Machine learning in materials design and discovery: Examples from the present and suggestions for the future

J. E. Gubernatis and T. Lookman

Phys. Rev. Materials **2**, 120301 — Published 20 December 2018

DOI: 10.1103/PhysRevMaterials.2.120301

LA-UR-18-29277

Use of Machine Learning in Materials Design and Discovery: Examples from the Present and Suggestions for the Future

J. E. Gubernatis and T. Lookman

Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545 U.S.A.

(Dated: November 16, 2018)

Abstract

We provide a brief discussion of "What is machine learning?" and then give a number of examples of how these methods have recently aided the design and discovery of new materials, such as new shape memory alloys, with enhanced targeted properties, such as lower hysteresis. These examples illustrate how discoveries can be made from large databases, for example, those generated by high throughput DFT calculations and also how they can be made from experimentally growing smaller databases in an active learning manner. Additionally, we discuss such advanced machine learning methods as multi-objective and multi-fidelity optimization that permit proposing new materials with the simultaneous optimization of more than one targeted property, such as a material with low hysteresis and high Curie temperature, and permit using fewer costly experiments and calculations by combining them with less costly ones to achieve modeling comparable to using only many costly ones. We conclude with a brief discussion of future machine learning opportunities in the context of high throughput experiment and on-the-fly adjustment of synthesis. More speculatively, we end by discussing how might we tailor material science more fittingly with machine learning.

I. INTRODUCTION

Using machine learning methods to aid the design and discovery of new materials is a rapid growth area in materials research and will continue to be so for the foreseeable future. Without doubt, this explosive growth was fueled by President Obama's 2011 announcement of the Materials Genome Initiative (MGI)[1] and is being sustained by the ensuing governmental infrastructure developed to co-ordinate this initiative's implementation by various agencies. The intent of the MGI is the maintenance of our country's manufacturing competitiveness by halving the time it takes to discover new materials. Ideally, we want to discover game-changing materials, those with one or more properties that lie far beyond those currently known, that enable the development of new technologies and lead to the marketing of new products based on these extraordinary properties. In other cases, alternative materials are sought, perhaps for reasons of cost or environmental friendliness. In MGI announcements, Kevlar and Li-based batteries, are oft-quoted examples of new materials marketed because they have exceptional properties or are desirable replacements.

The MGI has rapidly changed the way many material scientists do research, not only in this country but worldwide. Mandated is research that is much more data and computation driven than in the past. In practice, computations are creating large databases. Machine learning methods are the main tools being used to facilitate the exploration of these data. There is hope and reason to believe that these methods, when applied to properly crafted material databases and properly used, will augment and in many cases supplant the time consuming, intuition-based, trial-and-error experimentation that has been the traditional route to the design and discovery of new materials.

The intent of this article is not to review the progress and status of the MGI but rather to convey "some lessons learned" that illustrate the potential of machine learning methods in several perhaps not so obvious but in fact essential ways. For example, to date, much of the use of machine learning methods in materials science has been strongly coupled to the nearly simultaneous generation of large databases by high-throughput density functional theory (HT-DFT) calculations and in some cases the generation of databases by high throughput experiments. However, examples now exist where the use of less common machine learning tools makes it possible to grow small experimentally generated databases into larger ones and along the way to predict new materials. We review this approach and discuss how it

was also recently used in the context of non-high throughput DFT calculations: Instead of the present agenda of trying to compute all possibilities with computational accuracy sacrificed for speed, the alternative approach focuses on identifying a few possibilities at a time that will make the most difference with respect to what we presently know, computing these possibilities as accurately as possible, and verifying them experimentally.

Whether the databases are large or small, generated experimentally or computationally, the motivation for discovering new materials is coupled in an essential way with the need for ones with specific functionalities more advanced than those we currently have. Generally, improvements in more than one functionality are sought. For example, we might want new shape memory alloys with lower hystereses and higher Curie temperatures. Below, we discuss machine learning methods that optimize the search for materials needing the simultaneous enhancement of multiple properties (multi-objective optimization).

We also note the existence of multi-fidelity optimization methods that permit the combination of calculations with different levels of accuracy, experiments with different measurement precisions, and even calculations and measurements with the resulting precision approaching that of the more expensive calculations or measurements. Examples here include combining many HT-DFT calculations that use less accurate approximations for the electronic interactions with fewer DFT calculations that use more accurate ones to produce band gap predictions whose accuracies approach an all high fidelity analysis.

In short, instead of giving a somewhat standard summary of textbook machine learning ideas and methods, we focus on the broader picture, discuss some newer methods and more importantly reference their successes. Accordingly, we look more to the future than to the past. We are sharing lessons we learned. We acknowledge a number of very recent reviews of the field, for example,[2–12]. The active learning, multi-objective and multi-fidelity methods we discuss are not noted by them. Hence, this update of progress in the field complements the perspectives they provide.

In the next section, we present a brief historical perspective that gives a simple example of how data has been used to search for new materials to show how machine learning allows us to build upon it. We then give a somewhat philosophical description of machine learning. What is it? It is not physics, chemistry, or materials science. How does this domain knowledge enter? It is largely upon us. We then review several applications of different machine learning methods to the prediction of new materials, note cases where the predictions have

been validated experimentally, and thereby illustrate the broad spectrum of applications possible[13]. Finally, we conclude by noting other methods awaiting a chance to impact the design and discovery of new materials.

II. HISTORICAL PERSPECTIVE

Using data and in particular data generated from theoretical calculations to assist in the design and discovery of new material is not new. One of the earliest instances of such an endeavor was the search for new semiconductors, the game changing materials of the 70's, that resulted in the introduction into the materials design and discovery process of what are called structure maps (for example, Fig. 1). Initially applied to octet AB materials[14], these maps are simply scatter plots of two physical properties of the constituent A and B atoms, such as their ionization potentials, valences, ionic radii, etc. A pencil and ruler was used to draw boundaries between the plotted data that cluster known materials with the same crystal structure. The challenge was to identify the physical quantities to place on the x and y axes that promoted the greatest segregation of materials into the same structure. "Holes" in the resulting data clusters represented possible new materials and their likely crystal structures. One of the earliest and most effective structure maps was proposed by St. John and Bloch[15] who chose as the x and y co-ordinates the symmetric combinations

$$r_\sigma = |(r_p^A + r_s^A) + (r_p^B + r_s^B)| \tag{1}$$

$$r_\sigma \ = |r_p^A - r_s^A| + |r_p^B - r_s^B| \tag{2}$$

of the s and p orbital dependent radii of the A and B atoms estimated from an early pseudopotential, a concept then in its infancy. A pioneering paper by Chelikowsky and Philips[16] states the vision:

> "Structural energies are, for the most part, too small to be calculated quantum mechanically. ... However, if we consider the problem from the point of view of information theory, then the available structural data already contain a great deal of information. ... Thus one can reverse the problem, and attempt to extract from the available data quantitative rules for chemical bonding in solids."

In other words, the Periodic Table establishes trends in the chemical properties of the atoms as one moves across its rows. In the solid state, remnants of these trends persist. The

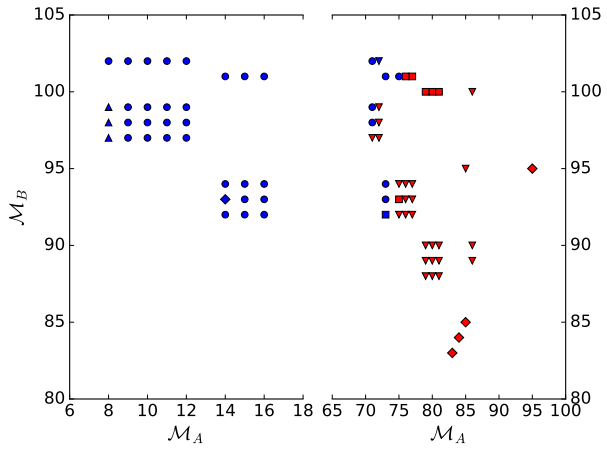

FIG. 1. Structure map of the octet AB compounds with Pettifor's Mendeleev numbers as the co-ordinates. The different symbols shapes denote the different crystal classes: the circles are rock salt; the down-pointing triangles, zinc blende; the squares, wurtzite; the up-pointing triangles, cesium chloride; and the diamonds, and the diamond, diamond. Symbols colored blue mark compounds expected to be ionically bonded (those in rock salt, wurtzite and cesium chloride structures); red, covalently bonded (zinc blende and diamond). For clarity, bounding boxes clustering the different crystal structures are not drawn.

problem is to extract these trends from the data and use them to predict new materials. This vision is also that of much of today's use of machine learning in the materials sciences.

The structure map approach was extended by Zunger[17] to all AB materials with the symmetric combinations of ionic radii but with r_A and r_B computed by a different pseudo-potential, and with a pencil and ruler he clustered the 574 then known AB materials in the observed 34 crystal structures. In the mean time, Pettifor[18] proposed a different set of co-ordinates for a structure map based on what he called Mendeleev numbers. This one-dimensional sequence of numbers relabels the elements in the two-dimensional Periodic Table for the most part by going down the columns. He showed doing this captures co-ordination tendencies of the elements and hence structural similarities between materials differing by the presence of one or more of these elements. These co-ordinates do not require computation or measurement and hence became readily used in structure maps for clustering physical properties other than crystal structure, such as melting temperatures.

