Chapter 1
A Perspective on Materials Informatics: State-of-the-Art and Challenges

T. Lookman, P.V. Balachandran, D. Xue, G. Pilania, T. Shearman, J. Theiler, J.E. Gubernatis, J. Hogden, K. Barros, E. BenNaim and F.J. Alexander

Abstract We review how classification and regression methods have been used on materials problems and outline a design loop that serves as a basis for adaptively finding materials with targeted properties.

T. Lookman (✉) · P.V. Balachandran · D. Xue · J.E. Gubernatis · E. BenNaim
Theoretical Division, T-4, Los Alamos National Laboratory, Los Alamos 87545, USA
e-mail: txl@lanl.gov

P.V. Balachandran
e-mail: pbalachandran@lanl.gov

D. Xue
e-mail: xdz@lanl.gov

J.E. Gubernatis
e-mail: jg@lanl.gov

E. BenNaim
e-mail: ebn@lanl.gov

G. Pilania
Materials Science Division, MST-8, Los Alamos National Laboratory,
Los Alamos 87545, USA
e-mail: gpilania@lanl.gov

T. Shearman
Program in Applied Mathematics, University of Arizona, Tucson 85721, USA
e-mail: toby.shearman@gmail.com

J. Theiler
ISR Division, Los Alamos National Laboratory, Los Alamos 87545, USA
e-mail: jt@lanl.gov

J. Hogden
CCS Division, CCS-3, Los Alamos National Laboratory, Los Alamos 87545, USA
e-mail: hogden@lanl.gov

K. Barros
Theoretical Division, T-1, Los Alamos National Laboratory, Los Alamos 87545, USA
e-mail: kbarros@lanl.gov

F.J. Alexander
CCS Division, Los Alamos National Laboratory, Los Alamos 87545, USA
fja@lanl.gov

T. Lookman et al. (eds.), *Information Science for Materials Discovery and Design*, Springer Series in Materials Science 225,
DOI 10.1007/978-3-319-23871-5_1

1.1 Introduction

There has been considerable interest over the last few years in accelerating the process of materials design and discovery. The Materials Genome Initiative (MGI) [1], Integrated Computational Materials Engineering (ICME) [2] and Advanced Manufacturing [3] initiatives have spurred considerable activity and brought new researchers into the nascent field of materials informatics which includes the accelerated design and discovery of new materials. The activity has also highlighted some of the open questions in this emerging area and our objective here is to provide a perspective of the field in terms of general problems and information science methods that have been used to study classes of materials, and point to some of the outstanding challenges that need to be addressed. We are guided here by our own recent work at the Los Alamos National Laboratory (LANL).

One of the earliest-studied problems in modern materials informatics relates to the classification of AB solids into their stable crystal structures, based on key attributes of the chemistry and properties of the individual A and B constituents. The emphasis was on finding features that can give rise to easily visualized two-dimensional structural maps by "drawing" boundaries between classes. The problem was first studied in the 1960s [4] but Chelikowski and Phillips [5], studying the same problem in 1978, recognized the connections to information science. Realizing that energy differences between structures were rather small, they observed that "from the point of view of information theory, …the available structural data already contain a great deal of information: about 120 bits, in the case of the AB octet compounds. Thus one can reverse the problem, and attempt to extract from the available data quantitative rules for chemical bonding in solids." They realized that suitable combinations of orbital radii of the individual A and B atoms were appropriate features for predicting the crystal structure of the AB solids. Over the last few years, this problem has been revisited with a variety of machine learning methods (decision trees,support vector machines, gradient boosting, etc.) [6–8] and there have been a number of studies that have classified different materials classes, such as perovskites [9]. Feature selection from data remains a fundamental exercise and here principal component analysis and correlation maps have been widely employed. Recently, high-throughput approaches have been utilized to form combinations of features from a given set and then certain key combinations are down-selected [6].

