Combinatorial Experimentation and Materials Informatics

Ichiro Takeuchi, Mikk Lippmaa, and Yuji Matsumoto

Abstract

High-throughput experimentation is effective in systematically producing large and diverse data sets. The marriage of combinatorial materials science and informatics is a natural one, and results are beginning to emerge from the integration of elements of materials informatics with data from combinatorial libraries. We discuss data management issues in high-throughput experimentation and highlight recent examples where data-mining tools are being implemented for extracting knowledge and predicting new compounds, with an emphasis on electronic materials.

Keywords: *catalytic, combinatorial methods, electronic material, ferroelectric, informatics, magnetic properties, thin film.*

Combinatorial Approach to Materials

Pioneered by the pharmaceutical industry, the combinatorial method is now widely accepted as a watershed in the discovery and optimization of new compounds.

The high-throughput combinatorial approach allows simultaneous synthesis and screening of large arrays of different materials. In recent years, the scope of the combinatorial strategy has rapidly expanded beyond drug discovery, and there have been significant advances in its application to the development of new polymers, catalysts, and materials for electronics[1,2] (see the April 2002 issue of *MRS Bulletin* on Combinatorial Materials Science).

In any materials synthesis procedure, there are a number of parameters, independent and interdependent, that govern the formation and properties of the resulting phases. When such experiments are carried out in an ordinary one-by-one manner, subtle experiment-to-experiment variations in process parameters could potentially lead to significant deviation from the expected behavior. Such a problem can be circumvented by the use of the high-throughput approach, which allows rapid and systematic navigation in a multidimensional parameter space, producing statistically higher-quality data.

In a combinatorial setting, up to thousands of different stoichiometries, encompassing a wide compositional landscape, are synthesized in parallel. New characterization and screening techniques capable of handling combinatorial libraries and composition spreads have emerged, and most of the physical properties that are critical for the development of new materials can now be quickly mapped with a throughput rate that matches the synthesis process. Depending on the exact nature of the measurement scheme, enormous amounts of physical property data can be amassed from a single combinatorial library.

Systematic materials characterization data can be used to formulate strategies for optimizing a product or constructing models of underlying physical mechanisms that govern the materials properties. Data-mining techniques can facilitate this process, so that one can visualize large and complex data sets, extract structure–property and process-parameter–property relationships, validate physical models, and so on.

Data Management and Data Mining

Until recently, mining of materials data has been a daunting task, because the only way to assemble data sets with sufficient volume has been to extract measurement results from a large number of publications. This task is time-consuming, and it is often impossible to evaluate the reliability of individual data points because measurement and sample preparation techniques may vary. High-throughput parallel synthesis and characterization of samples can, for the first time, solve this problem by creating model data sets of systematic materials data covering large composition and parameter spaces.

Data mining in materials science can be implemented on a number of levels. In the simplest case, visualization, analysis, and modeling are applied to data obtained from a single or a few libraries, covering a single family of compounds. This can be useful for optimizing materials properties. A more challenging task is to answer more fundamental questions by looking at wider collections of data from different families of compounds, bringing together data from many sample libraries. This can be done if the experimental data are stored in a way that facilitates data exchange and long-term storage. Similar requirements for handling experimental data in the fields of organic synthesis and drug development have resulted in the development of a wide variety of specialized laboratory information management systems (LIMS).[3] On the other hand, very few tools exist for managing data from high-throughput materials development experiments that would facilitate collecting and distributing large materials data sets.[4] These problems are being solved on two levels: developing standards for representing materials data, and building database tools for rapidly processing, storing, and distributing large data sets.[5] There are numerous projects underway to create standards for describing materials and characterization data using XML,[6] but none of them has gained the widespread acceptance necessary for successful general distribution of XML-encoded materials characterization data sets. This is clearly a field that urgently needs further development.

The sheer quantity of data produced by combinatorial experiments represents a difficult logistical challenge. For instance, the size of the raw data collection from synchrotron microdiffraction results for a single combinatorial library with hundreds of members can easily reach terabytes.

