Perspective: Composition–structure–property mapping in high-throughput experiments: Turning data into knowledge

Jason R. Hattrick-Simpers, John M. Gregoire, and A. Gilad Kusne

Citation: APL Materials **4**, 053211 (2016); doi: 10.1063/1.4950995
View online: http://dx.doi.org/10.1063/1.4950995
View Table of Contents: http://aip.scitation.org/toc/apm/4/5
Published by the American Institute of Physics

Perspective: Composition–structure–property mapping in high-throughput experiments: Turning data into knowledge

Jason R. Hattrick-Simpers,[1,2] John M. Gregoire,[3] and A. Gilad Kusne[4,5]
[1]*Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina 29208, USA*
[2]*SmartState™ Center for Economic Excellence in the Strategic Approaches to the Generation of Electricity, Columbia, South Carolina 29208, USA*
[3]*Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, California 91125, USA*
[4]*Materials Measurement Science Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA*
[5]*Materials Science and Engineering Department, University of Maryland, College Park, Maryland 20742, USA*

(Received 15 February 2016; accepted 9 May 2016; published online 26 May 2016)

With their ability to rapidly elucidate composition-structure-property relationships, high-throughput experimental studies have revolutionized how materials are discovered, optimized, and commercialized. It is now possible to synthesize and characterize high-throughput libraries that systematically address thousands of individual cuts of fabrication parameter space. An unresolved issue remains transforming structural characterization data into phase mappings. This difficulty is related to the complex information present in diffraction and spectroscopic data and its variation with composition and processing. We review the field of automated phase diagram attribution and discuss the impact that emerging computational approaches will have in the generation of phase diagrams and beyond.
[http://dx.doi.org/10.1063/1.4950995]

I. INTRODUCTION

Although their use can be traced back as far as Haber-Bosch catalyst development for ammonia production, the use of combinatorial techniques and their deployment for high-throughput experimentation (HTE) became widespread in the 1990's after personal computers became affordable.[1] Since its rediscovery, HTE has been adopted to search for new materials across a broad spectrum of fields including catalysis, functional materials, and materials for energy.[2,3] Here, HTE has made an impact in fundamental science, for instance validating theoretical predictions for minimal hysteresis in shape memory alloys, and the discovery of a shuttling mechanism for the creation of ordered olefin block copolymers.[4–6] Outside of traditional academic research, HTE also has been broadly accepted and implemented by companies such as BASF, DOW, Micron Technology, and General Electric as a means to "Fail Fast" thus reducing time spent researching dead-end materials. Examples of technologies utilizing materials developed via HTE include Dow InFuse™ polymers and the flash memory developed for the iPhone 5.[7] Since these techniques can be used to rapidly validate and improve models or to confirm the existence of theoretically predicted materials, the initial White House white paper and subsequent reports have highlighted the importance of HTE to realizing the goals of the Materials Genome Initiative (MGI).[8,9,37]

The increased rate of data generation obtained through HTE techniques led to the early adoption of data reduction and data mining techniques by the community to advance effective experimental design.[10–12] One particularly challenging topic has been that of the automated generation of structural phase diagrams from structural characterization data, which is typically produced via

scanning x-ray diffraction (XRD) or Raman spectroscopy experiments. Structural phase diagrams are the basis of materials science, representing the relationship between material structure and fabrication parameters—typically composition and temperature. Although the majority of binary phase diagrams have been reported, most ternary and quaternary phase diagrams remain unexplored. Being able to correlate observed properties to the phase and composition of a particular material enables the insights that can guide theoretical studies and also yields important leads into exciting new materials phenomena. Recent work has suggested that the expected relationship between structure and properties can be exploited by using any property measurement and data mining to propose a possible phase diagram.[13] In this case, structure-mapping experiments can be designed to test the proposed phase diagram rather than characterize the phase behavior *de novo*. Solving the phase behavior of a set of material samples through interpretation of XRD or Raman measurements remains the most prevalent method for establishing structure-property relationships in combinatorial experiments, and in this review, we focus on techniques for facilitating and automating this phase mapping problem.

Unfortunately, translating a single XRD or Raman pattern into knowledge can be a time consuming process and since an automated experiment can generate 1000's of these patterns as a function of composition, time, and temperature, manual data processing and interpretation becomes infeasible. This problem is exacerbated by a number of factors, which are difficult to control and that greatly complicate data analysis. First, HTE libraries are often synthesized by techniques such as magnetron sputtering or thermal processing of elemental precursors, which can result in variability in the microstructure (e.g., randomly oriented polycrystalline, fiber texturing, varying grain sizes, or epitaxial growth) across a HTE sample. XRD phase analysis is typically performed using powder diffraction patterns, but generating equivalent data for materials spanning this range of morphologies and microstructures is challenging and often impractical. A further complication is that a measured XRD signal may change across a materials library even for materials with the same crystal structure (e.g., changes of texturing from random polycrystalline to fiber-textured). Second, for solid state materials, different compositions may crystallize into the same structure with lattice parameter(s) that are composition-dependent, which is manifested in XRD patterns as peaks that shift as a function of composition, often called peak shift. Peak shifting in Raman spectra is more complex and is related to the strength of bonds, mass of the atoms, etc. As will be discussed below, many techniques applied by the community will misclassify a shifted peak as a new phase if the shift is an appreciable fraction of the peak width. The third type of challenge arises when attempting to index the XRD pattern of a candidate phase where reference patterns are not available. Patterns acquired through combinatorial deposition and structural characterization techniques typically lack the data quality required for standard crystal refinement techniques, and given the vast possibilities of crystal structures with composition-dependent lattice parameters that could give rise to a given XRD pattern, indexing known and especially newly discovered phases is quite challenging.