The problem of materials design is about predicting the composition and processing of materials with a more desired targeted property and therefore involves regression that leads to an inference model from training data. For example, for ferroelectrics one may wish to discover lead-based or lead-free piezoelectrics with a high transition temperature or high piezoelectric coefficient. For shape memory alloys, one may seek compounds with reduced dissipation or low hysteresis. Typically, such materials are usually found in an Edisonian fashion using intuition and time-consuming trial and error. In recent years, theory has become powerful enough to predict very accurately some material characteristics, for example, *ab initio* calculations predict elastic constants, inter-atomic distances, crystal structure, polarization, etc. However, the parameter space is just too large and there are too many

possibilities, and even if nature rules out many of the possible combinations, the numbers are still staggering. Moreover, physical and chemical constraints make the realization of many theoretically possible materials impossible. Thus, one needs to successively improve or learn from available data candidate materials for further experiments and calculations. Recently, a number of studies have utilized regression methods to predict materials with given properties. However, most research in materials design has been based on high throughput approaches using electronic structure calculations. Typically, a large database is assembled with calculated properties and this is successively screened for materials with desired properties. High-throughput experiments have also been undertaken more recently to screen for candidate materials for further experiments [10, 11]. When it comes to multicomponent alloys or solid solutions, these methods have limitations. Moreover, very few studies have combined statistical inference with the high-throughput approach.

1.2 Statistical Inference and Design: Towards Accelerated Materials Discovery

Figure 1.1 illustrates our vision for the overall materials informatics/design problem. This shows a feedback loop that starts with the available assembled data (box 5), which may be obtained from multiple sources, including experiments or calculations. Materials knowledge (box 1) is then key in selecting the features and prescribing the constraints amongst them. Our aim is to train a statistical inference model that estimates the property (regression) or classification label with associated uncertainties (box 2). Classification models answer categorical questions: Is a compound stable?

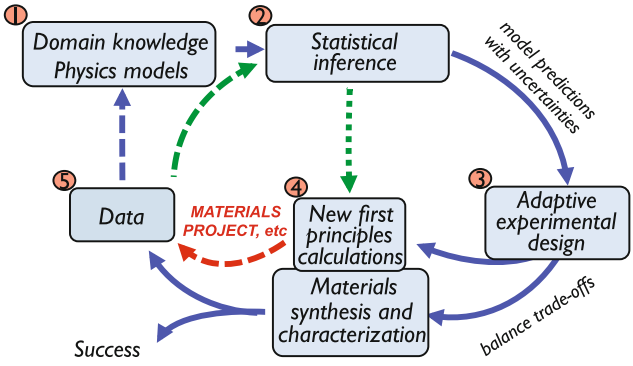

Fig. 1.1 Statistical Inference and design: A feedback loop to find a material with a desired targeted property. Prior or domain knowledge, including features, provide input to an inference model that predicts a label or a property with uncertainty. An experimental design or decision making module balances trade off between exploiting information or further exploring the high dimensional search space where the desired material may be found. A material is suggested for experimentation or calculation and the process repeats itself incorporating updated information

Is it a piezoelectric? What is its crystal symmetry? Regression models produce numerical estimates: What is the material's piezoelectric coefficient? What is its transition temperature? Because there usually is a limited quantity of training data, and because the space of possibilities is so high-dimensional, incorporation of domain knowledge is of potentially great value. Here explicitly Bayesian approaches, in which this knowledge is coded into prior probability distributions, and more traditional machine learning algorithms (such as support vector machines) in which case the domain knowledge could be incorporated as constraints or folded into the kernel design, become important [12].

Much existing work is essentially based on going from box 1 to box 4 in Fig. 1.1. A case in point are projects such as the Materials Project [13] and AFLOWLIB [14] focused on establishing databases using electronic structure calculations to make predictions. However, there are a few studies that use inference to make predictions. Examples include predictions of melting temperature [7, 8, 15] or piezoelectrics with high transition temperatures [16]. The search for piezoelectrics serves as a good example to contrast the two approaches. Extensive *ab initio* calculations were performed on a chemical space represented by $63^2 = 3969$ possible perovskite ABO_3 (up to Bi but excluding a few such as H and inert gases) end structures [17]. The number of possibilities were filtered down to 49 by discarding compounds that are nonmetallic or whose structures have small energy barriers to distortions across the morphotropic phase boundary (MPB) according to preset values. Almost no optimization or learning tools are used other than what may be involved in seeking an optimal minimum energy solution at zero temperature. All the physics is contained in this first-principles calculation, and we are not aware if any of this group's predictions of piezoelectricity have been verified experimentally. On the other hand, the approach of Balachandran et al. [16] on the same type of problem was to focus on a given subclass of piezoelectrics (e.g. Bi based) with known crystallographic and experimental data and use off-the-shelf inference tools to obtain candidates with high transition temperatures and that were formable. The tools included principal component analysis (PCA) for dimensionality reduction, partial least squares (PLS) regression for predicting transition temperatures and recursive partitioning (or decision trees) with a metric such as Shannon entropy for classification. The training data sets for PCA or regression studies were rather small (about 20 data points, 30 features) but data sets with 350 data points were also used to identify stable/formable perovskite compounds. Two new compounds were predicted, of which one has been synthesized [18], with the predicted transition temperature differing by 30–40 %. However, a key element lacking is the issue of uncertainties in predictions.

