric architectures, the presence of defects of various scales and types, and so on.

Machine learning may also be used to eliminate redundancies underlying repetitive but expensive operations, especially when interpolations in high-dimensional spaces are required, such as when properties across enormous chemical and/or configurational spaces are desired. An example of the latter scenario, i.e., an immense configurational space, is encountered in first principles molecular dynamics simulations, when atomic forces are evaluated repetitively (using expensive quantum mechanical schemes) for myriads of very similar atomic configurations. The area of machine learning force fields has burgeoned to meet this need. Yet another setting where large chemical and configurational spaces are encountered is the emerging domain of high-throughput materials characterization, where on-the-fly predictions are required to avoid data accumulation bottlenecks. Although materials informatics efforts so far have largely focused on model problems and the validation of the general notion of data-driven discovery, active efforts are beginning to emerge that focus on complex real-world materials applications, strategies to handle situations inaccessible to traditional materials computations, and the creation of adaptive prediction frameworks (through adequate uncertainty quantification) that build efficiencies within rational materials design efforts.

ACKNOWLEDGEMENTS

We acknowledge financial support from several grants from the Office of Naval Research that allowed them to explore many applications of machine learning within materials science, including N00014-14-1-0098, N00014-16-1-2580, and N00014-10-1-0944. Several engaging discussions with Kenny Lipkowitz, Huan Tran, and Venkatesh Botu are gratefully acknowledged. GP acknowledges the Alexander von Humboldt Foundation.

AUTHOR CONTRIBUTIONS

R.R. lead the creation of the manuscript, with critical contributions on various sections and graphics by G.P., R.B., A.M.K. and C.K. All authors participated in the writing of the manuscript.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing financial interests.

REFERENCES

1. Gopnik, A. Making AI more human. *Sci. Am.* **316**, 60–65 (2017).
2. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
3. Srinivasan, S. & Ranganathan, S. *India's Legendary Wootz Steel: An Advanced Material of the Ancient World* (National Institute of advanced studies, 2004).
4. Ward, G. W. R. *The Grove Encyclopedia of Materials and Techniques in Art* (Oxford University Press, 2008).
5. Hume-Rothery, W. Atomic theory for students of metallurgy. *J. Less Common Met.* **3**, 264 (1961).
6. Hall, E. O. The deformation and ageing of mild steel: III discussion of results. *Proc. Phys. Soc. B* **64**, 747–753 (1951).
7. Petch, N. J. The influence of grain boundary carbide and grain size on the cleavage strength and impact transition temperature of steel. *Acta Metall.* **34**, 1387–1393 (1986).
8. Van Krevelen, D. W. & Te Nijenhuis, K. *Properties of Polymers: Their Correlation with Chemical Structure; their Numerical Estimation and Prediction from Additive Group Contributions* (Elsevier, 2009).