To delineate composition–structure–property relationships, early HTE works made use of commercially available data analysis and imaging software. Such software was not designed with sufficient functionality to automatically perform phase attribution, as a consequence, simple stacked plots were used to correlate composition and structural phase transitions in binary phase diagram studies.[14] Higher-order systems (e.g., systems including temperature, ternary or more additives, etc.) are difficult to visualize directly using such software and thus extracting composition–structure relationships has been a substantial bottleneck for HTE studies. In this manuscript, we will discuss efforts by the HTE community to surmount this obstacle by incorporating advanced data minimization, visualization, and machine learning algorithms to automatically create these correlations. Here, we will succinctly review the state of the art in knowledge extraction for the HTE community. We will first discuss data types used for structural determination, followed by data preprocessing techniques typically used, how the data are represented, metrics used to differentiate different spectra, data reduction techniques, and then the state-of-the-art in phase diagram determination algorithms.

II. HTE LIBRARY SYNTHESIS AND STRUCTURAL CHARACTERIZATION

A variety of HTE synthesis techniques have been demonstrated in the literature including those employing physical vapor deposition (PVD) and solution synthesis techniques; there are several excellent reviews on this subject.[3,15,16] In PVD, compositional variation is achieved via either taking advantage of the natural spatial variation of the flux from a source or by using a shuttering system. Generally speaking, solution synthesis studies (e.g., ink-jet printing) control composition by dosing different compositions of precursors by pipetting. The resultant HTE libraries are then characterized as-synthesized or after post-processing under different temperatures/environments. This can be done either uniformly across the entire HTE library or as a separately varying parameter. The geometry of an individual point in fabrication parameter space (FPS) on a HTE library is determined by the figure of merit screening technique to be employed [e.g., discrete squares for scanning superconducting quantum interference device (SQUID) or scanning electrolytic characterization]. The result is that each HTE sample contains a well-defined mapping of composition-processing space onto the Cartesian geometry of the (typically planar) substrate, which can then be used to create composition–processing–structure linkages. In the following discussion, we use the term "HTE-library" to refer to an experiment containing multiple points in FPS, and the term "sample" to refer to an individual point in FPS.

In typical materials science studies, structural information includes composition, phase, microstructure (e.g., texturing, nanoparticle morphology), and macroscale structure (e.g., tensile testing specimen geometry). Microstructural information is gathered via x-ray diffraction (XRD), Raman spectroscopy, scanning electron microscopy (SEM), and transmission electron microscopy (TEM). In the thin-film HTE literature, emphasis is placed on the characterization of composition and crystal phase with few studies focusing on the degree of material texturing. Scanning XRD, scanning SEM, and scanning Raman spectroscopy are the tools most commonly reported in the HTE literature to correlate composition and structure.

Scanning XRD is performed in a reflection or transmission geometry, typically with a 2D area detector that captures scattering over a wide range of angles as a 2D diffraction image. Typically, the XRD measurement samples a large ensemble of crystal grains, and symmetry about the χ angle (random orientation of these grains) is assumed and the 2D image is integrated along this angle to produce a 1D diffraction pattern of intensity vs scattering angle (or scattering vector magnitude). Figure 1 provides a schematic flow diagram illustrating how data are captured from a wafer, processed for analysis, and subsequently analyzed using the techniques described below. Raman spectroscopy similarly produces a 1D response of Raman scattered light intensity versus light frequency. To the best of our knowledge, there are no currently existing examples of HTE TEM experiments but with improved automation in TEM sample preparation, column alignment,

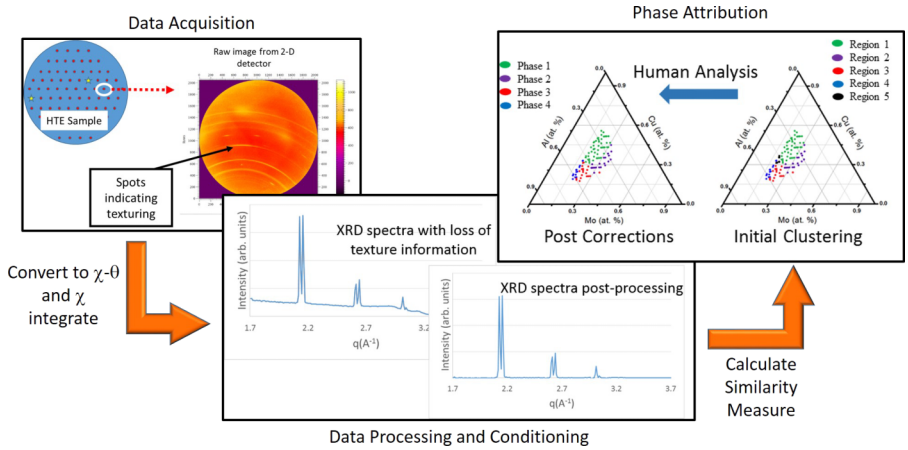

FIG. 1. Schematic overview of the process for converting diffraction/spectroscopic and compositional data into structural phase diagrams.

sample selection, and focusing this is an area with great potential for rapid progress. Both SEM and TEM instruments can generate electron diffraction images, which are analogous to the 2D XRD images and while these images are often analyzed prior to the χ–integration described above, the XRD analysis algorithms described below could be applied to the analysis of electron diffraction data.

