Current Opinion in Solid State and Materials Science 21 (2017) 121–128

Contents lists available at ScienceDirect

Current Opinion in Solid State and Materials Science

journal homepage: www.elsevier.com/locate/cossms

Statistical inference and adaptive design for materials discovery

Turab Lookman [a,*], Prasanna V. Balachandran [a], Dezhen Xue [a,d], John Hogden [b], James Theiler [c]

[a] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[b] Computer and Computational Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[c] Intelligence and Space Research, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[d] State Key Laboratory for Mechanical Behavior of Materials, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 12 January 2016
Revised 3 August 2016
Accepted 2 October 2016
Available online 10 October 2016

Keywords:
Experimental design
Adaptive learning
Statistical inference
Materials design

ABSTRACT

A key aspect of the developing field of materials informatics is optimally guiding experiments or calculations towards parts of the relatively vast feature space where a material with desired property may be discovered. We discuss our approach to adaptive experimental design and the methods developed in decision theory and global optimization which can be used in materials science. We show that the use of uncertainties to trade-off exploration versus exploitation to guide new experiments or calculations generally leads to enhanced performance, highlighting the need to evaluate and incorporate errors in predictive materials design. We illustrate our ideas on a computed data set of M_2AX phases generated using *ab initio* calculations to find the sample with the optimal elastic properties, and discuss how our approach leads to the discovery of new NiTi-based alloys with the smallest thermal dissipation.

1. Overview and need for design

There has been much interest recently in using information science tools for materials discovery and design, with various national initiatives (Office of Science and Technology Policy at the White House, Basic Energy Sciences and National Science Foundation [1,2]) helping to define field of "materials informatics". The focus of this article is to show how experiments or calculations can be guided optimally to enable the discovery of new materials with targeted properties in as few iterations as possible. The central premise in *experimental design* is that experiments and/or calculations are expensive and time-consuming and therefore desired is an efficient and rational approach to discovery so that the laborious trial-and-error efforts may be avoided. The field of experimental design using statistical methods has a rich and long history [3,4] and it has been applied in many areas including aspects of materials processing in chemical engineering [5–7] and the design of computer experiments [8]. Our focus will be on the problem of materials discovery and the use of methods based on the value of information and global optimization techniques [9–11], which have been successfully developed and applied in the aerospace and automobile industries [12].

A key element of the discovery approach we will use is recognizing how the role of *uncertainties* due to statistical inference or measurements should be used to explore the vast search space for materials with better properties than those that exist in the available training data set [13]. This is a departure from most of the activity in materials informatics field, which involves generating and screening relatively large amounts of computational data on specific materials and identifying correlations in the inputs (descriptors or features) [14,15]. A number of recent studies have also used regression methods to identify materials for further examination [16–19]. Having to deal with relatively small amounts of data is typical of many materials design problems that involve learning from experimental data. By applying methods developed in fields such as decision theory and global optimization, we show how an adaptive design loop can iteratively guide the next experiments or computations for materials with targeted properties, especially if the experiments and/or calculations are expensive to perform [11,20]. Such methods have been successfully applied in the automotive and aerospace industries where complex, expensive codes are in use and it is too time-consuming to use these to exhaustively search the high-dimensional feature space in a brute-force manner [21]. Instead, surrogate or inference models are used for the design. After a broad overview of the approaches so far utilized in the nascent and emerging field of materials informatics, we will illustrate our ideas with examples on materials problems using both computational and experimental data.

The materials databases in efforts such as materialsprojects.org [22], AFLOWLIB [23] and OQMD [24] contain hundreds of thousands of compounds taking up 10–few 100's of gigabytes (GB) of data. To

* Corresponding author.
E-mail address: txl@lanl.gov (T. Lookman).

http://dx.doi.org/10.1016/j.cossms.2016.10.002
1359-0286/

put this in context, Google and Facebook process 100s of petabytes (PB) of data in a year. Thus, the materials discovery problem from these materials databases is comparatively not a big data problem. Moreover, the notion of a materials or inorganic gene is itself not a new concept, it even predates the decoding of the human gene by about two decades. It was the English crystallographer Alan Mckay, then at Birbeck College, who suggested that "the crystal is a structure, the description of which is much smaller than the structure itself" and it serves as a "carrier of information". He proposed the construction of an inorganic gene as a biological approach to inorganic systems so one has a genomic paradigm, *i.e.* how fundamental pieces of information taken as bits of data collectively, describe a crystal [25]. The problem of materials informatics, in the way we think about it today, is also not new. Chelikowsky and Phillips [26] studied the classic problem of classifying AB *sp*-bonded octet solids with sixfold coordinated rocksalt or fourfold coordinated zincblende/wurtzite in the 1970s. They recognized that the energy differences between structures calculated using nonlocal pseudopotentials were often too small (0.1% of the cohesive energy) to be calculated in those days, and suggested an information theory point of view to learn rules on bonding from the data containing roughly 80 compounds. Following work of Mooser and Pearson [27] and St. John and Bloch [28], they went on to construct structural maps to classify the AB compounds. These were defined by the minimal number of features, in terms of symmetry-adapted combinations of *s* and *p*-orbitals of atoms A and B in the compound calculated using nonlocal pseudopotentials. Recently, several groups have revisited this problem from a statistical learning perspective using classifiers such as decision trees and support vector machines to estimate the average classification accuracy and the associated model variance where a decision boundary is learned in a supervised manner [29–31,19]. Today the use of elaborate machine learning tools allows us to classify and draw the decision boundaries with far greater accuracy than the use of pencil and paper approaches of yesteryears. The approach has suggested new features, such as the difference in the effective Born charges in the rocksalt and zincblende/wurtzite structures [32], as well as new combinations of orbital radii which allow us to classify with greater accuracy than original features [19].