What would happen if we were to use simultaneously both the ionic radii and the

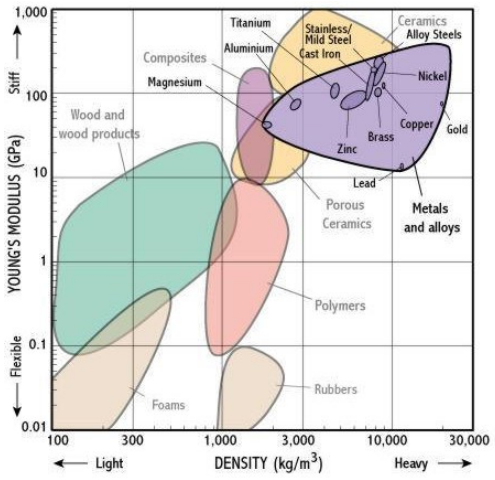

FIG. 2. An Ashby plot with Young's modulus and density as the co-ordinates. Multiple material classes are represented.

Mendeleev numbers to capture the two different types of trends? Obviously, we would need more than a ruler and pencil to identify and draw the boundaries between the data clusters because we do not know how to plot anything in four dimensions. In this particular case, machine learning is the replacement for the pencil and ruler: It allows us to extend the concept of a structure map to three or more dimensions, provides us with a method to remove the subjectivity in the decisions of where to draw the boundaries, and replaces boxy boundaries with more sophisticated and flexible manifolds[19].

While a structure map is useful for a first cut in identifying possible new materials from data, it is clearly limited. It is interesting to note another type of two-dimensional scatter plot, called the Ashby plot (Fig. 2), that is used in the materials engineering problem of selecting the best material to use in a particular application[20]. It displays two properties of many materials for multiple classes of materials simultaneously. Historically, an Ashby plot displayed the Young's modulus versus the density for overlays of metals, polymers, ceramics, foams, etc. Specific procedures evolved for choosing the material class and the material in that class for the application at hand, for example, depending on whether the material would be used as a rod in tension or as a plate subject to bending.

What is characteristic for each material class in an Ashby plot are limiting boundaries of high Young's modulus and low density, high Young's modulus and high density, etc. From the point of view of an Ashby plot, the search for new materials is about pushing one of these

boundaries in a favorable direction. For example, for alloys of the form $A_xB_yC_{1-x-y}$ with A, B and C fixed, for what values of x and y will we find a material with higher Young's modulus and lower density relative to the alloys we already know? As we discuss below (Section VI), these boundaries are called Pareto fronts, and machine learning methods exist that adaptively predict values of x and y whose Young's modulus and density are likely to push the front favorably. Used this way, machine learning proactively addresses a goal of MGI.

III. MACHINE LEARNING

We define machine learning as a collection of computational methods for using information in data we have to make predictions about data we currently do not have. Closely related to machine learning is data mining which uses machine learning and other methods to unveil information *already in the data we have* but is not apparent. Crudely speaking, in our AB material example, observing clusters in the structure plot is data mining; predicting new materials using the rule of filling in the holes in the clusters is machine learning. Better definitions of both fields exist; however, our experience has been that we are asked less often about what is machine learning than we are asked about where is the physics, chemistry, or material science. We already stated that there is no physical science in the machine learning methods. They are application neutral. Our example about structure maps gives a hint about how the science enters. In part, it is through the data. To understand the entry better requires that we discuss a bit more about what machine learning is and how it is validated.

Except for the biosciences (for example, the Human Genome Initiative), chemical sciences (Chem-informatics), and astronomy (for example, satellite data analysis), the other physical sciences, including the material sciences, are somewhat "Johnnies come lately" in their use of machine learning methods. The widespread use of these methods in the engineering sciences, social sciences, financial sciences, statistics, marketing, etc. have lead to a plethora of methods and techniques that are application independent, even though each originated in a particular field. For example, below we discuss the use of machine learning in identifying possible new perovskite materials. The gradient tree boosting method used there has been used to study the short-fin eel population in waters off New Zealand.

In machine learning, there are many generic tasks[21–23]. These tasks include modeling the data by some probability distribution function, clustering the data, classifying the data, regression analysis, feature reduction, etc. For each task, numerous methods exist. What defines the discipline using any of these methods is not just the source of the data and the information sought from it but also the parametrization of the data. The parameterization defines the variables we use in the machine learning. These variables are most often called the features but sometimes called descriptors. The features are ultimately the main portals for the science.

In a few more words, it is useful to think of a data point as a sample drawn from some probability distribution function that represents the complete description of the problem. Of course, we do not know this distribution. Further, we do not know the independent features on which it depends. For a given material, on the other hand, we can easily conjure up a host of physical properties relative to the constituents of the material and to the material as a whole that we believe are relevant and use all as features. In the Historical Perspectives section, we noted that different parameterizations of structure maps, that is, different choices of features, changed how well the data separated into crystal structures. In machine learning language, the search for the best structure map was a search for the best feature set. In short, using what we know *a priori* about what physical quantities control the properties of interest is currently where our domain knowledge mainly enters. Often, we can propose too many features with most not being independent of each other. Occam's razor controls our psyche: Less is more. While there are machine learning methods to aid in identifying from this set the ones most and least important, the burden is upon us to choose how we populate the set. We can depopulate or repopulate as needed.

In the physical sciences, we typically look at data to establish general principles and then use these principles to predict beyond what we know. On the one hand, we are trying to use machine learning to help us extend our knowledge of materials beyond what we know. On the other hand, we can ask what are the general principles for using these methods? Unfortunately, there are at best a few.

First, we point out the No Free Lunch Theorem (NFLT)[24] that says "A universal optimizer does not exist." Most machine learning methods optimize something, typically performing a *constrained* fit of a cost, loss, or utility function to the data. The NFLT suggests there is no best optimizer for doing this. In part, this is why there are so many methods. You

might have several methods that you like more than others. For a given application, one will work better than the rest. Typically, you will need to try all to find out which one. If you change the problem, for example, by adding a lot more data or by adding and removing some features, another of your favorite methods might rise to the top.

Likely while in high school or thereabouts, we all were told if you use enough parameters you can fit anything. Indeed this is the case. In machine learning, this dictum translates into the bias (the error of the fit) versus variance (the variations in its predictions) problem[22]. In standard applications, machine learning returns a statistics-based model built upon the data. In making predictions, we use this model mainly to interpolate between the data we know. If we fit the data too accurately (low bias), the risk is we become extremely limited in how far we can controllably interpolate (high variance) between data points.

How does machine learning address the bias versus variance problem? In most cases, some form of a technique called cross-validation is used[22,23]. In cross-validation, we machine learn on only part of the data and then test the resulting model by observing how well it predicts the held-out data. By doing this multiple times for different sets of hold-outs, we can compute, for example, the average accuracy of the model as well as the average's standard deviation. In most fields, a high accuracy with a small standard deviation from the cross-validation are the bases for confidence in the predictions of the model. In the physical sciences, we also have the luxury of asking how well do these predictions compare with experiment.

IV. LARGE DATABASE EXAMPLES

Since x-ray diffraction was discovered, databases of measured structures of molecules and solids have been assembled and grown large. Several well known structural databases are the Cambridge Structural Database (CSD)[25] and the Inorganic Chemistry Structural Database (ICSD)[26]. Today's databases combine to cover organic, metal-organic, purely inorganic compounds (including pure elements, minerals, and inter-metallic compounds), and alloys. The ICSD, for example, is devoted to inorganic compounds, has nearly 200,000 entries with about 6000 added per year, and returns for each entry such information as the unit cell, space group, atomic parameters, site occupations, Wyckoff positions, molecular formulas and weights, mineral groups, etc. Because of repetitions, incorrect entries, incom-

plete entries, etc., the effective sizes of these databases are smaller than the size advertised. Additionally, not all known materials lie in them. While giving detailed structural information of a material, the entries give little information about other physical properties and functionalities.

What are some of the uses made of this structural information of already fabricated materials? Returning to our example of AB materials, the central task for finding new such materials is finding from a known compound, say NaCl, what substitutes for Na or Cl. Here, we would naturally try atoms chemically similar to Na or Cl, expecting to keep the result in the same crystal structure. In short, the structural databases for fabricated materials contain information about the observed substitutions of one element for another for various crystal structures and chemical compositions. In particular, they give an empirical likelihood that element A can be replaced by another element B and remain in the same crystal structure. Using the information in the ICSD, Glawe *et al.*[27], for example, constructed a matrix where each entry (A,B) measured this likelihood, and then by using a sparse matrix method to reorder its rows and columns to reduce the bandwidth of the matrix, they generated blocks of structurally and chemically similar compounds along the diagonal of the reordered matrix to create a modified Pettifor chemical scale. This new scale was similar to the one found by Pettifor who used intuition and trial and error on a much smaller set of data. Thus, they reaffirmed and sharpened this classic work. Similar in spirit is the work of Hautier *et al.*[28]. Using machine learning methods, they constructed not a matrix but a probability function for substitution that fitted the known data very well, and from it re-established such relatively known trends as rare-earths substituting freely for each other as do the alkali elements. With the model, they can now predict the likelihood (probability) of substituting ions into known structures to produce compounds not yet in the ICSD; that is, with it, they can propose not yet formed prototype polymorphs.