9. Mueller, T., Kusne, A. G. & Ramprasad, R. In *Reviews in Computational Chemistry*, 186–273 (John Wiley & Sons, Inc, 2016).
10. Ward, L. & Wolverton, C. Atomistic calculations and materials informatics: a review. *Curr. Opin. Solid State Mater. Sci.* **21**, 167–176 (2017).
11. Green, M. L. et al. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Appl. Phys. Rev.* **4**, 011105 (2017).
12. Hattrick-Simpers, J. R., Gregoire, J. M. & Kusne, A. G. Perspective: composition–structure–property mapping in high-throughput experiments: turning data into knowledge. *APL Mater.* **4**, 053211 (2016).
13. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
14. Theodoridis, S. *Machine Learning: A Bayesian and Optimization Perspective* (Academic Press, 2015).
15. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2013).
16. Sanchez, J., Ducastelle, F. & Gratias, D. Generalized cluster description of multicomponent systems. *Phys. A: Stat. Mech. Appl.* **128**, 334–350 (1984).
17. Fontaine, D. Cluster approach to order-disorder transformations in alloys. *Solid State Phys.* **47**, 33–176 (1994).
18. Zunger, A. First-principles statistical mechanics of semiconductor alloys and intermetallic compounds, NATO Advanced Study Institute, Series B: Physics Vol. 319 (Turchi, P. & Gonis, A. eds), 361419 (Plenum, New York, 1994).
19. Laks, D. B., Ferreira, L. G., Froyen, S. & Zunger, A. Efficient cluster expansion for substitutional systems. *Phys. Rev. B* **46**, 12587–12605 (1992).
20. van de Walle, A. & Ceder, G. Automating first-principles phase diagram calculations. *J. Phase Equilib.* **23**, 348 (2002).
21. Mueller, T. & Ceder, G. Bayesian approach to cluster expansions. *Phys. Rev. B* **80**, 024103 (2009).
22. Cockayne, E. & van de Walle, A. Building effective models from sparse but precise data: application to an alloy cluster expansion model. *Phys. Rev. B* **81**, 012104 (2010).
23. Seko, A., Koyama, Y. & Tanaka, I. Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations. *Phys. Rev. B* **80**, 165122 (2009).
24. Mueller, T. & Ceder, G. Exact expressions for structure selection in cluster expansions. *Phys. Rev. B* **82**, 184107 (2010).
25. Lance, N. J., Hart, G. L. W., Zhou, F. & Ozolins, V. Compressive sensing as a paradigm for building physics models. *Phys. Rev. B* **87**, 24–32 (2015).
26. Sanders, J. N., Andrade, X. & Aspuru-Guzik, A. Compressive sensing for the fast computation of matrices: application to molecular vibrations. *ACS Cent. Sci.* **1**, 035125 (2013).
27. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
28. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
29. Ghiringhelli, L. M. et al. Learning physical descriptors for materials science by compressed sensing. *New. J. Phys.* **19**, 023017 (2017).
30. Lookman, T., Alexander, F. J. & Rajan, K. *Information Science for Materials Discovery and Design* (Springer, 2015).
31. Kim, C., Pilania, G. & Ramprasad, R. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. *Chem. Mater.* **28**, 1304–1311 (2016).

