1.18

DATA MINING IN MATERIALS DEVELOPMENT

Dane Morgan and Gerbrand Ceder
Massachusetts Institute of Technology, Cambridge MA, USA

1. Introduction

Data Mining (DM) has become a powerful tool in a wide range of areas, from e-commerce, to finance, to bioinformatics, and increasingly, in materials science [1, 2]. Miners think about problems with a somewhat different focus than traditional scientists, and DM techniques offer the possibility of making quantitative predictions in many areas where traditional approaches have had limited success. Scientists generally try to make predictions through constitutive relations, derived mathematically from basic laws of physics, such as the diffusion equation or the ideal gas law. However, in many areas, including materials development, the problems are so complex that constitutive relations either cannot be derived, or are too approximate or intractable for practical quantitative use. The philosophy of a DM approach is to assume that useful constitutive relations exist, and to attempt to derive them primarily from data, rather than from basic laws of physics.

As an example, consider what will likely stand forever as the greatest application of DM in the hard sciences, the periodic table. In 1869 Mendeleev organized the elements based on their properties, without any guiding theory, into the first modern periodic table [3]. With the advent of quantum theory it became possible to predict the structure of the periodic table and DM was no longer strictly necessary, but the results had already been known and used for many years. Even today, the easy organization of data made possible by the classifications in the periodic table make it an everyday tool for research scientists. Mendeleev established a simple ordering based on a relatively small amount of data, and so could do it on paper. However, today's data sets can be many orders of magnitude larger, and an impressive array of computational algorithms have been developed to automate the task of identifying relationships within data.

S. Yip (ed.),
Handbook of Materials Modeling, 395–421.

DM is becoming an increasingly valuable tool in the general area of materials development, and there are good reasons why this area is particularly fruitful for DM applications. There is an enormous range of possible new materials, and it is often difficult to physically model the relationships between constituents, and processing, and final properties. For this reason, materials are primarily still developed by what one might call informed trial-and-error, where researchers are guided by experience and heuristic rules to a somewhat restricted space of constituents and processing conditions, but then try as many combinations as possible to find materials with desired properties. This is essentially human DM, where one's brain, rather than the computer, is being used find correlations, make predictions, and design optimal strategies. Transferring DM tasks from human to computer offers the potential to enhance accuracy, handle more data, and allow wider dissemination of accrued knowledge. Other key drivers for growing DM use in materials development are ease of access to large databases of materials properties, new data being generated in large quantities by high-throughput experiments and quantitative computational models, and improved algorithms, computer speed, and software packages leading to more effective and easy to use DM methods. Note that DM is also used in other areas of materials science beside materials development, e.g., design and manufacturing [4, 5], but this work will not be discussed here.

The interdisciplinary nature of DM creates a special challenge, since a typical materials scientist's education does not provide an introduction to DM techniques, and the computer scientists and statisticians usually involved in developing DM methods are equally unlikely to be versed in materials science. The goal of this paper is to help foster communication between the disciplines and show examples of how they can be joined productively. We introduce DM concepts in a fairly general framework, discuss a few of the more common methods, and describe how DM is being used to tackle some materials development problems, including predicting physiochemical properties of compounds, modeling electrical and mechanical properties, developing more effective catalysts, and predicting crystal structure. The breadth of methods and applications makes a comprehensive discussion impossible, but hopefully this brief introduction will be enough to allow the interested reader to follow up on specific areas of interest.

2. Key Methods of Data Mining

Data Mining (DM) is a vast and rapidly changing topic, with many different techniques appearing in many different fields. Broad reviews of the issues, methods, and applications are given in Refs. [1, 2] and somewhat less comprehensively but more in depth in Refs. [6, 7]. There is some disagreement about exactly what constitutes DM, as opposed to, e.g., knowledge discovery or statistical analysis. We will not worry much about such distinctions, and give DM the rather all encompassing definition of using your data to obtain information. This essentially defines every discovery task as some kind of DM, but there is really a continuum. The more data one has, and the less physical modeling one includes, then the more time one will spend on data management, models, and investigation, and the more DM the task will be. If one has eight data points of force and acceleration, and one performs a linear regression to fit mass, it is silly to consider it DM. There is very little time spent on the data, and one is essentially just fitting an unknown parameter in the known physical law $F = ma$. However, if one is trying to predict what song can be a commercial hit based on a database of song characteristics and sales data, then the primacy of data, and the absence of any guiding theory, make it clearly a DM problem [8].

DM in materials development generally focuses on prediction. Relationships are established between desired dependent properties (e.g., melting temperature or catalytic activity) and independent properties that are easily controlled and measured (e.g., precursor concentrations or annealing temperatures). Once such a relationship is established, dependent properties can be quickly predicted from independent ones, without having to perform costly and time consuming experiments. It is then possible to optimize over a large space of possible independent properties to obtain the desired dependent property. In general, we will define X as the independent properties or variables, Y as the dependent properties or variables, F as the derived relationship between X and Y, and Y_{Pred} as the predicted values of Y based on F and X. The goal of a DM effort is usually to determine F such that Y_{Pred} represents Y as effectively as possible.

There are several key areas that need to be considered in a DM application such as the one described above: data management and preparation, prediction methods, assessment, optimization, and software.

2.1. Data Preparation and Management

Data preparation and management will not be discussed in detail since the issues are very dependent on the specific data being used. However, the tasks associated with cleaning and managing the data can often take up the bulk of a DM project, and should not be underestimated. Data must be stored so that it can be accessed efficiently, interfaced with equipment, updated, etc. Solutions can range from simple flat files to sophisticated database software. Issues often exist with the type and quality of the data, and it is frequently necessary to make significant transformations to bring the data into a universally comparable format, and to regroup data into appropriate new variables. There is sometimes erroneous or just missing data, which may need to be dealt with in some manner before or during the DM process. Finally, data must be adequately comprehensive to be amenable to DM. It may be necessary to obtain further data in key areas, perhaps guided by the DM results in an iterative procedure. These issues are described in many data mining books, e.g., Ref. [7].

2.2. Prediction Methods

Prediction methods form the heart of DM tools relevant for materials development. Although there are many DM approaches that can be used for prediction, here we focus only on three of the most popular, linear regression, neural networks, and classification methods.

Linear regression is often one of the first approaches to try in a DM project, unless one has reasons to expect nonlinear behavior. It is assumed that the relationship F is a linear function, and the unknown parameters are determined by multivariate linear regression to minimize the squared error between Y_{Pred} and Y (these methods are discussed in many textbooks, e.g., Refs. [9, 10]. Linear regression is generally performed by matrix manipulations and is very robust and rapid. There are many variations on strict regression, e.g., adding weights or transforming variables with logarithms. Some of the most useful regression tools are those for reducing the number of independent variables (X), sometimes called dimensional reduction.

It is frequently the case that there are many possible independent variables, but not all of them will be truly independent or important. Furthermore, the original data categories may not be optimal, and linear combinations of the variables, called latent variables, might be more effective. For example, alloy properties affected by strain will depend on the *differences* in atomic sizes, rather than the size of each constituent element separately. It is often difficult to have enough data to properly fit coefficients for a large number of variables (e.g., uniformly gridding a space of n variables with m points for each variable requires n^m data points, which rapidly becomes unmanageable. This is sometime called the "curse of dimensionality" and is a much more significant problem in nonlinear fitting methods, such as the neural networks described below). Having too many variables that are not well constrained can lead to overfitting and poor predictive ability of the function F. Ideally, the DM method will help the user define and include the most effective latent variables for prediction. One common method for defining latent variables is Principal Component Analysis (PCA), which yields latent variables that are orthogonal and ordered by decreasing variance [11]. Assuming that variance correlates well with the importance of the latent variable to the dependent variables, then the principal components are ordered in a sensible fashion and can be truncated at some point. Orthogonality assures that latent variables are independent and will represent different variations. A limitation of this approach is that no information about Y is used in picking the variables. Some improvement can often be obtained by using Partial Least Squares (PLS) regression [9, 12–14], which is similar in spirit to PCA, but constructs orthogonal latent variables that maximize the covariance between X and Y. PLS latent variables capture a lot of the variation of X, but are also well correlated with Y, and so are likely to provide effective predictions.