III. PHASE DIAGRAM DETERMINATION

Once structure and fabrication parameter data are collected, the samples are sorted into groups of similar structure, and in the case of samples with phase segregation, groups of shared phase composition. Samples that are grouped together describe connected regions of the FPS that result in similar phase or phase mixtures. For regions composed of previously known phases, the abundance of each phase and the lattice parameters can be identified using a structure refinement method, e.g., Rietveld refinement. For studies exploring new FPS, the resulting phases may be unknown or may not be easily identifiable with previously known structures, requiring significant amounts of analysis time from a crystallographer. However, HTE studies have resulted in large, rapidly growing volumes of structure data, making manual analysis by experts infeasible. As a result, the materials community has turned to the rapid data analysis techniques of machine learning to convert the large volumes of structure data into phase diagrams. While machine learning promises to offer high-throughput phase diagram determination, its success is dependent on a few factors—prior knowledge, data preprocessing, data representation, similarity or dissimilarity measures, and model choice. These will be discussed in detail below.

A. Data conditioning

1. Prior knowledge

Prior knowledge plays a key role in determining analysis performance, and this includes knowledge of both the structure measurement method as well as the material space under investigation. Knowledge of the structure measurement method is required in data preprocessing—allowing the analyst to identify and remove instrument related noise, background, and features from the data, all of which can result in poor analysis performance. Prior knowledge of the material space under investigation may exist in the literature or material property databases. Utilizing this information in addition to the study-based data can provide a fuller basis from which to make predictions. A set of machine learning algorithms, supervised and semi-supervised learning methods, allow for the incorporation of such prior knowledge. Additionally, knowledge of the structure measurement technique (XRD, Raman, etc.) can be used to reduce the space of potential phase diagram solutions to those that are physically realizable through the use of constraints.

2. Data preprocessing techniques

During structure measurements, information of material structure is convolved with instrument effects, resulting in added instrument-based noise, background signals, and other artifacts in the resulting data. These instrument-based artifacts can result in greatly reduced analysis performance or even analysis failure. Measurement noise can be reduced with the use of data filtering techniques such as convolving the data with a smoothing "mask"—e.g., running average smoothing. Alternatively, a curve fitting method such as cubic spline interpolation can be used to reduce noise effects. Background subtraction can easily be performed when the instrument background effect is known, such as $1/x$ low-angle air scattering. A collection of such functions is available in structure refinement packages. When the background is unknown, the background can be characterized using a well-known dummy sample (e.g., a blank silicon wafer) in place of the sample to be measured. The characterized background signal can then be subtracted from measurement data. An example of this was performed in Ref. 17 to remove background from XRD data.

3. Data representation

Various data representations of structure data have been used in the literature.[18–21] The integrated 1D diffraction pattern has been used most extensively as part of a variety of high throughput phase diagram determination techniques.[20,22,23] This is most likely due to its widespread use in manual phase identification and the development of expert intuition in interpreting 1D data. As previously mentioned, software tools exist to facilitate phase identification for 1D data, allowing for "sanity checks" when evaluating machine learning algorithm performance. 1D data have also been shown to work well as the data format for spectral decomposition using matrix factorization techniques. The identified constituent phase abundances can then also be used in determining a phase diagram.[20,22,23] Alternatively LeBras *et al.*[18,25] converted 1D XRD patterns into discrete lists of peak information containing information on peak position, amplitude, and width. Each diffraction pattern was converted into a data list using a wavelet-based peak detection algorithm[26] and the discrete data lists were then used in a phase diagram determination method that utilizes integer linear programming based reasoning. By converting the 1D data into data lists, the data are reduced in dimension and complexity, removing extraneous instrument-based information from the data, such as peak shape warping due to the X-ray source and instrument-based noise. However, the reduced dimensionality and complexity of the data comes at the cost of an added pre-processing step with significant computational cost. Another field of data dimension reduction is called latent variable analysis[27] and contains some high speed methods that can facilitate phase diagram determination.

4. Measures for comparing XRD/Raman patterns

A similarity or dissimilarity measure is used to quantify the relationship between two pieces of structure data and is used to identify whether two samples originate from the same phase region. The dissimilarity is often represented as

$$D_{ij} = d\left(y_i, y_j\right),$$

where D_{ij} is the dissimilarity between the structure data y for samples i and j. A dissimilarity measure increases as the samples become less similar while a similarity measure increases with increasing similarity. The common dissimilarity measure used in 2 or 3 dimensions is the Euclidean metric used to determine the distance between two samples. However, the Euclidean metric does not perform well in the much higher dimensional space of structure data, which is typically hundreds to thousands of dimensions for 1D data and millions of dimensions for image data, each dimension corresponding to an intensity as a function of frequency or spacing. These high dimensional data exhibit effects such as feature shifts, where a feature shifts from one dimension to another, e.g., XRD peak shifting. The selected measure should be able to identify two samples with a feature shift as similar. Also, the high dimensional nature of the data means that it succumbs to the "curse of dimensionality"—the higher the dimension of the data, the larger the volume of space in which the data reside and the sparser the data, making all data seem dissimilar. More complex measures exist for high dimensional data, but they often come with significantly greater computation cost.

