Materials Informatics: The Materials “Gene” and Big Data

Krishna Rajan

Department of Materials Design and Innovation, University at Buffalo, State University of New York, Buffalo, New York 14260; email: krajan3@buffalo.edu

Annu. Rev. Mater. Res. 2015. 45:153–69

The *Annual Review of Materials Research* is online at matsci.annualreviews.org

This article’s doi:
10.1146/annurev-matsci-070214-021132

Keywords

uncertainty, statistical inference, information theory, fuzzy logic, rough sets

Abstract

Materials informatics provides the foundations for a new paradigm of materials discovery. It shifts our emphasis from one of solely searching among large volumes of data that may be generated by experiment or computation to one of targeted materials discovery via high-throughput identification of the key factors (i.e., “genes”) and via showing how these factors can be quantitatively integrated by statistical learning methods into design rules (i.e., “gene sequencing”) governing targeted materials functionality. However, a critical challenge in discovering these materials genes is the difficulty in unraveling the complexity of the data associated with numerous factors including noise, uncertainty, and the complex diversity of data that one needs to consider (i.e., Big Data). In this article, we explore one aspect of materials informatics, namely how one can efficiently explore for new knowledge in regimes of structure-property space, especially when no reasonable selection pathways based on theory or clear trends in observations exist among an almost infinite set of possibilities.

BIG DATA AND THE MATERIALS GENE

In a special issue of the *Philosophical Transactions of the Royal Society* in 2012 entitled "Beyond Crystals," Cartwright & Mackay (1) posed, in their paper, the provocative question, "How do structure and information interact?" This discussion is an extension of Mackay's seminal work over decades on the topic of generalized crystallography, wherein the crystal is a structure whose description is much smaller than the structure itself and carries information about the structure on larger length scales (2). Although the phrase materials genome has entered the lexicon of materials science in the last few years, as Rajan and colleagues (3, 4) noted in earlier publications, Mackay had in fact been suggesting for decades in the context of crystallography that one may be able to construct an "inorganic gene" by exploring how fundamental pieces of information, treated as discrete bits of data, can collectively characterize the stability and properties of a given crystal chemistry. Mackay and colleagues, however, have cautioned against simple extrapolation of the concept of a biological gene to materials science. Although local rules determining properties such as stability of a crystal structure emerge from our understanding of fundamental electronic structure, the broader questions of what determines other and larger length scales of structure, and the overall complexity of structure-property relationships, have to account for how the critical genes of characteristics of information are distributed. Mapping the complexity of a material requires understanding algorithmically how these genes of information self-organize and self-assemble. In this review, we define informatics as the assembly of tools and strategies that help elucidate this information self-assembly process and extract patterns and uncover relationships between data that would otherwise not be easily seen or be left undiscovered.

This discussion raises the issue as to what characteristics of data we need to extract the complexity that Mackay and his colleagues have so elegantly described. At present, we have sophisticated models and experimental techniques that address specific segments of engineering design such as developing new materials (e.g., first-principles calculations), refining legacy materials (e.g., through processing and microstructural modification), and systems design and manufacturing. The linking of information across multiple scales can of course be done in many ways, but no single tool can account for the interaction of the myriad of parameters that govern materials development and assess the complexity of interactions of these parameters in defining engineering performance. Current approaches that utilize informatics tools such as data mining, evolutionary algorithms, and other statistical methods do so in conjunction with physically based and/or heuristically driven models, in which the primary focus is to search for information from large data sets generated by computations and/or experiments (5–8).

These advances have helped establish means for rapidly generating large databases and, when coupled with machine learning methods, allow us to efficiently search and share information from these large databases. However, the sheer size of databases is insufficient for materials discovery and design. These approaches can help guide materials design, provided that we can interrogate these databases with a clear a priori rationale or a query, on the basis of existing heuristic and/or theoretical understanding of materials behavior. However, exploring the full translational potential from fundamental materials discovery to the development of manufacturing processes and engineering design will require us to look for connections for which explicit theories or heuristic phenomena may not currently exist. The key challenges are the development and incorporation of mathematical and statistical techniques involving the study of high-dimensional, sparse, and noisy data (Big Data) and methods to extract patterns from them that can provide us with Mackay's inorganic gene.

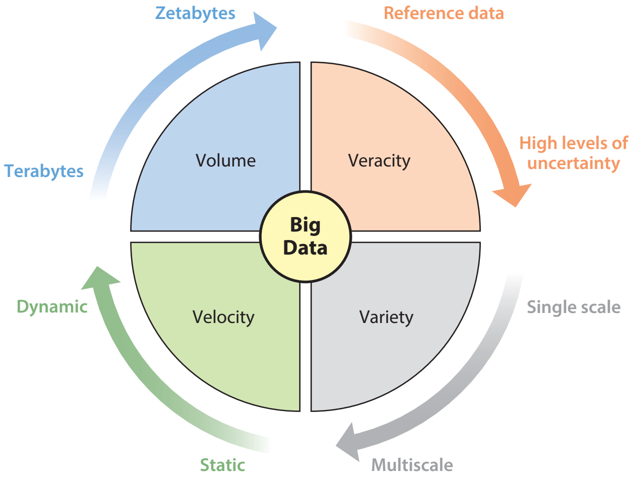


Figure 1

Big Data is characterized as the intersection of the primary attributes of volume, veracity, variety, and velocity.

HARNESSING THE CHARACTERISTICS OF BIG DATA

The characteristics of Big Data, identified in the previous section, are effectively summarized by the phraseology of the four V's: volume, variety, veracity, and velocity (9). Much of the effort in the last decade in the materials science community has focused on ways to increase the volume of data via high-throughput computation or high-throughput combinatorial experiments (10–19). In this article, we wish to highlight those aspects of materials informatics that will help one to harness all the V's of Big Data, and not just volume (**Figure 1**).

We summarize the characteristics of Big Data in the context of materials science as follows:

1. Data volume refers to the size of a data set; what constitutes a large size varies from subject to subject. For instance, the size of crystallographic data sets in terms of the number of compounds in a crystallographic database runs into the hundreds of thousands, whereas the number of experimental and computationally derived thermochemical metallurgical phase diagrams runs in the few tens of thousands.
2. With regard to data veracity, in the numerous efforts to create and compile databases, data are ultimately of value if they can be used repeatedly for the common good of the community. The value of reference data is based on the confidence one has in their accuracy. Veracity in data refers to the broad issue of uncertainty. The source of that uncertainty can vary due to the practical reality in materials science that we have many missing data and from the fact that, when we do have data, there is associated uncertainty due to experimental and/or computational limitations.
3. Data variety is concerned with the fact that, in materials science, data take all forms, ranging from discrete numerical values to qualitative descriptions of materials behavior and imaging data. In the case of materials science problems, the source of this variety can be associated with the nature of the length scale of the data, e.g., elastic properties described as a singular scalar quantity such as modulus or described in terms of a tensor quantity.
4. Data velocity refers to the dynamics of the evolution of data for which the analysis of that motion allows us to reveal new information that would otherwise not be easily seen. This concept can also relate to the harnessing of real-time data acquisition (e.g., data from

dynamic experiments such as in situ deformation studies or time-of-flight spectroscopy). Additionally, the concept of data velocity in materials science is not limited to time-series data but also encompasses data that are systematically perturbed or changed due to some other metric (e.g., chemistry or processing).

In advancing data-driven or data-intensive research, computational materials science, for instance, relies on advancing many modeling approaches ranging from ab initio calculations to kinetic and systems-level modeling. Applying each of these models and finding ways to integrate the output from these models are major research foci and are critical to harnessing the data we have. However, informatics not only can be an efficient search strategy for large data sets but also can allow us to learn from those data by deriving information that may be outside the models on which such data are based, and we can use this learning process to efficiently and robustly explore the information space in a way that is not possible by using existing models. In this context, the tools of machine learning coupled to statistics need to be judiciously linked to the foundations of materials science, namely theory, modeling, and experiments, to make databases a laboratory for generating new information and not just a repository for retrieving known or expected information (**Figure 2**).