However one defines the latent variables, it is important to test their effectiveness, and there are a number of methods to identify statistically significant variables in a regression (e.g., ANOVA) [7, 9]. Another popular method is to make use of cross validation, which is discussed below, to exclude variables that are not predictive.

Neural Network (NN) methods [15] are more general than linear approaches and have become a popular prediction tool for many areas. NNs loosely model the functioning of the brain, and consist of a network of neurons that can take inputs, sum them with weights, operate on the sum with a transfer function, and then emit an output. The NN is generally viewed as having layers, the first takes input from outside the NN, and the last outputs the final results to the user, while layers in between are called hidden and communicate only with other layers. For the problems considered here, the NN plays the role of the relationship F between X and Y. The weights of the neurons are unknown and must be determined by training based on known input X and output Y, where the goal is generally to minimize $|Y_{\mathrm{Pred}} - Y|$. The training process is analogous to a linear regression, except that the unknown weights are much more difficult to determine and many different training methods exist. Similar problems occur with excessive numbers of independent variables, and some dimensional reduction, e.g., by PCA, may be necessary.

The strength of NNs is that they are very flexible, and with enough training can in principle represent any function, making them more powerful than linear methods. However, this increased power comes at a price of increased complexity. NNs have many choices that must be made correctly for optimal performance, including the number of layers, the number of neurons in each layer, the type of transfer function for each neuron, and the method of training the neural network. In general, training a NN is orders of magnitude slower than a linear regression, and convergence to the optimal parameters is by no means assured. NNs also have the drawback that it is less obvious how the X and Y variables are related than in a linear regression, making intuitive understanding more challenging.

The problems of inadequate training and overfitting data are quite serious with NN's. Some NN's make use of "Bayesian regularization" [16–19], which includes uncertainty in the NN weights and provides some protection against overfitting. Another common solution is combining predictions from a number of differently trained NN's (prediction by "committee") (this approach is used in, e.g., Refs. [20, 21]). Another interesting approach, which can only be used in cases where one if faced with many similar problems, is to retrain NNs on related problems, making use of the information already gained in their previous training (this is done in, e.g., Ref. [22]).

Classification maps data into predefined classes rather than continuous variables, where the classes are defined based on the dependent properties Y. For example, if Y is conductivity, one could classify materials into metals and insulators, and try to predict to which class a material should belong based on X, rather than performing a full regression of Y on X to predict the continuous conductivity values. Another example is predicting crystal structure, where each different structure type can be considered a class, and the goal is to be able to predict class (assign a structure type) based on the independent data X. In classification DM the relation F maps X onto categories Y_{Pred}, rather than continuous values.

There are a range of different classification methods, as described in most standard textbooks (we found Ref. [6] particularly lucid on these issues). The only classification scheme that will be discussed here is the K-nearest neighbor method, which is one of the simplest. This approach requires that one can define a distance between any two samples, d_{ij} = distance between $X_i$and X_j. Classification for a new X_i is performed by calculating its K nearest neighbors in the existing data set, and then assigning X_i to the class that contains the most items from the K neighbors. The spirit of this approach underlies structure maps for crystal structure prediction, discussed in more detail below. Other classification approaches use Bayesian probabilistic methods, decision trees, NNs, *etc*. but will not be described here [1, 6, 7].

There are some issues with defining a metric of success for classifications. Since Y_{Pred} and Y represent class occupancies, there is not necessarily any way to measure a distance between them. One way to view the results is what is rather wonderfully called a confusion matrix, where matrix element m_{ij} gives the number of times a sample belonging in class $C_i$was assigned to C_j. In order to define a metric for success it is important to realize that when assigning samples to a class there are two parameters that characterize the accuracy, the fraction of samples correctly placed into the class (true positives), and the fraction of samples incorrectly placed into the class (false positives). These can vary independently and their importance can be very dependent on the problem (for example, in classifying blood as safe, it is important to get as many true positives as possible, but absolutely essential not to allow any false positives, since that would allow unsafe blood into the blood supply). Therefore, the metric for success in classification must be chosen with some care.

Note that *clustering*, which is similar to classification, is differentiated by the fact that clustering groups data without the data clusters being predefined. This is sometimes called "unsupervised" learning and will not be discussed further here, but can be found in most DM references.

2.3. Assessment

Cross-validation (CV) [23, 24] is a technique to assess the predictive ability of a fit and reduce the danger of overfitting. In a CV test with N data points, $N - n$ data points are fit and used to predict the n points excluded from the fit. The predicted error of the excluded points is the CV score. This process can be averaged over many possible subsets of the data, which is called "leave n out CV". The key concept behind CV is that the CV score is based on data not used in the fit. For this reason, the CV score will decrease as the model becomes more predictive, but will start to increase if the model under- or overfits the data. This in contrast to predicted errors in data that is included in the fit, which will always decrease with more fitting degrees of freedom.

For example, consider a linear regression on a set of latent variables. The root mean square (RMS) error in the fit data will be a monotonically decreasing function of the number of latent variables used in the regression. However, the CV score will generally decrease for the initial principal components, and then start to increase again as the number of principal components gets large. The initial decrease in the CV score occurs because statistically meaningful variables are being added and the regression model is becoming more accurate. The increasing CV score signals that too many variables are being used, the regression is fitting noise, and that the model is overfit. By minimizing the CV score it is therefore possible to select an optimal set of latent variables for prediction. This idea is illustrated schematically in Fig. 1.

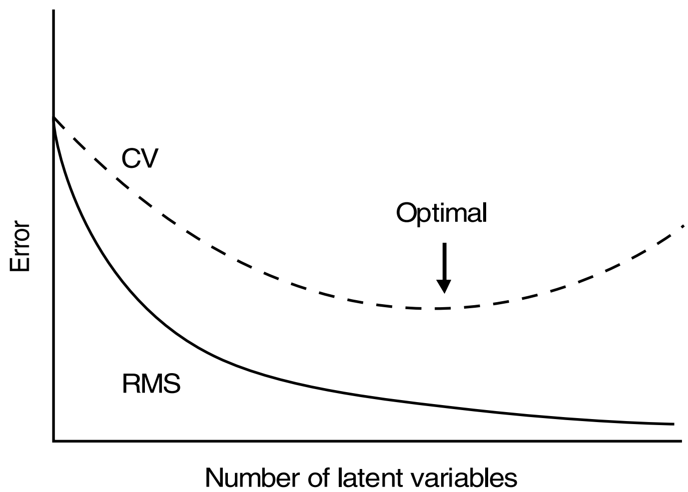

Figure 1. A schematic comparison of the error calculated with data included in the fit (normal RMS fitting error – solid line) and excluded from the fit (CV score – dashed line).

Test data is another important assessment tool, and simply refers to a set of data that is excluded from working data at the beginning of the project and then used to validate the model at the end of model building. To some extent, the CV method does this already, but in the common case where the model is altered to optimize the CV score, it will overestimate the true predictive accuracy of the model [23]. It is only by testing on an entirely new data set, which the model has not previously encountered, that a reliable estimate of the predictive capacity of the model can be established. Sometimes there is not enough data to create an effective test data set, but it is certainly advisable to do so if at all possible.

2.3.1. Optimization

Optimization methods [25, 26] are not usually considered DM, but they are an essential tool of many DM projects. For example, once a predictive model has been established, one frequently wants to optimize the inputs to give a desired output. This usually cannot be done with local optimization schemes (e.g., conjugate gradient methods) due to a rough optimization surface with many local minima. It is therefore frequently necessary to use an optimization method capable of finding at least close to the global minimum in a landscape with many local minima. A detailed discussion of these methods is beyond the scope of this article, but common approaches include simulated annealing Monte Carlo, genetic algorithms, and branch and bound strategies. Genetic algorithms seem to be the most popular in the DM applications discussed here, and work by "evolving" toward an optimal sample population through operations such as mixing, changing, and removing samples.

2.3.2. Software

Many DM algorithms are fairly simple, and can be programmed relatively quickly. Often the underlying numerical operations involve no more than standard matrix operations, and access to widely available basic linear algebra subroutines (BLAS) is adequate. However, DM is generally very explorative, and it is common to try many different approaches. Coding everything from scratch becomes prohibitive, and will lock the user into the few things they can readily implement. Fortunately, there are a large number of both free and commercial DM tools available for users. Some tools, like the Neural Net Toolbox in Matlab, are implemented in languages likely to be familiar to the materials scientist, and are readily accessible. An impressive list of possible tools is given in Appendix A of Refs. [6, 7]. It should also be remembered that for the academic user many companies will have special rates, so it is worth exploring commercial software.