Visualization is usually the first step in data mining. It is an important step because it enables transformation of large data sets into a form that can be handled efficiently by the human brain. The task at hand is to generate effective means of presenting multidimensional data sets in such a way that relevant parameter correlations become visible, even when there is no *a priori* knowledge as to what types of correlations to look for. It is a particularly nontrivial task to represent spectral type data for a two-dimensional combinatorial library simultaneously. An example of how such a problem may be approached is shown in Figure 1, where x-ray microdiffraction data (for 2θ from 42.5° to 47°) from 536 diffraction patterns taken from a composition-spread library that covers the entire ternary-composition phase space of the Ni-Fe-Co system[7] are plotted. To view all the diffraction data together, a three-dimensional volumetric contour representation is used so that diffraction peak positions and intensities are displayed simultaneously as a function of composition.[8] This plot underscores the utility of an added dimension in properly conveying all the information. Inspection of such a figure together with plots of other properties is instrumental in finding correlations between physical parameters. In a similar study of the Ni-Mn-Al ternary system, evidence that a particular x-ray peak shift had a strong anticorrelation with the remanent magnetization revealed the origin of magnetism in a particular composition region.[8,9]

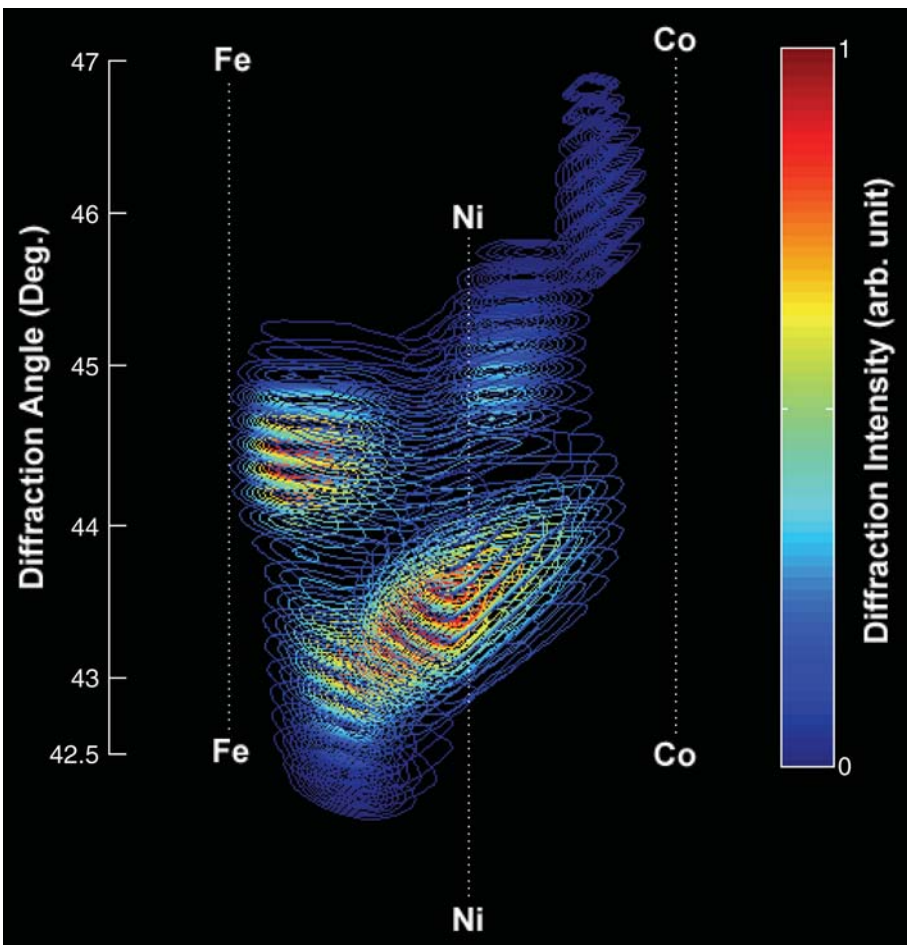

Figure 1. Three-dimensional contour representation of all the x-ray diffraction spectra from a Ni-Fe-Co composition-spread library. From the large high-intensity region in the 43–44.5° range, one can readily discern how the x-ray peak angle continuously shifts as a function of composition in the fcc phase. The other high-intensity peak in the Fe-rich region at around 44.5–45° is the bcc phase. The plot is best viewed from different angles or in a rotating animation (viewable at www.combi.umd.edu under FeCoNi XRD visualization).