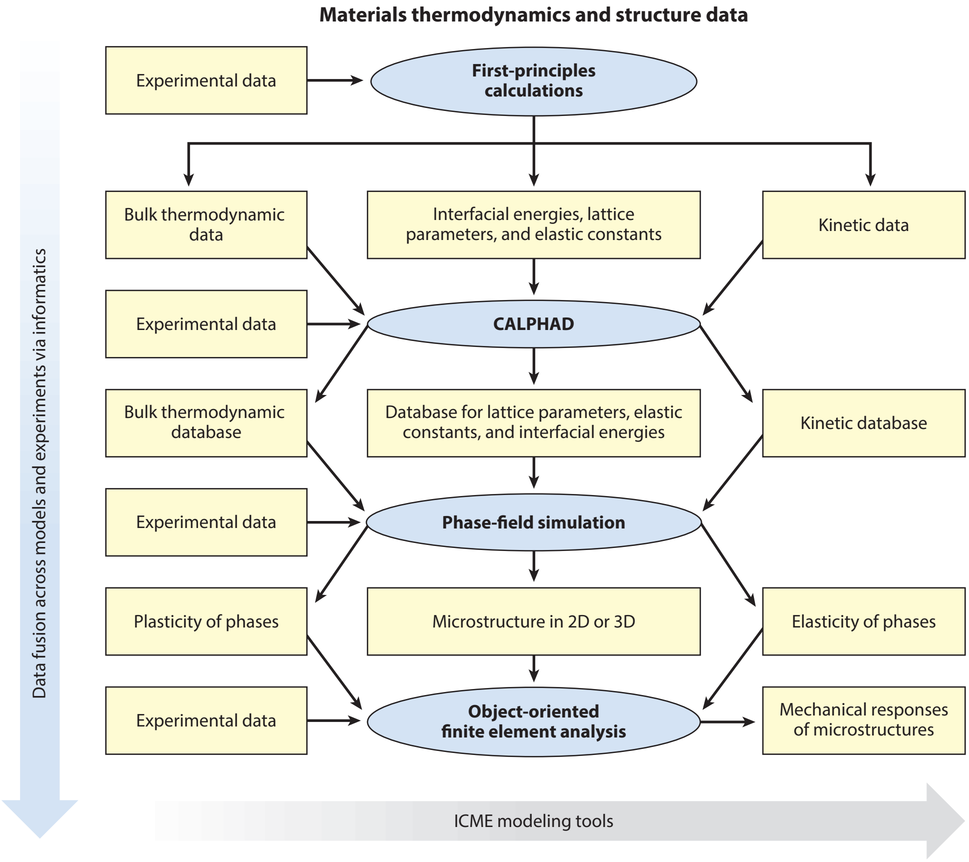

Materials thermodynamics and structure data
Experimental data
First-principles calculations
Bulk thermodynamic data
Interfacial energies, lattice parameters, and elastic constants
Kinetic data
Experimental data
CALPHAD
Bulk thermodynamic database
Database for lattice parameters, elastic constants, and interfacial energies
Kinetic database
Experimental data
Phase-field simulation
Plasticity of phases
Microstructure in 2D or 3D
Elasticity of phases
Experimental data
Object-oriented finite element analysis
Mechanical responses of microstructures
Data fusion across models and experiments via informatics
ICME modeling tools

Figure 2

An example of the role of informatics to enhance the fusion of data across multiscale models. ICME refers to integrated computational materials engineering. Adapted with permission from Reference 20. Copyright 2006, Springer.

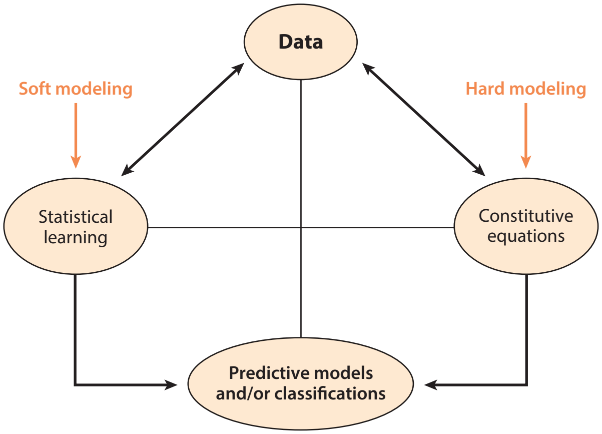


Figure 3

Comparison of complementary strategies for the application and utilization of data by using soft and hard modeling. Adapted with permission from Reference 21. Copyright 2014, Elsevier.

Modeling seeks to utilize data defined by the constitutive equations one needs to solve. However, informatics allows one to get into the realm of soft modeling, which seeks to harness the data with no explicit a priori assumption as to those constitutive relationships. If done properly, the predictions of the constitutive models should be recovered as well as allowing us to refine and/or identify new models that are beyond the foundations of traditional hard models.

The tools of soft computing, when coupled with both experiments and computational methods based on hard or theory/heuristic-based modeling (which forms the core of most computational materials science), allow us to move into a knowledge space that may not be readily feasible (**Figure 3**) (22–24). Zadeh (22) defined the principal constituents of soft computing as

> fuzzy logic, neural network theory and probabilistic reasoning, with the latter subsuming belief networks, genetic algorithms, parts of learning theory and chaotic systems. In the triumvirate of fuzzy logic, neural networks, and probabilistic reasoning, fuzzy logic is mainly concerned with imprecision; neural networks with learning and probabilistic reasoning with uncertainty.

Hence, algorithmically, there are a wide array of methods to include in our computational arsenal, and the reader is directed toward some of the references cited at the end of this article as a guide to the computational approaches that are available (**Figure 4**) (25–38). In the following discussion, we provide some examples of how, by utilizing these soft computing methods, we can harness the multiple V's of Big Data for knowledge discovery by taking advantage of the tools of machine learning and inference to help build an unsupervised learning framework for knowledge discovery, even when the volume or data size is relatively small.

APPLICATIONS

In the field of crystallography, empirical observations have long guided materials scientists to find combinations of parameters associated with crystal geometry and bonding that are signatures for why a given compound takes on a given crystal structure. Even today, with the tremendous advancements in high-throughput electronic structure calculations, these design rules are guideposts in the interpretation of what, from a data science perspective, is a classification problem. Examples

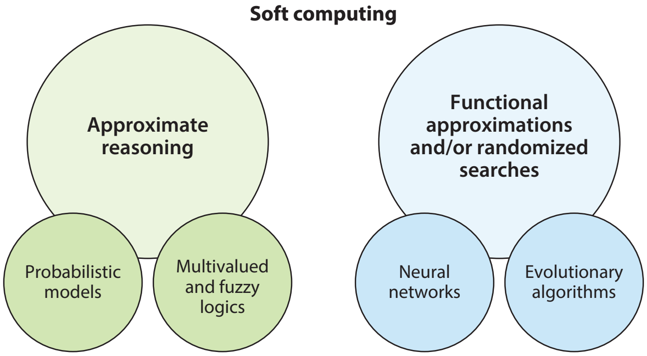


Figure 4

A schematic of the computational taxonomy used in soft computing methods. Adapted with permission from Reference 32. Copyright 2008, Elsevier.

include the classical Hume-Rothery rules for substitutional alloys and the Philips–Van Vechten rules for classifying compounds on the basis of their ionicity (4, 39–43). Such design rules, although useful for retrospectively creating groupings among chemistry-structure relationships, have had a lesser impact on structure-property prediction. This reality is apparent from the facts that the conditions for classification are not universal and that the boundaries between these classifications are often diffuse. Again, in terms of the semantics of informatics, the membership of a given compound in a given class of structure types is fuzzy. In the following subsection, we provide some examples of how we exploit that fuzzy membership, which is a reflection of not only the veracity or uncertainty of data but also the variety of groupings or clusters in which data may reside.