Much of the recent interest in the field has been catalyzed by the Materials Genome Initiative (MGI) with the overarching goal to cut in half the time and costs of bringing new materials to market. Thus, the aims of MGI span materials discovery and property optimization all the way to deployment via systems engineering. How exactly this is to be done and what is the appropriate framework or paradigm is the key question. Why is there a need to accelerate the process? If we look at the time it has taken for various materials to be deployed, it is roughly of the order of 25–30 years. The discovery and optimization is thus a key challenge; we need to know the appropriate materials, with targeted properties, to be deployed. For example, the III-V GaAs semiconductors had enormous impact between 1965 and 1985, especially with Si paving the way for Very Large-Scale Integration (VLSI) technology due to the transformational impact of the Czochralski process for fabricating single crystals. Similarly, after 200 years of lighting technology the efficiencies gained barely approach 30–40%, but the discovery of wide band gap materials has paved the way for their applications as energy-saving light emitting diodes (LEDs) and in high power and high temperature electronics. The theme of driving innovation through new chemistries and structural motifs could not be more true than in the case of photovoltaics to harness energy. We have seen a tremendous rise over the last 3–4 years in photovoltaic efficiency (to ~22%) with the use of hybrid (organic molecule at A-site) perovskites [33]. The perovskite is a very different structure and shows the importance of how the chemistry and structure influence the property and raises the question of whether there are other structural arrangements to try other than perovskites. Thus, the challenge we have is to combine chemical and structural complexity which gives rise to rich, emerging behavior. The chemical space of even simple perovskites can be quite extensive, in the case of the perovskite structural motif there are over 3000 possible chemistries and numerous combinations of the basic structural motif. Only about 20% of the chemistries/structure are experimentally investigated and reported in the Inorganic Crystal Structure Database [34].

So how do we accelerate the discovery process in a rational manner? Materials design is an optimization problem with the goal of maximizing (or minimizing) some desired property of a material, denoted by y, by varying some features that characterize the material chemistry, structure, composition, processing conditions and/or microstructure, denoted by x. Optimizing a material generally proceeds by making predictions about y and then selecting or computationally/intuitively designing an x at which y is measured and the result (x, y) is added to the database of known properties. The primary hurdle in material design is measuring y because it requires synthesis and characterization of new materials, which can be expensive and time-consuming. For this reason it is necessary to have an optimization approach to minimize the number of new materials that need to be experimentally tested. A key aspect is feature selection, identifying features that characterize the material composition and which help optimize the desired material property in terms of which one wants to optimize a given property. This can an be done using domain knowledge where the meaning and importance of the selected features is clear, or by the use of high-throughput approaches in which a certain number of features are initially chosen and various binary or ternary combinations of new features from this initial set are screened for their relative importance. Thus, finding targeted properties is an optimization, control and learning problem. It is important to have forward models that are physics-driven (e.g. Ginzburg-Landau or phase-field theory, finite element), but these are often complex and difficult to use for design. Thus, surrogate or inference models are essential for optimizing a targeted property. In addition, we want to glean certain aspects of the physics by learning the inter-relationships among the features.

2. Materials by adaptive design

The state-of-the-art in materials informatics consists of (a) assembling a library of crystal structures, chemistries relevant to the problem and (b) defining the training space with a given number of samples and features. The features can be bond angles, bond lengths, energetics from first principles calculations, as well as experimental data, such as thermodynamics from experiments. This is used to build an inference model using off-the-shelf pattern recognition tools, such as classifiers and regressors based on linear or kernel ridge regression, least squares regression, decision trees, Gaussian process modeling or support vectors. There are very few examples one can cite of new materials synthesized, characterized after prediction through this approach. Part of the difficulty is that the data for real materials is very limited in size (~10–100 samples), the materials are multicomponent with defects and there are uncertainties that can arise from sampling or measurement errors. In addition, the search space of materials that are missing or yet to be synthesized is often very large (~1 million). Thus, high-throughput approaches using computational data can be limited in how well they can do. For example, the search for Pb-free piezoelectrics often involves much more than screening a large number of chemistries in the perovskite ABO_3 structure based on size of energy band gap and energy differences between distorted

phases across the morphotropic phase boundary (MPB) [35]. The phase diagram of Pb-free piezoelectrics, such as the Barium Titanate based $(1-x)Ba(Ti_{1-m}Ca_m)\,O_3$-$x(Ba_{1-n}Zr_n)\,TiO_3$, is very sensitive to changes in dopant concentrations Ca (m) and Zr (n). The curvature of the MPB, and hence its temperature sensitivity and response, depend sensitively on (m, n, x). For this design problem, the data set consists of 10–15 phase diagrams with a search space of $\sim 10^3$ and the challenge is to solve the inverse problem for (m, n, x) for a targeted response, such as a large piezoelectric coefficient at the MPB.

The problem of limited sample size and a very large feature space (in the thousands) is well known in bioinformatics where classifiers have been trained to predict different types of breast cancers using gene signatures as features. In that field it has been increasingly realized that a probabilistic Bayesian framework, incorporating prior knowledge, has much predictive merit for developing classifiers and regressors and their associated errors. In genomics this has subsequently led to the design of optimal experiments for improving drug intervention in genetic regulatory networks [36,37].

In contrast to cancer genomics and bioinformatics, in materials informatics we have only recently begun to learn how different regressors and classifiers perform on a variety of materials data sets [31]. This is essential in order to have guiding principles for materials design. The importance of uncertainties in predictions is also only beginning to be appreciated [38]. An aspect that has been little addressed in materials discovery is how do we guide the next experiments (or calculations), especially if they are expensive in terms of costs and/or time? [13]. This goes beyond mere regression or classification studies, which are important in training inference models from data but do not address the problem of designing the next experiments that need to be performed towards finding a desired material with a given label or property. In the following we review the approach we have taken and discuss our results using computational and experimental data.