Today, the biggest databases of structural and other material information are those recently created by HT-DFT calculations. Three such databases are the Materials Project[29], AFLOW[10], and the Open Quantum Materials Database (OQMD)[30], plus the NOMAD Repositiory which contains information from the previous three and more[31]. The AFLOW database has nearly 2 million DFT calculations based on 60,000 ICSD entries and nearly 2 million prototypes. The OQMD has information generated by over 470,000 DFT calculations based on 40,000 compounds in the ICSD (compounds already formed) plus 430,000

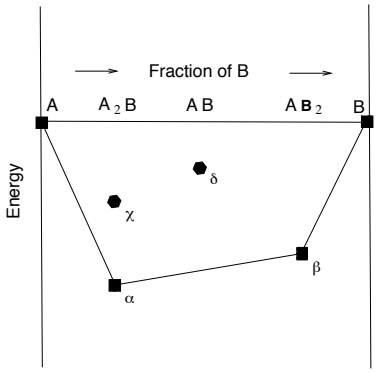

FIG. 3. A simple illustration of a $T = 0$ and $P = 0$ convex hull. The energy is plotted vertically and compositions of various compounds formed from the A and B atoms are plotted horizontally. The solid lines connect points on the hull. Phases α and β lie on the hull and hence are in some stable crystal structure. Phase χ, at composition A_2B, lies above the α phase and hence is a metastable A_2B of a different structure (a polymorph). δ denotes a case where the composition decomposes into the stable α and β phases. After[29], supplementary material.

prototypes (compounds not yet formed) generated by decorating the known crystal structures in the ICSD with the combinatorial replacement with all ions in the periodic table, observing such constraints as charge neutrality. For each compound, the database has the compound's space group, formation energy, number of atoms per unit cell, lattice vectors, density of states, visualization of the crystal structure, etc. For a given chemical composition, it calculates its zero temperature and zero pressure phase diagram based on a convex hull (CH) analysis using all the compounds in the database (Fig. 3). To determine in which ground state a chemical composition is stable, the analysis finds the set of phases which have a ground state energy lower than all other structures or a linear combination of structures. These ground states are linked to form a convex hull. The stability of a chemical composition in a given structure is measured by the energy difference between it and the CH. A stable compound sits on the hull and hence has an energy difference of zero. A metastable compound sits above the hull. The analysis allows for a composition to decompose (phase separate) into two or more ground state compositions in appropriate crystal structures. The energy of a decomposing compound also sits above the hull. Solid solutions

are not considered.

While the databases help identify prototypes, the challenge is understanding which of them are likely to be formed at non-zero temperatures and pressures. Part of the problem is many compounds recorded in the ICSD are in non-zero temperature and pressure structures that are thermodynamically metastable. These structures can differ from the structures which are stable in the ground state, or they might not have stable ground states and, for instance, decompose. In fact, the same ground state uncertainties could occur if the material is thermodynamically stable. Simply put, many compounds in the ICSD sit above the DFT computed convex hull. Using the Materials Project database, Sun *et al.*[32] studied this distance as a function of chemistry and composition. For predicting formable compounds, they proposed including those within a certain distance above the hull, which we call the degree of metastability, in addition to prototypes on the CH.

Different HT-DFT databases can give different predictions depending on how the DFT is done and the CH is constructed. The latter depends on the type and number of crystal structures in which the chemical composition is found or is assumed found and the number and types of structures into which it is assumed to decompose. With respect to the former, Legrain *et al.*[33] recently illustrated these differences in *ab initio* approaches in a paper on materials screening for the discovery on new half-heuslers. They found that the *ab initio* approaches showed significant inconsistencies among themselves about whether a given chemical composition would be found in a half-heusler structure or not. The chemical compositions they considered were taken from the ICSD and a set of prototypes generated by a combinatorial decoration of the three allowed crystal structures. They found the *ab initio* predictions were also significantly inconsistent with the experimentally observed structures, but machine learning models trained on the experimental data had cross-validated predictions of accuracy of 91% in predicting the known half-heuslers and 99% in predicting those that were not.

The work of Balachandran *et al.*[34] on predicting new perovskites with OQMD found a similar degree of inconsistencies with known data and a similar degree of success with the cross-validated predictions of machine learning models. They emphasized that part of the reason for inconsistencies is that the CH analysis and machine learning are predicting two different substantive quantities. The CH analysis is predicting the zero temperature mechanical stability of the crystal structure; the machine learning models are predicting

formability as they are trained on data of materials formed and measured. If a degree of metastability is added to the CH analysis, as suggested by Sun *et al.*, the OQMD stability predictions show more agreement with compounds known to have been formed. Still this agreement is only 67%. It was proposed that if machine learning predicts a compound is formable and OQMD predicts it is stable, then this compound is more likely to be formable experimentally as a perovskite than the other machine learning predicted prototypes. In this way, a total of 87 new perovskites, including 6 in the cubic phase, were proposed for synthesis. By and large, these proposed perovskites have a lanthanide or actinide occupying the A and B sites of the perovskite structure or A being an alkali, alkali rare earth or transitional metal, and B being a *p*-block element. This proposal awaits experimental verification.

What were the machine learning analyses of Legrain *et al.* and Balachandran *et al.*? Both combed the ICSD and recent literature for known formable compounds with the ABC chemical composition of half-heuslers or ABO_3 composition for oxide perovskites. Each material class is defined by a limited set of space groups. For half-heuslers, it is 216; for perovskites, fifteen[35]. If the compound belongs to the right space group or groups, it was given the label "half-heusler" or "perovskite." If it was anything else, for example, a stable compound in another space group or compound decomposing into a number of compositions in other structures, it was given a label "not-half-heusler" or "not-perovskite," thereby reducing the prediction problem to what in machine learning is called a binary classification problem. From existing experimental information, Legrain *et el.* identified 164 half-heuslers, 11,022 not-half-heuslers, and 71,178 prototypes; Balachandran *et al.* identified 254 perovskites, 136 not-perovskites, and 625 prototypes. The tasks are now to specify a set of features, choose a machine learning classifier, and cross-validate the data. The result is a model into which features of prototype compounds are inputed and outputted are the predictions of being a half-heusler or a perovskite or something else which is known only to be not a half-heusler or a perovskite.

For the binary classifiers, Legrain *et al.* used random forests[36], and Balachandran *et al.* used random forests and gradient tree boosting[37]. These machine learning methods are examples of ensemble methods, meaning that their results are combinations of the results of more than one model. A random forest method is a linear combination of the application of the decision tree method (a strong classifier) to many random samples of the data and sub-sets of its features. A decision tree method[21–23] classifies the data by scanning the

feature sub-sets and splitting the data into two pieces based on the values of the one feature that maximizes a function that defines a decision boundary. Often, this function is the information theory entropy so that the split pieces represent an information gain relative to the unspilt piece. Each piece is then split into two and so on. Usually, the splitting is stopped before each piece is one data item. The depth of the tree determines the tightness of the fit. Generally, a random forest method is well suited to control variance. A gradient tree boosting method uses an ensemble of weak classifiers built iteratively. Weights are assigned to the data. If at a given step a datum is misclassified, its weight is increased before moving to the next step. The current classifier is modified by adding to it the gradient of a function that refines the classification boundary. The weak classifier in the gradient tree boosting method used by Balachandran *et al.*[38] was, roughly speaking, a random forest of shallow tree depth (weak classifiers). Generally, accuracy is the advantage of the gradient tree boosting method.

Both groups started with a relatively long list of features. One characteristic of a decision tree method is its ability to return an estimate of the relative importance of a given feature to the fit. Roughly, this is a measure of the frequency with which a feature triggered a split. Eventually, Legrain *et al.* used a recursive method to reduce their list of relative importances to six features that included such quantities as ratios of the C and B ionic radii and C and A electronegativities, plus several statistical covariances computed between various pairs of radii and eletronegativities. Balachandran *et al.* borrowed from prior experience[39] and selected four sets of two feature pairs: the tolerance and octahedral factors, Shannon's ionic radii for the A and B atoms[40], the bond valence theory distances between the A and O ions and the B and O atoms[41], and Villars's simple choice of Mendeleev numbers[42]. The first feature pair has a long standing use in structure maps for classifying perovskite materials. The other three have been receiving more recent use. They found that the accuracies and variances of predictions of the first two classifiers agreed on the average, and using Mendeleev numbers or bond valence distances produced a slightly less accurate model than using the other two feature pairs. Besides classifying perovskite or not, they also classified a perovskite whether it is cubic or not. Beyond their utility as repositories of information, these examples illustrate how databases of DFT calculations, in conjunction with experimental databases, can be employed to discover possible new materials.

V. SMALL DATABASE EXAMPLES

In contrast to the combinatorial approaches in high throughput calculations that generate large data sets on ideal systems at $T = 0$ and $P = 0$, real materials problems typically involve multi-components, solid solutions and defects at $T \neq 0$ and $P \neq 0$. While for crystals, DFT often returns reliable structural information and useful estimates of such physical quantities as dielectric and elastic constants, in general DFT cannot return information on important functionalities such as whether the material is a superconductor, anti-ferromagnetic, etc. Most functionalities are established and quantified by experiment. As experiments can be quite time consuming and expensive, often data on only relatively few well-characterized samples (between 10 and 100) are available. Hence, it is important to consider approaches that we can tailor and apply to small data sets. The important question becomes how can we learn from the existing limited data and guide the next experiments in a way that minimizes the number of new materials and measurements needed to find a material with an enhanced targeted property[43]. Industry, as well as application areas such as drug design and cancer genomics, are very much at the forefront in developing iterative feedback methods, that is, adaptive learning methods, to reduce the number of experiments or calculations needed[44,45]. Such methods fall within the scope of what is known as Bayesian Global Optimization (BGO)[46,47]. Here, we will describe how these methods are now starting to be used in materials science.