One of the core objectives in materials science is the search for structure-function relationships. The challenge in discovering those relationships increases exponentially as one tries to sort through large quantities of data (keeping in mind our definition of Big Data). The tools of machine learning can help facilitate the process of discovering classifications within large and diverse data sets. The best analogy to consider is that of putting together a puzzle, which has large numbers of pieces, with each piece also having complex geometries, and where the differences between adjacent pieces are not easily discernible in contrast due to the complexity of the final picture of that puzzle. One can add another level of difficulty by noting that we may not have all the pieces of the puzzle and that the pieces that are available may not fit perfectly. Thus, the grand challenge is whether we can still efficiently discover what the pieces of the puzzle add up to. This metaphor of putting together a puzzle with imperfect information is an apt analogy to many problems in materials science. Hence in this article we discuss approaches that go far beyond the issue of efficient searching of data among large data sets to one of learning from existing data sets. A critical step in this process is the need to establish criteria to initiate the sorting of the pieces in the puzzle and help us on the way to systematically finding ways of identifying the best-fitting pieces. In the following sections, different approaches to this challenge are presented in addressing a variety of problems in materials discovery.

Ranking Genes from Uncertain and Diverse Information: Statistical Inference Methods

This subsection explores how, by harnessing the variety and veracity of descriptors, one can uncover information by finding groups or classifications that would not have been difficult to assess. Descriptors are the parameters that carry or possess information that influences the structure-property relationships we wish to discover. The quantification of how much useful information any descriptor may have can be assessed in many ways and is a foundational question in the field of information theory and statistical inference (35, 44). Having information can guide decision making, and the more precise that information, the more accurate that decision is. The quantification of that information uncertainty can serve as a powerful means of establishing a strategy for sorting data in a way that allows one to discover relationships and hence discover the criteria for that sorting process.

Alloy design rules are an example of such sorting or classification rules embedded in the field of materials science. In the context of the present discussion, despite the long history of such rules, the data explicitly defined in these rules of alloy theory, although useful in providing a retrospective view of linking crystal structure to crystal chemistry, are not easily amenable to being used to predict the design of new materials. To address this problem, Kong et al. (45, 46) applied the metrics of information entropy and more specifically mutual information to quantitatively assess the information impact of each descriptor associated with all the models on the observed phase stability of compounds. Information entropy is a probabilistic measure of the relatedness between data sets. Kong et al. discovered new groupings of binary compounds with the stoichiometry AB_2 by using the concept of information entropy and by classifying a data set (in this case crystal structure data of AB_2 compounds) in such a way as to minimize the entropy (uncertainty) of the data. As a result, sets of if-then design rules were ascertained. For example, if the condition of any attribute is smaller or larger than a specific value, then the instance (i.e., a particular compound) of interest is included in this or that class (i.e., a structure type). A brief summary of the approach is as follows. The information entropy function, H, of a given data set is defined as

$$\begin{aligned} H &= -\sum_{i=i}^{n} p(x_i)\log p(x_i), \\ & p(x_i) \geq 0, \\ & \sum_{i=1}^{n} p(x_i) = 1. \end{aligned}$$

In the above equation, $p(x_i)$ is the probability of the occurrence (on the basis of known observations; see **Figure 5**) of a particular structure type x_i in the crystal structure database, and K is the constant that corresponds to a choice of the unit of measure.

If one begins from the unclassified data set, the compounds are subdivided into two smaller groups at each classification step according to the optimal splitting value. The numeric constraints that make the best bisection at each step of partitioning maximize the reduction of H after each partition step. Thus, the goodness, ΔH, or information gain IG (i.e., the reduction of H), that is, the mutual information (the amount of information one parameter contains about another) that can be achieved by a partition step, is defined as

$$IG(X\,|Y) = \Delta H = -\sum_{i=1}^{v} p(x_i)\log p(x_i) - \sum_{i=1}^{v} p(x_i, y_i)\log p(y_i/x_i),$$

where $p(x_i)$ is the probability of the occurrence of a structure type x_i. The relative contribution of each descriptor to influencing the membership in each class of crystal-structure type can be quantitatively evaluated by calculating the information gain. The entropy change at each step of partitioning is calculated and summed up for both the respective parameters and crystal-structure types to quantitatively measure the contributions of the respective parameters to the partitioning process (**Figure 6**).

The new information uncovered by this process is summarized in **Table 1**, which shows that the classically developed heuristic design rules have been expanded and refined by our informatics techniques.

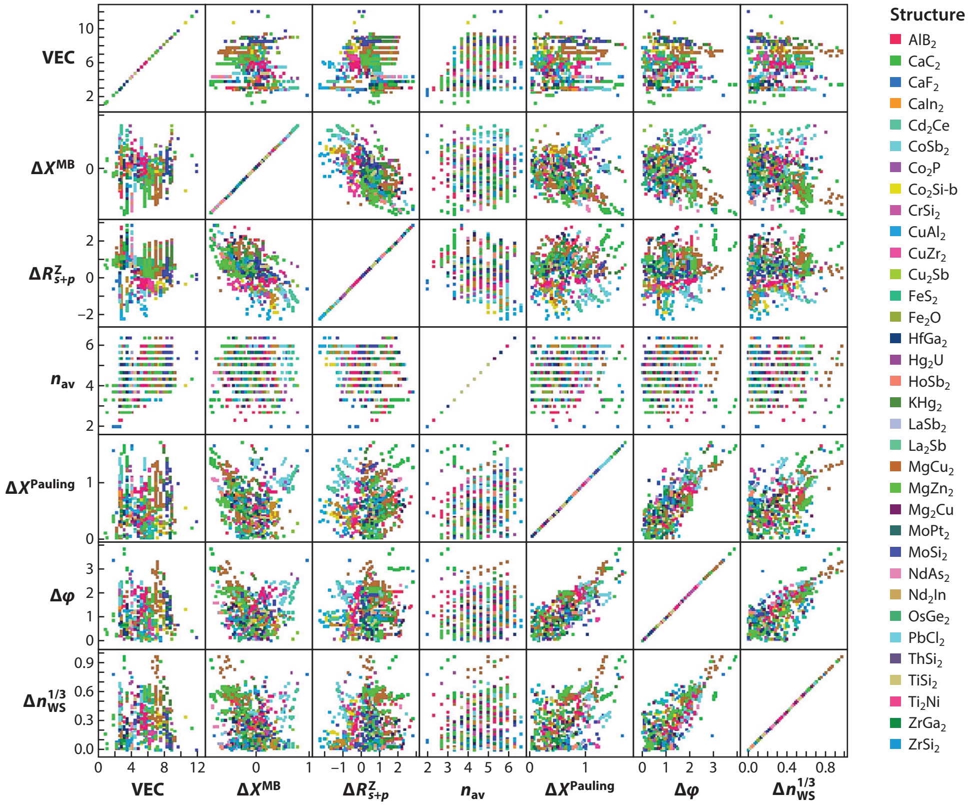


Figure 5

Two-dimensional feature space of AB_2-type inorganic compounds. The coordinates of each plot are atomic and physical parameters governing phase stability on the basis of classical theories of alloys for the compounds originating from their constituent elements. The color-coded points on the plot represent 840 AB_2 compounds that are classified into 34 different crystal structure types. Each subplot indicates a two-dimensional structure-map feature space constructed by different pairwise permutations of classical rules from numerous alloy design theories. Parameter notation is as follows: VEC, average number of valence electrons per atom; ΔX^{MB}, weighted difference of electronegativities (using the Martynov-Batsanov scale); ΔR^{Z}_{s+p}, weighted difference in pseudopotential radii sum (based on Zunger theory); n_{av}, average principal quantum number; $\Delta X^{\mathrm{Pauling}}$, electronegativity difference based on Pauling theory; $\Delta\varphi$, chemical potential difference for electronic charges; $\Delta n^{1/3}_{\mathrm{WS}}$, electron density difference in a Wigner-Seitz atomic cell. Co_2Si-b refers to a specific crystallographic version of cobalt silicide that is modified from the normal crystal structure by having what is known as twin orientation. Reproduced with permission from Reference 45. Copyright 2012, American Chemical Society.