Our approach towards accelerated discovery is illustrated in Fig. 1a. Prior or domain knowledge (1 in Fig. 1a) enters into constraining an inference model (2) that is trained on available data and predicts a label or a property with associated uncertainties. Directly using the best score from these predictions for design is not necessarily optimal in guiding the next experiments or calculations. An experimental design or decision making step or module (3) is key in balancing the trade off between regions of the search space where data exists and other regions with limited data, yet significant potential for finding materials with enhanced properties. Thus, uncertainties from (2) are used in (3) to balance the trade-off between exploiting the model predictions and exploring the search space to suggest the next material for experimentation or calculation (4). Existing training data is augmented by the new results (5) and the adaptive process repeats itself. Although it is desirable to find a material with the target property in as few iterations of Fig. 1a as possible, some of the issues that need to be addressed include the acquisition function for the resulting material, stopping criteria, the quality of the inference model and the size of the training data. Our motivation here is really codesign – how do we seamlessly couple experiments, available knowledge in the form of theory and physics models as well as informatics – in much the same way they have been used in biology and medical research over the last 20–30 years that has given rise to the field of bioinformatics – to iteratively "learn" a "discovery model" by guiding and using the results from experiments and calculations. Our approach is unique in that most of the work done in materials discovery essentially involves one or two of the green steps in Fig. 1, let alone incorporating uncertainties and doing adaptive design.

The methods we discuss have been extensively developed and studied in areas related to formulating measurement policies for sequential online measurements or "multi-bandit" problems, and offline problems under ranking and selection where several alternatives can exist and where choices or decisions need to be made [20]. Similarly, the interest in global optimization, with its relevance to solving complex problems of interest to industry, has catalyzed much work on the development of acquisition functions within the Bayesian approach to design [39,21]. Of particular note is the work of Kushner [9] on probability of improvement and Mockus [10] and Jones [11] on expected improvement. More recent acquisition functions used in various fields include Gaussian Process-Upper Confidence bound [40], Thompson sampling [41,42], which maximizes the expected reward relative to a randomly drawn belief, and the Mean Objective Cost of Uncertainty [43]. The latter chooses sampling or infill points directly based

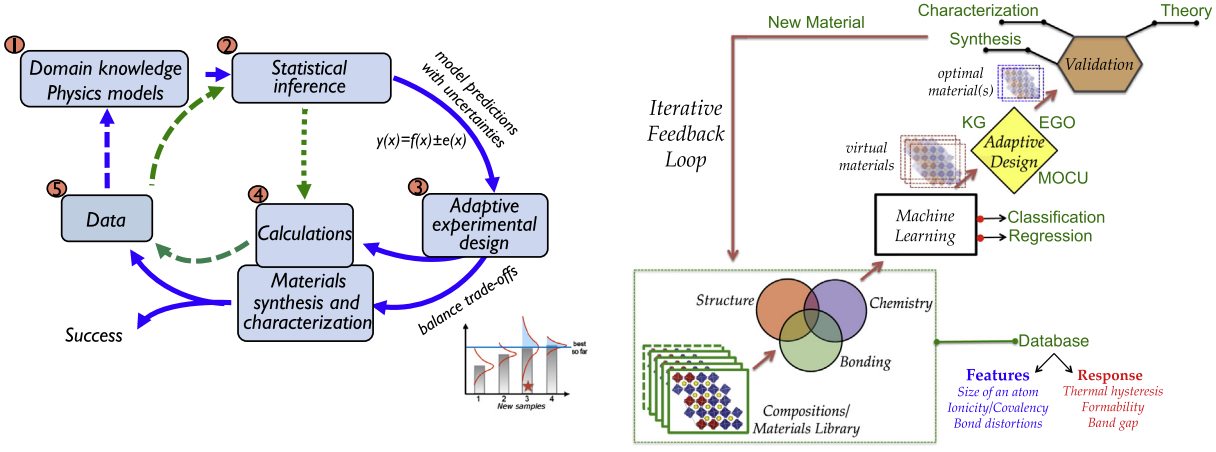

Fig. 1. (a) The adaptive design loop where the key challenge is to minimize the number of iterations it takes to discover a new material with desired properties by finding a reliable surrogate to the true unknown function $f(x)$. Existing work on materials design is largely based on following one or two of the green arrows. Our approach introduces the adaptive experimental design step that uses uncertainties, $e(x)$, to balance the trade-off between exploration and exploitation in suggesting the next experiment or calculation to be performed. (b) The loop in practice where EGO (Efficient global optimization), KG (Knowledge gradient) and MOCU (Mean Objective Cost of Uncertainty) are acquisition functions or *selectors* that choose the next optimal experiment or calculation.

on minimizing the mean objective uncertainty cost with each iteration. Fig. 1b spells out how our loop of Fig. 1a looks in practice with the elements above included.

The interest from industry is in part due to the fact that optimization using complex codes that solve ordinary differential equations puts the same amount of effort in high and poor performance regions and thus the costs rise exponentially. As a result, analysis from such high-fidelity codes in early design stages where uncertainties are higher, is not an efficient use of resources. Engineers thus opt for adaptive sampling (or judicious selection of new or infill points at which to call the true function) over traditional design of experiments methods. The adaptive sampling involves the global accuracy of the model to assure global search (exploration), and the use of the model to identify the optimum (exploitation). It allows bounding a region of high performance that reduces the design time in later stages of the process. In order to reduce the concept evaluation time, which involves a multi-criteria decision-making process with large amounts of data and expert knowledge which is often imprecise and subjective, engineers employ cheap surrogate models (hence surrogate-based optimization [39,21]) using training data, such as Gaussian Process Models, for each concept where the tool to model the concept is computationally intense. The surrogate models are also often referred to as meta models or response surfaces and judiciously choose the infill points to sample the objective function in promising areas based on a constantly changing surrogate. The success or failure of surrogate-based optimization rests on the correct choice of model and infill criteria.

3. Accelerated search: finding materials with optimal elastic properties using computed data

We apply the ideas reviewed above to a test problem. Before we discuss in Section 4 the relevance to experiments, our objective is to demonstrate that the framework and methods proposed above are robust enough to find the sample with the optimal elastic properties in a data set of M_2AX phases generated using *ab initio* calculations [38]. M_2AX phases belong to a class of ternary ceramics that show unique mechanical, elastic, thermal and electrical properties. In the M_2AX phases, the X atoms reside in the edge-connected M octahedral cages and the A atoms reside in slightly larger right prisms [44]. Fig. 2a, illustrates the layered hexagonal crystal structure and the chemical search space of M_2AX phases.