A few measures have been used in the literature to compare high dimensional structure data. The geometry-based L1 norm was used in Ref. 28 and provides a high speed measure for defining dissimilarity between 1D diffraction data. The geometry-based cosine metric and the statistics-based Pearson's correlation coefficient have also been used successfully.[17,20] Both provide high speed dissimilarity measures and improved performance in identifying similar structures that differ in only peak height, a result of differing structure-order lengths in the samples. However, the L1 norm, cosine metric, and Pearson's correlation coefficient perform poorly in the presence of peak shifts. Alternatively, the dynamic time warping measure has been shown to perform well in the presence of peak shifting, when the magnitude of peak shifting is known, but comes at a computational cost that is orders of magnitude larger.[18,29] A thorough discussion of measures for structure comparison is forthcoming.[30]

B. Data visualization and manual analysis

Humans are adept at interpreting XRD patterns and sets of XRD patterns with the practical limitation being creating a data representation that is amenable to human pattern recognition capabilities. For a 1-D composition space (a binary composition library or linear slice through higher-order composition space), the XRD intensity can be mapped in false color as a function of the scattering angle and composition. By inspecting so-called "heat maps," humans can generally identify collections of peaks belonging to a phase and phase mixtures, essentially solving the phase diagram manually. Software for generating this type of data visualization has been reported, for example, by Long *et al.*[10,17] and LeBras *et al.*[31] The latter software has also been implemented into a web interface UDiscover. It[25] and utilized for crowdsourcing data analysis by training non-experts to recognize patterns in heat map images. Combined with the capability to generate a comprehensive set of heat maps for XRD data sets in high-order composition spaces, the phase map problem could in principle be solved by brute force human computation.

To provide expert users with more information-rich representations of large XRD data sets, dimensionality reduction algorithms have been successfully employed. Dimensional reduction techniques take advantage of the fact that high dimensional data can often be described by a lower dimensional manifold in the high dimensional data space. A common technique is principle component analysis (PCA) now used for a variety of applications in materials informatics.[32,33] PCA performs well when samples are a linear mixture of constituent phases. However, peak shifting can result in a non-linear mapping. Multidimensional data scaling (MDS) methods can identify such curved manifolds by allowing for non-linear dissimilarity measures. For example, MDS in combination with the Pearson's correlation coefficient was used to visualize peak shifting in the Fe-Ga-Pd material system in 3 dimensions.[17] This method is also used to visualize in 2D the effect of measure choice on phase diagram determination via clustering.[30] Non-negative matrix factorization (NMF), a method used to identify constituent phases, has also been used to identify a 2D linear manifold described by non-negative basis vectors.[33]

While these visualization and dimensional reduction algorithms can help harness the power of human computation and assist with the formulation of automated algorithms, purely computational algorithms that produce objective solutions to the phase mapping problem are preferred to remove subjectivity and increase overall throughput.

C. Rapid phase diagram determination models

While a tremendous variety of machine learning methods exist,[27] only a handful have been reported as successful in determining phase diagrams. Machine learning methods are often grouped by their data inputs and their models. Broadly many machine learning methods fall into the categories of supervised, unsupervised, or semi-supervised learning techniques. For phase diagram determination, supervised techniques require as input a set of samples with fabrication parameters and structure data as well as phase region labels. The relationship between fabrication parameters, structure, and phase region label is then learned by "training" on the input data and extrapolated to identify the phase region labels for samples without labels. These methods are often used to identify phase region labels, which have been seen in the input, though they can also aid in discovering novel phases through the use of novelty and anomaly detection methods. Unsupervised methods do not require phase region labels as input, they sort samples into clusters associated with phase regions by identifying data structure in the combined frame parameter and structure data spaces. A further description of these general methodologies can be found in Ref. 27. We will discuss four methods with proven success in phase diagram determination: feature learning, clustering, matrix factorization, and constraint reasoning. The feature learning described here is part of a supervised learning method while various implementations of clustering, matrix factorization, and constraint reasoning can be deployed with either an unsupervised or semi-supervised approach.