More recently, Kong et al. (45) built on the work described above to go from just the discovery of new or refined classifications of materials to the next step of prediction of new materials chemistries and their associated properties. Once rules and the range over which the rules are valid are identified, this information can be harnessed to guide searches for other new data that may fit the same rules. Using that logic, Kong et al. built on newly discovered chemical design rules for alloys accounting for the uncertainty of information to search for combinations of elements that may meet these design rules and hence suggest the identification of new materials along with the targeted new properties.

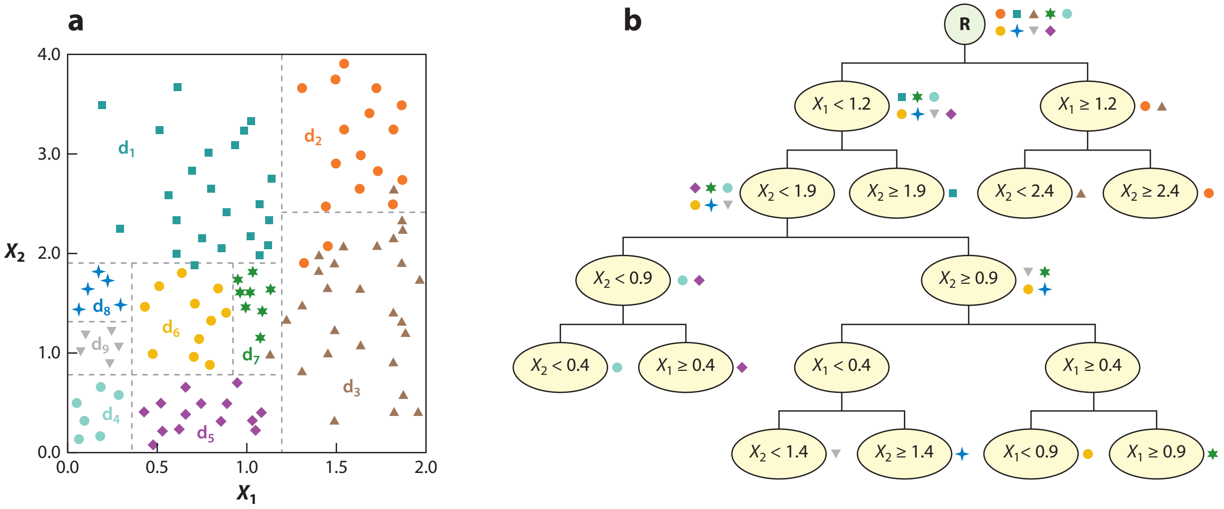


Figure 6

A schematic of the logic of extracting design rules from the partitioning of data. (*a*) Schematic illustration of the partitioning of data into different groupings (each group is identified by a common symbol). The gray dashed lines are best-fit demarking boundaries that group data (denoted by d) in terms of parameters X_1 and X_2. In some cases (e.g., d_1–d_6, d_7–d_3 and d_3–d_2), the boundaries do not provide a perfect separation or classifications. The information shown in panel *a* is captured in terms of a decision tree, whereby each node in the tree identifies the range of X_1 and X_2 where a given data set resides. (*b*) The root node (R) of the classification tree corresponds to the data before the classification, and the respective branches show the criteria for the partitioning of the parameter space into nine different structure domains (d_1–d_9). The partitioning steps can be represented by the corresponding tree structure. This tree structure diagram provides information on the classified groups at each level of the tree along with the splitting criteria for the boundaries between them. Adapted with permission from Reference 46. Copyright 2012, Institute of Physics.

As an example, Kong et al. (45) developed so-called QSPRs (quantitative structure-property relationships) between the charge density and the elastic constants for B_2 intermetallics. Using a combination of informatics techniques for screening all potentially relevant charge density descriptors, they found that elastic constants C_{11} and C_{44} are determined solely from the magnitude of the charge density at its critical points, whereas C_{12} is determined by the shape of the charge density at its critical points. From this reduced charge density selection space, Kong et al. developed simple polynomial expansion–type models for predicting the elastic constants of an expanded

Table 1 New crystal chemistry design rules derived from statistical inference measures. From Reference 46

Phases	Empirically known rules	Newly observed rules
Laves phase (e.g., $MgCu_2$, $MgZn_2$)	Atomic size factor compounds	The valence electron factor and atomic size factor are dominant; the electrochemical factor is less dominant.
AlB_2 structure types	Valence electron factor and atomic size factor	The valence electron factor and atomic size factor are dominant; the electrochemical factor is less significant. However, for KHg_2-type compounds, the electrochemical factor is a competing factor.
Zintl compounds	Electrochemical factor compounds	The electrochemical factor is dominant; the atomic size factor is less significant. For CaF_2 structure types, the valence electron factor is more significant than the electrochemical factor. For $PbCl_2$ types, however, the electrochemical factor is more dominant than the valence electron factor.

number of intermetallic systems. These models were then used to predict the mechanical stability of new systems. The data-driven model based on charge density descriptors disclosed hidden structure-property relationships that could not be identified from crystallographically derived structural information. From these relationships, Kong et al. developed chemical design maps and stability design rules for accelerated chemical selection for targeted elastic behaviors. By using a decision tree framework through following a pathway according to the corresponding descriptor condition, the stability of a new material is estimated, and by using an informatics approach, complicated, apparently hidden relationships between the electron charge density distribution and the mechanical behavior of intermetallic compounds can be identified. The geometry of the charge density at its critical points was used to develop a direct relationship between the charge density and elastic constants as a predictor of the mechanical behavior of intermetallic systems.

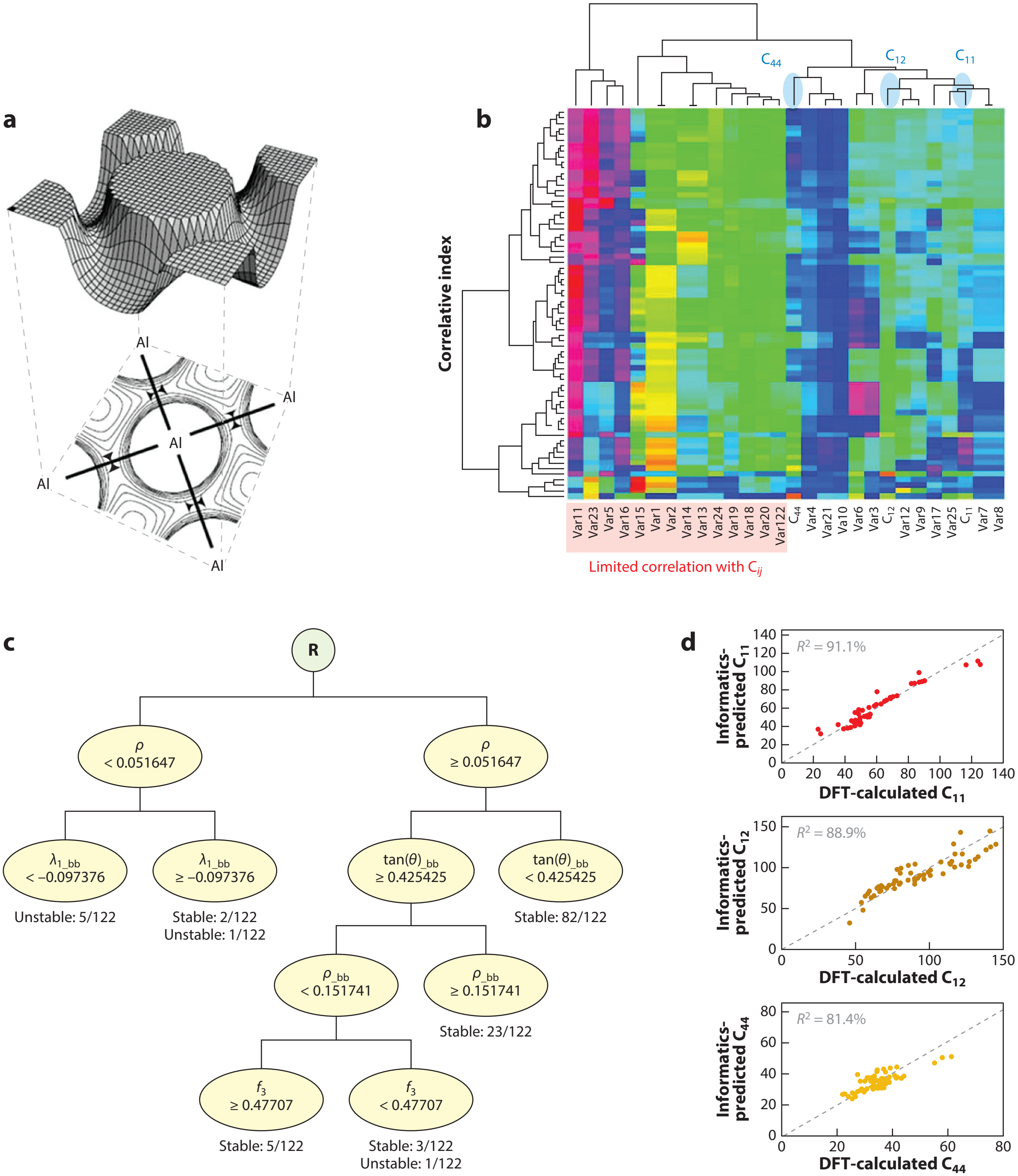
Importantly, in this study Kong et al. (45) embedded statistical inference methods (known as variable-importance projections) into an ensemble of soft computing methods to identify which descriptors primarily impact properties. This approach permitted these researchers to enhance the models' robustness by avoiding overfitting of the data while not excluding any governing physics (**Figure 7**).