We consider a data set of 223 compounds belonging to the family of M_2AX phases estimated using density functional theory (DFT) calculations and taken from the literature. Details may be found in Cover et al. [45] who compared calculated elastic constants with experimental measurements and report variations up to 30 GPa. Although a total of 240 chemical compositions can be exhaustively enumerated, 17 have negative elastic constants and so were removed from the data set due to stability issues. We consider the problem of maximizing the elastic modulus bulk (B) using our adaptive design strategies. As features, x, we employed orbital radii of M, A, and X-atoms from the Waber-Cromer scale [46]. These included the s-, p- and d-orbital radii for M and s- and p-orbital radii for the A and X atoms. These serve as a good starting point as they capture the fundamental relationship between the electronic charge density and elastic response of these materials.

Among the regressor or surrogate models (2 in Fig. 1a) that we used are the Gaussian Process Model (GPM), which as outlined in Section 2 serves as the basis for Bayesian Optimization. It assumes that $y = f(x) + \mathcal{N}(0, \sigma^2)$ where f is the function that must be estimated. Unlike most regression, f is treated as a point in an infinite-dimensional Gaussian distribution (that is, a *Gaussian Process*) over the set of functions. The Gaussian Process is completely defined by its mean and covariance, which are functions of x. We consider the mean function, $f(x) = 0$ and the covariance to be the square exponential function: $\mathrm{cov}(f(x), f(x')) = \exp(-\theta\|x - x'\|^2) + \delta(x, x')\sigma_n$, where θ and σ_n are free parameters. This function decays smoothly with separation between data points and thus encodes this prior knowledge. The parameters θ and σ_n (sometimes called the nugget) are obtained using maximum likelihood and cross-validation, respectively. It can easily be shown that the posterior distribution of $f(x)$ given the training samples is Gaussian with a mean and standard deviation that can be calculated. This mean and standard deviation serve as the predicted y value and estimated error around the prediction.

We will also show results using Support Vector Regressors (SVR), which share the characteristic with kernel ridge regression of employing kernels to learn the nonlinear function $f(x)$. They do so by learning a linear function in the space induced by the respective kernels, corresponding to a nonlinear function in the original space. In particular, $f(x) = \sum_i a_i \kappa(x, x_i)$, where the a_i's are coefficients that are fit to the data, and $\kappa(x, x_i)$ is the kernel function. For $\mathrm{SVR_{lin}}$, the kernel function is a dot product of x and x_i, leading to an $f(x)$ linear in x. For $\mathrm{SVR_{rbf}}$, the kernel is a radial basis function $\kappa(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$, where σ is determined by cross-validation. To obtain uncertainties with the SVR models we use a bootstrap approach. Using randomly chosen samples (with replacement) of the training data, we re-trained the regressors multiple times, and maintained an ensemble of these regressors. For a new material, we applied each member of the ensemble to obtain the mean and standard deviation of various predictions in the ensemble, which then serve as input to the design.

In our choice of deciding what new sample to test, the range of acquisition functions, heuristics or *selectors* we will employ to balance exploration and exploitation are based on the idea of *expected improvement* or efficient global optimization (EGO) as discussed by Jones [11]. This assumes for each value of x, a probability density function, $P(y|x)$, of possible y values given that a compound with features x is measured. EGO suggests measuring the x that maximizes the expected improvement in y over our current best value. If μ^* is the value of the best material measured so far, then the expected improvement obtained by testing x' is given by $E(I) = E[\max(y, \mu^*) - \mu^*] = E[\max(y - \mu^*, 0)] = \int_{\mu^*}^{\infty} (y - \mu^*) P(y|x') dy$. If $P(y|x')$ is assumed to obey a Gaussian distribution with mean μ and variance σ^2, the expected improvement can be written as $E(I) = \sigma[\phi(z) + z\Phi(z)]$, where $z = (\mu - \mu^*)/\sigma$ and $\phi(z)$ and $\Phi(z)$ are the standard normal density and cumulative distribution functions, respectively [11]. It is instructive to consider the limiting behavior of $E(I)$. If all the samples have the same, small σ, then $E(I) \to \mu - \mu^*$, thus maximizing $E(I)$ is obtained by choosing the largest μ (exploitation). Similarly, for large σ, $E(I) \to \sigma$, so that the samples of choice are those with the largest σ (exploration). At intermediate values of σ there is a tradeoff between the two. The other functions related to EGO which we also consider are:

- Knowledge Gradient, KG [47]: Choose μ^* to be the best of the predicted compounds.
- Max: This just chooses the highest expected score from the regressor.
- Max-A: Alternates between choosing the material with the highest expected score and the material with the most uncertain estimated score.
- Max-P: Maximizes the probability that a material will be an improvement, without regard to the size of the improvement.
- Random: Randomly choose an unmeasured compound.

In applying our loop of Fig. 1a, we would normally form an inference model in step 2 on the given data and then perform