1. Supervised learning

Bunn *et al.* used Adaboost feature learning, as part of a multi-step machine learning algorithm in their analysis of phase formation and oxidation of Ni-Al thin films using 1D XRD patterns, Raman, and Luminescence spectroscopy and composition data.[21] The three sets of phase characterization data were combined in this study as each provided unique abilities for phase discrimination: XRD was used to characterize metallographic phase, Raman was used to distinguish non-passivating oxides, and luminescence spectroscopy provided a method to identify the formation of α-Al_2O_3. The samples were partitioned into two sets: one used to train the algorithm and the other to test algorithm performance. The training samples were labelled for phase content by an expert and the algorithm was trained on the labels and structure data to identify diffraction pattern and spectroscopic features that were predictive of phase content. The algorithm was limited to simple features such as diffraction intensity and the first derivative of the diffraction pattern. In this study, unprocessed and processed data were considered independently. The algorithm was then applied to the testing samples data with a minimum accuracy of 88.9% for unprocessed data and 98.4% for post-processed data. A sliding window approach helped to mitigate issues with peak shift. Disadvantages of this approach include the need for a fairly large training set (~50–100 spectra depending on complexity) and that all features of all future spectra be contained within the training set. A diagram of the workflow used in this work can be found in Figure S1 of the supplementary material.[36]

2. Unsupervised learning

The ultimate goal in the phase mapping problem is to generate phase diagrams from combinatorial XRD data sets through automated execution of unsupervised learning algorithms. To describe the different approaches to solving the phase mapping problem and to appreciate the ever-increasing sophistication of the algorithms, we employ the visualization tool of Figure 2, where 3 categories of algorithms are noted: Clustering, Matrix Factorization, and Reasoning. We first describe algorithms that exclusively use one of these approaches and then describe the state-of-the-art hybrid algorithms.

a. Clustering. Clustering techniques sort samples into groups ("clusters") such that samples within the same group are more similar than samples in different groups. Under the expectation that XRD or Raman patterns from samples with the same phase will be more similar than patterns from different phases, clustering offers a straightforward mechanism for partitioning a large number of patterns into a small number of clusters that may represent individual phases or phase fields. Clustering can be generally represented by the mapping

$$C_k : \{s_1, s_2, \ldots, s_N\} \rightarrow \{1 \ldots k\},$$

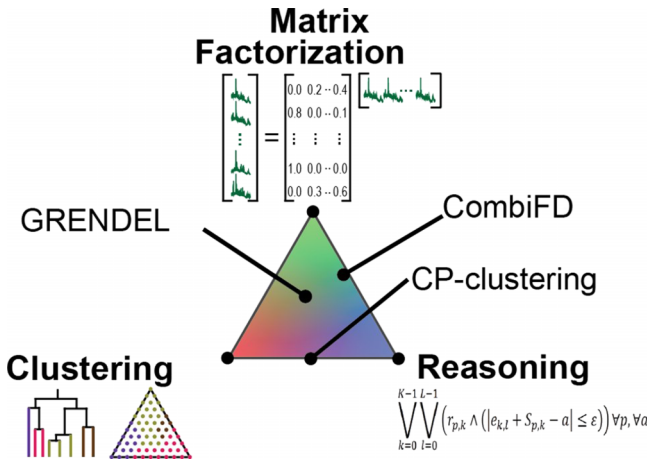

FIG. 2. Foundational unsupervised learning algorithms have been developed using Clustering, Matrix Factorization, and Reasoning techniques. State-of-the-art algorithms hybridize these complementary approaches.

where for our case, each sample s_i is mapped to one of k clusters. A variety of clustering algorithms exist, each with its benefits and disadvantages. Identifying the optimal clustering method for a specific clustering problem can be considered an art.[27]

A variety of clustering techniques have been used to identify phase diagrams from structure data. Hierarchical clustering, a method that clusters based purely on dissimilarity, was used with the Pearson metric to cluster 1D XRD data into phase regions.[17] Spectral clustering is a clustering method that uses the dissimilarity matrix to identify manifolds of sample connectivity and then assigns samples to clusters based on shared manifolds. Spectral clustering was used in Ref. 20 with the cosine metric to perform an initial clustering of structure data (either XRD or Raman spectra data) before performing a more complex hybrid technique (described below). Both implementations of hierarchical and spectral clustering focused on clustering the samples by using only the structure data and do not take into account dissimilarities in material composition, a factor of prior knowledge of the material space. By ignoring sample composition, samples with similar structure but significantly different composition can be clustered together—i.e., grouped together into the same phase region, resulting in non-realistic disconnected phase regions.

The use of composition data to cluster samples was introduced in an on-the-fly analysis algorithm,[28] which uses the mean shift theory clustering method to tie cluster identification to both structure data space and composition data space. While this approach provides a soft constraint for cohesive phase regions in composition space, reasoning methods (described below) allow for hard constraints.

b. Spectral decomposition/matrix factorization. Phase diagram determination is a combination of identifying fabrication parameter space phase regions as well as identifying the constituent phases of these regions. Constituent phase identification is performed using spectral decomposition methods such as matrix factorization. Matrix factorization was first introduced into phase identification by Long *et al.* in Ref. 22 and has become a key part of phase diagram determination. Matrix factorization assumes that the structure data y_i (XRD, Raman spectra, etc.) for each sample i can be described by a mixture of structure data associated with a set of phases $X = x_j, j \in \{1 \ldots M\}$ with phase abundances β and an error term ϵ. Also, as both XRD and Raman spectra data are required to be non-negative (minimum of zero) and the abundances must also be non-negative, a non-negative constraint is applied

$$\boldsymbol{y}_i = \sum_{j=1}^{M} \boldsymbol{\beta}_j \boldsymbol{x}_j + \epsilon \quad \beta, x \in \mathbb{R}_{\geq 0},$$

where β and x are the scalar values that make up the vectors $\boldsymbol{\beta_j}$ and $\boldsymbol{x_j}$.