32. Kim, C., Pilania, G. & Ramprasad, R. Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX3 perovskites. *J. Phys. Chem. C* **120**, 14575–14580 (2016).
33. Goldsmith, B. R. et al. Uncovering structure-property relationships of materials by subgroup discovery. *New. J. Phys.* **19**, 013031 (2017).
34. Bialon, A. F., Hammerschmidt, T. & Drautz, R. Three-parameter crystal-structure prediction for sp-d-valent compounds. *Chem. Mater.* **28**, 2550–2556 (2016).
35. Pearson's crystal data. Crystal structure database for inorganic compounds. *Choice Rev. Online* **45**, 45–3800–45–3800 (2008).
36. Oliynyk, A. O. et al. High-throughput machine-learning-driven synthesis of Full-Heusler compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
37. ASM international the materials information society–ASM international. http://www.asminternational.org/. Accessed 23.06.2017.
38. Dey, P. et al. Informatics-aided bandgap engineering for solar materials. *Comput. Mater. Sci.* **83**, 185–195 (2014).
39. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *NPJ Comput. Mater.* **2**, 201628 (2016).
40. Lee, J., Seko, A., Shitara, K., Nakayama, K. & Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B Condens. Matter* **93**, 115104 (2016).
41. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
42. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).
43. Faber, F. A., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).
44. Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B Condens. Matter* **89**, 094104 (2014).
45. Deml, A. M., O'Hayre, R., Wolverton, C. & Stevanović, V. Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression. *Phys. Rev. B Condens. Matter* **93**, 085142 (2016).
46. Legrain, F., Carrete, J., van Roekeghem, A., Curtarolo, S. & Mingo, N. How the chemical composition alone can predict vibrational free energies and entropies of solids. *Chem. Mater.* **29**, 6220–6227 (2017).
47. Medasani, B. et al. Predicting defect behavior in B2 intermetallics by merging ab initio modeling and machine learning. *NPJ Comput. Mater.* **2**, 1 (2016).
48. Seko, A., Maekawa, T., Tsuda, K. & Tanaka, I. Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids. *Phys. Rev. B Condens. Matter* **89**, 054303 (2014).
49. Pilania, G., Gubernatis, J. E. & Lookman, T. Structure classification and melting temperature prediction in octet AB solids via machine learning. *Phys. Rev. B Condens. Matter* **91**, 214302 (2015).
50. Chatterjee, S., Murugananth, M. & Bhadeshia, H. K. D. H. δ TRIP steel. *Mater. Sci. Technol.* **23**, 819–827 (2007).
51. De Jong, M. et al. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. *Sci. Rep.* **6**, 34256 (2016).
52. Aryal, S., Sakidja, R., Barsoum, M. W. & Ching, W.-Y. A genomic approach to the stability, elastic, and electronic properties of the MAX phases. *Phys. Status Solidi* **251**, 1480–1497 (2014).
53. Seko, A. et al. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Phys. Rev. Lett.* **115**, 205901 (2015).
54. Li, Z., Ma, X. & Xin, H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catal. Today* **280**, 232–238 (2017).
55. Hong, W. T., Welsch, R. E. & Shao-Horn, Y. Descriptors of oxygen-evolution activity for oxides: a statistical evaluation. *J. Phys. Chem. C* **120**, 78–86 (2016).
56. Pilania, G. et al. Using machine learning to identify factors that govern amorphization of irradiated pyrochlores. *Chem. Mater.* **29**, 2574–2583 (2017).
57. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
58. Xue, D. et al. Accelerated search for BaTiO3-based piezoelectrics with vertical morphotropic phase boundary using bayesian learning. *Proc. Natl Acad. Sci. U. S. A* **113**, 13301–13306 (2016).
59. Ashton, M., Hennig, R. G., Broderick, S. R., Rajan, K. & Sinnott, S. B. Computational discovery of stable M_2AX phases. *Phys. Rev. B. Condens. Matter* **94**, 20 (2016).
60. Pilania, G., Balachandran, P. V., Kim, C. & Lookman, T. Finding new perovskite halides via machine learning. *Front. Mater.* **3**, 19 (2016).
61. Fernandez, M., Boyd, P. G., Daff, T. D., Aghaji, M. Z. & Woo, T. K. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO_2 capture. *J. Phys. Chem. Lett.* **5**, 3056–3060 (2014).