In Fig. 1.2, we demonstrate using an example, where we have used bootstrap methods (i.e. sampling with replacement) to estimate prediction uncertainties. Here, we took the same Bi-based piezoelectrics data set as that utilized in the work of Balachandran et al. [16] We generated a large number of bootstrapped samples (as opposed to using just one in the earlier work of Balachandran et al.) and utilized support vector regression (SVR) for predicting the Curie temperature (T_C). Our results with uncertainties are shown in Fig. 1.2. On average, we obtained a standard deviation of 37 °C from the mean value of predicted T_C. More importantly, we also

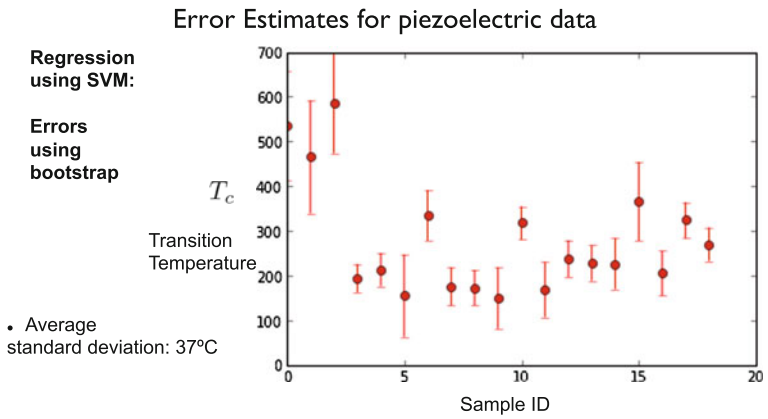

Fig. 1.2 Predictions using support vector regression (SVR) with uncertainties from bootstrap method. The piezoelectric data set of Bi-based $PbTiO_3$ solid solutions was used for machine learning. T_C (in °C, y-axis) is the predicted ferroelectric Curie temperature at the morphotropic phase boundary (MPB). We use the SVR model and predicted T_C for two new compounds, $BiLuO_3$-$PbTiO_3$ and $BiTmO_3$-$PbTiO_3$, to be 552.5 ± 79 and 564.2 ± 97 °C, respectively. Experimentally, T_C for $BiLuO_3$-$PbTiO_3$ was measured as 565 °C [18]

predicted the T_C for two new compounds, $BiLuO_3$-$PbTiO_3$ and $BiTmO_3$-$PbTiO_3$, to be 552.5 ± 79 and 564.2 ± 97 °C, respectively with 95 % confidence. Experimentally, T_C for $BiLuO_3$-$PbTiO_3$ was measured as 565 °C [18], in close agreement with the current results from SVR. On the other hand, PLS predicted the T_C for $BiLuO_3$-$PbTiO_3$ to be 705 °C. The merit of this example is that it shows in a rather modest manner that the informatics approach, even if manual and piecemeal, is potentially capable of predicting new materials.

A key aspect of our design loop is the uncertainty associated with the properties predicted from inference (box 2). These play a role in the adaptive experimental design (box 3) which suggests the next material to be chosen for further experiments or calculation (box 4) by balancing the tradeoffs between "exploration and exploitation". That is, at any given stage a number of samples may be predicted to have given properties with uncertainties. The tradeoff is between exploiting the results by choosing to perform the next experiment on the material predicted to have the largest property or further improving the model by performing the experiment or calculation on a material where the predictions have the largest uncertainties. By choosing the latter, the uncertainty in the property is expected to (given the model, statistiscs) decrease, the model will probably improve and this will influence the results of the next iteration in the loop. While there is a considerable literature on error estimation methodologies, accurate and reliable error estimation with limited data is harder than simple prediction, and there is an even a stronger case for incorporating domain knowledge. [8, 19, 20]