The essential idea is that we first use machine learning methods to construct a "surrogate" representing the data. The problem can be one of binary classification, in which case we can use the methods described above, or optimizing a property, such as a Curie temperature, in which case to construct the surrogate we need a regression method, such as support vector regression (SVR)[21–23], to fit the features parameterizing the data to the property. If all we were doing is machine learning, we would then use the learned model to make a prediction. However, this prediction is not necessarily optimal as it merely "exploits" the model's prediction. If we think of a cost function landscape for the property in the space of features and if the cost function is not convex, such a "best value" prediction would likely correspond to a local minimum[48]. To minimize the number of experiments, we need a means to "explore" this landscape by choosing a better next experiment than merely using the best prediction from the model[49]. In some sense, the surrogate alone allows us

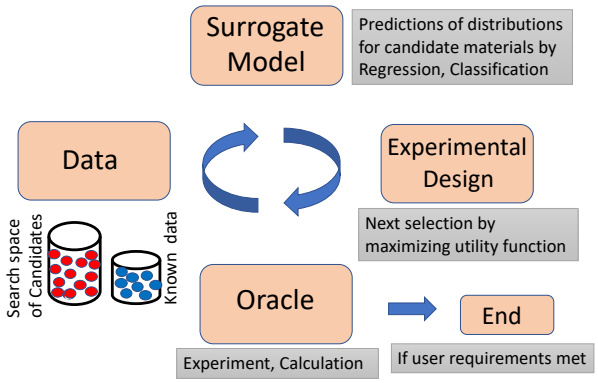

FIG. 4. An active learning loop for finding optimal targets includes a surrogate model learned from data and an experimental design component with a utility function that encodes the basis for selection of the next experiment or calculation. Possible compounds are ranked based on maximizing the expected utility and the idea is for the iterative loop to successively improve the search.

to interpolate between data points, but what we are looking for is a means to explore (extrapolate) beyond the points where we are most certain.

A Bayesian approach for the surrogate model, sampled from a prior distribution that incorporates smoothness and locality, has been found to be a powerful means to find the extrema of functions. Moreover, assuming the sampling to be a Gaussian process[50] has been demonstrated to be efficient in guiding the next experiment. The predicted mean and variance of this process subsequently serve as the input to a utility function, which prioritizes the basis of the decision to be made of what to do next in terms of doing a particular experiment or calculation[51]. Thus, the next experiment is chosen based on maximizing the expected utility function from the list of the many allowed possibilities. This aspect of choosing the optimal next experiment, or experimental design as its known in statistics[52,53], based on the prediction of the surrogate model is part of the active learning loop (Fig. 4), which is repeated until the desired outcome, for example, a material with an optimal property, is met.

Since a Gaussian process is frequently encountered in physics and statistics, and we will reuse it in later Sections, we briefly describe it here. It is a collection of random variables such that any finite collection of them has a multi-variate Gaussian distribution. In terms of machine learning algorithms, they belong to the class called Bayesian methods.

These methods do not target a best fit but instead compute something called a posterior (probability) distribution over models. The latter provides a quantification of the uncertainty in the predictions of the model. The significance of assuming Gaussian distributions for regression problems is that we can exactly perform many of the required algebraic operations and integrations by executing simple linear algebra operations on vectors and matrices. Here, we will just summarize a few results of lengthy analyses[50].

We assume our data, $\mathcal{D} = \{(\vec{x}_i, y_i)|i = 1, N\}$, is drawn from some underlying probability distribution $f(\vec{x})$ where $\vec{x}$ is a vector whose components represent the known values of the features chosen for the problem. A Gaussian noise model $\mathcal{N}(0, \sigma_n)$ represents the error η_i associated with the measurements or predictions of y_i, our targeted physical observable,

$$y_i = f(\vec{x}_i) + \eta_i. \tag{3}$$

Given the data, what we want is an estimate of the mean value y^* of y for a proposed set of features $\vec{x}^*$. In terms of probability theory, we seek the conditional probability $P(y^*|\vec{x}^*, \mathcal{D})$. For a Gaussian process, we can show that

$$P(y^*|\vec{x}^*, \mathcal{D}) = \mathcal{N}(\mu, \Sigma) \tag{4}$$

where the mean μ and covariance matrix Σ of this Gaussian distribution are

$$\mu = K_{xx^*} K_{xx}^{-1} y$$
$$\Sigma = K_{x^*x^*} - K_{xx}^T K_{xx^*}$$

The K's in turn are block matrices computed from a covariance matrix function

$$K = \begin{pmatrix} K_{xx} & K_{xx^*} \\ K_{xx^*}^T & K_{x^*x^*} \end{pmatrix} \tag{5}$$

with the matrix K_{xx} expressing the covariances among the observed values of the features, the matrix K_{xx^*} between the observations and the proposed feature, and the matrix, $K_{x^*x^*}$ between the proposed features. In essence,

$$\begin{bmatrix} \vec{y} \\ y^* \end{bmatrix} \sim \mathcal{N}(\vec{0}, K). \tag{6}$$

The covariance matrix K, usually called the kernel, is some assumed function that usually depends on the displacement $\vec{d} = \vec{x} - \vec{x}'$ in the space of features. For example, a simple,

natural, but not always the most common choice is

$$K(\vec{d}) = \sigma^2 \exp\left(-\frac{|\vec{d}|^2}{2\ell^2}\right) + \sigma_n \delta(\vec{x} - \vec{x}'). \tag{7}$$

After the parameters of the kernel, in the above case σ, σ_n and ℓ, are adjusted to the data by a procedure called the maximum likelihood method, we can insert the kernel and the proposed $\vec{x}^*$ into (4), find μ, our estimate for the new y^*, and use Σ as the estimate of its uncertainty.

Thus, using a Gaussian process for the surrogate model, maximizing a utility function, and feeding back the new information to augment the data set, retrains the GPs and forms an active learning loop that is the basis of BGO. We can choose the utility function from a range of functions, including minimal or maximal variances in the predictions for allowed possibilities, upper or lower confidence bounds indicating a linear combination of exploitation and exploration but with the linear coefficient varying with the number of measurements[54], relative entropy (Kullbach-Leibler divergence), modified objective cost of uncertainty[55–57], the well known criteria of improvement from the current best in the training data, such as probability of improvement $P[I]$[58] or expected improvement $E[I]$[43,59], and extension of expected improvement to nonzero measurement noise known as sequential kriging optimization[60] or knowledge gradient[61], which aims to find the sample that most improves the model rather than maximize expected improvement. The relative performance of several of these functions have been recently compared for the case when the surrogate model is well matched to the data[62]. If μ^* is the best value of the material property y obtained at some point, then the expected improvement $E[I]$ possible by testing the material with proposed features $\vec{x}'$ is given by

$$E[I] = E[\max[(y - \mu^*), 0]] = \int_{-\infty}^{\mu^*} (y - \mu^*) p(y|\vec{x}') dy, \tag{8}$$

where $I = \max[(y - \mu^*), 0]$ is the improvement and the possible values of y are Gaussian distributed according to $p(y|\vec{x}')$. Simple manipulations lead to the result

$$E[I] = \sigma[\phi(z) + z\Phi(z)], \tag{9}$$

where $z = (\mu - \mu^*)/\sigma$ and $\phi(z)$ and $\Phi(z)$ are the standard normal density and cumulative distribution functions. In spite of its simplicity, the expected improvement is found to be an excellent performer for extrema problems; however, maximum variance performs well for other types of problems.

These methods have been utilized in computational codes to rapidly and efficiently learn features to attain targeted properties. An example is the design of light emitting diodes (LEDs) using APSYS, an industry standard code in the field of semiconductor physics. One finds that an efficiency of 75% is obtained within 10% of the total iterations that could be performed[63]. Thus, these methods show enormous promise in minimizing computational costs incurred in running simulations designed to find optimal solutions.

In materials science, these ideas recently accelerated the discovery of a number of new alloys and ceramics with a feedback loop that involves experiments instead of calculations. An example is the search for NiTi-based shape memory alloys with very small thermal hysteresis. Thermal hysteresis governs fatigue, and the idea is to find chemistries and compositions which will minimize thermal hysteresis. The approach of Xue *et al.*[64] assumes a family of alloys defined by $Ni_{50-x-y-z}Ti_{50}Cu_xFe_yPd_z$, where x, y and z are compositions constrained by $50 - x - y - z \leq 30$, $x \leq 20$, $y \leq 5$ and $z \leq 20$ to avoid undesirable solid-solutions. The number of components here is less than five, but in principle we can include more components at the expense of having a larger search space of allowed possibilities. For this example, 22 well-characterized samples from the same laboratory comprised the training data out of a possible space of $800,000$ allowed compositions. The best compound that had a thermal hysteresis as small as $1.84K$ was found on the sixth iteration. Of the 36 compounds synthesized and characterized, 14 had better performance than those in the initial training data. One can ask if these findings are the result of random occurrence. In statistics, a P-value (due to the eminent statistician R. A. Fisher) is often used as a measure of statistical significance. For the alloy problem, it can be shown that $P < 0.001$, implying that the probability that the results are based purely on random chance is very small. The study compared the performance of several surrogate models and utility functions and found the combination of a support vector regressor (SVR) for the surrogate model and the $E[I]$ for the utility function had the best performance on the training data. In contrast, Yuan *et al. experimentally* compared the performance of four utility functions in their search for $BaTiO_3$-based piezoelectrics with large electrostrains. Their objective was to find a solid solution constrained to the family of compounds given by $Ba_{1-x-y}Ca_xSr_yTi_{1-u-v}Zr_uSn_vO_3$ with a large strain at an electric field of 20 kV/cm. Here x, y, u and v are the mole fractions of specific dopants that obey $1 - x - y > 0.6$, $x < 0.4$, $y < 0.3$, $1 - u - v > 0.6$, $u < 0.3$ and $v < 0.3$. There are potentially about $605,000$ possible compositions (controlled

within 0.01%). Yuan *et al.*'s training data was 61 compounds that they synthesized under controlled conditions in their laboratory[65]. Clearly, this problem's large search space cannot be explored experimentally by just trial and error. The BGO design strategy with $E[I]$ found the optimal compound with a Sn composition of 3% in the third iteration. On either side of this Sn composition, the strain decreases. These examples illustrate the efficacy of discovering new compounds with targeted properties in an active learning strategy involving experiments. Other types of examples include work on guiding DFT calculations towards targeted regions[66,67], use of probability of improvement to identify grain boundaries[68], optimization of graphene nano-ribbons to identify configurations with high thermoelectrical properties[69], maximizing the intermolecular binding energy for lead ion solvation in hybrid inorganic-organic perovskites[70], use of knowledge gradient function to study the stability of an emulsion as well as to maximize the output current in an optoelectronic device[71], maximizing the expected utility of the KL divergence to select parameters of a Cahn-Hilliard model for thin films on a substrate[72] and minimizing the difference between predicted properties and experiments of a polymer nanocomposite[73] using GP and expected improvement. The influence of different means of estimating model uncertainty for a number of materials data sets has also been examined[74]. The increasing use of high throughput synthesis and characterization methods in materials science will make BGO type approaches even more relevant as decisions are made on-the-fly.