62. Emery, A. A., Saal, J. E., Kirklin, S., Hegde, V. I. & Wolverton, C. High-throughput computational screening of perovskites for thermochemical water splitting applications. *Chem. Mater.* **28**, 5621–5634 (2016).
63. Kalidindi, S. R. et al. Role of materials data science and informatics in accelerated materials innovation. *MRS Bull.* **41**, 596–602 (2016).
64. Brough, D. B., Kannan, A., Haaland, B., Bucknall, D. G. & Kalidindi, S. R. Extraction of process-structure evolution linkages from x-ray scattering measurements using dimensionality reduction and time series analysis. *Integr. Mater. Manuf. Innov.* **6**, 147–159 (2017).
65. Kalidindi, S. R., Gomberg, J. A., Trautt, Z. T. & Becker, C. A. Application of data science tools to quantify and distinguish between structures and models in molecular dynamics datasets. *Nanotechnology* **26**, 344006 (2015).
66. Gupta, A., Cecen, A., Goyal, S., Singh, A. K. & Kalidindi, S. R. Structure–property linkages using a data science approach: Application to a non-metallic inclusion/steel composite system. *Acta Mater.* **91**, 239–254 (2015).
67. Brough, D. B., Wheeler, D., Warren, J. A. & Kalidindi, S. R. Microstructure-based knowledge systems for capturing process-structure evolution linkages. *Curr. Opin. Solid State Mater. Sci.* **21**, 129–140 (2017).
68. Panchal, J. H., Kalidindi, S. R. & McDowell, D. L. Key computational modeling issues in integrated computational materials engineering. *Comput. Aided Des. Appl.* **45**, 4–25 (2013).
69. Brough, D. B., Wheeler, D. & Kalidindi, S. R. Materials knowledge systems in python—a data science framework for accelerated development of hierarchical materials. *Integr. Mater. Manuf. Innov.* **6**, 36–53 (2017).
70. Kalidindi, S. R. Computationally efficient, fully coupled multiscale modeling of materials phenomena using calibrated localization linkages. *International Scholarly Research Notices* **2012**, 1–13 (2012).
71. Adamson, G. W. & Bush, J. A. Method for relating the structure and properties of chemical compounds. *Nature* **248**, 406–407 (1974).
72. Adamson, G. W., Bush, J. A., McLure, A. H. W. & Lynch, M. F. An evaluation of a substructure search screen system based on bond-centered fragments. *J. Chem. Doc.* **14**, 44–48 (1974).
73. Judson, P. *Knowledge-Based Expert Systems in Chemistry: Not Counting on Computers* (Royal Society of Chemistry, 2009).
74. Huan, T. D. et al. A polymer dataset for accelerated property prediction and design. *Sci. Data* **3**, 160012 (2016).
75. Mannodi-Kanakkithodi, A. et al. Rational co-design of polymer dielectrics for energy storage. *Adv. Mater.* **28**, 6277–6291 (2016).
76. Treich, G. M. et al. A rational co-design approach to the creation of new dielectric polymers with high energy density. *IEEE Trans. Dielectr. Electr. Insul.* **24**, 732–743 (2017).
77. Huan, T. D. et al. Advanced polymeric dielectrics for high energy density applications. *Prog. Mater. Sci.* **83**, 236–269 (2016).
78. Sharma, V. et al. Rational design of all organic polymer dielectrics. *Nat. Commun.* **5**, 4845 (2014).
79. Lorenzini, R. G., Kline, W. M., Wang, C. C., Ramprasad, R. & Sotzing, G. A. The rational design of polyurea & polyurethane dielectric materials. *Polymer* **54**, 3529 (2013).
80. Liu, C. S., G, P., C, W. & R, R. How critical are the van der waals interactions in polymer crystals? *J. Phys. Chem. A* **116**, 9347 (2012).
81. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* **6**, 20952 (2016).
82. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
83. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B Condens. Matter* **92**, 014106 (2015).
84. Mannodi-Kanakkithodi, A., Huan, T. D. & Ramprasad, R. Mining materials design rules from data: the example of polymer dielectrics. (Under Review). *Chem. Mat.* **29**, 9001–9010 (2017)
85. PolymerGenome. http://polymergenome.org.
86. Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22**, 3762–3767 (2010).
87. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
88. Behler, J., Martonák, R., Donadio, D. & Parrinello, M. Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential. *Phys. Rev. Lett.* **100**, 185501 (2008).
89. Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys. Condens. Matter* **26**, 183001 (2014).

90. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
91. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
92. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
93. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B Condens. Matter* **87**, 184115 (2013).
94. Szlachta, W. J., Bartók, A. P. & Csányi, G. Accuracy and transferability of gaussian approximation potential models for tungsten. *Phys. Rev. B. Condens. Matter* **90**, 104108 (2014).
95. Bartók, A. P. & Csányi, G. Gaussian approximation potentials: a brief tutorial introduction. *Int. J. Quantum Chem.* **115**, 1051–1057 (2015).
96. Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B Condens. Matter* **95**, 094203 (2017).
97. Jindal, S., Chiriki, S. & Bulusu, S. S. Spherical harmonics based descriptor for neural network potentials: structure and dynamics of Au_{147} nanocluster. *J. Chem. Phys.* **146**, 204301 (2017).
98. Thompson, A., Swiler, L., Trott, C., Foiles, S. & Tucker, G. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).
99. Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **115**, 1058–1073 (2015).
100. Li, Z., Kermode, J. R. & De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
101. Botu, V. & Ramprasad, R. Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B. Condens. Matter* **92**, 094306 (2015).
102. Glielmo, A., Sollich, P. & De Vita, A. Accurate interatomic force fields via machine learning with covariant kernels. *Phys. Rev. B. Condens. Matter* **95**, 214302 (2017).
103. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.* **115**, 1074–1083 (2015).
104. Botu, V., Chapman, J. & Ramprasad, R. A study of adatom ripening on an al (111) surface with machine learning force fields. *Comput. Mater. Sci.* **129**, 332–335 (2017).
105. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *J. Phys. Chem. C.* **121**, 511–522 (2017).
106. Feynman, R. P. Forces in molecules. *Phys. Rev.* **56**, 340–343 (1939).
107. Bianchini, F., Kermode, J. R. & De Vita, A. Modelling defects in Ni–Al with EAM and DFT calculations. *Modell. Simul. Mater. Sci. Eng.* **24**, 045012 (2016).
108. Ercolessi, F. & Adams, J. B. Interatomic potentials from first-principles calculations: the force-matching method. *Europhys. Lett.* **26**, 583–588 (1994).
109. Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R. & Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).
110. Snyder, J. C. et al. Orbital-free bond breaking via machine learning. *J. Chem. Phys.* **139**, 224104 (2013).
111. Snyder, J. C., Rupp, M., Müller, K.-R. & Burke, K. Nonlinear gradient denoising: Finding accurate extrema from inaccurate functional derivatives. *Int. J. Quantum Chem.* **115**, 1102–1114 (2015).
112. Fancher, C. M. et al. Use of bayesian inference in crystallographic structure refinement via full diffraction profile analysis. *Sci. Rep.* **6**, 31625 (2016).
113. Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 6367 (2014).
114. Kusne, A. G., Keller, D., Anderson, A., Zaban, A. & Takeuchi, I. High-throughput determination of structural phase diagram and constituent phases using GRENDEL. *Nanotechnology* **26**, 444002 (2015).
115. Hattrick-Simpers, J. R., Gregoire, J. M. & Kusne, A. G. Perspective: composition? structure?property mapping in high-throughput experiments: turning data into knowledge. *APL Mater.* **4**, 053211 (2016).
116. Bunn, J. K., Hu, J. & Hattrick-Simpers, J. R. Semi-Supervised approach to phase identification from combinatorial sample diffraction patterns. *JOM* **68**, 2116–2125 (2016).
117. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
118. Lookman, T., Balachandran, P. V., Xue, D., Hogden, J. & Theiler, J. Statistical inference and adaptive design for materials discovery. *Curr. Opin. Solid State Mater. Sci.* **21**, 121–128 (2017).
119. Felsenstein, J. Bootstrap condence levels for phylogenetic trees. In *The Science of Bradley Efron*, Springer Series in Statistics (eds Morris, C. N. & Tibshirani, R.) 336–343 (Springer, New York, NY, 2008).
120. Powell, W. B. et al. *Optimal learning*. (Wiley, Oxford, 2012).
121. Powell, W. B. et al. The knowledge gradient for optimal learning. In *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, Inc., 2010).
122. Ryzhov, I. O., Powell, W. B. & Frazier, P. I. The knowledge gradient algorithm for a general class of online learning problems. *Oper. Res.* **60**, 180–195 (2012).
123. Micchelli, C. A. & Pontil, M. On learning vector-valued functions. *Neural Comput.* **17**, 177–204 (2005).
124. Álvarez, M. A., Rosasco, L. & Lawrence, N. D. *Kernels for Vector-valued Functions: A Review* (Now Publishers Incorporated, 2012).
125. Forrester, A. I. J., Sóbester, A. & Keane, A. J. Multi-fidelity optimization via surrogate modelling. *Proc. R. Soc. A* **463**, 3251–3269 (2007).
126. Perdikaris, P., Venturi, D., Royset, J. O. & Karniadakis, G. E. Multi-fidelity modelling via recursive co-kriging and Gaussian-Markov random fields. *Proc. Math. Phys. Eng. Sci.* **471**, 20150018 (2015).
127. Dudiy, S. V. & Zunger, A. Searching for alloy configurations with target physical properties: impurity design via a genetic algorithm inverse band structure approach. *Phys. Rev. Lett.* **97**, 046401 (2006).