Extracting measures of confidence, while at the same time encoding prior knowledge, is not an easy task but recent research in cancer genomics has demonstrated that increasing confidence in classification analysis built on small databases benefits

significantly from using prior knowledge [21, 22]. Prior domain knowledge constrains statistical outcomes by producing classifiers that are superior to those designed from data alone. How to use prior knowledge in classification and regression is a problem not only for materials and cancer genomics but for machine learning generally. Developing ways of constructing and using prior domain knowledge will distinguish the materials machine learning approach to classification and regression. The lesson learned from high-throughput genomics concerning classification is that, in high dimensional, small-sample settings, model-free classification is virtually impossible. The reason is that the salient property of any classifier is its error rate because the error rate quantifies its predictive capacity, which is the essential issue pertaining to scientific validity. Since the error rate must be estimated, there must be an estimation procedure and, with small samples, this procedure must be applied to the same data as that used for designing the classifier. In cancer genomics, Dalton and Dougherty [19, 20] addressed the problem by formulating error estimation as an optimization problem in a model-based framework and leads to a minimum-mean-square-error (MMSE) estimate of the classifier error. They formulate a prior probability distribution over a class of possible distributional models governing the features to be measured and the possible decisions to be made, each such model being known as a feature-label distribution. They then design a classifier from the data and an optimal MMSE error estimate is derived from the data. How well this approach will work for materials problems remains an open question.

In Figs. 1.3 and 1.4 we provide more details of our loop. Figure 1.3 shows how the loop would actually work in practice, and some of the algorithms that may be

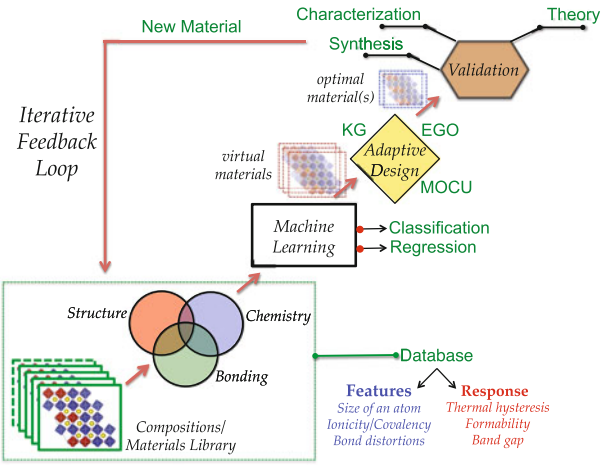

Fig. 1.3 The design loop in practice showing different stages of machine learning and adaptive design strategies with an iterative feedback loop. For completeness, we have also included experiments (synthesis and characterization), which are vital for validation and feedback. KG, EGO and MOCU stand for knowledge gradient, effective global optimization and mean objective cost of uncertainty, respectively

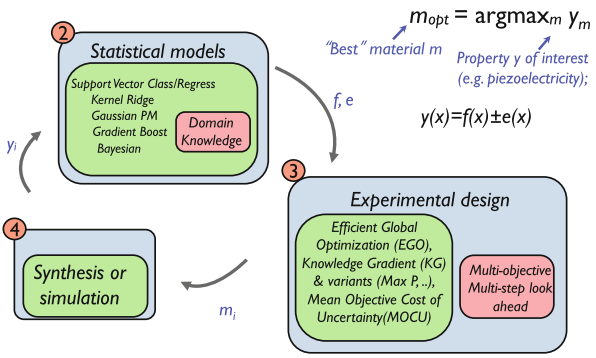

Fig. 1.4 A sub-component of our adaptive design loop showing the synergy between statistical models (box 2), experimental design (box 3) and validation (typically via experimental synthesis or simulation as shown in box 4). Statistical models use the available data to fit a regression model (f) along with an uncertainty measure (e). The experimental design component then evaluates the tradeoff between exploitation and exploration and suggests the "best" material (y_i) for validation. Here the term "best" need not correspond to a material with the optimal response. Alternatively, it refers to the choice of a material that would reduce the overall uncertainty in our model. Different statistical learning (including Bayesian learning) and adaptive design methods are given

used as part of the statistical inference and design tools, are shown in greater detail in Fig. 1.4. The green emphasize algorithms that can be utilized today and the red represent areas requiring further study and development. Design algorithms include well known exploitation-exploration strategies such as efficient global optimization (EGO) [23], and the closely related knowledge gradient(KG) [24] based on single-step look ahead.