It is important to recognize that the BGO strategy is heuristic, and an important unmentioned consideration is the stopping criterion. In the previous examples, the iterative loop was performed until a material was discovered with acceptable performance superior to any in the training data. However, there is no assurance that in subsequent iterations the performance will not degrade. In addition, a number of examples suggest that even when the surrogate is not particularly good (poorly fits the data), the resultant BGO scheme performs quite well in finding good solutions. Thus, these observations raise questions about how good the surrogate model needs to be. Answering these questions represent active areas of research.

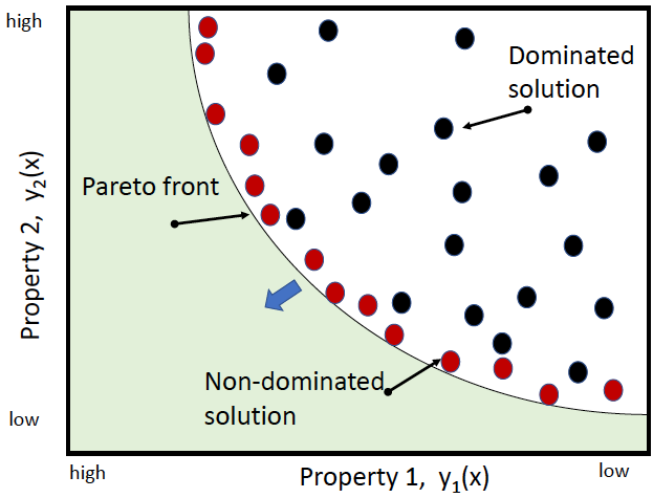

FIG. 5. A Pareto front for two objectives or properties, y_1 and y_2. A solution is a material with a given combination of y_1, y_2, which is viable and meets the constraints but is not necessarily optimum in either criterion. A dominated solution (black dot) is one where there exists some other solution better in both or either of y_1 and y_2, and a non-dominated solution (red dot) is one where no other solution is better in y_1 and y_2. The trade-off line or surface of non-dominated solutions is the Pareto front. The arrow is the direction in which the PF needs to move in order to find materials with large y_1 and small y_2.

VI. MULTIOBJECTIVE OPTIMIZATION

It is often the case in materials science that more than one property is of interest so that we may wish, for example, to maximize one property but minimize the other with the aim of finding the material with the best trade-off in properties. We often display the data for two properties in a *Pareto* plot (Fig. 5) in which the axes correspond to the properties, and we then identify a special boundary in the plot on which lie materials where none of the properties can be improved without deteriorating the value of the other property. These materials define a Pareto front (PF) that represents the best trade-off, and well known examples of such fronts include the Ashby plots.

The method of Section V generalizes to more than one property so that we may select data or samples such that the candidate material after measurement improves the existing PF (in the green-shaded area of Fig. 5. With two properties, we need to train a surrogate model for both properties independently with their requisite features. Within the BGO method, we can

use a Gaussian Process or another predictor, such as SVR, or an ensemble-based model, such as Gradient Tree Boosting, for the surrogate model. The GP naturally gives a distribution with its mean and variance. For the latter predictors, we generate an ensemble of model predictions from which we can then estimate the mean of the objectives and their variances using the statistical method of bootstrapping[21–23] by sampling a large number (typically 1000-5000) models with subsets of the training data selected randomly with replacement. An important question in building a machine-learned model is ensuring that we do not underfit or overfit the data. The use of n-fold cross validation, hyper-parameter tuning, as well as ensemble methods, such as bagging and boosting[22,23], help minimize the risks of underfitting and overfitting. Bagging uses complex models (*e.g.*, random forest) to smooth predictions, whereas boosting uses simpler models (*e.g.*, gradient tree boosting) to boost their aggregate complexity.

The utility function that encodes the decision making process is now a two-dimensional generalization of the one dimension case[75]. So in the case of improvement, we write the probability $P[I]$ that a new sample of a so far unsynthesized compound is an improvement over existing data as the total probability of a candidate data-point with $P[I] = \int_{\text{green}} \phi(y_1, y_2) dy_1 d_{y2}$, where the integration is over the green-shaded region in Fig. 5 and y_1 and y_2 are the properties, and $\phi(y_1, y_2)$ is the uncorrelated Gaussian probability distribution function formed from the mean and variance of y_1 and y_2 distributions, that is, $\phi(y_1, y_2) = \phi(y_1)\phi(y_2)$. We have therefore assumed a Gaussian distribution for the predicted values with a mean and variance. Similarly, the expected improvement $E[I]$ is the first moment of I, the improvement from the best so far, of the joint probability distribution $\phi(y_1, y_2)$ over the green area in Fig. 5. It has been shown[75,76] how to geometrically calculate $E[I] = P[I(x)] * L$ (akin to a moment) in a couple of ways depending on whether the "length" L is evaluated using the Centroid or Maximin approaches, and their relative performance has also been compared[77]. For the Centroid, L is the distance between the centroid at a candidate data point, x, and closest point on the Pareto front. Thus, for possible candidate points in the region of improvement, $E[I]$ is calculated by taking the product of $P[I]$ with the minimum distance between points on the known PF and centroid of the probability distribution within the region of improvement. The candidate point with the largest $E[I]$ is then the choice for the next measurement. Similarly, for the Maximin, L is maximum of the minimum distance of either of the means (μ_1, μ_2) of a particular candidate point from

individual PF points. The former considers improvement over both the properties combined, whereas $E[I]$ for Maximin considers each property separately, takes the one which is smaller from a particular PF point, and then maximizes that amongst all the PF points. Both strategies select a data-point such that its measurement gives the maximum change to the PF. As for the one property case, the process of updating the model predictions and testing new materials tries to ensure that the model is reasonably accurate throughout the whole space ("exploration") and that it also converges to the global minimum rapidly ("exploitation"). Thus, there is competition amongst these goals to accurately learn the model.

It is instructive to add another property to the NiTi-based shape memory alloy example considered previously[64,78]. In addition to searching for a composition to minimize the thermal hysteresis, we also add minimizing the transition temperature. Starting with over 100 well-characterized alloys, each being described in terms of one or more features representing aspects of structure, chemistry, and bonding. The features are selected based on prior materials knowledge. For example, it is known the martensitic transition temperatures (which affect thermal hysteresis) are strongly correlated with the valence electron concentration (fraction of valence electrons) and electron number per atom. In addition, the thermal hysteresis is affected by the atomic size of the alloying chemistries and so the features include a number of different types of radii, electronegativities, and valence electron numbers. Figure 6a shows the optimal Pareto Front with seven data points in this data set of 100 points, and Fig. 6b compares the different strategies and shows that employing multi-objective optimization design strategies decreases the number of measurements required to find the optimal PF by nearly 20% compared to random selection. The Centroid-based design strategy and pure exploration perform similarly; however, the Maximin approach shows superior performance compared to all other strategies, particularly if the prior datasets are smaller. These results are quite general and other data sets, including those from DFT codes, such as for band gaps and dielectric constants, behave similarly[79,80]. Recently a related approach, the hypervolume indicator, formulated the expected improvement in the context of multiobjective optimization[81–83]. The hypervolume is a measure of the size of the space enclosed by all solutions on the Pareto front and a user-defined reference point. The expected improvement in hypervolume is the gain for a given input point and measures the closeness of an approximation or candidate point to the Pareto front. Although not as well

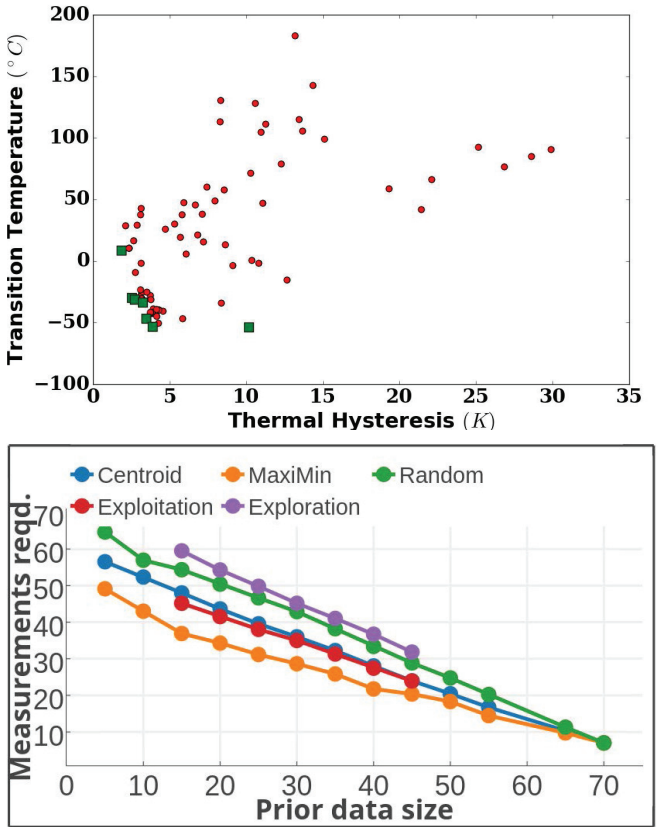

FIG. 6. Upper: The seven points (green) in the optimal PF of a shape memory alloy data set with over 100 points. These were obtained using a design process starting from a set of subset of data-points which are considered as known with the goal to find all the optimal PF points in as few design cycles as possible. Each red colored point is dominated by at least one point in the PF. Lower: A comparison of several selection strategies. For the same data set, the size of the prior training dataset is plotted against the average number of cycles required to find all the points in optimal PF. The design strategies using expected improvement, $E[I]$, for multiobjectives in which the exploration and exploitation of data are more balanced, perform well.

studied, the approach is quite competitive compared to the usual expected improvement criterion on the objectives.