PCA and QSAR

Data mining is particularly useful for dealing with data systematically generated to cover materials process descriptor space: analysis of relationships between material properties and process parameters can reveal hidden factors that are difficult or impossible to measure directly. For example, correlations between selected materials properties and process parameters are shown as a matrix plot in Figure 2. This experiment was designed to explore the mechanism of lattice defect formation in thin $Nb:SrTiO_3$ films grown by pulsed laser deposition. Approximately 60 thin films were grown at various process conditions, where the ablation laser parameters, background gas pressure, and target–source distance were varied while the growth was monitored by electron diffraction. The lattice parameters, optical absorption, transport, and dielectric properties of all samples were also measured.

These correlation plots immediately showed that the deviation of the lattice constant of the film from the bulk value and optical transmittance at 1000 nm have a minimum at a specific energy density of the ablation laser (highlighted in the figure). The optical transmittance also had a clear linear correlation with the lattice parameter of the film and the intensity of the film x-ray diffraction peak. In this material, the optical absorption in the near-infrared range probes the free carrier density and, thus, film conductivity. Simple scatter plots of the raw measurement results were sufficient to show that deviations of the laser energy from a critical value result in defect-rich films (lattice parameter deviates from bulk value, diffraction peak intensity drops) with a reduced carrier activation ratio (increased optical transmittance). More detailed analysis of all experimental results allowed the defect types to be identified and traced back to minute cation stoichiometry changes in the films.

Another common technique that can be used to find the dominant interactions in a multidimensional data set is to reduce the dimensionality of the descriptor space by principal component analysis (PCA). In one example, PCA was used to look for correlations between descriptors in the synthesis process of ferroelectric $Bi_4Ti_3O_{12}$ thin films that were deposited using a vapor-phase flux-mediated epitaxy technique.[10] The purpose of the combinatorial experiment was to rapidly identify the optimum flux composition for the growth of single-crystalline $Bi_4Ti_3O_{12}$. In total, six pseudo-ternary composition spreads, spanning compositions of a variety of impurity fluxes, were synthesized. Based on scanning x-ray diffraction and subsequent microstructural characterization, it was found