1.3 Progress and Concluding Remarks

Our work at LANL has involved studying a number of materials problems along the lines of the approach described. These include problems involving classification learning and regression, which essentially involve an inner loop of Fig. 1.1 with boxes 2, 4 and 5. We have examined the role of features in classifying AB octet solids [8] and perovskites [9], as well as predicting new ductile RM intermetallics, where R and M are rare earth and transition metal elements, respectively [25]. These studies have suggested new features that led to better classification as well as new materials. In the case of RM intermetallics, we have shown that machine learning methods naturally uncover the functional forms that mimic most frequently used features in the literature, thereby providing a mathematical basis for feature set construction without *a priori* assumptions [25]. Our classification models (Fig. 1.5) that use orbital radii as features predicted that ScCo, ScIr, and YCd should be ductile, whereas each was previously proposed to be brittle. These results show it is possible to design

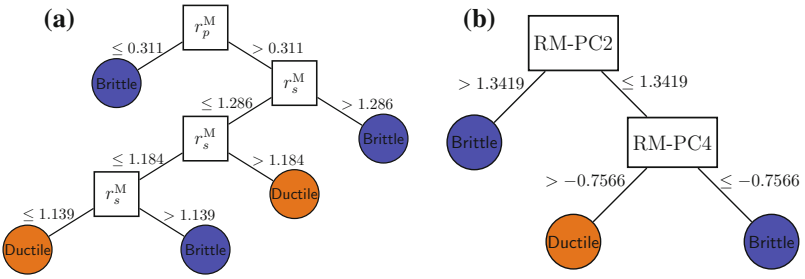

Fig. 1.5 Classification learning using decision trees to predict whether a given RM intermetallic, where R and M are rare earth and transition metal elements, respectively, is brittle and ductile. **a** Decision tree that uses the orbital radii as features and (b) Decision tree that uses the principal components (RM-PC2 and RM-PC4) that automatically extracts features in the form of the linear combinations of orbital radii. For example, RM-PC2 is defined as $-0.70r_p^M + 0.08r_s^M - 0.71r_d^M$. Features r_p^M, r_s^M and r_d^M are the p-, s- and d-orbital radii of atom-M, respectively

targeted mechanical properties in intermetallic compounds, which has significant implications for next-generation multi-component alloy discovery.

Our on-going work on multi-objective regression includes predicting functional polymers with large band gaps, as well as large dielectric constants for energy storage applications. Similarly, we are also performing high-throughput density functional theory (DFT) calculations to generate large data sets, which are subsequently mined using machine learning methods to identify new and previously unexplored candidate water splitting compounds for catalysis.

In the area of adaptive design, our focus has been on demonstrating the feedback loop of Figs. 1.1 or 1.3 with tight coupling to an "oracle", which can be experiments (synthesis and characterization) or calculations. Specific materials studies include discovering new low thermal dissipation shape memory alloys, as well as Pb-free piezoelectric solid-solutions starting from experimental data on specific multicomponent systems. The search spaces can be well defined, for example, they can be a factor of 10^5 greater than the size of the training data. In addition, extensive databases from *ab initio* calculations become invaluable in benchmarking the various algorithms. For example, elastic moduli data for the hexagonal layered M_2AX phases consist of a library of 240 compounds. The *ab initio* data of the elastic constants and moduli were taken from the literature [26] with results well calibrated to experiments. In the M_2AX phases, X-atoms reside in the edge-connected M octahedral cages and the A atoms reside in slightly larger right prisms [27]. These M_2AX phases represent a unique family of materials with layered crystal structure and both metallic- and ceramic-like properties. We used orbital radii of M, A, and X atoms from the Waber-Cromer scale [28] as features, which include the s-, p-, and d-orbital radii for M, while the s- and p-orbital radii were used for A and X atoms. With the M_2AX data, we benchmarked our adaptive design strategy, i.e. explored different training set sizes, regressors, regressor/optimization combinations, etc., and uncovered invaluable guidelines that were eventually useful for real materials design problems.

Implementing the loop using simulation codes allows us to optimize the use of these codes in seeking a well defined set of parameters or constraints for given targeted outcomes. For example, an industry standard code for simulating semiconducting materials is APSYS (Advanced Physical Models of Semiconductor Devices). It is based on 2D/3D finite element analysis of electrical, optical and thermal properties of compound semiconductor devices, with silicon as a special case with an emphasis on band structure engineering and quantum mechanical effects. Inclusion of various optical modules allows one to configure applications involving photosensitive or light emitting diodes (LEDs). We have been recently using APSYS to investigate how to optimize the LED structure (number of quantum wells, indium concentration) of GaAs based systems for highest internal quantum efficiencies at high currents.