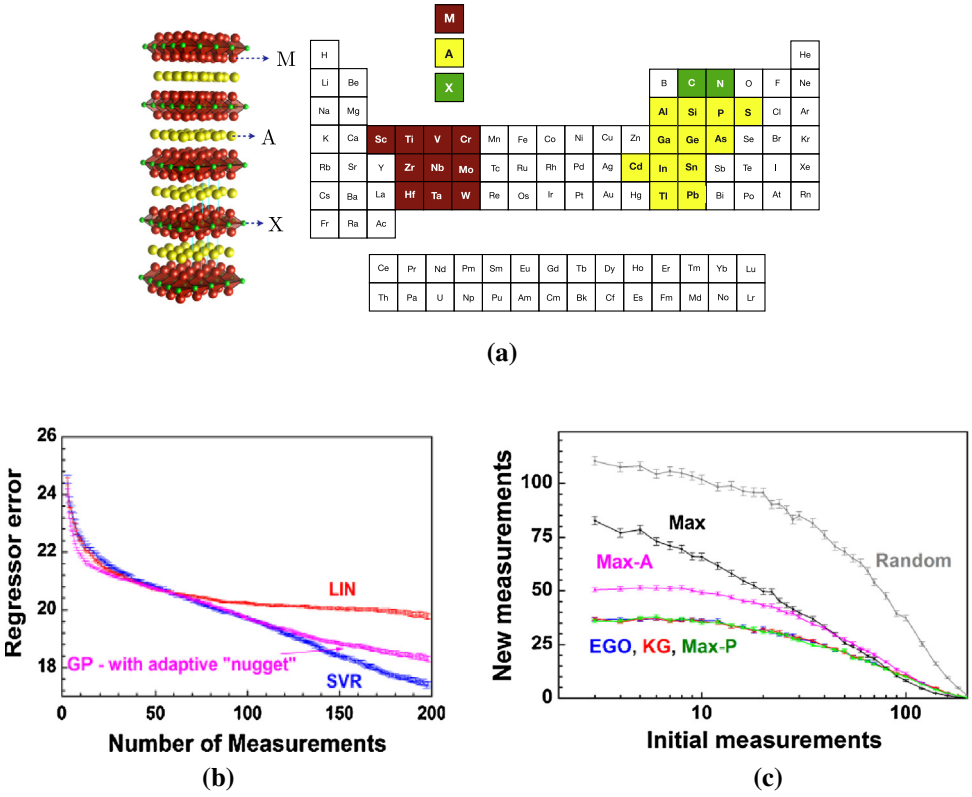

Fig. 2. (a) The M_2AX phases served as our test problem with a computed data set of 223 compounds. The crystal structure and chemical search space of M_2AX phases. (b) The relative performance of three regressors as a function of the number of measurements or size of the training data set, showing that the regressor SVR_{rbf} performs much better than SVR_{lin} and better than GPM if the size of the training set exceeds 100. (c) For SVR_{rbf} the number of new measurements (or iterations of the loop of Fig. 1) is plotted as a function of the number of initial measurements in the training set using different acquisition functions or *selectors*. With fewer initial measurements, EGO, KG and Max-P discover the optimal composition with fewer new measurements than selectors Max and Max-A. Choosing measurements at random gives the worst performance.

the design in step 3 to pick the material to measure *y*. Here we will exercise the loop by using the data set with the properties determined, without the need to perform calculations (step 4). Fig. 2b shows the relative performance of our three regressors as a function of the number of measurements, *M*, or size of the training data set. We performed 1000 trials, each initialized with a different randomly chosen set of $M = 20$ compositions. For each trial an ensemble of regressors were trained from random-with-replacement resampling of the data so that the regressor predicts a mean and variance for the property, which in this case is the shear modulus, *G*. The regressor error is the cross validation error and Fig. 2b shows that SVR_{rbf} performs much better than SVR_{lin} and better than GPM if the size of the training set exceeds 100. We can similarly compare the acquisition functions or selectors discussed above by determining the number of new measurements that hypothetically need to be performed to find the sample with the largest *G*. This is shown in Fig. 2c for SVR_{rbf} where the number of new measurements (or iterations of the loop of Fig. 1) is plotted as a function of the number of initial measurements in the training set. We see that when the initial number of random measurements is small, SVR_{rbf} with functions EGO, KG and Max-P all discover the optimal composition with fewer new measurements (about 35) than the other selectors (Max, Max-A). Merely choosing measurements at Random is clearly unsatisfactory. Also, Max, which chooses the highest score and is used most often in the literature, does not perform well. Some insight into the effectiveness of our directed search is shown in Fig. 3. The performance of our loop can also be measured in terms of an *opportunity cost* defined as the modulus difference between the current-best and the overall-best. This is shown as a function of the total number of measurements in Fig. 3a. Here we see that Max initially performs better and EGO, KG and Max-P are only preferred after a certain number of measurements. This apparently different behavior between Max and EGO, KG or Max-P selectors may be understood if we conjecture that the SVRrbf regressor is a continuous, nonlinear function in the feature space with several local maxima (Fig. 3b). Initially there are fewer data points to train the regressor and so the algorithm is constrained to a local maximum. Even after a relatively large number of iterations (50), the opportunity cost for Max does not reach zero, suggesting that the algorithm repeatedly selects candidate compositions in the vicinity of the local maximum. The Max selector does not take into account uncertainties in the predicted values and therefore is less efficient at exploring the search space.

4. Experiments with NiTi-based alloys

Learning how the design loop performs on computed data is a step towards assembling measurement data and guiding experiments. Our approach is to consider problems amenable to relatively straightforward experiments and where the physics is reasonably well understood, but the challenge is to optimize the properties. Ferroic or adaptive materials such as shape memory alloys and ferroelectrics have been our preferred choice of problems. Recently, we have considered the problem of finding NiTi-based shape memory alloys with the smallest thermal dissipation [48]. These alloys undergo a structural transition from the high symmetry cubic B2 to Rhombohedral (R) or monoclinic (B19′) phase and therefore care must be exercised in distinguishing the transformations, both of which have niche applications in industry. We considered the following alloy chemistry, $Ti_{50}Ni_{x-y-z}Cu_xPd_yFe_z$,

