ACS Combinatorial Science

Research Article

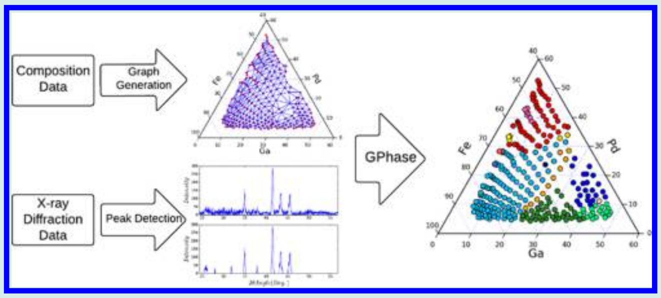
pubs.acs.org/acscombsci

Automated Phase Segmentation for Large-Scale X-ray Diffraction Data Using a Graph-Based Phase Segmentation (GPhase) Algorithm

Zheng Xiong,[†] Yinyan He,[†] Jason R. Hattrick-Simpers,[‡] and Jianjun Hu*,[†,§]

[†]Department of Computer Science and Engineering and [‡]Department of Chemical Engineering, University of South Carolina, Columbia, South Carolina 29208, United States

[§]School of Mechanical Engineering, Guizhou University, Guiyang, Guizhou 550025, China



ABSTRACT: The creation of composition−processing−structure relationships currently represents a key bottleneck for data analysis for high-throughput experimental (HTE) material studies. Here we propose an automated phase diagram attribution algorithm for HTE data analysis that uses a graph-based segmentation algorithm and Delaunay tessellation to create a crystal phase diagram from high throughput libraries of X-ray diffraction (XRD) patterns. We also propose the sample-pair based objective evaluation measures for the phase diagram prediction problem. Our approach was validated using 278 diffraction patterns from a Fe−Ga−Pd composition spread sample with a prediction precision of 0.934 and a Matthews Correlation Coefficient score of 0.823. The algorithm was then applied to the open Ni−Mn−Al thin-film composition spread sample to obtain the first predicted phase diagram mapping for that sample.

KEYWORDS: *graph segmentation, phase segmentation, high-throughput experiments, phase diagram, X-ray diffraction*

■ INTRODUCTION

The Materials Genome Initiative (MGI) has the stated purpose of shortening the time between the discovery and commercialization of advanced materials[1] by seamlessly linking computational modeling and experimentation with advanced data analytics to extract underlying trends from data sets. The availability of curated large-scale experimental materials property databases and effective analytical tools to mine them represent vital components to the aspirations of the MGI. Of particular importance to the goal of the MGI is the mapping of composition−processing−structure−property relationships across a wide spectrum of technologically important materials systems. High-throughput experimental (HTE) studies, where thousands of samples are synthesized, processed, and screened for their figure of merit in a single experiment, represent the key enabling experimental technology to the MGI.

Over the past decade, HTE approaches have been widely adopted by academic and industrial groups. Technologies discovered using this approach include the dielectric gate stack used in the iPhone 5 and InFuse Polymers.[2] Characterizing the crystal structure of a material as a function of its composition and environment is a general problem for the entire HTE community and can provide important experimental validation of theoretical predictions. Further, tools developed that can automatically analyze high-dimensional structural data would be of general use to the wider materials community, especially those performing structural determinations. Of particular interest to the community are tools that can use X-ray diffraction data to automatically perform phase attribution and create composition-structure-processing phase diagrams. A typical HTE X-ray diffraction experiment involves the acquisition of anywhere from 500 to 40 000 diffractograms, far too many for scientists to parse through individually. A large variety of synthesis methods for HTE samples exist but broadly they can be composed of discrete compositions or as continuously varying compositional spreads or composition wedges.