In summary, the use of classification and regression methods, in combination with optimization strategies, has the potential to impact discovery and design in materials science. What is needed is to establish how these tools perform on an array of materials classes with differing physics in order to distill some guiding principles for use by the materials community at large.

Acknowledgments We acknowledge funding support from a Laboratory Directed Research and Development (LDRD) DR (#20140013DR) at the Los Alamos National Laboratory (LANL).

References

1. *Materials Genome Initiative for Global Competitiveness* (2011)
2. S.R. Kalidindi, M. De Graef, Materials data science: current status and future outlook. Ann. Rev. Mater. Res. **45**(1), 171–193 (2015)
3. T.D. Wall, J.M. Corbett, C.W. Clegg, P.R. Jackson, R. Martin, Advanced manufacturing technology and work design: towards a theoretical framework. J. Organ. Behav. **11**(3), 201–219 (1990)
4. E. Mooser, W.B. Pearson, On the crystal chemistry of normal valence compounds. Acta Crystallogr. **12**, 1015–1022 (1959)
5. 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**, 2453–2477 (1978)
6. 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**, 105503 (2015)
7. 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**, 104104 (2012)
8. G. Pilania, J.E. Gubernatis, T. Lookman, Structure classification and melting temperature prediction of octet AB solids via machine learning. Phys. Rev. B **91**, 124301 (2015)
9. G. Pilania, P.V. Balachandran, J.E. Gubernatis, T. Lookman, Predicting the formability of ABO3 perovskite solids: a machine learning study. Acta Crystallogr. B **71**, 507–513 (2015)
10. S.M. Senkan, High-throughput screening of solid-state catalyst libraries. *Nature* **394** (6691), 350–353, 07 (1998)
11. H. Koinuma, I. Takeuchi, Combinatorial solid-state chemistry of inorganic materials. Nat. Mater. **3**, 429–438 (2004)
12. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2008)

13. 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)
14. 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**, 227–235 (2012)
15. A. Seko, T. Maekawa, K. Tsuda, I. 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)
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), 2271–2290 (2011)
17. R. Armiento, B. Kozinsky, M. Fornari, G. Ceder, Screening for high-performance piezoelectrics using high-throughput density functional theory. Phys. Rev. B **84**, 014103 (2011)
18. W. Hu, Experimental search for high Curie temperature piezoelectric ceramics with combinatorial approaches. Ph.D. dissertation, Iowa State University (2011)
19. L.A. Dalton, E.R. Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework–Part I: discrete and Gaussian models. Pattern Recognit. **46**(5), 1301–1314 (2013)
20. L.A. Dalton, E.R. Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework—Part II: properties and performance analysis. Pattern Recognit. **46**(5), 1288–1300 (2013)
21. K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, B.K. Mallick, Gene selection: a Bayesian variable selection approach. Bioinformatics **19**(1), 90–97 (2003)
22. E.R. Dougherty, A. Zollanvari, U.M. Braga-Neto, The illusion of distribution-free small-sample classification in genomics. Curr genomics **12**(5), 333–341 (2011)
23. D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions. J. Glob. Optim. **13**(4), 455–492 (1998)
24. W. Powell, I. Ryzhov, *Optimal Learning*, Wiley Series in Probability and Statistics (Wiley, Hoboken, 2013)
25. P.V. Balachandran, J. Theiler, J. M. Rondinelli, T. Lookman, Materials Prediction via Classification Learning Sci. Rep. **5**, 13285 (2015)
26. M.F. Cover, O. Warschkow, M.M.M. Bilek, D.R. McKenzie, A comprehensive survey of M_2AX phase elastic properties. J. Phys.: Condens. Matter **21**(30), 305403 (2009)
27. M.W. Barsoum, M. Radovic, Elastic and mechanical properties of the MAX phases. Ann. Rev. Mater. Res. **41**, 195–227 (2011)
28. J.T. Waber, D.T. Cromer, Orbital radii of atoms and ions. J. Chem. Phys. **42**(12), 4116–4123 (1965)

http://www.springer.com/978-3-319-23870-8

Information Science for Materials Discovery and Design
Lookman, T.; Alexander, F.J.; Rajan, K. (Eds.)
2016, XVII, 307 p. 134 illus., 88 illus. in color.,
Hardcover
ISBN: 978-3-319-23870-8