3. Applications

There are far too many studies using DM methods to offer a comprehensive revue. Therefore, we focus on a few key areas where DM techniques are highlighted and seem to be playing an increasingly important role.

3.1. Quantitative Structure–Property Relationships (QSPR)

Quantitative Structure–Property Relationships (QSPR), and the closely related techniques of Quantitative Structure–Activity Relationships (QSAR), are based on the fundamental tenet that many molecular properties, from boiling point to biological activity, can be derived from basic descriptors of molecular structure. For some examples, see the general review of using NNs to predict physiochemical properties in Ref. [27] QSPR/QSAR are generally considered methods of chemistry, but are closely related to the activities of a DM material scientist.

QSPR/QSAR is a large field and here we consider only one particularly illustrative example, the work of Chalk *et al.*, predicting boiling points for molecules [20]. The boiling point for any given compound is not a particularly hard measurement, but the ability to quickly predict boiling points for many compounds, particularly ones that only exist as computer models, can be useful for screening in, e.g., drug design. Computing the boiling point of a compound directly from physical principles requires a very accurate model of the energetics and significant computation. Therefore, researchers have generally turned to DM applications in this area.

Chalk *et al.* have a database of 6629 molecular structures and boiling points. The dependent variables Y are taken as the boiling points. A set of descriptors, X_0, are developed based on structural and electronic characteristics (derived from semiempirical atomistic models). A technique called formal inference-based recursive modeling (FIRM) is then used to asses the relevance of each variable (this technique will not be described here but allows the influence of a variable to be tested). A set of 18 descriptors are settled on as likely to be significant and they are used for the independent variables X. A test data set of 629 molecules that span the whole range of boiling temperatures is removed. The remaining 6000 molecules are then used to find the optimal model function F to map X to Y.

F is represented by a NN, and after some initial testing one is chosen with 18 first layer nodes, 10 nodes in the hidden second layer, and a single node in the third layer. The transfer functions are all sigmoids ($\text{sig}(x) = 1/(1 + \exp(-x))$) and trained with a back-propagation algorithm. In order to control for overfitting the data is broken up into 10 disjoint subsets and a "leave

600 out" cross validation is performed. This trains 10 distinct NNs on 5400 molecules each. The NN training is stopped when the CV score reaches a minimum. The prediction function F is taken to be a committee, and uses the mean result of the values predicted by all 10 NNs. The final test for F is done by comparing the predicted and true boiling points for the 629 molecule test set, giving errors with a standard deviation of only 19 K (the predicted vs. true melting temperatures for the test set are shown in Fig. 2). The predictive capacity is good enough that for many of the largest prediction errors it was possible to go back to the experimental data and show that the input data itself was in error. One could now imagine using a genetic algorithm and the predicting function F to search the space of molecular structures to find, e.g., a very high melting temperature molecule, although no such work was performed by the authors.

It is worth noting that computation plays an important role in providing the basic input data in the study. All of the structural and electrostatic descriptors were generated by semi-empirical atomistic models. Using computational methods can be an efficient way to generate large amounts of descriptor information, greatly reducing the amount of experimental work required.

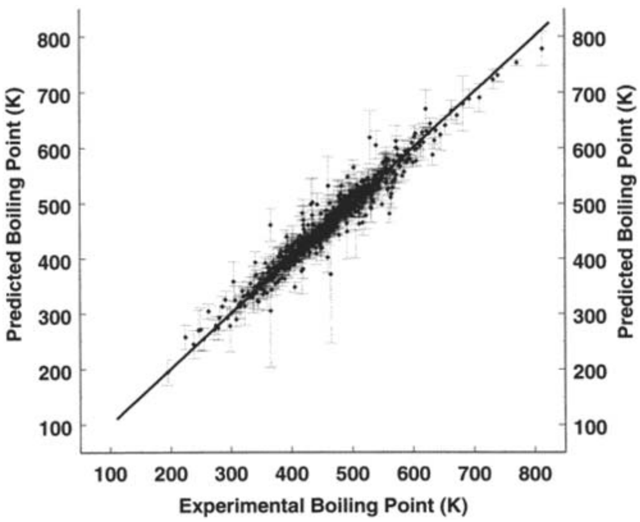

Figure 2. Predicted vs. true boiling points for 629 compounds. Prediction is done by neural networks fit to 6000 boiling points that did not include the 629 shown here. (After [20], reproduced with permission).

3.2. Processing–Structure–Property Relationships

Processing–Structure–Property (PSP) relationships refer to the challenging materials problem of connecting the processing parameters of a material to its structure and properties. Processing conditions might include such things as initial composition of reactants and annealing schedule, while structural aspects might be crystal structure or grain size, and final properties are such characteristics as yield stress and corrosion resistance. PSP relationships are very important because they allow processing parameters to be adjusted to create optimal materials. PSP relationships tend to involve many different phenomena, with widely varying length and time scales, making direct modeling extremely challenging. However, analogous to QSPR's reliance on the fact that properties must be a function of the structure of the molecules involved, in PSP relationships we know that properties must follow from structure in some manner, and that structure is somehow determined by processing. The assurance that PSP relationships exist, combined with the challenge of directly modeling them, makes this a good area for DM applications.

One of the most active groups in this area has been Bhadeshia and co-workers. Bhadeshia's review in 1999 [21] covers a lot of the material's work that had been done up to that time in neural network (NN) modeling, and he and co-workers have continued to apply NN techniques in PSP applications to such areas as creep modeling [28, 29], mechanical weld properties [30, 31], and phase fractions in steel [32]. In general, these studies follow the DM framework used in QSPR above. Many of the data and codes used by Bhadeshia *et al.*, as well as many others, can be found online as part of the Materials Algorithm Project [33].

Malinov and co-workers have also done extensive work with DM tools in PSP relationships, and have developed a code suite, complete with graphical user interface, to make use of their models [34]. Their work has focused primarily on Ti alloys [35–37] and nitrocarburized steels [38, 39]. The NN software they developed uses a cross validation (CV)-like strategy to assess the effectiveness of different NN architectures, training methods, and trainings, so that the best network can be obtained by optimization, rather than intuitive choice. It is a general trend in DM applications to try to automatically optimize as many choices as possible, since this gives the best results with the least user intervention. Many apparent DM choices, such as which latent variables or NN architectures to use, can in fact be determined by performing a large number of tests. Implementing this type of automation is generally limited by the user's willingness to code the required tests, the time it takes to perform the optimization, and the amount of data required for sufficient testing. Also, one should ideally have a test set that is entirely excluded from all the optimization processes for final testing.

A particularly interesting application by Malinov *et al.* is the prediction of time–temperature-transformation (TTT) diagrams for Ti alloys [34, 35, 37]. TTT diagrams give the time to reach a specified fraction of phase transformation at each temperature, and for a given phase fraction they are a curve in time–temperature space. They can be modeled to some extent directly with Johnson–Mehl–Avrami theory, but Malinov *et al.* chose to use a NN model so as to be able to predict for many systems and composition variations. The details discussed here are all from Ref. [35]. The data set was 189 TTT diagrams for Ti alloys, and the independent variables were taken to be the compositions of the 8 most common alloying elements and oxygen. Some additional elements that were not prevalent enough in the data set for accurate treatment had to be removed or mapped onto a Mo equivalent. It should be noted that the authors are careful to identify the ranges of the concentrations of alloying elements present in the test set. This is very important, since given the limited data, it is not clear that this NN would give accurate predictions outside the concentration ranges used in training.

The dependent variables represented more of a problem, since TTT diagrams are curves, not single values. Malinov *et al.* solved this problem by representing the TTT diagram as a 23-tuple. Two entries gave the position of the TTT graph nose, its time and temperature. Ten entries gave the upper portion of the curve, where each entry was the fractional change in time for a fixed change in temperature, and ten more the lower portion. Finally, one entry was reserved for the martensite start temperature. These considerations, for both the independent and dependent variables, demonstrate some of the data processing that can be required for successful DM. The final predictions are quite accurate for test sets, and allowed exploration of the dependence of TTT curves on alloy composition. A number of TTT diagram predictions for (at that time) unmeasured materials were given, and some of these have since been measured, demonstrating reasonably good predictive ability for the NN model (see Fig. 3) [37].

