Current Opinion in Solid State and Materials Science xxx (2016) xxx–xxx

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 xxxx

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

Please cite this article in press as: T. Lookman et al., Statistical inference and adaptive design for materials discovery, Curr. Opin. Solid State Mater. Sci. (2016), http://dx.doi.org/10.1016/j.cossms.2016.10.002