For a set of samples from the same material space, e.g., a set of samples from a composition spread, the matrix representation is given by

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \epsilon \quad \beta, x \in \mathbb{R}_{\geq 0}.$$

Here, row i of Y and X corresponds to the sample data y_i and phase data x_i, respectively. Due to the non-negative constraint, this method is called non-negative matrix factorization (NMF). Various techniques exist for evaluating $\boldsymbol{\beta}$ and $\boldsymbol{X}$ with the shared goal of minimizing error ϵ. Matrix factorization can be used when the phase data X are completely unknown (unsupervised learning) and when partial information is known (semi-supervised learning). For example, phases known to exist in the material system can be used to populate X with fixed valued rows while additional rows are evaluated for the remaining unknown phases. The dimensionality reduction algorithms described in the Data Visualization and Manual Analysis section produce basis spectra that are analogous to the components X, but while basis spectra produced by algorithms such as PCA make no direct connection to crystalline phases, the NMF components may produce a single basis spectrum for each phase in the system. Pure-NMF techniques reported in the literature have largely been unable to track peak-shift arising from alloying. New approaches to appropriately model peak shifting are needed with recent progress in this area discussed in the Hybrid Algorithms section below.

c. Constraint reasoning. When expert analysts formulate phase diagrams from XRD patterns, they ensure that several well-defined criteria are met. For example, every XRD peak must be

associated with a phase and the composition region in which a phase is observed must be connected. The shortcomings of purely clustering or purely matrix factorization techniques are that a resulting phase diagram (the optimal solution from a given algorithm) may violate the logical requirements that are rooted in the underlying materials physics. To formally adhere to the physical constraints,[23,31,34] expressed the phase mapping problem using a set of logical requirements within a constraint reasoning[24] and then a satisfiability modulo theory (SMT) reasoning framework. The SMT algorithm was demonstrated using synthetic data and reproduced the ground truth phase diagrams with excellent fidelity, but the implementation of this approach suffers from practical limitations. The algorithm is computationally intensive for large data sets and perhaps more limiting is the reliance on accurate peak detection in every XRD pattern, which creates sensitivity to experimental noise. While experimental data sets have yet to be solved by this purely reasoning approach, the formal logical statement of the problem and the demonstrated utility of constraint reasoning provide a critical departure from the purely Clustering and Matrix Factorization algorithms.

d. Hybrid algorithms. Recently reported algorithms, which constitute the state of the art, have made fundamental modifications to clustering and matrix factorization techniques by encoding prior knowledge (typically physical constraints) such that a resulting phase diagram is the optimal solution that satisfies specific requirements. To date, the encoded constraints are a subset of those outlined in the pure-reasoning approach, motivating the depiction of these algorithms in Figure 2 as a mixture of clustering, matrix factorization and reasoning.

e. Constraint programming (CP)-Clustering. An early approach for correcting the shortcomings of clustering algorithms was developed by LeBras *et al.*[18] in which a constraint programming (CP) model was developed to represent the reasoning logic described above. Like the SMT reasoning framework, directly solving the CP model proved to be problematic, prompting the development of a hybrid approach which iterated between clustering XRD patterns based on their similarity and applying the CP formulation of the physical constraints. The clustering step was used to rationally define data subsets, which were analyzed using the CP framework, and then these data subsets were stitched together using as global CP model to arrive at global solution.

f. GRENDEL. GRENDEL is a hybrid approach that combines clustering and non-negative matrix factorization, allowing information from each method to improve results for the other.[20,35] This hybrid method also allows for additional constraints to be imposed. Specifically, a material system in thermal equilibrium will have discrete phase regions, which bound the existence of individual phases and vice versa, the existence of particular phases describe the boundaries of phase regions. However, NMF assumes that each phase can appear anywhere throughout the fabrication parameter space. GRENDEL combines a graph-cut clustering method with NMF to ensure that phase regions are cohesive in composition space and that phases are bound to phase regions (clusters) and that samples are placed in the phase region (cluster) whose constituent phases best describe the sample. The combination of methods is performed using an objective function, which permits for additional constraints through regularization terms, such as ensuring that the volume described by each cluster's constituent phases is minimized, thus imposing a soft constraint that the identified phases look as much like the structure data as possible. This method has been shown to work with both XRD and Raman spectra data with speeds that allow for on-the-fly analysis.[20] A diagram of the workflow used in GRENDEL can be found in Figure S2 of the supplementary material.[36]

g. CombiFD. While GRENDEL uses objective function regularization to introduce soft physical constraints, the CombiFD[23] method represents physical constraints as a set of linear inequality equations, much like the Reasoning method of Ref. 18. This approach allows for tremendous flexibility with a wide range of potential constraints. CombiFD also allows for the use of commercial state-of-the-art optimization software. Ermon used the CombiFD framework to encode both phase region cohesivity and to impose an upper limit on the number of phases present in each sample, a requirement of Gibb's phase rule for materials in equilibrium. Ermon *et al.* also demonstrated that by appropriately setting parameters for the minimum peak width (typically known for a given XRD experiment) and maximum peak shifting (typically between 1% and 10% depending on the chemical system), monotonic peak shifting (all peaks shift in the same direction, as is the case for typical metal alloying) can also be appropriately modeled to avoid false identification of "shifted"

phases. For a system with hundreds of samples, evaluating the phase diagram can take minutes to a few hours.