A set of studies using DM techniques to model Al alloys recently came out of Southampton University [40–44]. The work by Starink *et al.* [44] summarizes studies on strength, electrical conductivity, and toughness. These studies are particularly interesting since they directly compare different DM methods as well as more physically based modeling, based on known constitutive relations. Starink *et al.* make use of linear regression and Bayesian NN models like those discussed above, but also apply neurofuzzy methods and support vector machines. We will not discuss these further except to point out that the latter is a relatively new development that seems to have some improved ability to give accurate predictions over the more common NN methods, and will likely grow in importance [45–47]. For the cases of direct comparison, Starink *et al.* find that physically based modeling performs slightly better. However, these examples involve very small data sets (around 30 samples),

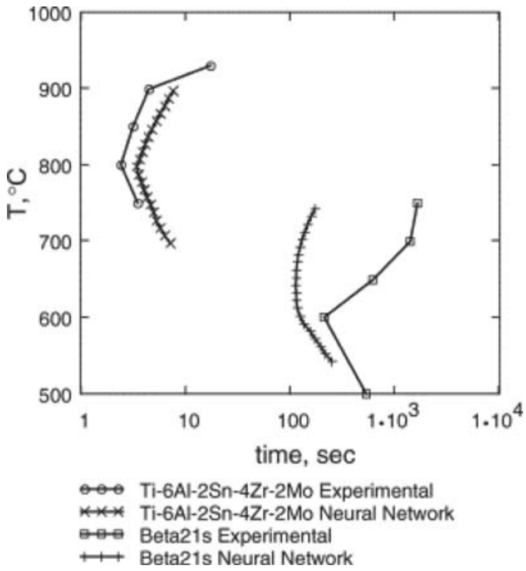

Figure 3. Comparison of predicted and measured TTT diagrams for different Ti alloys. These predictions were made and published before the experimental measurements were taken. (After Ref. [37], reproduced with permission.)

so one expects there to be significant undertraining in DM methods. Also of interest is the over three-fold decrease in predictive error for conductivity when going from linear to nonlinear DM methods, demonstrating why nonlinear NN methods have become the dominant tool for many applications.

Starink *et al*. make some use of the concept of hybrid physical and DM approaches. This is a very natural idea, but worth mentioning explicitly. The spirit of DM is often one of using as little physical knowledge as possible, and allowing the data to guide the results. However, by introducing a certain amount of physical knowledge, a DM effort can be greatly improved. As summarized by Starink *et al*., this can be done through initially choosing independent variables based on known physics, using functional forms that are physically motivated in the DM, and using DM to fit remaining errors after a physical model has been used.

3.3. Catalysis

A particularly exciting area of DM applications at present is in catalysis. A lot of recent activity in this field has been driven by the advent of high-throughput experiments, where the ability to rapidly create large data sets has created a new need for data mining concepts to interpret and guide experiment. Some reviews in this area can be found in Refs. [48–50].

Some authors have taken approaches similar to those used in QSPR/QSAR applications and the PSP modeling described above – finding a NN model to connect the properties of interest to tractable descriptors, and then exploring that model to understand dependencies or optimize properties [22, 50–56]. The input independent variables are generally the compositions of possible alloying materials in the catalyst, and the output is some measure of the catalytic activity. Note that it is quite possible to have multiple final nodes in the network to output multiple measures of interest, such as conversion of the reactants and percentages of different products [51, 52]. It is also possible to look at catalytic behavior for a fixed catalyst under different reactor conditions, where the reactor conditions become the independent variables [22]. Once a NN has been trained, the best catalyst can be found through optimization of the function defined by the NN. This is generally done with a genetic algorithm [51, 54, 56], but other methods have also been explored [55].

Baerns *et al*. have done influential work in using a genetic algorithm to design new catalysts, but have skipped the step of fitting a model altogether, directly running experiments on each new generation of catalysts suggested by the genetic algorithm [57–59]. For example, Baerns *et al*. studied oxidative dehydrogenation of propane to propene using metal oxide catalysts with up to eight metal constituents, and found a general trend toward better catalytic activity with each generation, as shown in Fig. 4. Although optimizing the direct experimental data limits the number of samples that can be examined (Baerns *et al*. generally look at only a few hundred) the results have been very encouraging, e.g., leading to an effective multicomponent catalyst for low-temperature oxidation of low-concentration propane [58]. Further success

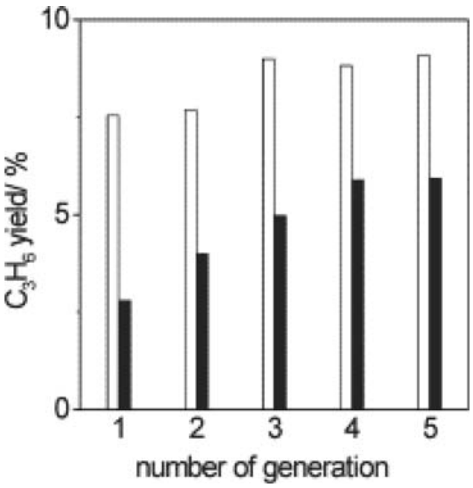

Figure 4. The best (open bar) and mean (solid bar) yield of propene at each generation of catalysts created by genetic algorithm. (After [57], reproduced with permission.)

was obtained in studying oxidative dehydrogenation of propane to propene by following up on materials suggested by the combinatorial genetic algorithm search with further noncombinatorial "fundamental" studies [57].

Baerns *et al*.'s work demonstrates that the best results are sometimes obtained by combining DM and more traditional approaches. Further improvements in high-throughput methods will make direct iterative optimization of the experiments increasingly effective, but a fitted model will likely always be able to explore more samples and provide more rigorous optimization. The choice to use a fitted model is then a balance between the advantage of being able to optimize more accurately and the disadvantage of having a less accurate function to optimize. Umegaki *et al*. suggest that, in direct comparisons, a combined NN and genetic algorithm approach is more effective than direct optimization of experimental results, but this is a complex issue and will be problem dependent [56].

Despite many encouraging successes, DM in catalysis still faces a number of challenges. As pointed out by Hutchings and Scurrell [49] extending the independent variables to include more preparation and processing variables might significantly broaden the search for optimal materials. In addition, issues related to lifetime, stability, and other aspects of long-term performance are often difficult to predict and need to be addressed. Finally, Klanner *et al*. point out that there are different challenges for optimizing a library over a well known space of possible compositions and designing a discovery program for development in areas where there is essentially no precedent [50]. In the case of development of truly new materials, the problem of using a QSPR/QSAR approach in catalysis design is complicated because of the inherent difficulties of characterizing heterogeneous solids to build diverse initial libraries. Structure is a good metric for measuring diversity of molecular behavior, and therefore allows relatively easy assembly of diverse libraries for exploration. However, the very nonlinear behavior of solid catalysts, where activity is often dependent on such subtle details as surface defects, means that at this point there is no metric for measuring, *a priori*, the diversity of solid catalysts. Klanner *et al*. therefore suggest that development work will have to take place through building a large initial set of descriptors, based on synthesis data and properties of the constituent elements, and then use dimensional reduction to get a manageable number. Finally, no effort has been made here to make comparisons of DM to direct kinetic equation modeling in catalysis design. Some comments with regards to theses methods, and how they can be integrated with DM approaches, are given in Ref. [60].

It should be noted that the above issue of assembling diverse libraries, along with using genetic algorithms for intelligent searching, can be viewed as parts of the general problem of optimized experimental design. This is not a new area, but has become increasingly important due to the advent of high-throughput methods. It also encompasses such well developed fields as statistical Design of Experiments. This is a fruitful area for statistical and DM methods, and many of the relevant issues have already been mentioned, but we will not discuss it further here. The interested reader can consult the review by Harmon and references therein [48]. Another DM area that has been receiving increased attention due to high-throughput experiments is correlating the results of cheap and fast experimental measurements with properties of interest. This becomes particularly important when it is necessary to characterize large numbers of samples quickly, and careful measurement of the desired properties is not practical. For a discussion of this issue in high-throughput polymer research see Refs. [61, 62] and a number of rapid screening tools and detection schemes used in high-throughput catalysis development are described in Ref. [63].