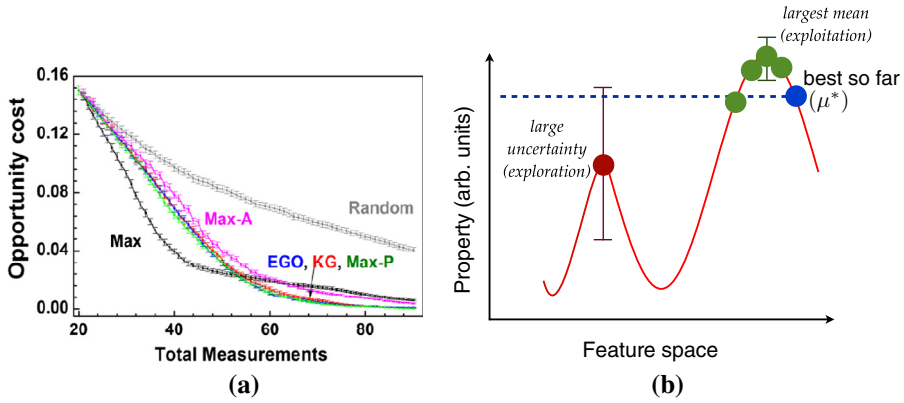

Fig. 3. (a) A measure of the performance of the loop. The *opportunity cost* (difference between the current-best and the overall-best) as a function of the total number of measurements. Max initially performs better but EGO, KG and Max-P are preferred after a sufficient number of measurements. (b) Schematic of feature space with local maxima showing that training the regressor with relatively few data points can lead to a local maximum so that initially Max, which does not allow for exploration, can perform better.

which leads to a search space of almost 800,000 compounds with different (x, y, z) values within measurement error. Our data set included approximately 22 alloys, for which the thermal hysteresis (ΔT) was known. Heating and cooling across the martensitic transformation temperature results in hysteresis (ΔT) as the transformation temperatures do not coincide, giving rise to fatigue. Our objective was to design a new alloy that has very low ΔT. Additionally, our experimental setup permitted making four alloys at a time. Therefore, we predicted four compounds, using the prediction from each iteration to drive the next result.

The loop as used in practice is shown in Fig. 4 for this alloy design and the six features, including valence electron number and atomic radii which are known from materials knowledge to influence the thermal hysteresis, are also listed. After iterating over the loop 9 times, we discovered 14 new alloys with lower ΔT than present in our training set after synthesis and characterization of 36 alloys. The alloy with the smallest thermal dissipation was found in the iteration 6 and had composition $Ti_{50}Ni_{46.7}Cu_{0.8}Pd_{0.2}Fe_{2.3}$ with a thermal hysteresis of 1.84 K. This was a 42% improvement in the hysteresis over the alloy with the smallest hysteresis (ΔT = 3.15 K) in the training set. Fig. 5 shows the heat flow under cooling and heating over 60 cycles from differential calorimetry measurements. The shift from cycle 1 to 60 is 0.02 K for this B2 to R transformation compared to 25 K for NiTi, which undergoes a B2 to B19′ transformation. We are not aware of any methodology which has demonstrated a comparable level of performance; our discovery is statistically significant with a *p*-value <0.001 (the chances that our samples were discovered by chance is less than 1 in 1000). As far as we are aware, this is the first study to show that a guided design loop can achieve in a direct and systematic way what is normally accomplished by intuition and trial-and-error. Although we only designed for the smallest ΔT, a number of the new compounds had martensite start temperatures not too far from room temperature. This was fortuitous but emphasizes the need to perform multiobjective optimization on hysteresis and temperature so that both objectives can be simultaneously optimized and the appropriate sample chosen from the

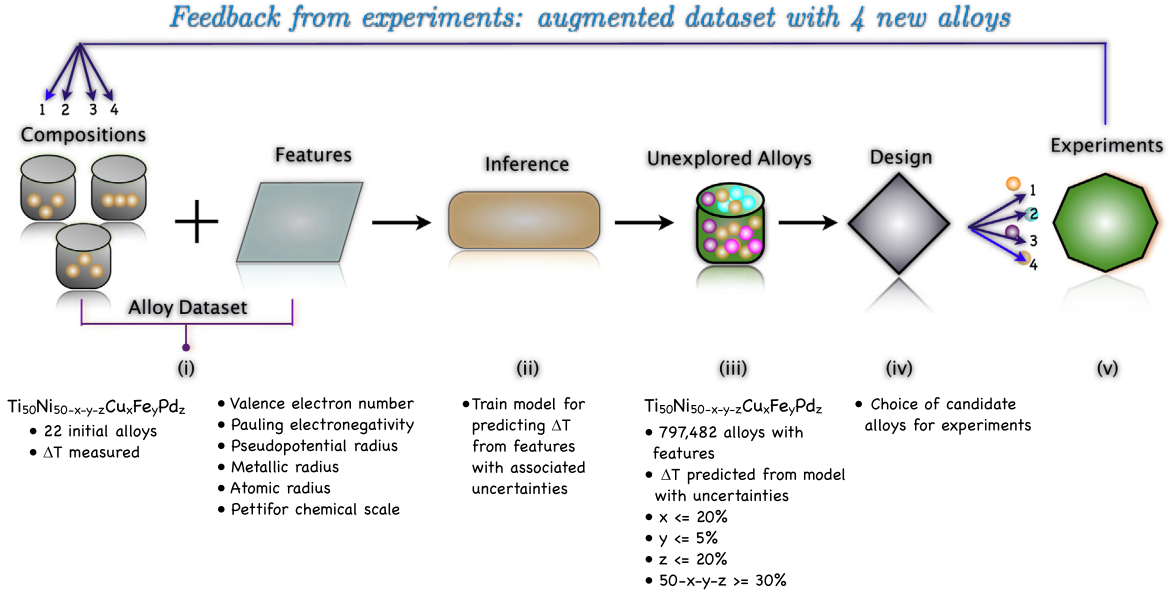

Fig. 4. The adaptive design loop in practice used for the accelerated search for new NiTi alloys with minimum thermal hysteresis. The list of steps involved in the experimental design is shown.