Looking for a Robust Gene

In the preceding subsection, we show how recursive partitioning methods driven by an information entropy metric helped to identify, with a relatively small number of compounds, a set of design rules that govern the stability of compounds. Although the concept of mutual information helps us to assess the information impact of descriptors on the final properties of materials, it does not necessarily permit us to explore the dynamics of those interactions. So now we can add another level of complexity into our analysis by considering how the strength of their influence may actually change as we add more data. Agarwal et al. (48) explored this issue by developing a hybrid algorithmic approach that carefully linked a variety of techniques, including evolutionary neural nets and multiobjective genetic algorithms. Of particular interest is the embedding of predator-prey algorithms with genetic algorithmic methods, whereby these researchers explored how the descriptors (genes) can evolve and interact dynamically. The predator-prey approach, as the name suggests, is to explore how stronger, more influential parameters annihilate weaker ones. The final output was represented in terms of a decision tree as before (**Figure 8**) (49). Taking an unsupervised learning approach and interrogating this decision tree, these investigators discovered that new subclassifications of compounds are associated with commonalities or groupings on the basis of differences in stacking fault energy among these different compounds. The clustering in stacking fault energy in turn is manifested by the fact that each subclass of compounds can be associated with a specific type of stacking sequence and atomic packing in their respective crystal structures (see **Figure 8**). In this example, uncovering what we term a robust gene(s) results in the discovery of new subclassifications that were not targeted for discovery.

Figure 7

Data analysis workflows. (*a*) Raw data: a two-dimensional cut from charge density landscape associated with fcc Al. (*b*) Heat map and dendogram showing statistical correlation plots between all the descriptors derived from the computation of the charge density and associated properties. (*c*) Classification tree. The numbers shown at the end of each branch indicate the number of compounds classified in each branch in a manner similar to that in **Figure 6*b***. (*d*) Predictive models based on descriptor development. DFT denotes density functional theory. Adapted with permission from Reference 47. Copyright 2015, Elsevier.


a
Al
Al
Al
Al
Al
b
C44
C12
C11
Correlative index
Var11
Var23
Var5
Var16
Var15
Var1
Var2
Var14
Var13
Var24
Var19
Var18
Var20
Var122
C44
Var4
Var21
Va10
Var6
Var3
C12
Var12
Var9
Var17
Var25
C11
Var7
Var8
Limited correlation with Cij
c
R
ρ < 0.051647
ρ ≥ 0.051647
λ1_bb < −0.097376
λ1_bb ≥ −0.097376
tan(θ)_bb ≥ 0.425425
tan(θ)_bb < 0.425425
Unstable: 5/122
Stable: 2/122
Unstable: 1/122
Stable: 82/122
ρ_bb < 0.151741
ρ_bb ≥ 0.151741
Stable: 23/122
f3 ≥ 0.47707
f3 < 0.47707
Stable: 5/122
Stable: 3/122
Unstable: 1/122
d
R² = 91.1%
Informatics-predicted C11
DFT-calculated C11
0
20
40
60
80
100
120
140
R² = 88.9%
Informatics-predicted C12
DFT-calculated C12
0
50
100
150
R² = 81.4%
Informatics-predicted C44
DFT-calculated C44
0
20
40
60
80