3.4. Crystal Structure

The prediction of crystal structure is a classic materials problem that has been an area of ongoing research for many years. Now that modeling efforts have made computational materials design a real possibility in many areas, the problem of predicting crystal structure has become more practically pressing, since it is usually a prerequisitie for any extensive materials modeling. Crystal structure prediction is an area well suited for DM efforts, since there is no generally reliable and tractable method to predict structure, and there is a lot of structural data collected in crystallographic databases (e.g., ICSD [64], Pauling files [65], CRYSTMET [66], ICDD [67]).

Some of the most successful methods for crystal structure prediction are what are known as *structure maps*, reviewed at length in Refs. [68, 69]. Structure maps exist primarily for binary and ternary compounds, and the best known examples are probably the Pettifor maps [70]. To understand how Pettifor maps work, consider the map designed for AB binary alloys. Each possible element is assigned a number, called the Mendeleev number. Then each alloy AB can be plotted on a Cartesian axis by assigning it the position (x, y), where x is the Mendeleev number for element A and y is the Mendeleev number for element B. At position (x, y) one places a symbol representing the structure type for alloy AB. When enough data is plotted the like symbols tend to cluster – in other words, alloys with the same structure type tend to be located near each other on the map. This can be clearly seen in the Pettifor map in Fig. 5. The probable structure type for a new alloy can simply be found by locating where the new alloy should reside in the map and examining the nearby structure types.

Structure maps were not originally introduced as an example of DM, but can be understood within that framework. One can extend the idea of using Mendeleev number to a general "vector map," which maps each alloy to a

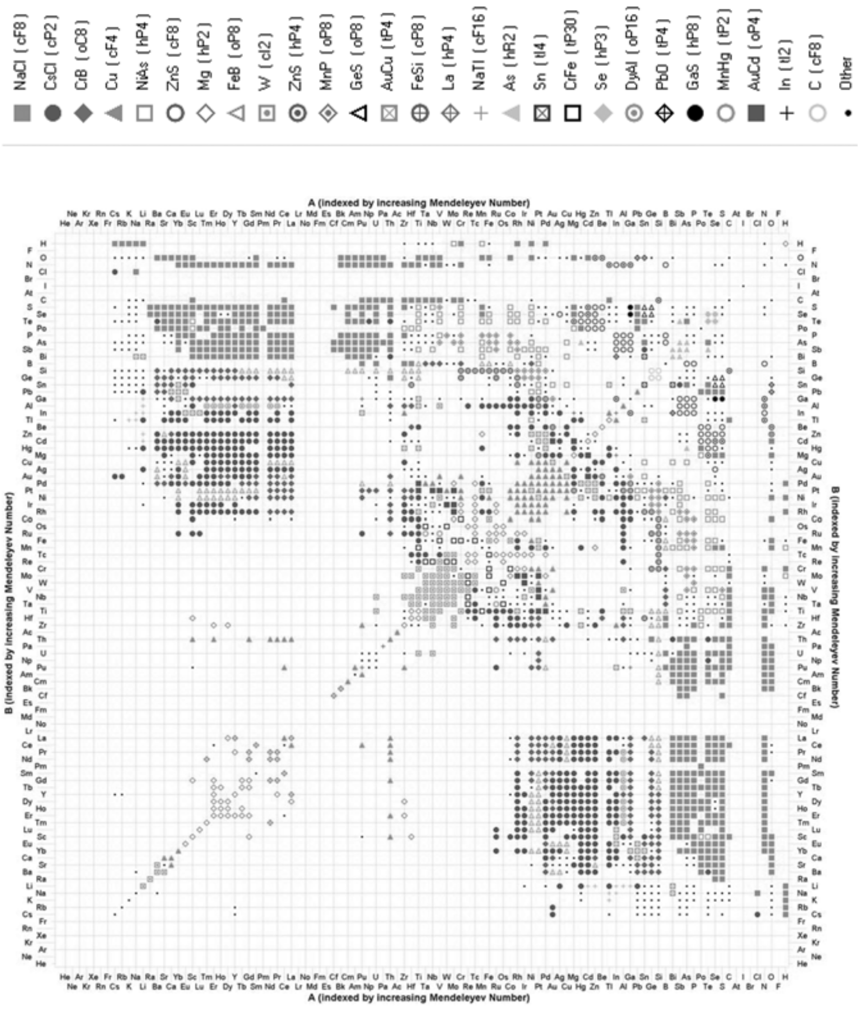

Figure 5. An AB binary alloy Pettifor map. Notice that like structure types show a clear tendency to cluster near one another. Provided by John Rodgers using the CRYSTMET database [66].

multicomponent vector. The vector components might be any set of descriptors for the alloy, such as Mendeleev numbers, melting temperatures, or differences in electronegativities. Once the alloys have been mapped to representative vectors they are amenable to different DM schemes. Since crystal structures are discrete categories, not continuous values, some sort of classification DM is going to be required.

Structure maps work by defining a simple Euclidean metric on the alloy vectors and making the assumption that alloys with the same structure types will be close together. When a new alloy is encountered its crystal structure is predicted by examining the neighborhood of the new alloy in the structure map. Structure types that appear frequently in a small neighborhood of the new alloy are good candidates for the alloy's structure type. This is a geometric classification scheme, along the lines of K-nearest-neighbors described above. There is no unique way to define the vectors that create the structure map, and many different physical quantities, such as electronegativities and effective radii, have been proposed for constructing structure maps. Ref. [64] lists at least 53 different atomic parameters that could be used as descriptors to define a structure map. The most accurate Pettifor maps are built by mapping alloys to vectors using a specially devised chemical scale [71]. The chemical scale was motivated by many physical concerns, but is fundamentally an empirical way to map alloys to vectors, chosen to optimize the clustering of alloys with the same crystal structures.

A number of new ideas are suggested by viewing crystal structure prediction from a DM framework. First, it is clear that many standard assessment techniques have only recently begun to be incorporated. It was not until about 20 years after the first Pettifor maps that an effort was made to formalize their clustering algorithm and assess their accuracy using cross validation techniques (the accuracy was found to very good, in some cases giving correct predictions for non-unique structures 95% of the time) [72]. Also, the question of how to assess errors can be fruitfully thought of in terms of false positives (predicting a crystal structure that is wrong) and false negatives (failing to predict the crystal structure that is right). For many situations, e.g., predicting structures to be checked by *ab initio* methods or used as input for Rietveld refinements, a false positive is not a large problem, since the error will likely be discarded at a later stage, but a false negative is critical, since it means the correct answer will not be found with further investigation. This leads to the idea of using maps to suggest a candidate structure list, rather than a single candidate structure [72]. Using a list creates many false positives, but greatly reduces the chance of false negatives.

A DM perspective on structure prediction encourages one to think of moving beyond present structure map methods. For example, different metrics, other classification algorithms, or mining on more complex alloy descriptors, might yield more accurate results. Some work along these lines has already occurred, including machine learning based structure maps [73] and NN and clustering predictions of compound formation [74]. A similarly spirited application used partial least squares to predict higher level structural features of zeolites in terms of simpler structural descriptors [75], and is part of a more general center focused on DM in materials science [76].

The structure maps have at least two severe limitations. As described above, they predict structure type given that the alloy has a structure at a given stoichiometry, but do not consider the question of whether or not an alloy will have an ordered phase at that stoichiometry. This is not a problem when a structure is known to exist and one wants to identify it, but in many cases that information is not available. There are some successful methods for identifying alloys as compound forming versus having no-compounds, e.g., Meidema's rules [77] or Villar's maps for ternary compounds [68], but the problem of identifying when an alloy will show ordering at a given composition has not been thoroughly investigated in the context of structure maps. However, it is certainly possible that further DM work could be of value solving this problem, and some potentially useful methods are discussed below.

Another serious limitation on structure maps is that classification DM is only effective when an adequate number of samples of each class are available. There are already thousands of structure types, the number is still increasing, and only a small percentage of possible multicomponent alloy systems have been explored [68]. Therefore, it seems unlikely that sufficiently many examples of all the structure type classes will ever be available for totally general application of structure maps. Infrequent structure types are less robustly predicted with structure maps, and totally new structure types cannot be predicted at all. The problem of limited sampling can be alleviated by restricting the area of focus, e.g., considering only the most common structure types, which are likely to be well sampled, or only a subset of alloys, where all the relevant structure types can be discovered. However, the very significant challenge of sampling all the relevant structure types creates a need for other methods.