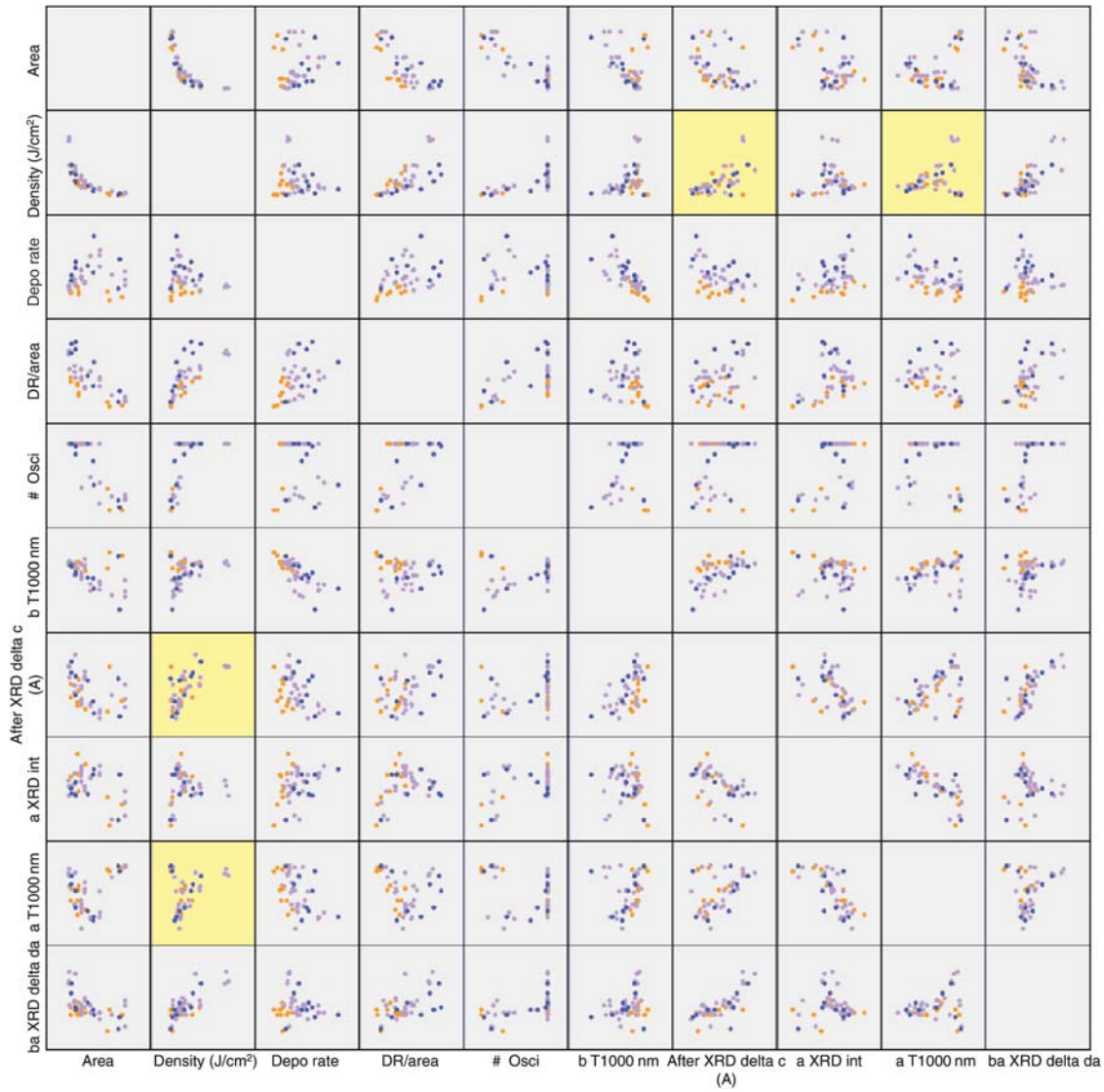

Figure 2. A correlation matrix plot of optical, structural, and transport properties, together with several process parameters of Nb-doped $SrTiO_3$ thin films grown by pulsed laser deposition. The table rows from top to bottom represent the ablation laser spot area (Area), energy density (Density), film deposition rate (Depo rate), deposition rate scaled by laser spot size (DR/area), number of observed electron diffraction oscillations (# Osci), transmission at 1000 nm before annealing (b T1000 nm), lattice parameter deviation from bulk value as measured by x-ray diffraction (After XRD delta c), x-ray peak intensity after annealing (a XRD int), transmission at 1000 nm after annealing (a T1000 nm), and lattice-constant deviations from bulk value before annealing (ba XRD delta da). Different colors correspond to larger (orange), intermediate (violet), or smaller (blue) target–sample deposition distances. The boxes highlighted in yellow show the trends discussed in the text.

that CuO as the flux material resulted in single-crystalline $Bi_4Ti_3O_{12}$ with a nearly perfect structure. Three hundred composition points from the experiment were used to run PCA with 24 descriptors, including various chemical parameters such as acidity (i.e., a generalized electron negativity of metal ions),[11] formation enthalpy, Pauling bonding ionicity, and the Dietzel parameter (which measures the likelihood of formation of intermediate phases).[12] The first and second principal components were found to predominantly represent the crystallinity of the *c*-axis oriented film and the bonding ionicity in the flux, respectively. The loads plot in Figure 3 indicates that crystallinity has anticorrelations with the acidity and the Dietzel parameter and no strong correlation with the bonding ionicity. The finding that the smaller Dietzel parameter leads to higher crystallinity is in agreement with a known fact from earlier bulk synthesis of similar compounds that CuO serves as a good flux in general due to the reduced chance of formation of intermediate phases.[13]

While composition spreads are powerful for mapping phase diagrams, continuous variation of the composition translates to an inherent lack of diversity in the data set, even if a large number of sampling points are used across a ternary phase space as in the above case. Because physical attributes frequently vary linearly as a function of continuously changing composition, mining processes relying solely on composition-spread data often do not lead to significant revelations beyond locating local maxima of physical properties,