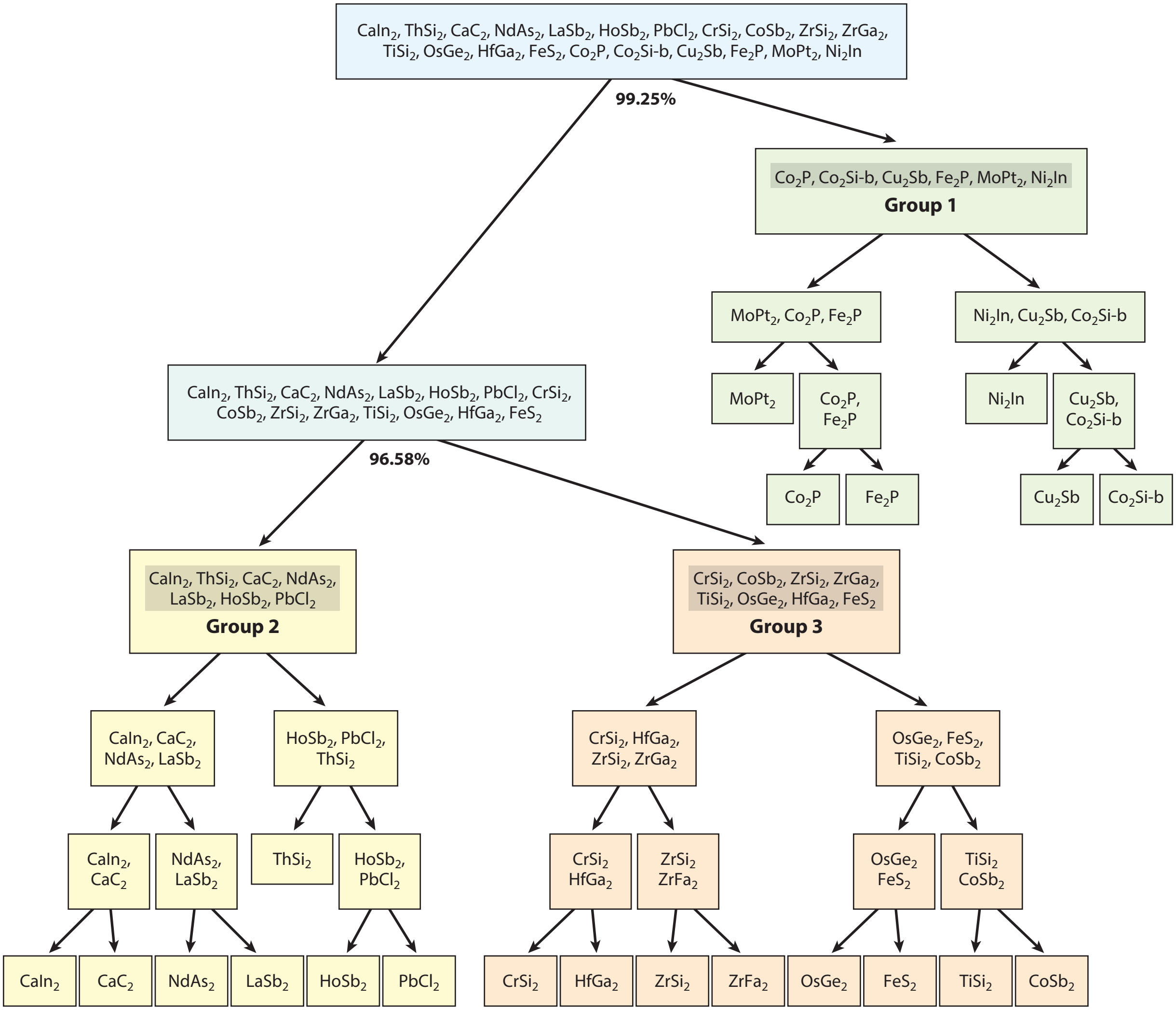

$CaIn_2$, $ThSi_2$, CaC_2, $NdAs_2$, $LaSb_2$, $HoSb_2$, $PbCl_2$, $CrSi_2$, $CoSb_2$, $ZrSi_2$, $ZrGa_2$, $TiSi_2$, $OsGe_2$, $HfGa_2$, FeS_2, Co_2P, Co_2Si-b, Cu_2Sb, Fe_2P, $MoPt_2$, Ni_2In
99.25%
Co_2P, Co_2Si-b, Cu_2Sb, Fe_2P, $MoPt_2$, Ni_2In
Group 1
$MoPt_2$, Co_2P, Fe_2P
Ni_2In, Cu_2Sb, Co_2Si-b
$MoPt_2$
Co_2P, Fe_2P
Ni_2In
Cu_2Sb, Co_2Si-b
Co_2P
Fe_2P
Cu_2Sb
Co_2Si-b
$CaIn_2$, $ThSi_2$, CaC_2, $NdAs_2$, $LaSb_2$, $HoSb_2$, $PbCl_2$, $CrSi_2$, $CoSb_2$, $ZrSi_2$, $ZrGa_2$, $TiSi_2$, $OsGe_2$, $HfGa_2$, FeS_2
96.58%
$CaIn_2$, $ThSi_2$, CaC_2, $NdAs_2$, $LaSb_2$, $HoSb_2$, $PbCl_2$
Group 2
$CrSi_2$, $CoSb_2$, $ZrSi_2$, $ZrGa_2$, $TiSi_2$, $OsGe_2$, $HfGa_2$, FeS_2
Group 3
$CaIn_2$, CaC_2, $NdAs_2$, $LaSb_2$
$HoSb_2$, $PbCl_2$, $ThSi_2$
$CrSi_2$, $HfGa_2$, $ZrSi_2$, $ZrGa_2$
$OsGe_2$, FeS_2, $TiSi_2$, $CoSb_2$
$CaIn_2$, CaC_2
$NdAs_2$, $LaSb_2$
$ThSi_2$
$HoSb_2$, $PbCl_2$
$CrSi_2$ $HfGa_2$
$ZrSi_2$ $ZrFa_2$
$OsGe_2$ FeS_2
$TiSi_2$ $CoSb_2$
$CaIn_2$
CaC_2
$NdAs_2$
$LaSb_2$
$HoSb_2$
$PbCl_2$
$CrSi_2$
$HfGa_2$
$ZrSi_2$
$ZrFa_2$
$OsGe_2$
FeS_2
$TiSi_2$
$CoSb_2$

Figure 8

Discovery of new subclassifications in binary intermetallics (*shaded gray*). Unsupervised learning techniques revealed that by using the hybrid data-mining methods whereby the survival or stability of the descriptors was tracked, the three groupings are associated with distorted packing sequences. This type of classification via data mining has never been shown before and demonstrates the ability to extract subtle but very important structural classifications that otherwise would not be readily seen by simply examining basic space group information. Co_2 Si-b (in group 1) is a compound facilitated by the presence of a class of planar defects known as twins. Significantly, classification groupings detected by machine learning algorithms that explore the robustness of descriptors permit one to discover nuances associated with defects and not just crystal structure chemistry. Data from Reference 46; reprinted with permission from Taylor & Francis.

One can also enhance these classification models to become predictive models by applying the decision rules in our searching of new elemental constituents within our crystal structure. Andriotis et al. (50), for example, not only integrated the methods discussed above with a ranking of how descriptors impact properties but also discussed how that ranking can change with different potential changes in mechanisms governing those properties. By accounting for these uncertainties, they rapidly provided a new and virtual chemical library of binding energies and

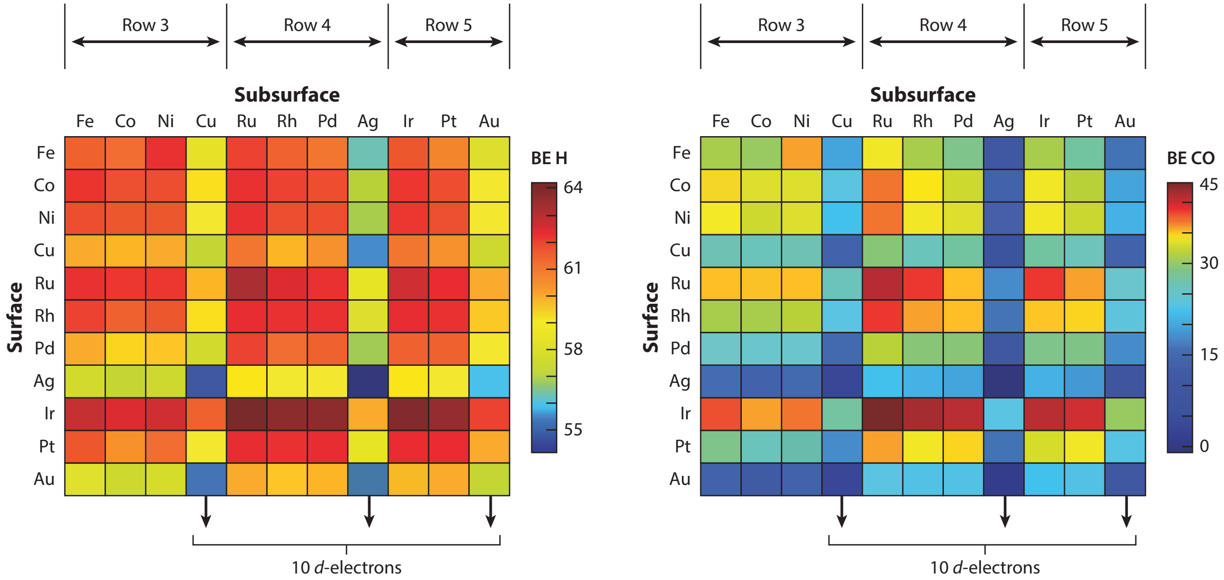


Figure 9

A virtual combinatorial library linking chemistry (binary alloys)–property [binding energy (BE)] relationships with spatial variations associated with atomic coordination (surface and subsurface coordination). BE values for H (*left*) and CO (*right*) from quantitative structure activity relationship (QSAR)-type models for various surface and subsurface chemistries. Such an information-rich combinatorial library would not have been possible solely through traditional experimental combinatorial libraries. From a computational perspective, the use of informatics methods incorporating the statistically derived assessment of the impact of descriptors on properties as discussed in the text provides a promising alternative to large, expensive, high-throughput combinatorial electronic structure calculations. Reproduced with permission from Reference 50. Copyright 2014, American Institute of Physics.

simultaneously identified new correlations between electronic structure and catalytic activity of metal surfaces (**Figure 9**).