D. Conclusions

The use of HTE approaches to rapidly map composition–structure–property relationships provides a great opportunity to expedite the rate at which new materials are discovered and commercialized. Solving the underlying phase behavior of a set of material samples enables the direct establishment of structure–property relationships in combinatorial experiments, and the development of algorithms that generate phase diagrams from combinatorial structural characterization data remains a prolific challenge for both materials and computer scientists. Here we have provided a summary of the state-of-the-art methods employed to overcome this hurdle by the HTE community. The development of faster and more reliable methods that appropriately model peak shifting, impose physical constraints (such as Gibbs Phase Rule), and reject instrumental data artifacts will help to realize prolific gains in the rate and efficiency of materials discovery. These high throughput capabilities are well-complemented by ongoing advances in the fields of high-throughput density functional theory and on-the-fly design of experiments, providing the key components for a materials discovery engine to propel a wide range of technologies, particularly emerging technologies with deep societal impact.

ACKNOWLEDGMENTS

J.H.S gratefully acknowledges support from Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award No. DE-AR0000492 and the SouthCarolina SmartState™ Center for Strategic Approaches to the Generation of Electricity (SAGE). J.M.G. acknowledges support from the Joint Center for Artificial Photosynthesis, a DOE Energy Innovation Hub, supported through the Office of Science of the U.S. Department of Energy Award No. DE-SC0004993, and the Computational Sustainability Network, supported through National Science Foundation Expeditions in Computing Grant No. 1521687. The authors thank Carla Gomes and Ronan LeBras for insightful discussions.

[1] P. Chen, "Electrospray ionization tandem mass spectrometry in high-throughput screening of homogeneous catalysts," Angew. Chem., Int. Ed. **42**, 2832 (2003).
[2] W. F. Maier, K. Stöwe, and S. Sieg, "Combinatorial and high-throughput materials science," Angew. Chem., Int. Ed. Engl. **46**, 6016 (2007).
[3] M. L. Green, I. Takeuchi, and J. R. Hattrick-Simpers, "Applications of high throughput (combinatorial) methodologies to electronic, magnetic, optical, and energy-related materials," J. Appl. Phys. **113**, 231101 (2013).
[4] J. Cui, Y. S. Chu, O. O. Famodu, Y. Furuya, J. Hattrick-Simpers, R. D. James, A. Ludwig, S. Thienhaus, M. Wuttig, Z. Y. Zhang, and I. Takeuchi, "Combinatorial search of thermoelastic shape-memory alloys with extremely small hysteresis width," Nat. Mater. **5**, 286 (2006).
[5] D. Kan, L. Pálová, V. Anbusathaiah, C. J. Cheng, S. Fujino, V. Nagarajan, K. M. Rabe, and I. Takeuchi, "Universal behavior and electric-field-induced structural transition in rare-earth-substituted $BiFeO_3$," Adv. Funct. Mater. **20**, 1108 (2010).
[6] D. J. Arriola, E. M. Carnahan, P. D. Hustad, R. L. Kuhlman, and T. T. Wenzel, "Catalytic production of olefin block copolymers via chain shuttling polymerization," Science **312**(5774), 714 (2006).
[7] M. L. Green, J. R. Hattrick-Simpers, C. Choi, I. Takeuchi, A. M. Joshi, S. C. Barron, T. Chiang, A. Davydov, S. Empedocles, J. M. Gregoire, and A. Mehta, Fulfilling the Promise of the Materials Genome Initiative via High-Throughput Experimentation, MRS, 2014, available at http://www.mrs.org/mgi-workshop-2014/.
[8] T. Kalil and C. Wadia, Materials Genome Initiative for Global Competitiveness, 2011, available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf.
[9] Materials Genome Initiative National Science and Technology Council Committee on Technology Subcommittee on the Materials Genome Initiative, 2014, available at https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf.
[10] I. Takeuchi, C. J. Long, O. O. Famodu, M. Murakami, J. Hattrick-Simpers, G. W. Rubloff, M. Stukowski, and K. Rajan, "Data management and visualization of x-ray diffraction spectra from thin film ternary composition spreads," Rev. Sci. Instrum. **76**, 62223 (2005).
[11] Y. Watanabe, T. Umegaki, M. Hashimoto, K. Omata, and M. Yamada, "Optimization of Cu oxide catalysts for methanol synthesis by combinatorial tools using 96 well microplates, artificial neural network and genetic algorithm," Catal. Today **89**, 455 (2004).
[12] U. Rodemerck, M. Baerns, M. Holena, and D. Wolf, "Application of a genetic algorithm and a neural network for the discovery and optimization of new solid catalytic materials," Appl. Surf. Sci. **223**, 168 (2004).