One promising idea is to abandon the use of structure types as the most effective way to classify structures and replace it with a scheme easier to sample. An idea along these lines is to classify alloys by the local environments around each atom [68, 78]. Local environments may in fact be a more relevant method of classification than structure type for understanding physical properties, and there seem to be far fewer local environments than different structure types. This is analogous to classifying proteins by their different folds, which are essential to function and come in limited variety [79].

Computational methods, using different Hamiltonians, offer an increasingly practical route toward crystal structure prediction. Given an accurate Hamiltonian for an alloy, the stable crystal structures can be calculated by minimizing the total energy. These techniques can also predict entirely new structures never seen experimentally, since the prediction is done on the computer. Unfortunately, the structural energy landscape has many local minima, and it cannot be explored quickly or easily. Researchers in this area therefore are forced to make a tradeoff between the speed and accuracy of the energy methods, and the range of possible structures that are explored. For example, Jansen has used simple pair potentials to explore the energy landscape, and then applied more accurate *ab initio* methods for likely structural candidates [80]. This is a common approach, to optimize with simplified expressions and then use slower and more accurate *ab initio* energy methods on only the more promising areas. A similar approach was taken to predict a range of inorganic structures from a genetic algorithm [81]. If one restricts the possible structures, then direct optimization of *ab initio* energies can be performed. For example, low cohesive energy structures for 32 possible alloying elements were found on a four atom, face centered cubic unit cell by optimizing *ab initio* energies using a genetic algorithm [82]. Although these approaches are quite promising, optimizing the energy over the space of all possible atomic arrangements is generally not practical. It is necessary to find some approach to guide the calculations to regions of structure space that are likely to have the lowest energy structures and can be explored effectively.

A practical and common method to guide calculations is sometimes colloquially referred as the “round up the usual suspects” approach, borrowing a quotation from Captain Louis Renault in the end of Casablanca. This approach simply involves calculating structures one thinks are likely to be ground states and is another example of human DM, where the scientist is drawing on their own experience to guide the calculations toward the correct structure. As mentioned in the introduction, formalizing human DM on the computer offers many advantages in accuracy, verification, portability, and efficiency. An improvement can be made by limiting the human component to suggesting a few likely parent lattices, and then fitting simplified Hamiltonians on each parent lattice to predict stable structures. This approach, called cluster expansion, has been well validated in many systems [83, 84] and has been successful in predicting some structures that had not been previously identified experimentally [85, 86]. However, choosing the correct parent lattice and performing the fitting required for cluster expansion is at present still difficult to automate, although efforts along these lines are being made [87].

Ideally, the process of guiding computational crystal structure prediction would be entirely automated by DM methods. A step in this direction has been taken by Curtarolo *et al.* who have demonstrated how one might combine experimental data, high-throughput computation, and DM methods to guide new calculations toward likely stable crystal structures [88]. Experimental information is used to get a list of commonly occurring structure types, and then these are calculated using automated scripts for a large number of systems. Mined correlations between structural energies are then used to guide calculations on new systems toward stable regions, reducing the number of calculations required to predict crystal structures. This approach can, in theory, be expanded to totally new structure types, since these can be generated on the computer, and work in this direction is under development.

4. Conclusions

We have seen here a number of different examples of DM applications in different areas, and it is valuable to step back and note some overall features. In general, DM applications in materials development still need to prove themselves, and relatively few new discoveries have been made using them. Many of the results in this field consist primarily of exploring new models to demonstrate that such modeling is possible, that accurate predictions can be made, and that useful understanding of dependencies on key variables can be obtained. This will inevitably cause some skepticism about the final utility of the methods, but it is appropriate for a field which is still relatively young and finding its place. A similar evolution has been taken by, e.g., *ab initio* quantum mechanical techniques. It is only recently that these methods have moved out the stage where the accuracy of the model was the key issue to the stage where the bulk of papers focus on the materials results, not the techniques. All the drivers for using DM methods identified in the introduction, more data, databases, and DM tools, will only become increasingly forceful with continuing advances in experiment, computation, algorithms, and information technology. For these reasons, we believe that DM approaches are going to be increasingly important tools for the modern materials developer.

A number of the above examples showed the necessity of combining DM methods with more traditional physical approaches. Whether it is microstructural modeling in the area of processing–structure–property prediction or kinetic equation modeling in catalysis design, physical modeling is by no means standing still, and its utility will continue to expand. In the few cases where authors make direct comparisons, it is not clear that DM applications have been more effective [44, 89]. It is already true that DM approaches, although more data focused, are deeply intertwined with traditional physical modeling. A researchers knowledge of the physics of the problem strongly influences such things as choices of descriptors (e.g., exponentiating parameters where thermal activation is expected), choices in the predictive model (e.g., using linear models when linear relationships are expected), and many unwritten small decisions about how the DM is done. DM and physical modeling, despite an apparent conflict, are really best used collaboratively, and effective materials researchers will need to combine both tools to have maximal impact.

Another important feature to note is the difference between DM in materials science and the more established areas of drug design and QSPR/QSAR. Although the overall framework is very similar, establishing effective descriptors for independent variables seems to be harder in materials applications. Bulk materials, more common in traditional materials science applications, often have atomic-, nano-, and micro-structural features that are hard to characterize and quantify with effective descriptors. In their absence, further progress on many problems will require additional descriptors relating to processing choices.

Finally, we would like to stress the natural synergy between DM and other kinds of computational modeling. High-throughput computation can help provide the wealth of data needed for robust data mining, as was illustrated above in the use of computationally optimized structures for boiling point modeling [20] and crystal structure prediction [80–82, 88]. Impressive examples of high-throughput *ab initio* computation providing large amounts of accurate materials data can be found in Refs. [90–92]. High-throughput computation not only increases the effectiveness of DM methods, but extends the reach of computational modeling, since DM methods can help span the challenging range of length and time scales involved in materials phenomena. The growing power of DM and other computational methods will only increase their interdependence in the future.

Finally, on a more personal note, we have found that one of the most valuable contributions of DM to our research has been to expand how we think about problems. DM encourages one to ask how one can make optimal use of data and to look deeply for patterns that might provide valuable information. DM makes one think on a large scale, thereby encouraging the automation of experiment, computation, and data analysis for high-throughput production. DM also encourages a culture of careful testing for any kind of fitting, through cross validation and statistical methods. Finally, DM is inherently inderdisciplinary, encouraging materials scientists to learn more about analogous problems and techniques from across the hard and soft sciences, thereby enriching us all as researchers.

References

[1] W. Klosgen and J.M. Zytkow, *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press, Oxford, 2002.
[2] N. Ye, *The Handbook of Data Mining*, Lawrence Erlbaum Associates, London, 2003.
[3] D. von Mendelejeff, "Ueber die Bezlehunger der Eigenschaften Zu den Atomgewichte der Elemente," *Zeit. Chem.*, 12, 405–406, 1869.
[4] M.F. Ashby, *Materials Selection in Mechanical Design.*, Butterworth-Heinemann, Boston, 1999.
[5] D. Braha, *Data Mining for Design and Manufacturing*, Kluwer Academic Publishers, Boston, 2001.
[6] M.H. Dunham, *Data Mining: Introductory and Advanced Topics*, Pearson Education, Inc., Upper Saddle River, New Jersey, 2003.
[7] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, Wiley-Interscience, IEEE Press, Hoboken, New Jersey, 2003.
[8] PolyphonicHMI, (http://www.polyphonichmi.com/technology.html).
[9] M.H. Kutner, C.J. Nachtschiem, W. Wasserman, and J. Neter, *Applied Linear Statistical Models*, McGraw-Hill, New York, 1996.
[10] A.C. Rencher, *Methods of Multivariate Analysis*, Wiley-Interscience, New York, 2002.