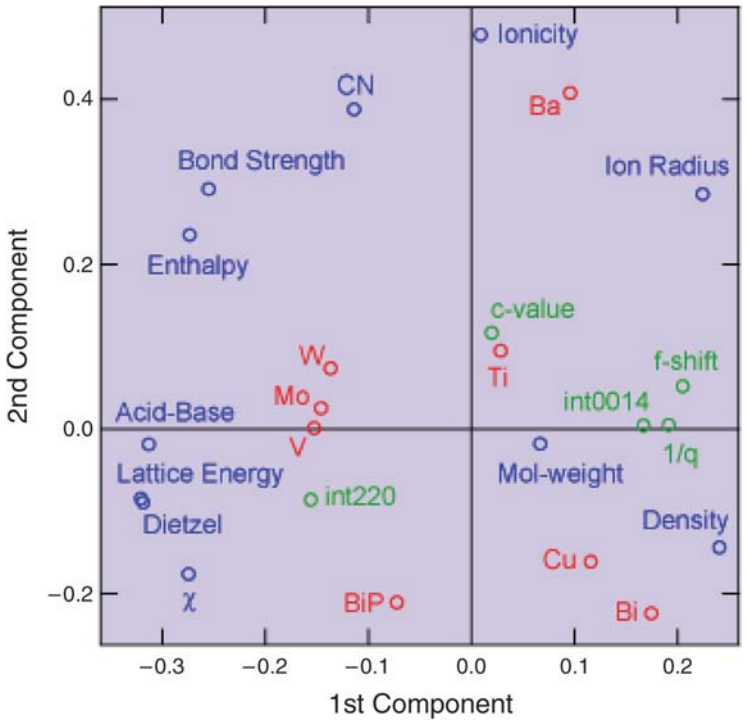

Figure 3. Loads plot from a principal component analysis of a combinatorial experiment to search for an effective flux material to be used for the growth of $Bi_4Ti_3O_{12}$ thin films. Twenty-four descriptors include various chemical parameters (blue), the coordination number (CN), and flux compositions (red). They also include measured experimental results (green): x-ray peak intensities int0014 and int220 for the (0014) and (220) peaks, respectively; c-axis lattice parameter and measures of dielectric properties; and frequency shift f and 1/q obtained from microwave microscopy. Parameters f and q are the resonant frequency and the quality factor of the microscope cavity, respectively. The principal components are linear combinations of correlated descriptors that capture the most dominant trends. The loads plots show the weight of each parameter on the principal component and display the correlations between parameters. The relationship between the crystallinity of the film and the chemical parameters was deduced from the analysis.

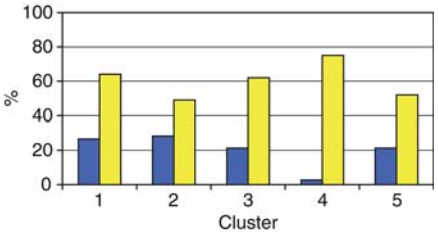

Figure 4. Comparison of prediction rates (percentage of correct predictions on the y axis) for different catalyst classes derived from neural network analysis (yellow bars) with statistical expectation values (blue bars). Based on this, it is possible to generate a library of catalysts exhibiting, for instance, high selectivity to partial oxidation products (cluster 3) with 60% chance, whereas it is only 20% on the basis of random selection. (Described in References 17 and 18.)

unless data from a large number of different systems are combined.

These examples underscore the utility of simple mining techniques to reveal underlying mechanisms of physical properties and trends. However, the ultimate goal of data mining is the prediction of new compounds. The notion of quantitative structure–activity relationship (QSAR) plays a prevalent role in designing libraries for the discovery of new drugs. Whereas there are effective descriptors to characterize biological molecules and drugs in general, there are no well-established QSAR models in materials science due to the complexity of solid materials. Yet the sheer diversity and complexity of solid materials call for intelligent optimization strategies in high-throughput experimentation.[14–16]

Recently, a QSAR model was developed for catalysis. Farrusseng et al. synthesized a diverse library consisting of 467 different multicomponent catalysts, which were tested in the oxidation of propene.[17,18] The resulting catalytic performances were statistically studied and separated into five types of properties such as catalysis exhibiting low conversion or high CO_2 selectivity. A set of 3200 attributes was generated as combinations of parameters including enthalpies of formation of different oxides, coordination numbers of atoms, ionization energies, and electronegativities. Neural networks and hierarchical classification techniques were then used to establish a QSAR model between the catalytic property and the descriptors of the solids. As a result, attributes were reduced to an essential set of 23, which can be used to predict whether a catalyst would fall into one of the five performance classes with a predictive power substantially exceeding random statistical selection (Figure 4). This information can now be used to design compounds with targeted catalytic properties.