Identifying Fuzzy Genes

In this final subsection, we describe approaches that allow us to assess the level of confidence in our prediction as well as to uncover trends in data when we are dealing with the issues of approximation, vagueness, and uncertainty that are inherent in a small database. Ganguly et al. (51), in a study of designing new materials chemistries for scintillator applications, faced a problem common to most materials discovery programs. Specifically, the size of these data sets of known compounds seems like a good starting point for exploration, but the few known materials are often insufficient for typical data mining analyses. From the approximate knowledge contained in the relatively small databases, consideration should be given to issues related to vagueness, uncertainty, approximation, and fuzziness in the data. Hence Ganguly et al. developed a methodology by integrating a number of approaches, including the use of rough sets, fuzzy logic, and genetic algorithms, to determine the range of descriptors for optimizing light-yield new materials chemistries for scintillators. The concept of rough sets is briefly highlighted here, as this approach was explored many years ago in materials science but was not fully exploited (52, 53).

Rough sets is a generalization of a branch of mathematics known as set theory, which grants an element the freedom to possibly belong to a set, thereby addressing the uncertainty in the data. The theory has shown immense usefulness in feature selection and ranking of variables from data

Table 2 Fuzzy structure-property relationships for different descriptors. From Reference 51

Descriptors	Relative importance of light yield	Range for high light yield
Density	0.05	7.0–8.0
Stokes shift	0.33	100–2,500
Valence electron factor	0.20	4.0–5.0
Atomic size factor	0.20	1.75–2.0
Electrochemical factor	0.23	1.25–1.45

or information systems (54–56). Our group has been incorporating rough-set methods to extract general patterns in the data (57, 58); continuous variables need to be discretized into intervals by proper selection of variable values (or cuts) that demarcate the boundary of two consecutive intervals. In other words, if a variable contributes a larger number of cuts or subclassifications to the prediction of a property than does another variable, then that first variable is considered to be more significant in determining that property.

In a study of the design and prediction of new scintillator compounds, Ganguly et al. (51) used output from the rough sets as the input in a sequential fashion into other soft computing methods, including fuzzy logic and genetic algorithms. Both the rough set and fuzzy logic accommodate uncertainty in data, and therefore a description of the uncertainty is included in the input into the genetic algorithm. The input data were then converted into a series of if-then rules for predicting a class of light yield, taking uncertainty into account. The output from the fuzzy logic was subsequently used in the genetic algorithm so that the optimization of light yield was accomplished with uncertainty defined in the input data. **Table 2** shows a typical result in which we are able to not only identify key attributes but assess their relative importance for property estimates.

Ganguly et al. (51) also demonstrated that, by building regression models using these fuzzy metrics, one can uncover new and unexpected subsets of correlations that capture both global and local trends that conventional data analysis would have been unable to find. Thus, soft computing methods can be effectively applied to the consideration of the approximation, vagueness, and uncertainty inherent in a relatively small database based on the rough-set results, as demonstrated in this case for new scintillator compounds (see **Figure 10**).

CONCLUSIONS AND FUTURE ISSUES

A clear theme of this review is the need to harness the concepts of Big Data and to link these data to the identification of key parameters (i.e., the genes associated with structure-property relationships), which, as suggested above, is more than simply large volumes of data. The collective impact of the different characteristics of Big Data—known as the four V's: volume, velocity, variety, and veracity—is what we ultimately need to understand how to harness informatics as a discovery tool in materials informatics.

Just as experimental strategies for high-throughput experimentation have evolved over the last decade, there have been significant advances in generating large data libraries via high-throughput computation. Although this approach has offered some very valuable opportunities to lay out a vast array of potentially new data to explore, there is a concurrent need to identify new structure-property correlations by fully harnessing the power of statistical learning methods to discover knowledge with the existing data, with all its uncertainty and apparent imperfections. The examples discussed in this article show the power of harnessing the imperfections in data to make new and accelerated discoveries. Such harnessing permits one to fully exploit the power of materials

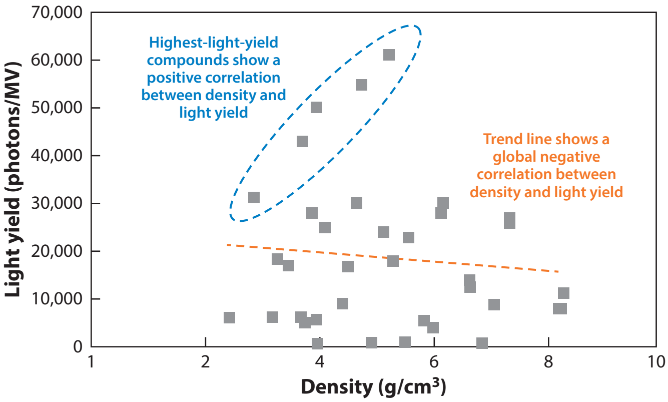


Figure 10

A plot linking variations in light yield with materials density for newly predicted compounds using fuzzy descriptors. The design rules based on the rough-set metrics can capture a wide cross section of property trends. In this example, we show the ability to capture local trends as opposed to global trends, which permits one to identify outliers in terms of data systematics (*blue dashed oval*), which are the highest-light-yield compounds. The rough-set methods are able to identify a very different set of design rules for these compounds whereby the functional relationship between light yield and density is completely different from that for compounds with a low light yield (*orange dashed line*). Adapted from Reference 51 with permission from Taylor & Francis.

informatics not simply to search for information among existing data but also to uncover new knowledge beyond what is apparent within a data library.

SUMMARY POINTS

1. This article highlights the role of ensemble soft computing methods that can overcome the limitations of hard modeling, especially with regard to relatively small data sets.
2. The definition of a materials gene can be formalized through the framework of soft computing methods and Big Data characteristics.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The author gratefully acknowledges the support of the NSF under grants DMR-13-07811, DMS-11-25909, and NSF/DHS grant CMMI-09-389018; support from AFOSR grants FA9550-11-1-0158 and FA9550-12-1-0456; support from ARO grant W911NF-10-1-0397; and support from the Wilkinson Professorship of Interdisciplinary Engineering at Iowa State University.

LITERATURE CITED

1. Cartwright JHE, Mackay AL. 2012. Beyond crystals: the dialectic of materials and information. *Philos. Trans. R. Soc. A* 370:2807–22