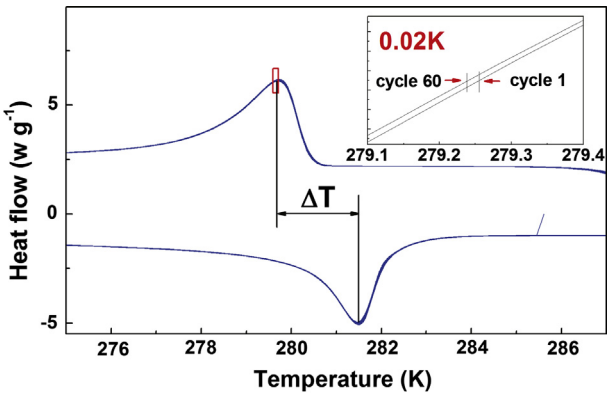

Fig. 5. Experimental differential scanning calorimetry (DSC) curves for the predicted $Ti_{50.0}Ni_{46.7}Cu_{0.8}Fe_{2.3}Pd_{0.2}$ alloy, whose peak-to-peak ΔT is measured as 1.84 K, which is the lowest among related NiTi-based SMAs. Thermal cycles (60 heating and cooling cycles) also show a small shift in the transition temperatures (~0.02 K in the inset), indicating excellent thermal fatigue resistance.

Pareto front for the problem. Methods such as EGO have been generalized to address this problem [49].

5. Conclusions

Although we used a GPM among our regressors, our approach is still very much a data-driven approach where the use of the domain knowledge is primarily in the feature selection. It is prudent to remind ourselves of the *no free lunch theorem* [50], which essentially states that there is no universal optimizer – what will work on one particular data set with a given model may not necessarily work with another data set. Thus, a challenge will continue to be the incorporation of materials knowledge in the form of constitutive and scaling relations, and known theoretical and empirical results, to constrain the regressor to make improved predictions than those achieved via a purely data-driven approach. One way to do this is to formulate a prior within a Bayesian type approach, beyond using a naive Bayesian prior to enable robust predictions. This has been studied to a greater degree within cancer genomics where the knowledge of metabolic pathways is encoded into prior distributions [51]. Our objective has been to show and emphasize that many of the methods available in the information sciences, especially those using uncertainties to explore the search space, can be applied to guide experiments and materials discovery. However, what is needed is to study how these methods perform on a variety of materials problems and data so that we can distill certain key principles which can apply across materials classes.

Acknowledgements

We acknowledge funding support via the Los Alamos National Laboratory (LANL) Laboratory Directed Research and Development (LDRD) DR (#20140013DR) project on Materials Informatics.

References

[1] Materials Genome Initiative for Global Competitiveness, 2011.
[2] Designing Materials to Revolutionize and Engineer our Future (DMREF), NSF 15-608, 2015.
[3] R. Fisher, The arrangement of field experiments, in: K. Samuel, N.B. Josnson (Eds.), Breakthroughs in Statistics: Methodology and Distribution, Springer, New York, 1992.
[4] R. Fisher, The Design of Experiments, Oliver and Boyd, 1937.
[5] P.J. Wissmann, M.A. Grover, A new approach to batch process optimization using experimental design, AIChE J. 55 (2) (2009) 342–353, http://dx.doi.org/10.1002/aic.11715.
[6] P.J. Wissmann, M.A. Grover, Optimization of a chemical vapor deposition process using sequential experimental design, Indust. Eng. Chem. Res. 49 (12) (2010) 5694–5701, http://dx.doi.org/10.1021/ie901055e.
[7] M.J. Casciato, S. Kim, J.C. Lu, D.W. Hess, M.A. Grover, Optimization of a carbon dioxide-assisted nanoparticle deposition process using sequential experimental design with adaptive design space, Indust. Eng. Chem. Res. 51 (11) (2012) 4363–4370, http://dx.doi.org/10.1021/ie2028574.
[8] T. Santner, B. Williams, W. Notz, The Design and Analysis of Computer Experiments, Springer Verlag, New York, 2003.
[9] H.J. Kushner, A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise, J. Basic Eng. 86 (1) (1964) 97–106.
[10] J. Mockus, On Bayesian methods of extremum search, Autom. Comp. Tech. 72 (1972) 53–62.
[11] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, J. Global Optimiz. 13 (4) (1998) 455–492, http://dx.doi.org/10.1023/A:1008306431147.
[12] A. Forrester, A. Sobester, A. Keane, Engineering Design via Surrogate Modelling: A Practical Guide, John Wiley and Sons, 2008.
[13] T. Lookman, P. Balachandran, D. Xue, G. Pilania, T. Shearman, J. Theiler, J. Gubernatis, J. Hogden, K. Barros, E. BenNaim, F. Alexander, A perspective on materials informatics: state-of-the-art and challenges, in: T. Lookman, F.J. Alexander, K. Rajan (Eds.), Information Science for Materials Discovery and Design, Springer Series in Materials Science, vol. 225, Springer International Publishing, 2016, pp. 3–12, http://dx.doi.org/10.1007/978-3-319-23871-5_1.
[14] S. Curtarolo, G. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nat. Mater. 12 (2013) 191–201, http://dx.doi.org/10.1038/nmat3568.
[15] R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T.O.L. S, D. Chon, K.R. Poeppelmeier, A. Zunger, Prediction and accelerated laboratory discovery of previously unknown 18-electron ABX compounds, Nat. Chem. 7 (4) (2015) 308–316.
[16] P.V. Balachandran, S.R. Broderick, K. Rajan, Identifying the inorganic gene for high–temperature piezoelectric perovskites through statistical learning, Proc. R. Soc. A: Math., Phys. Eng. Sci. 467 (2132) (2011) 2271–2290, http://dx.doi.org/10.1098/rspa.2010.0543.
[17] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, Accelerating materials property predictions using machine learning, Scient. Rep. 3 (2013) 2810, http://dx.doi.org/10.1038/srep02810.
[18] B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B 89 (2014) 094104, http://dx.doi.org/10.1103/PhysRevB.89.094104.
[19] P.V. Balachandran, J. Theiler, J.M. Rondinelli, T. Lookman, Materials prediction via classification learning, Scient. Rep. 5 (2015) 13285, http://dx.doi.org/10.1038/srep13285.
[20] W. Powell, I. Ryzhov, Optimal Learning, Wiley Series in Probability and Statistics, Wiley, 2013.
[21] A.I. Forrester, A.J. Keane, Recent advances in surrogate-based optimization, Prog. Aerosp. Sci. 45 (1–3) (2009) 50–79.
[22] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: the materials project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002, http://dx.doi.org/10.1063/1.4812323.
[23] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, AFLOWLIB.ORG: a distributed materials property repository from high-throughput *ab initio* calculations, Comput. Mater. Sci. 58 (1) (2012) 227–235, http://dx.doi.org/10.1016/j.commatsci.2012.02.002.
[24] J. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), JOM 65 (11) (2013) 1501–1509, http://dx.doi.org/10.1007/s11837-013-0755-4.
[25] A.L. MacKay, Generalized crystallography, Comp. Maths. Appl. B12 (1966) 21–37.
[26] J.R. Chelikowsky, J.C. Phillips, Quantum-defect theory of heats of formation and structural transition energies of liquid and solid simple metal alloys and compounds, Phys. Rev. B 17 (1978) 2453–2477, http://dx.doi.org/10.1103/PhysRevB.17.2453.
[27] E. Mooser, W.B. Pearson, On the crystal chemistry of normal valence compounds, Acta Cryst. 12 (1959) 1015–1022.
[28] J. John, A.N. Bloch, Quantum-defect electronegativity scale for nontransition elements, Phys. Rev. Lett. 33 (1974) 1095–1098, http://dx.doi.org/10.1103/PhysRevLett.33.1095.
[29] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J.R. Chelikowsky, W. Andreoni, Data mining for materials: computational experiments with *AB* compounds, Phys. Rev. B 85 (2012) 104104, http://dx.doi.org/10.1103/PhysRevB.85.104104.
[30] L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science: critical role of the descriptor, Phys. Rev. Lett. 114 (2015) 105503, http://dx.doi.org/10.1103/PhysRevLett.114.105503.
[31] G. Pilania, J. Gubernatis, T. Lookman, Structure classification and melting temperature prediction of octet AB solids via machine learning, Phys. Rev. B 91 (2015) 124301.
[32] G. Pilania, J. Gubernatis, T. Lookman, Classification of octet AB-type binary compounds using dynamical charges: a materials informatics perspective, Scient. Rep. 5 (2015) 17504, http://dx.doi.org/10.1038/srep17504.
[33] G. Hodes, Perovskite-based solar cells, Science 342 (6156) (2013) 317–318, http://dx.doi.org/10.1126/science.1245473.