[11] J.E. Jackson, *A User's Guide to Principal Components*, John Wiley & Sons, New York, 1991.
[12] S.d. Jong, "Simpls: an alternative approach to partial least squares regression," in *Chemometrics and Intelligent Laboratory Systems*, 18, 251–263, 1993.
[13] B.M. Wise and N.B. Gallagher, *PLS_Toolbox 2.1 for Matlab*, Eigenvector Reseach, Inc., Manson, WA, 2000.
[14] S. Wold, A.H.W. Ruhe, and W.J. Dunn, "The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses," *SIAM J. Sci. Stat. Comput.*, 5, 735–743, 1984.
[15] M.T. Hagan, H.B. Demuth, and M.H. Beale, *Neural Network Design*, Martin Hagan, 2002.
[16] D.J.C. Mackay, "Bayesian interpolation," *Neural Comput.*, 4, 415–447, 1992.
[17] D.J.C. Mackay, "A practical bayesian framework for backpropagation networks," *Neural Comput.*, 4, 448–472, 1992.
[18] D.J.C. Mackay, "Probable networks and plausible predictions – a review of practical bayesian methods for supervised neural networks," *Network-Comput. Neural Syst.*, 6, 469–505, 1995.
[19] D.J.C. MacKay, "Bayesian modeling with neural networks," In: H. Cerjack (ed.), *Mathematical Modeling of Weld Phenomena*, vol. 3. The Institute of Materials, London, pp. 359–389, 1997.
[20] A.J. Chalk, B. Beck, and T. Clark, "A quantum mechanical/neural net model for boiling points with error estimation," *J. Chem. Inf. Comput. Sci.*, 41, 457–462, 2001.
[21] H. Bhadeshia, "Neural networks in materials science," *ISIJ Int.*, 39, 966–979, 1999.
[22] J.M. Serra, A. Corma, A. Chica, E. Argente, and V. Botti, "Can artificial neural networks help the experimentation in catalysis?," *Catal. Today*, 81, 393–403, 2003.
[23] K. Baumann, "Cross-validation as the objective function for variable-selection techniques," *Trac-Trend Anal. Chem.*, 22, 395–406 2003.
[24] A.S. Goldberger, *A Course in Econometrics*, Harvard University Press, Cambridge, MA, 1991.
[25] E.K.P. Chong and S.H. Zak, *An Introduction to Optimization*, John Wiley & Sons, New York, 2001.
[26] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, 1992.
[27] J. Taskinen and J. Yliruusi, "Prediction of physicochemical properties based on neural network modelling," *Adv. Drug Deliv. Rev.*, 55, 1163–1183, 2003.
[28] H. Bhadeshia, "Design of ferritic creep-resistant steels," *ISIJ Int.*, 41, 626–640, 2001.
[29] T. Sourmail, H. Bhadeshia, and D.J.C. MacKay, "Neural network model of creep strength of austenitic stainless steels," *Mater. Sci. Technol.*, 18, 655–663, 2002.
[30] S.H. Lalam, H. Bhadeshia, and D.J.C. MacKay, "Estimation of mechanical properties of ferritic steel welds part 1: yield and tensile strength," *Sci. Technol. Weld. Joining* 5, 135–147, 2000.
[31] S.H. Lalam, H. Bhadeshia, and D.J.C. MacKay, "Estimation of mechanical properties of ferritic steel welds part 2: Elongation and charpy toughness," *Sci. Technol. of Weld. Joining*, 5, 149–160, 2000.
[32] M.A. Yescas, H. Bhadeshia, and D.L. MacKay, "Estimation of the amount of retained austenite in austempered ductile irons using neural networks," *Mater. Sci. Eng. A*, 311, 162–173, 2001.
[33] S. Cardie and H.K.D.H. Bhadeshia, "Materials algorithms project (map): Public domain research software & data," In: *Mathematical Modelling of Weld Phenomena IV*, Institute of Materials, London, 1998.

[34] S. Malinov and W. Sha, "Software products for modelling and simulation in materials science," *Comput. Mater. Sci.*, 28, 179–198, 2003.

[35] S. Malinov, W. Sha, and Z. Guo, "Application of artificial neural network for prediction of time-temperature-transformation diagrams in titanium alloys," *Mater. Sci. Eng. Struct. Matter Properties Microstruct. Process*, 283, 1–10, 2000.

[36] S. Malinov, W. Sha, and J.J. McKeown, "Modelling the correlation between processing parameters and properties in titanium alloys using artificial neural network," *Comput. Mater. Sci.*, 21, 375–394, 2001.

[37] S. Malinov and W. Sha, "Application of artificial neural networks for modelling correlations in titanium alloys," *Mater. Sci. Eng.*, A365, 202–211, 2004.

[38] T. Malinova, S. Malinov, and N. Pantev, "Simulation of microhardness profiles for nitrocarburized surface layers by artificial neural network," *Surf. Coat. Technol.*, 135, 258–267, 2001.

[39] T. Malinova, N. Pantev, and S. Malinov, "Prediction of surface hardness after ferritic nitrocarburising of steels using artificial neural networks," *Mater. Sci. Technol.*, 17, 168–174, 2001.

[40] S. Christensen, J.S. Kandola, O. Femminella, S.R. Gunn, P.A.S. Reed, and I. Sinclair, "Adaptive numerical modelling of commercial aluminium plate performance," *Aluminium Alloys: Their Physical and Mechanical Properties, Pts 1–3*, 331–3, 533–538, 2000.

[41] O.P. Femminella, M.J. Starink, M. Brown, I. Sinclair, C.J. Harris, and P.A.S. Reed, "Data pre–processing/model initialisation in neurofuzzy modelling of structure-property relationships in Al–Zn–Mg–Cu alloys," *ISIJ Int.*, 39, 1027–1037, 1999.

[42] O.P. Femminella, M.J. Starink, S.R. Gunn, C.J. Harris, and P.A.S. Reed, "Neurofuzzy and supanova modelling of structure–property relationships in Al–Zn–Mg–Cu alloys," *Aluminium Alloys: Their Physical and Mechanical Properties, Pts 1–3*, 331–3, 1255–1260, 2000.

[43] J.S. Kandola, S.R. Gunn, I. Sinclair, and P.A.S. Reed, "Data driven knowledge extraction of materials properties," In: *Proceedings of Intelligent Processing and Manufacturing of Materials*, Hawaii, USA, 1999.

[44] M.J. Starink, I. Sinclair, P.A.S. Reed, and P.J. Gregson, "Predicting the structural performance of heat-treatable al-alloys," In: *Aluminum Alloys - Their Physical and Mechanical Properties, Parts 1-3*, vol. 331–337, pp. 97–110, Trans Tech Publications, Switzerland, 2000.

[45] H. Byun and S.W. Lee, "Applications of support vector machines for pattern recognition: A survey," *Pattern Recogn. Support Vector Machines, Proc.*, 2388, 213–236, 2002.

[46] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

[47] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[48] L. Harmon, "Experiment planning for combinatorial materials discovery," *J. Mater. Sci.*, 38, 4479–4485, 2003.

[49] G.J. Hutchings and M.S. Scurrell, "Designing oxidation catalysts – are we getting better?," *Cattech*, 7, 90–103, 2003.

[50] C. Klanner, D. Farrusseng, L. Baumes, C. Mirodatos, and F. Schuth, "How to design diverse libraries of solid catalysts?," *QSAR & Combinatorial Science*, 22, 729–736, 2003.

[51] T.R. Cundari, J. Deng, and Y. Zhao, “Design of a propane ammoxidation catalyst using artificial neural networks and genetic algorithms,” *Indust. & Eng. Chem. Res.*, 40, 5475–5480, 2001.

[52] T. Hattori and S. Kito, “Neural-network as a tool for catalyst development,” *Catal. Today*, 23, 347–355, 1995.

[53] M. Holena and M. Baerns, “Feedforward neural networks in catalysis - a tool for the approximation of the dependency of yield on catalyst composition, and for knowledge extraction,” *Catal. Today*, 81, 485–494, 2003.

[54] K. Huang, X.L. Zhan, F.Q. Chen, and D.W. Lu, “Catalyst design for methane oxidative coupling by using artificial neural network and hybrid genetic algorithm,” *Chem. Eng. Sci.*, 58, 81–87, 2003.

[55] A. Tompos, J.L. Margitfalvi, E. Tfirst, and L. Vegvari, Information mining using artificial neural networks and “holographic research strategy,” *Appl. Catal. A*, 254, 161–168, 2003.