Although there has been considerable work devoted to finding Pareto fronts using techniques as diverse as Monte Carlo sampling[77] and swarm optimization, guiding experiments and calculations towards choosing appropriate materials or features likely to enhance properties in as few iterations as possible, is the ultimate aim of accelerated discovery. We have

therefore focused on this problem and the challenges indicated previously in Sec. V for a single objective also apply here, including the assumption that the data on the random surface follow a Gaussian distribution.

VII. MULTIFIDELITY OPTIMIZATION

A multi-fidelity optimization method, sometimes called co-kriging or sequential kriging, combines many inexpensive lower accuracy calculations with fewer higher accuracy calculations to make predictions whose accuracies are comparable to those produced by higher accuracy computations alone. The greater the difference in the costs of the computations, the greater is the cost advantage of the method. These methods are also based on the mathematics and assumptions of Gaussian processes[50].

In multi-fidelity optimization, using just two accuracies for simplicity, our data becomes $\mathcal{D} = \{\mathcal{D}_{\text{cheap}}, \mathcal{D}_{\text{expensive}}\}$. A requirement is that the set of features in the expensive data is a subset of the features in the cheap data. Then, we assume the expensive data are produced by Gaussian processes that are the differences between the cheap and expensive observations for the shared set of features. With more than two levels of accuracy, we would simply define more Gaussian processes to account for the new data and differences between the accuracies of shared feature subsets at the different accuracy levels. Analysis eventually leads to a conditional probability analogous to (4) but where the kernel K (5) is a 3×3 block matrix instead of a 2×2 block matrix with the blocks indexed by cheap, expensive and proposed.

Without simplifying assumptions, the covariance matrices become large and expensive to manipulate. The main assumption is that for any feature $\vec{x}' \neq \vec{x}$, the Gaussian processes for the cheap calculations do not add any information to the Gaussian process of the expensive calculation for $\vec{x}$. This assumption zeros some off-diagonal blocks of the co-variance matrix. A recent breakthrough was made by Le Gratiet and Garnier[84,85] who proved that with this assumption they could decouple any Gaussian process scheme with s-levels of fidelity and solve the multi-fidelity optimization problem as a sequence of Gaussian processes that successively involve smaller co-variance matrices.

The title of the likely seminal paper in the field, "Predicting the output from a complex computer code when fast approximations are available," by Kennedy and O'Hagan[86] gives a huge hint on the obvious uses of these methods in materials science. For DFT calcula-

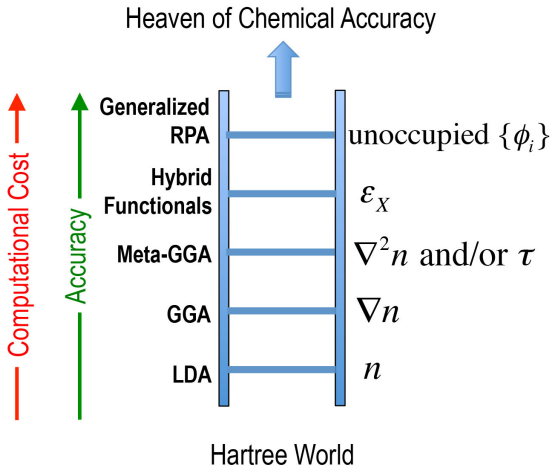

FIG. 7. Jacob's ladder of density functional approximations to the exchange-correlation energy that specifies a prototypical opportunity for multi-fidelity optimization of density functional theory calculations.

tions, for example, a cost-accuracy hierarchy of functional approximations for the exchange-correlation energy is well established (Fig. 7). In fact, Pilania *et al.*[87] used this hierarchy to test two-level multi-fidelity predictions of the band gaps of double-perovskite materials (the elpasolites). Accurate band gap calculations are often a big problem for DFT calculations with the best methods being very expensive. For a set of 600 compounds and semi-local (cheap) and hybrid exchange-correlation (expensive) functionals as the two fidelity levels for DFT calculations, they computed all the higher and lower accuracy band gaps. They then studied how well the multi-fidelity optimization method predicted the higher accuracy band gaps as a function of the number of low and high fidelity data points used in cross-validation. Not surprisingly, if the number of low fidelity data is fixed, the accuracy improves as the number of high fidelity data is increased. More interestingly, if the number of high fidelity data is fixed, the accuracy improves as the number of low fidelity data are increased.

The above study points to the promise of using these machine learning methods for materials science, where we have many complex computer codes and measurements and even more fast approximations. We can envision, for example, using these methods to upgrade HT-DFT databases, which as discussed above can have accuracy problems for certain classes of materials. We can also use them with experimental data or with a combination of experimental and theoretical data. For example, solid state cooling uses the ordering-disordering

in ferroic materials undergoing phase transitions. The temperature change induced by the presence or absence of an external field can be directly measured *in situ* but is time consuming and hence will yield few data points. On the other hand, this change is less accurately but more easily inferred by approximating the differentials in Maxwell relations and using simpler experiments than measuring the changes in the order parameter versus the field. The multi-fidelity methods easily extend to multi-objective predictions: What we called y in the above equations we simply replace by $\vec{y}$ and adjust the dimensions of various matrices accordingly.

VIII. OTHER OPPORTUNITIES

Advances in combinatorial synthesis and characterization are allowing us to explore larger parameter spaces for rapid screening of possible new materials. For example, thin film synthesis by sputtering can create composition gradients in three or more elements to hone in on promising compositions for targeted properties, such as lattice parameters and local co-ordination[88]. This analysis can then be followed by in-depth bulk synthesis on just a few select compounds. This approach is a down-selection strategy rather than an active learning one. In future, advances in measurement techniques will allow more high throughput experiments[89] to be conducted, and then the ability to make on-the-fly decisions as to what next to synthesize and test by using, for example, the methods we discussed, becomes crucial to save cost and time. Feature sets that include processing conditions, such as laser power, travel speeds, and cooling rates, will become important in predicting and controlling the resultant material microstructure in manufacturing, for example, by using laser processing.

National user facilities, such as the Advanced Photon Source at Argonne and the Linac Coherent Light Source at the Stanford Linear Accelerator Center, are the sources of big data in materials science that generate up to 100TBs of data per sample. Reconstructing the real-space macrostructural image from the data is a challenging task. For example, in High Energy Diffraction Microscopy, a forward model needs to be run at every finite element to find the orientation of the "grain" that matches the Bragg spots on the detector pattern[90]. This task requires substantial computational resources, let alone causing a time delay in performing the next experiment. However, machine learning and optimization methods, and in particular convolution neural nets (CNNs)[91], provide a vehicle to construct

models, depending on the numbers of neurons and hidden layers, that directly fit to the large amounts of data to hundreds and thousands of parameters. CNNs are also being increasingly utilized in scientific problems, including in the analysis of the output from large scale computer simulations that generate vast amounts of data. In this activity, speed-up relies on constructing reliable surrogate models whose predictions replace the need to carry out complex calculations associated with the original problem. There is much current debate on the merits of deep learning models, especially because they often appear as "black-boxes." Recent work by Lin et al.[92] employs information theory and the renormalization group to discuss the reasons why deep learning works so well and how "cheap learning" may be crafted with far fewer parameters for functions of practical interest.

IX. SOME CLOSING THOUGHTS

Up to here, we have given admittedly selective glimpses of how machine learning has, can, and is being used to assist the design and discovery of new materials. In closing, we return to the question, "Where is the material science?" In the Machine Learning section, we answered by saying it currently was mainly in the choice of features used to parameterized the data. We now note several other ways it can enter. Materials science, as many other sciences, is based on fundamental principles. We thus know a lot about the data and physical behavior or possible data or behavior before we look at them. How might we take better advantage of this domain knowledge?

In several places, we stated that a particular technique is Bayesian. By this we mean the method was developed and is stated in terms of products and integrals over products of probabilities. For two sets of events A and B, a fundamental result of probability theory is Bayes's Theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{10}$$

where $P(A|B)$ and $P(B|A)$ are their conditional probabilities and $P(A)$ and $P(B)$ are their individual probabilities. For data analysis[93], we usually write this theorem as

$$P(\text{Model}|\text{Data}) = \frac{P(\text{Data}|\text{Model})P(\text{Model})}{P(\text{Data})}. \tag{11}$$

$P(\text{Model}|\text{Data})$ is called the posterior distribution. It represents the probability of the model after taking the data into account. Knowing it constitutes knowing the complete

probabilistic solution to the problem. P(Data|Model) is the likelihood function. It is the probability of the data given the model. P(Model) is called the prior. It represents what we know about the model before the data is taken into account. Lastly, P(Data) is called the evidence. It is the normalization of the posterior, that is, $P(\text{Data}) = \int_{\text{Model}} P(\text{Model}|\text{Data})$.