[34] A. Belsky, M. Hellenbrandt, V.L. Karen, P. Luksch, New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design, Acta Crystall. Sect. B 58 (3 Part 1) (2002) 364–369, http://dx.doi.org/10.1107/S0108768102006948.
[35] R. Armiento, B. Kozinsky, M. Fornari, G. Ceder, Screening for high-performance piezoelectrics using high-throughput density functional theory, Phys. Rev. B 84 (2011) 014103, http://dx.doi.org/10.1103/PhysRevB.84.014103.
[36] E.R. Dougherty, A. Zollanvari, U.M. Braga-Neto, The illusion of distribution-free small-sample classification in genomics, Curr. Genom. 12 (5) (2011) 333–341, http://dx.doi.org/10.2174/138920211796429763.
[37] E.R. Dougherty, M.L. Bittner, Epistemology of the Cell: A Systems Perspective on Biological Knowledge, IEEE Press, 2011.
[38] P.V. Balachandran, D. Xue, J. Theiler, J. Hogden, T. Lookman, Adaptive strategies for materials design using uncertainties, Scient. Rep. 6 (2016) 19660.
[39] A. Booker, J.E. Dennis, P. Frank, D. Serafini, V. Torczon, M. Trosset, A rigorous framework for optimization of expensive functions by surrogates, Struct. Optimiz. 17 (1) (1999) 1–13, http://dx.doi.org/10.1007/BF01197708.
[40] N. Srinivas, A. Krause, S. Kakade, M. Seeger, Information-theoretic regret bounds for gaussian process optimization in the bandit setting, IEEE Trans. Inf. Theory 58 (5) (2012) 3250–3265, http://dx.doi.org/10.1109/TIT.2011.2182033.
[41] W.R. Thompson, On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, Biometrika 25 (3–4) (1933) 285–294, http://dx.doi.org/10.1093/biomet/25.3-4.285.
[42] O. Chapelle, L. Li, An empirical evaluation of Thompson sampling, in: J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 24, Curran Associates, Inc., 2011, pp. 2249–2257.
[43] R. Dehghannasiri, B.-J. Yoon, E.R. Dougherty, Efficient experimental design for uncertainty reduction in gene regulatory networks, BMC Bioinf. 16 (suppl. 13) (2015) S2.
[44] M.W. Barsoum, M. Radovic, Elastic and mechanical properties of the MAX phases, Ann. Rev. Mater. Res. 41 (2011) 195–227, http://dx.doi.org/10.1146/annurev-matsci-062910-100448.
[45] M.F. Cover, O. Warschkow, M.M.M. Bilek, D.R. McKenzie, A comprehensive survey of M_2AX phase elastic properties, J. Phys.: Cond. Matter 21 (30) (2009) 305403.
[46] J.T. Waber, D.T. Cromer, Orbital radii of atoms and ions, J. Chem. Phys. 42 (12) (1965) 4116–4123, http://dx.doi.org/10.1063/1.1695904.
[47] P. Frazier, W. Powell, S. Dayanik, The knowledge-gradient policy for correlated normal beliefs, INFORMS J. Comput. 21 (4) (2009) 599–613, http://dx.doi.org/10.1287/ijoc.1080.0314.
[48] D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nat. Commun. 7 (2016) 11241, http://dx.doi.org/10.1038/ncomms11241.
[49] J. Svenson, T. Santner, Multiobjective optimization of expensive-to-evaluate deterministic computer simulator models, Comput. Statist. Data Anal. 94 (2016) 250–264.
[50] D.H. Wolpert, The lack of a priori distinctions between learning algorithms, Neural Comput. 8 (7) (1996) 1341–1390, http://dx.doi.org/10.1162/neco.1996.8.7.1341.
[51] L.A. Dalton, E.R. Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework—part I: discrete and Gaussian models, Pattern Recog. 46 (5) (2013) 1301–1314.