2. Mackay AL. 1995. Generalized crystallography. *J. Mol. Crystallogr.* 336:293–303
3. Balachandran P, Broderick SR, Rajan K. 2011. Identifying the "inorganic gene" for high temperature piezoelectric perovskites through statistical learning. *Proc. R. Soc. A* 467:2271–90
4. Rajan K, ed. 2013. *Informatics for Materials Science and Engineering: Data-Driven Discovery for Accelerated Experimentation and Application*. Oxford, UK: Elsevier
5. Datta S, Pratihar DK, Bandyopadhyay PP. 2012. Modeling of input-output relationships for a plasma spray coating process using soft computing tools. *Appl. Soft Comput.* 12:3356–68
6. Farrusseng D, Clerc F, Mirodatos C, Rakotomalala R. 2009. Virtual screening of materials using neuro-genetic approach: concepts and implementation. *Comput. Mater. Sci.* 45:52–59
7. Schoolin J, Brown M, Reed PAS. 1999. An example of the use of neural computing techniques in materials science—the modelling of fatigue thresholds in Ni-base superalloys. *Mater. Sci. Eng. A* 260:222–39
8. Yang X, Deng W, Zou L, Zhao H, Liu J. 2013. Fatigue behaviors prediction method of welded joints based on soft computing methods. *Mater. Sci. Eng. A* 559:574–82
9. Zikapoulous PC, Eaton C, deRoos D, Deutsch T, Lapis G. 2012. *Understanding Big Data—Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw Hill
10. Potyrailo R, Rajan K, Stoewe K, Takeuchi I, Chisholm B, Lam H. 2011. Combinatorial materials libraries: review of state of the art. *ACS Comb. Sci.* 13(6):579–633
11. Rajan K. 2008. Combinatorial materials sciences: experimental strategies for accelerated knowledge discovery. *Annu. Rev. Mater. Res.* 38:299–322
12. Wales DJ. 2005. The energy landscape as a unifying theme in molecular science. *Philos. Trans. R. Soc. A* 363:357–75
13. Jain A, Ping Ong S, Hautier G, Chen W, Richards WD, et al. 2013. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* 1:011002
14. Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O. 2013. The high-throughput highway to computational materials design. *Nat. Mater.* 12(3):191–201
15. Hummelshøj JS, Abild-Pedersen F, Studt F, Bligaard T, Nørskov JK. 2012. CatApp: a web application for surface chemistry and heterogeneous catalysis. *Angew. Chem. Int. Ed.* 51(1):272–74
16. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera RS, et al. 2011. The Harvard Clean Energy Project: large-scale computational screening and design of organic photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* 2:2241–51
17. Castelli IE, Huser F, Pandey M, Li H, Thygesen KS, et al. 2015. New light harvesting materials using accurate and efficient bandgap calculations. *Adv. Energy Mater.* 5:14000915
18. Seko A, Takahashi A, Tanaka I. 2014. Sparse representation for a potential energy surface. *Phys. Rev. B* 90:024101
19. Hansen K, Montavon G, Biegler F, Fazli S, Rupp M, et al. 2013. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* 9:3404–19
20. Liu ZK, Chen LQ, Rajan K. 2006. Linking length scales via materials informatics. *JOM* 58(11):42–50
21. Rajan K. 2014. Nanoinformatics: data-driven materials design for health and environmental needs. In *Nanotechnology Environmental Health and Safety: Risks, Regulation, and Management*, ed. M Hull, D Bowman, pp. 173–98. Waltham, MA: Elsevier. 2nd ed.
22. Zadeh LA. 1994. Fuzzy logic, neural networks and soft computing. *Commun. ACM* 37:77–84
23. Zadeh LA. 1965. Fuzzy sets. *Inf. Control* 8:338–53
24. Zadeh LA. 1994. Soft computing and fuzzy logic. *IEEE Softw.* 11(6):48–56
25. Das SK, Kumar A, Das B, Burnwall AP. 2013. On soft computing techniques in various areas. In *Computer Science and Information Technology* (*Proc. Conf. ACER-13*), ed. R Bhattacharyya, A Kr Bhaumik, pp. 56–98. doi: 10.5121/csit.2013.3206. Krishnagar, India: ACER-13
26. Jang H, Topal E. 2014. A review of soft computing technology applications in several mining problems. *Appl. Soft Comput.* 222014:638–51
27. Yager RR. 1978. Fuzzy decision making including unequal objectives. *Fuzzy Sets Syst.* 1:87–95
28. Hipel KW. 1982. Fuzzy set methodologies in multicriteria modeling. In *Fuzzy Information and Decision Processes*, ed. MM Gupta, E Sanchez, pp. 279–88. New York: North-Holland
29. Wilk T, Wozniak M. 2012. Soft computing methods applied to combination of one-class classifiers. *Neurocomputing* 75:184–93

30. Perner P. 2014. Mining sparse and big data by case based reasoning. *Proc. Comput. Sci.* 35:19–33
31. Saridikas KM, Dentsoras AJ. 2008. Soft computing in engineering design—a review. *Adv. Eng. Inform.* 22:202–21
32. Verdegay JL, Yager RR, Bonissone PP. 2008. On heuristics as a fundamental constituent of soft computing. *Fuzzy Sets Syst.* 159:846–55
33. Murata N, Park H. 2009. Model selection and information criteria. In *Information Theory and Statistical Learning*, ed. F Emmert-Strieb, M Dehmer, pp. 333–54. New York: Springer
34. Oduguwa V, Roy R, Farrugia D. 2007. Development of a soft computing–based framework for engineering design optimisation with quantitative and qualitative search spaces. *Appl. Soft Comput.* 7:166–88
35. MacKay DJC. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge Univ. Press
36. Klir GJ, Wierman MK. 1998. *Uncertainty-Based Information*. Heidelberg, Ger.: Springer
37. Ghosh A, Jain LC, ed. 2005. *Evolutionary Computation in Data Mining*. Berlin: Springer
38. Inuiguichi M, Hirano S, Tsumoto S. 2003. *Rough Set Theory and Granular Computing*. Berlin: Springer
39. Zhang R, Rajan K. 2014. Statistically based assessment of formation enthalpy for intermetallic compounds. *Chem. Phys. Lett.* 12:177–81
40. Bucholz EW, Kong CS, Marchman KR, Sawyer WG, Phillpot SR, et al. 2012. Data-driven model for estimation of friction coefficient via informatics methods. *Tribol. Lett.* 47(2):211–21
41. Kong CS, Rajan K. 2012. Rational design of binary halide scintillators via data mining. *Nucl. Inst. Methods Phys. Res. A* 680:145–54
42. Broderick SR, Rajan K. 2011. Classification of oxide compounds through data mining density of states spectra. *J. Am. Ceram. Soc.* 94(9):2974–80
43. Broderick S, Rajan K. 2011. Data mining Ti–Al semi-empirical parameters for developing reduced order models. *Physica B* 406(11):2055–60
44. Haste T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning, Data Mining, Inference and Prediction*. Berlin: Springer. 2nd ed.
45. Kong CS, Luo W, Arapan S, Villars P, Iwata S, et al. 2012. Information theoretic approach for the discovery of design rules for crystal chemistry. *J. Chem. Inf. Model.* 52:1812–20
46. Kong CS, Villars P, Iwata S, Rajan K. 2012. Mapping the 'materials gene' for binary intermetallic compounds—a visualization schema for crystallographic databases. *J. Comput. Sci. Discov.* 5:015004
47. Kong CS, Broderick SR, Jones TE, Loyola C, Eberhart ME, Rajan K. 2015. Mining for elastic constants of intermetallics from the charge density landscape. *Physica B* 458:1–7
48. Agarwal A, Pettersson F, Singh A, Kong CS, Saxén H, et al. 2009. Identification and optimization of AB_2 phases using principal component analysis, evolutionary neural nets, and multiobjective genetic algorithms. *Mater. Manuf. Process.* 24(3):274–81
49. Pettersson F, Suh C, Saxen H, Rajan K, Chakraborti N. 2009. Analyzing sparse data for nitride spinels using data mining, neural networks and multiobjective genetic algorithms. *Mater. Manuf. Process.* 24:2–9
50. Andriotis AN, Mpourmpakis G, Broderick S, Rajan K, Datta S, et al. 2014. Discovering surface structure–chemistry relationships in catalysts via statistical learning methods. *J. Chem. Phys.* 140:094705
51. Ganguly S, Kong CS, Broderick SR, Rajan K. 2013. Informatics based uncertainty quantification in the design of inorganic scintillators. *Mater. Manuf. Process.* 28:726–32; doi: 10.1080/10426914.2012.736660
52. Jackson AG, Leclair SR, Ohmer MC, Ziarko W, Al-Kamhwi H. 1996. Rough sets applied to material data. *Acta Metall. Mater.* 44:4475–84
53. Jackson AG, Pawlak Z, LeClair SR. 1998. Rough sets applied to the discovery of materials knowledge. *J. Alloys Compd.* 279:14–21
54. Pawlak Z. 1997. Rough set approach to knowledge-based decision support. *Eur. J. Oper. Res.* 99:48–57
55. Pawlak Z, Skowron A. 2007. Rudiments of rough sets. *Inf. Sci.* 177:3–27
56. Walczak B, Massart DL. 1999. Rough sets theory. *Chemom. Intell. Lab. Syst.* 47:1–16
57. Dey P, Bible J, Datta S, Broderick S, Jasinkski J. 2014. Informatics-aided band gap engineering for solar materials. *Comput. Mater. Sci.* 83:185–95
58. Broderick S, Rajan K. 2015. Informatics derived materials databases for multifunctional properties. *Sci. Technol. Adv. Mater.* 16:013501

Annual Review of Biophysics

Volume 44, 2015

Contents

Index

Errata

An online log of corrections to *Annual Review of Biophysics* articles may be found at http://www.annualreviews.org/errata/biophys