[13] S. K. Suram, J. A. Haber, J. Jin, and J. M. Gregoire, "Generating information-rich high-throughput experimental materials genomes using functional clustering via multitree genetic programming and information theory," ACS Comb. Sci. **17**, 224 (2015).
[14] S. Kumar, V. Gupte, and K. Sreenivas, "Structural and optical properties of magnetron sputtered $Mg_XZn_{1-X}O$ thin films," J. Phys.: Condens. Matter **18**, 3343 (2006).
[15] K. Rajan, "Combinatorial materials sciences: Experimental strategies for accelerated knowledge discovery," Annu. Rev. Mater. Res. **38**, 299 (2008).
[16] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, "Combinatorial and high-throughput screening of materials libraries: Review of state of the art," ACS Comb. Sci. **13**, 579 (2011).
[17] C. J. Long, J. Hattrick-Simpers, M. Murakami, R. C. Srivastava, I. Takeuchi, V. L. Karen, and X. Li, "Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis," Rev. Sci. Instrum. **78**, 072217 (2007).
[18] R. LeBras, T. Damoulas, J. M. Gregoire, A. Sabharwal, C. P. Gomes, and R. B. van Dover, "Constraint reasoning and kernel clustering for pattern decomposition with scaling," in *Principles and Practice of Constraint Programming - CP 2011* (Springer Berlin Heidelberg, 2011), pp. 508–522.
[19] J. P. MacSleyne, J. P. Simmons, and M. De Graef, "On the use of 2-D moment invariants for the automated classification of particle shapes," Acta Mater. **56**, 427 (2008).
[20] A. G. Kusne, D. Keller, A. Anderson, A. Zaban, and I. Takeuchi, "High-throughput determination of structural phase diagram and constituent phases using GRENDEL," Nanotechnology **26**, 444002 (2015).
[21] J. K. Bunn, S. Han, Y. Tong, Y. Zhang, J. Hu, and J. R. Hattrick-Simpers, "Generalized machine learning technique for automatic phase attribution in time variant high-throughput experimental studies," J. Mater. Res. **30**, 879 (2015).
[22] C. J. Long, D. Bunker, X. Li, V. L. Karen, and I. Takeuchi, "Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization," Rev. Sci. Instrum. **80**, 103902 (2009).
[23] S. Ermon, R. LeBras, S. K. Suram, J. M. Gregoire, C. P. Gomes, B. Selman, and R. B. van Dover, "Pattern decomposition with complex combinatorial constraints: Application to materials discovery," in *Proceedings of the 29th AAAI International Conference on Artificial Intelligence* (AAAI, 2014).
[24] J. Lee, "Principles and practice of constraint programming–CP 2011," in *Proceedings of the 17th International Conference on CP 2011, Perugia, Italy, 12–16 September 2011* (Springer Science & Business Media, 2011).
[25] R. LeBras, R. Bernstein, J. M. Gregoire, S. K. Suram, C. P. Gomes, B. Selman, and R. B. van Dover, "A computational challenge problem in materials discovery: Synthetic problem generator and real-world datasets," in *Proceedings of the 28th AAAI International Conference on Artificial Intelligence* (AAAI, 2014).
[26] J. M. Gregoire, D. Dale, and R. B. van Dover, "A wavelet transform algorithm for peak detection and application to powder x-ray diffraction data," Rev. Sci. Instrum. **82**, 015105 (2011).
[27] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York, 2009).
[28] A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long, and I. Takeuchi, "On-the-fly machine-learning for high-throughput experiments: Search for rare-earth-free permanent magnets," Sci. Rep. **4**, 6367 (2014).
[29] L. A. Baumes, M. Moliner, N. Nicoloyannis, and A. Corma, "A reliable methodology for high throughput identification of a mixture of crystallographic phases from powder x-ray diffraction data," CrystEngComm **10**, 1321 (2008).
[30] Y. Iwasaki, A. G. Kusne, and I. Takeuchi, "Comparison of various metrics for cluster analysis of combinatorial x-ray diffraction data," Nature Computational Materials (submitted).
[31] R. LeBras, R. Bernstein, C. P. Gomes, B. Selman, and R. B. van Dover, "Crowdsourcing backdoor identification for combinatorial optimization," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence, International Joint Conference on Artificial Intelligence* (AAAI, 2012).
[32] C. Suh, A. Rajagopalan, X. Li, and K. Rajan, "The application of principal component analysis to materials science data," Data Sci. J. **1**, 19 (2002).
[33] T. Mueller, A. G. Kusne, and R. Ramprasad, "Machine learning in materials science: Recent progress and emerging applications," Rev. Comput. Chem. **29**, 186 (2016).
[34] S. Ermon, R. LeBras, C. P. Gomes, B. Selman, and R. B. van Dover, "Smt-aided combinatorial materials discovery," in *Theory and Applications of Sustainability Testing–SAT 2012* (Springer, 2012), p. 172.
[35] P. Massoudifar, A. Rangarajan, A. Zare, and P. Gader, "An integrated graph cuts segmentation and piece-wise convex unmixing approach for hyperspectral imaging," *6th IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)* (IEEE, 2014).
[36] See supplementary material at http://dx.doi.org/10.1063/1.4950995 for the AutoPhase workflow diagram in Figure S1 and the GRENDEL workflow diagram in Figure S2.
[37] Certain commercial equipment, instruments, or materials are identified in this report in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.