[56] T. Umegaki, Y. Watanabe, N. Nukui, E. Omata, and M. Yamada, “Optimization of catalyst for methanol synthesis by a combinatorial approach using a parallel activity test and genetic algorithm assisted by a neural network,” In: *Energy Fuels*, 17, 850–856, 2003.

[57] O.V. Buyevskaya, A. Bruckner, E.V. Kondratenko, D. Wolf, and M. Baerns, “Fundamental and combinatorial approaches in the search for and optimisation of catalytic materials for the oxidative dehydrogenation of propane to propene,” *Catal. Today*, 67, 369–378, 2001.

[58] U. Rodemerck, D. Wolf, O.V. Buyevskaya, P. Claus, S. Senkan, and M. Baerns, “High-throughput synthesis and screening of catalytic materials – case study on the search for a low-temperature catalyst for the oxidation of low-concentration propane,” *Chem. Eng. J.*, 82, 3–11, 2001.

[59] D. Wolf, O.V. Buyevskaya, and M. Baerns, “An evolutionary approach in the combinatorial selection and optimization of catalytic materials,” *Appl. Catal. A*, 200, 63–77, 2000.

[60] J.M. Caruthers, J.A. Lauterbach, K.T. Thomson, V. Venkatasubramanian, C.M. Snively, A. Bhan, S. Katare, and G. Oskarsdottir, “Catalyst design: knowledge extraction from high-throughput experimentation,” *J. Catal.*, 216, 98–109, 2003.

[61] A. Tuchbreiter and R. Mulhaupt, “The polyolefin challenges: catalyst and process design, tailor-made materials, high-throughput development and data mining,” *Macromol. Symp.*, 173, 1–20, 2001.

[62] A. Tuchbreiter, J. Marquardt, B. Kappler, J. Honerkamp, M.O. Kristen, and R. Mulhaupt, “High-output polymer screening: exploiting combinatorial chemistry and data mining tools in catalyst and polymer development,” *Macromol. Rapid Comm.*, 24, 47–62, 2003.

[63] A. Hagemeyer, B. Jandeleit, Y.M. Liu, D.M. Poojary, H.W. Turner, A.F. Volpe, and W.H. Weinberg, “Applications of combinatorial methods in catalysis,” *Appl. Catal. A*, 221, 23–43, 2001.

[64] G. Bergerhoff, R. Hundt, R. Sievers, and I.D. Brown, “The inorganic crystal-structure data-base,” *J. Chem. Compu. Sci.*, 23, 66–69, 1983.

[65] P. Villars, K. Cenzual, J.L.C. Daams, F. Hullinger, T.B. Massalski, H. Okamoto, K. Osaki, and A. Prince, *Pauling File*, ASM International, Materials Park, Ohio, USA, 2002.

[66] P.S. White, J. Rodgers, and Y. Le Page, “Crystmet: a database of structures and powder patterns of metals and intermetallics,” *Acta Cryst. B*, 58, 343–348, 2002.

[67] S. Kabekkodu, G. Grosse, and J. Faber, "Data mining in the icdd's metals & alloys relational database," *Epdic 7: European Powder Diffraction, Pts 1 and 2*, 378–3, 100–105, 2001.

[68] P. Villars, Factors governing crystal structures. In: J.H. Westbrook and R.L. Fleischer (eds.), vol. 1, John Wiley & Sons, New York, pp. 227–275, 1994.

[69] J.K. Burdett and J. Rodgers, "Structure & property maps for inorganic solids," In: R.B. King (ed.), *Encyclopedia of Inorganic Chemistry*, vol. 7, John Wiley & Sons, New York, 1994.

[70] D.G. Pettifor, "The structures of binary compounds: I. Phenomenological structure maps," *J. Phys. C: Solid State Phys.*, 19, 285–313, 1986.

[71] D.G. Pettifor, "A chemical scale for crystal-structure maps," *Solid State Commun.*, 51, 31–34, 1984.

[72] D. Morgan, J. Rodgers, and G. Ceder, "Automatic construction, implementation and assessment of Pettifor maps," *J. Phys. Condens. Matter*, 15, 4361–4369, 2003.

[73] G.A. Landrum, *Prediction of Structure Types for Binary Compounds*, Rational Discovery, Inc., Palo Alto, pp. 1–8, 2001.

[74] Y.H. Pao, B.F. Duan, Y.L. Zhao, and S.R. LeClair, "Analysis and visualization of category membership distribution in multivariate data," *Eng. Appl. Artif. Intell.*, 13, 521–525, 2000.

[75] A. Rajagopalan, C.W. Suh, X. Li, and K. Rajan, "Secondary" descriptor development for zeolite framework design: an informatics approach, *Appl. Catal. A*, 254, 147–160, 2003.

[76] K. Rajan, *Combinatorial materials science and material informatics laboratory (COSMIC)*, (http://www.rpi.edu/∼rajank/materialsdiscovery/).

[77] F.R. de Boer, R. Boom, W.C.M. Matten, A.R. Miedema, and A.K. Niessen, *Cohesion in Metals: Transition Metal Alloys*, North Holland, Amsterdam, 1988.

[78] J.L.C. Daams, "Atomic environments in some related intermetallic structure types," In: J.H. Westbrook and R.L. Fleischer (eds.), *Intermetallic Compounds, Principle and Practice*, vol. 1, John Wiley & Sons, New York, pp. 227–275, 1994.

[79] S. Dietmann, J. Park, C. Notredame, A. Heger, M. Lappe, and L. Holm, "A fully automatic evolutionary classification of protein folds: Dali domain dictionary version 3," *Nucleic Acids Res.*, 29, 55–57, 2001.

[80] M. Jansen, "A concept for synthesis planning in solid-state chemistry," *Angew. Chem. Int. Ed.*, 41, 3747–3766, 2002.

[81] S.M. Woodley, P.D. Battle, J.D. Gale, and C.R.A. Catlow, "The prediction of inorganic crystal structures using a genetic algorithm and energy minimisation," *Phys. Chem. Chem. Phys.*, 1, 2535–2542, 1999.

[82] G.H. Johannesson, T. Bligaard, A.V. Ruban, H.L. Skriver, K.W. Jacobsen, and J.K. Norskov, "Combined electronic structure and evolutionary search approach to materials design," *Phys. Rev. Lett.*, 88, pp. 255506-1–255506-5, 2002.

[83] D. de Fontaine, "Cluster approach to order-disorder transformations in alloys," In: *Solid State Physics*, H. Ehrenreich and D. Turnbull (eds.), vol. 47, Academic Press, pp. 33–77 1994.

[84] A. Zunger, "First-principles statistical mechanics of semiconductor alloys and intermetallic compounds," *Statics and Dynamics of Alloy Phase Transformations*, New York, 1994.

[85] V. Blum and A. Zunger, "Structural complexity in binary bcc ground states: The case of bcc Mo–Ta," *Phys. Rev. B*, 69, pp. 020103-1–020103-4, 2004.

[86] G. Ceder, "Predicting properties from scratch," *Science*, 280, 1099–1100, 1998.

[87] A. van de Walle, M. Asta, and G. Ceder, "The alloy theoretic automated toolkit: A user guide," *Calphad-Computer Coupling of Phase Diagrams and Thermochemistry*, 26, 539–553, 2002.

[88] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, "Predicting crystal structures with data mining of quantum calculations," *Phy. Rev. Lett.*, 91, 2003.

[89] B. Chan, M. Bibby, and N. Holtz, "Predicting 800 to 500 Degrees C Weld Cooling Times by using Backpropagation Neural Networks," *Trans. Can. Soc. Mech. Eng.*, 20, 75, 1996.

[90] T. Bligaard, G.H. Johannesson, A.V. Ruban, H.L. Skriver, K.W. Jacobsen, and J.K. Norskov, "Pareto-optimal alloys," *Appl. Phys. Lett.*, 83, 4527–4529, 2003.

[91] S. Curtarolo, D. Morgan, and G. Ceder, "Accuracy of *ab initio* methods in predicting the crystal structures of metals: Review of 80 binary alloys," *submitted for publication*, 2004.

[92] A. Franceschetti and A. Zunger, "The inverse hand-structure problem of finding an atomic configuration with given electronic properties," *Nature*, 402, 60–63, 1999.