Generally unstated is that much of machine learning is equivalent to maximizing the log posterior

$$\log[P(\text{Model}|\text{Data})] \propto \log[P(\text{Data}|\text{Model})] + \log[P(\text{Model})] \tag{12}$$

with respect to the parameters of the model with the evidence ignored. For the log prior, assumed functions satisfy general constraints about smoothness, non-negativity, etc. From this point of view, we see that instead of seeking the complete solution, machine learning instead generally settles for estimates of just two moments of that solution, the mean (the location of the maximum of the posterior) and variance about the mean (the width of the posterior distribution about its maximum). The log likelihood is the cost, loss, or utility function for the problem. Standard least-squares fitting of a model to data is equivalent to simply maximizing

$$\log[P(\text{Model}|\text{Data})] \propto \log[P(\text{Data}|\text{Model})] \tag{13}$$

Adding the log of the prior allows the machine learning methods to "regularize" the fitting of the model to the data, making the analysis more robust. Noting the connection of a machine learning method to Bayes's Theorem is not done as generally little insight is gained by doing so. However, with more specific physics-based assumptions for the prior, such as required bounds on model parameters, correlations to be favored by the model, etc., the machine learning would achieve a more specific material science character.

While the Bayesian approach is a natural and standard avenue to ingrain prior domain knowledge, we now speculate about a possible approach that focuses on what we can say *a priori* about the data instead of what we can say about the model. This approach mainly applies to clustering, classification and regression problems.

All the machine learning we discussed starts with a data matrix that typically has materials as its rows and features as its columns. Instead of expressing the data in this manner, we could instead express it for use in a multi-relational learning method[94–97] whereby the data is mapped to, or better yet simply assembled as, multiple matrices of the same dimensions where each matrix groups data among a set of one or two entities according to different

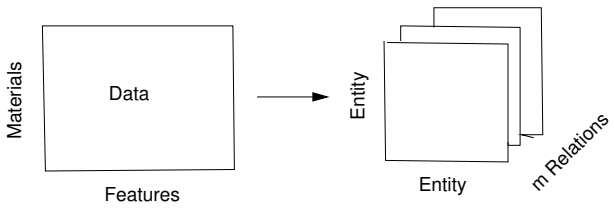

FIG. 8. Schematic mapping of the data matrix to a m-multi-relational representation. The entities labelling the rows and columns may differ. The relation changes from table to table.

relations based on our prior domain knowledge. This mapping is depicted in Fig. 8. The data matrix $\mathcal{D}_{mn}$ maps to a tensor χ_{ijk}. The tasks are to choose the entities and state the relations.

To be a bit less abstract, as an example of multi-relational learning, suppose we have lists the names of our past presidents and vice-presidents (the entities) and partial knowledge of the their political party, who they were president of, and who they were vice-president for (three relations). Mindful that in our history, we have had more political parties than Democrat or Republican and the vice-presidents were not always of the same party as the president, the task is to build a model that predicts the parties of each president and vice-president.

Retuning to materials science, one or both entities defining the dimensions of matrices χ_k for relation k in Fig. 8 most naturally would be a material, but the two entities could also be a pair of features, two quantities which are not features, etc. Presently, our data tables lump different types of features together. Instead, we could group in separate tables those naturally associated, say those describing crystal structure, alloy concentration, functionality, etc. The relations might be functional, F_k(value of i-th entity) = value of j-th entity, or logical, (i-th entity, k-th predicate, j-th entity) with $\chi_{ijk} = 1$ if a relation exists or 0 if it is does not. As implied by our "presidential" example, multi-relational learning does not require values for all entity pairs in the tables, something that allows us to use more available data. We are now learning not only about relations within a table and but also about those among the tables. This yields models that "fill-in" the missing entries.

ACKNOWLEDGMENTS

Our machine learning understanding and work has greatly benefited from conversations and collaborations with P. V. Balachandran, G. Pilania, J. Hogden, J. Teiler, D. Wolpert, D. Xue and R. Yuan. We thank G. Pilania and J. Lashley for helpful comments about the manuscript.

---

[1] https://obamawhitehouse.archives.gov/mgi.

[2] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).

[3] Yue Liu, Tianlu Zhao, Wangwei Ju, and Siqi Shi, "Materials discovery and design using machine learning," J. Materiomics **3**, 159 – 177 (2017).

[4] Anubhav Jain, Geoffroy Hautier, Shyue Ping Ong, and Kristin Persson, "New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships," Journal of Materials Research **31**, 977 (2016).

[5] Martin A. Mosquera, Bo Fu, Kevin L. Kohlstedt, George C. Schatz, and Mark A. Ratner, "Wave functions, density functionals, and artificial intelligence for materials and energy research: Future prospects and challenges," ACS Energy Letters **3**, 155 (2018).

[6] Keisuke Takahashi and Yuzuru Tanakab, "Materials informatics: a journey towards material design and synthesis," Dalton Trans. **45**, 10497 (2016).

[7] Y. Lyu, Y. Lkiu, and B. Guo, J. Materiomics **3**, 221 (2017).

[8] W. Lu, R. Xiao, J. Yang, H. li, and W. Zhang, J. Materiomics **3**, 191 (2017).

[9] X. Zhang and Y. Xiang, J. Materiomics **3**, 200 (2017).

[10] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingoand S. Sanvito, and O. Levy, Nat. Matr. **12**, 191 (2013).

[11] J. M. Rondinelli, N. A. Benedek, D. E. Freredman, A. Kovner, E. E. Rodriguez, E. S. Toberer, and L. W. Martin, Am. Ceramics Soc. Bull. **92**, 14 (2013).

[12] T. Mueller, A. G. Kusne, and R. Ramprasad, Rev. Comp. Chem. **29**, 188 (2016).

[13] Turab Lookman, Francis J. Alexander, and Krishna Rajan, eds., *Information Science for Materials Discovery and Design*, Springer Series in Materials Science, Vol. 225 (Springer, Heidelberg,

2016).

[14] E. Mooser and W. B. Pearson, Acta Crystallogr. **12**, 1015 (1959).

[15] J. St. John and A. N. Block, Phys. Rev. Lett. **33**, 1095 (1974).

[16] J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B **17**, 2453 (1978).

[17] A. Zunger, Phys. Rev B **22**, 5839 (1980).

[18] D. Pettifor, J. Less Common Metals **114**, 7 (1985).

[19] G. Pilania, J. E. Gubernatis, and T. Lookman, Phys. Rev. B **91**, 214302 (2015).

[20] M. Ashby, *Material Selection in Mechanical Design* (Butterworth-Heinemann, Burlingham, 2008).

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2008).

[22] P. Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data* (Cambridge University Press, New York, 2012).

[23] Z. Ivezić, A. J. Connolly, J. T. VanderPlas, and A. Gray, *Statistics, Data Mining and Machine Learning in Astronomy* (Princeton University Press, Princeton, 2014).

[24] D. Wolpert, IEEE Trans. Evolutionary Comp. **1**, 67 (1997).

[25] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, I. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moreck, R. T. Downs, and A. Le Ball, Nucleic Acid Research **40**, D420 (2012).

[26] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, Acta Crystallogr. B **58**, 364 (2002).

[27] H. Glawe, A. Sanna, E. K. U. Gross, and M. A. L. Marques, New J. Phys. **18**, 093011 (2016).

[28] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, Inorg. Chem. **50**, 656 (2011).

[29] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, Skinner D, and G. Ceder, APL Matr. **1**, 011002 (2013).

[30] J. E. Saal, S. Kirklin, A. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).

[31] https://repository.nomad-coe.eu.

[32] W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson, and G. Ceder, Sci. Adv. **2**, e1600225 (2016).

[33] F. Legrain, J. Carrete, A. van Roekeghem, G. K. H. Madsen, and N. Mingo, J. Phys. Chem. **122**, 625 (2018).

[34] P. V. Balachandran, A. E. Emory, J. E. Gubernatis, T. Lookman, C. Wolverton, and A. Zunger, Phys. Rev. M **2**, 043802 (2018).

[35] M. W. Lufaso and P. M. Woodward, Acta Crystallogr. B **57**, 725 (2001).

[36] L. Breiman, Mach. Learn. **45**, 5 (2001).

[37] J. H. Friedman, Ann. Stat. **29**, 1189 (2001).

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Machine Learning Research **12**, 2825 (2011).

[39] G. Pilania, P. V. Balachandran, J. E. Gubernatis, and T. Lookman, Acta Crystallogr. B **71**, 507 (2015).

[40] R. D. Shannon, Crystallogr. Sect. A **32**, 751 (1976).

[41] I. D. Brown, Chemical Reviews **109**, 6858 (2009).

[42] P. Villars, K. Cenzuak, J. Daams, Y. Chen, and S. Iwata, J. Alloys and Compounds **367**, 167 (2004).

[43] Donald R. Jones, Matthias Schonlau, and William J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," J. of Global Optimization **13**, 455–492 (1998).

[44] A. I. J. Forrester, A. Sóbester, and A. J. Keane, *Engineering Design via Surrogate Modelling: A Practical Guide* (John Wiley & Sons, Ltd., 2008).

[45] J. Liepe, S. Filippi, M. Komorowski, and M.P.H. Stumpf, "Maximising the information content of experiments in systems biology," PLOS Computational Biology (2013).

[46] E. Brochu, V. M. Cora, and N. de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning," ArXiv e-prints (2010), arXiv:1012.2599.

[47] E.G. Ryan, C. C. Drovandi, J.M. McGree, A. N. Pettitt, and I. Verdinelli, "A review of modern computational algorithms for bayesian optimal design," International Statistical Review **84** (2016).