Summary

Integration of combinatorial experimentation with mining and modeling techniques provides a powerful platform for the development of new materials. In principle, such a method is applicable to any materials science problem. Future challenges include interfacing the experiments with various existing materials databases and generating new databases, in turn, from combinatorial libraries. The successful incorporation of informatics techniques into combinatorial methodology can further enhance the efficiency of the overall approach and lead to accelerated innovation in materials research.

Acknowledgments

We acknowledge key contributions by our colleagues, including T. Ohnishi, R. Takahashi, and C.J. Long. We also acknowledge valuable discussions with D. Farrusseng, W. Maier, T. Chikyow, M. Kawasaki, K. Rajan, H. Koinuma, and X.-D. Xiang. The author is supported by NSF DMR 0231291, DMR 0094265, MRSEC DMR 0520471, and NIST.

References

1. X.-D. Xiang and I. Takeuchi, eds., *Combinatorial Materials Syntheses* (Marcel Dekker, New York, 2003).
2. H. Koinuma and I. Takeuchi, *Nat. Mater.* **3** (2004) p. 429.
3. For a comparison of LIMS products, see, for example, www.limsfinder.com or www.limsource.com (accessed November 2006).
4. Symyx Discovery Tools, www.symyx.com (accessed November 2006).
5. S. Meguro, T. Ohnishi, M. Lippmaa, and H. Koinuma, *Meas. Sci. Technol.* **16** (2005) p. 309.
6. MatML, schema for materials information, www.matml.org (accessed November 2006).

7. Y.K. Yoo, Q. Xue, Y.S. Chu, S. Xu, U. Hangen, H.-C. Lee, W. Stein, and X.-D. Xiang, *Intermetallics* **14** (2006) p. 241.
8. I. Takeuchi, C.J. Long, O.O. Famodu, M. Murakami, J. Hattrick-Simpers, G.W. Rubloff, M. Stukowski, and K. Rajan, *Rev. Sci. Instrum.* **76** 062223 (2005).
9. O.O. Famodu, J. Hattrick-Simpers, M. Aronova, K.-S. Chang, M. Murakami, M. Wuttig, T. Okazaki, Y. Furuya, and I. Takeuchi, *Mater. Trans., JIM* **45** (2004) p. 173.
10. R. Takahashi, Y. Yonezawa, M. Ohtani, M. Kawasaki, K. Nakajima, T. Chikyow, H. Koinuma, and Y. Matsumoto, *Adv. Funct. Mater.* **16** (2006) p. 485.
11. K. Tanaka and A. Ozaki, *J. Catal.* **8** (1967) p. 1.
12. R.E. Newnham, *Phase Diagrams: Material Science and Technology, Vol. 5* (Academic Press, London, UK, 1978).
13. G.P. Kelkar and A.H. Carim, *J. Am. Ceram. Soc.* **76** (1993) p. 1815.
14. J. Scheidtmann, D. Klar, J.W. Saalfrank, T. Schmidt, and W.F. Maier, *Quant. Struct.-Act. Relat. Combi. Sci.* **24** (2005) p. 203.
15. J. Scheidtmann, A. Frantzen, G. Frenzer, and W.F. Maier, *Meas. Sci. Technol.* **16** (2005) p. 119.
16. R.A. Potyrailo, *Angew. Chem. Int. Ed.* **45** (2004) p. 702.
17. D. Farrusseng, C. Klanner, L. Baumes, M. Lengliz, C. Mirodatos, and F. Schuth, *Quant. Struct.-Act. Relat. Combi. Sci.* **24** (2005) p. 78.
18. C. Klanner, D. Farrusseng, L. Baumes, M. Lengliz, C. Mirodatos, and F. Schuth, *Angew. Chem. Int. Ed.* **43** (2004) p. 5347. □