[48] D. Bhattacharjya, J. Eidsvik, and T. Mukerji, "The vaue of information in spatial decision making," Math Geosci **42**, 141 163 (2010).

[49] Prasanna V. Balachandran, Dezhen Xue, James Theiler, John Hogden, and Turab Lookman, "Adaptive Strategies for Materials Design using Uncertainties," Scientific Reports **6**, 19660 (2016).

[50] C. E. Rasmussen and K. J. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, 2006).

[51] D. V. Lindley, "On a Measure of the Information Provided by an Experiment," The Annals of Mathematical Statistics **27**, 986–1005 (1956).

[52] Merlise A. Clyde, in *International Encyclopedia of the Social and Behavioral Sciences*, edited by N. J. Smelser and P. B. Baltes (Elsevier, New York, 2001) p. 5075.

[53] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," Statistical Science **10** (1995).

[54] N. Srinivas, A. Krause, S.M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: no regret and experimental design,".

[55] B.J. Yoon, X. Qian, and E.R. Dougherty, "Quantifying the Objective Cost of Uncertainty in Complex Dynamical Systems," IEEE Trans. Signal Process **61**, 2256–2266 (2013).

[56] R. Dehghannasiri, B.J. Yoon, and E.R. Dougherty, "Optimal experimental design for gene regulatory networks in the presence of uncertainty," IEEE/ACM Trans. Comput. Biol. Bioinformatics **12**, 938–950 (2015).

[57] R. Dehghannasiri, D. Xue, X. Qian, L. Dalton, T. Lookman, and E.R. Dougherty, "Optimal experimental design for materials discovery," Computational Materials Science **129**, 311–322 (2017).

[58] H.J. Kushner, "A new method of locating the maximum of an arbitrary multi-peak curve in the presence of noise," Journal of Basic Engineering **86**, 97–106 (1964).

[59] J. Mockus, *Bayesian Approach to Global Optimization: Theory and Applications* (Kluwer Academic, Dordrecht, 1989).

[60] D. Huang, T.T. Allen, W.I. Notz, and N. Zeng, "Global optimization and stochastic black-box systems via sequential kriging meta-models," J. Global Optim. **34**, 441–466 (2006).

[61] P. I. Frazier, W. B. Powell, and S. Dayanik, "A knowledgegradient policy for sequential information collection," SIAM J. Control Optim. , 2410–2439 (2008).

[62] J. Theiler and B.G. Zimmer, "Selecting the selector: Comparison of update rules for discrete global optimization," Stat Anal Data Min: The ASA Data Sci. Journal **10**, 211–229 (2017).

[63] Bertrand Rouet-Leduc, Kipton Barros, Turab Lookman, and Colin J. Humphreys, "Optimisation of GaN LEDs and the reduction of efficiency droop using active machine learning," Scientific Reports **6**, 24862 (2016).

[64] Dezhen Xue, Prasanna V. Balachandran, John Hogden, James Theiler, Deqing Xue, and Turab Lookman, "Accelerated search for materials with targeted properties by adaptive design," Nat.

Commun. **7**, 11241 (2016).

[65] Ruihao Yuan, Zhen Liu, Prasanna V. Balachandran, Deqing Xue, Yumei Zhou, Xiangdong Ding, Jun Sun, Dezhen Xue, and Turab Lookman, "Accelerated Discovery of Large Electrostrains in $BaTiO_3$-Based Piezoelectrics Using Active Learning," Advanced Materials **30**, 1702884 (2018).

[66] Atsuto Seko, Tomoya Maekawa, Koji Tsuda, and Isao Tanaka, "Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids," Phys. Rev. B **89**, 054303 (2014).

[67] Atsuto Seko, Atsushi Togo, Hiroyuki Hayashi, Koji Tsuda, Laurent Chaput, and Isao Tanaka, "Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization," Phys. Rev. Lett. **115**, 205901 (2015).

[68] Shin Kiyohara, Hiromi Oda, Koji Tsuda, and Teruyasu Mizoguchi, "Acceleration of stable interface structure searching using a kriging approach," Japanese Journal of Applied Physics **55**, 045502 (2016).

[69] Masaki Yamawaki, Masato Ohnishi, Shenghong Ju, and Junichiro Shiomi, "Multifunctional structural design of graphene thermoelectrics by bayesian optimization," Science Advances **4**, earr4192 (2018).

[70] Henry C. Herbol, Weici Hu, Peter Frazier, Paulette Clancy, and Matthias Poloczek, "Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization," npj Computational Materials **4**, 51 (2018).

[71] S. Chen, K. Reyes, M. Gupta, M. McAlpine, and W. Powell, "Optimal learning in experimental design using the knowledge gradient policy with application to characterizing nanoemulsion stability." SIAM/ASA Journal on Uncertainty Quantification **3**, 320 (2015).

[72] R. Aggarwal, M. J. Demkowicz, and Y. M. Marzouk, in *Information Science for Materials Discovery and Design*, edited by Turab Lookman, Francis J. Alexander, and Krishna Rajan (Springer International Publishing, 2016) p. 13.

[73] Yixing Wang, Yichi Zhang, He Zhao, Xiaolin Li, Yanhui Huang, Linda S. Schadler, Wei Chen, and L. Catherine Brinson, "Identifying interphase properties in polymer nanocomposites using adaptive optimization," Composites Science and Technology **162**, 146 – 155 (2018).

[74] Julia Ling, Maxwell Hutchinson, Erin Antono, Sean Paradiso, and Bryce Meredig, "High-Dimensional Materials and Process Optimization Using Data-Driven Experimental Design with

Well-Calibrated Uncertainty Estimates," Integrating Materials and Manufacturing Innovation **6**, 207–217 (2017).

[75] Andy J. Keane, "Statistical Improvement Criteria for Use in Multiobjective Design Optimization," AIAA Journal **44**, 879–891 (2006).

[76] Joshua Svenson and Thomas Santner, "Multiobjective optimization of expensive-to-evaluate deterministic computer simulator models," Computational Statistics & Data Analysis **94**, 250–264 (2016).

[77] Abhijith M. Gopakumar, Prasanna V. Balachandran, Dezhen Xue, James E. Gubernatis, and Turab Lookman, "Multi-objective Optimization for Materials Discovery via Adaptive Design," Scientific Reports **8**, 3738 (2018).

[78] Dezhen Xue, Deqing Xue, Ruihao Yuan, Yumei Zhou, Prasanna V. Balachandran, Xiangdong Ding, Jun Sun, and Turab Lookman, "An informatics approach to transformation temperatures of NiTi-based shape memory alloys," Acta Materialia **125**, 532 –541 (2017).

[79] A. Mannodi-Kanakkithodi, G. Pilania, R. Ramprasad, T. Lookman, and J.E. Gubernatis, "Multi-objective optimization techniques to design the pareto front of organic dielectric polymers," Comp. Mat. Sci. **125**, 92 (2016).

[80] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, "Machine learning strategy for accelerated design of polymer dielectrics," Sci. Reports **6**, 20952 (2016).

[81] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler, "Hypervolume-based multiobjective optimization: theoretical foundations and practical implications," Theoret. Comput. Sci. **425**, 75–103 (2012).

[82] L. Lu and C.M. Anderson-Cook, "Adapting the hypervolume quality indicator to quantify trade-offs and search efficiency for multiple criteria decision making using pareto fronts," Qual. Reliab. Eng. Int. **29**, 1117–1133 (2012).

[83] Yongtao Cao, Byran J. Smucker, and Timothy J. Robinson, "On using the hypervolume indicator to compare pareto fronts: Applications to multi-criteria optimal experimental design," Journal of Statistical Planning and Inference **160**, 60 – 74 (2015).

[84] L. Le Gartiet, SIAM/ASA J. Uncertain. Quantif. **1**, 244 (2013).

[85] L. Le Gartiet and J. Garnier, Int. J. Uncertain. Quantif. **4**, 365 (2014).

[86] M. C. Kennedy and A. O'Hagan, Biometrika **87**, 1 (2000).

[87] G. Pilania, J. E. Gubernatis, and T. Lookman, Comp. Mat. Sci. **129**, 156 (2017).

[88] Hideomi Koinuma and Ichiro Takeuchi, "Combinatorial solid-state chemistry of inorganic materials," Nature Materials **3**, 429 (2004).

[89] Aaron Gilad Kusne, Tieren Gao, Apurva Mehta, Liqin Ke, Manh Cuong Nguyen, Kai-Ming Ho, Vladimir Antropov, Cai-Zhuang Wang, Matthew J. Kramer, Christian Long, and Ichiro Takeuchi, "On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets," Scientific Reports **4**, 6367 (2014).

[90] S.F. Li and R.M. Suter, "Adaptive reconstruction method for three-dimensional orientation imaging," J. Appl. Crystallogr. **46**, 512 (2013).

[91] Nicholas Lubbers, Turab Lookman, and Kipton Barros, "Inferring low-dimensional microstructure representations using convolutional neural networks," Phys. Rev. E **96**, 052111 (2017).

[92] Henry W. Lin, Max Tegmark, and David Rolnick, "Why does deep and cheap learning work so well?" Journal of Statistical Physics **168**, 1223–1247 (2017).

[93] D. S. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, Oxford, 2006).

[94] L. Getoor and B. Taskar, *Introduction to Statistcal Relational Learning* (MIT Press, Cambridge, 2007).

[95] L. De Raedt, *Logical and Relational Learning* (Springer, New York, 2008).

[96] T. G. Kolda and B. W. Bader, SIAM Review **51**, 455 (2009).

[97] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three way model for collective learning on multi-relational data," in *Proceedings of the 28th International Conference on Machine Learning*, edited by L. Getoor and T. Scheffer (ACM, Bellvue WA, 2011) p. 809.