Phases present in a sample are characterized by the peak distribution from X-ray diffraction patterns. In general, the peak position and intensity within the single phase region is a continuous function of the composition of the solid solution, shifting left or right depending upon the relative size of the substituted atoms to those it displaces. At the edge of a phase boundary, abrupt changes occur in the diffraction pattern. First, peaks that had been decreasing in intensity as a function of composition can vanish, indicating that a phase is no longer present in X-ray diffraction measurable quantities in the sample. Second, new peaks can abruptly appear, which are indicative of the formation of new phases in the sample. Finally, the peaks may stop shifting at a phase boundary, as it is no longer capable of incorporating the new substituent into its crystal lattice. The high complexity of the peak distribution of phase mixtures makes it extremely difficult and tedious for humans to conduct high-throughput and accurate phase attribution manually.

Over the past ten years, an increasing number of groups have been working to develop effective data visualization, minimization and automated data analytics tools to address this problem.[3−18]

Received: August 4, 2016
Revised: January 23, 2017
Published: January 26, 2017

DOI: 10.1021/acscombsci.6b00121

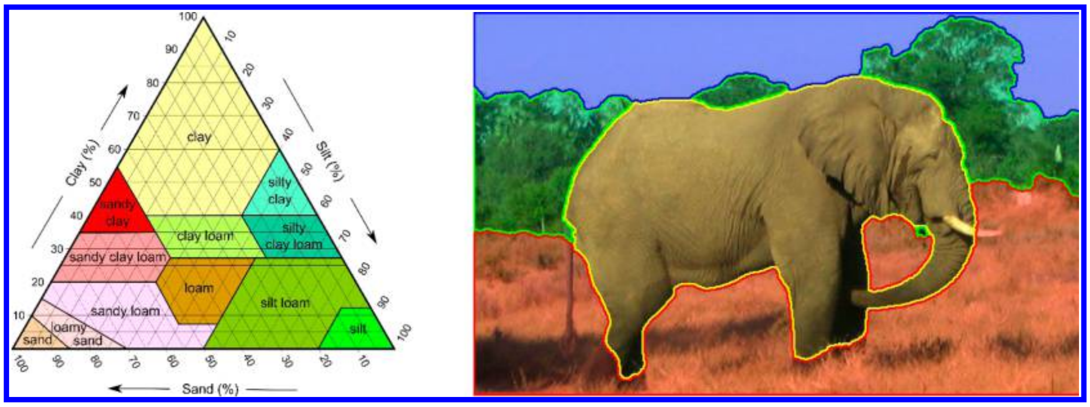


Figure 1. The similarity between phase diagram mapping and image segmentation: (a) Ternary phase diagram sample.[31] (b) Image segmentation sample.

The literature to date can be broadly divided into unsupervised, supervised and semisupervised approaches. Unsupervised techniques can involve the identification of basis vectors via non-negative matrix factorization, calculation of average distances using weighting functions to form dendrograms, or combinations thereof.[19−23] These techniques often suffer from their inability to track peak shift due to alloying and require human expertise to determine if the algorithm was effective in discriminating between patterns. We recently developed AutoPhase, a supervised machine learning algorithm using AdaBoost to identify the presence of different metal and oxide phases in combined X-ray diffraction and Raman spectroscopy data during high temperature oxidation of Ni−Al bond coats.[24] More recently, semisupervised techniques meant to overcome issues with peak shift and selection of training sets have been developed.[21,25] A recent review discusses the strengths and weaknesses of the respective techniques in greater detail.[26]

One method for mitigating the issue of peak shift and creating more physically intuitive phase mappings is including the constraint that small change of composition parameters will lead to a small change of X-ray diffraction. Suram[27] recently proposed a clustering algorithm based on information potential criterion and genetic programming, which may ensure the continuity of the composition regions to some extent. Although here, phase regions were defined via the scalar figure of merit measurements, rather than multidimensional diffraction patterns. Kusne[18] performed phase attribution by incorporating mean shift theory, which partially overcomes the issue of peak shifting, into a machine learning algorithm that was supervised using existing diffraction patterns. In another approach,[25] a conventional spectral clustering is used to generate a seed phase diagram, which is then refined based on a graph-cut algorithm to improve the cluster connectivity by minimizing an objective function. However, the seed phase diagrams with disconnect phase regions may put severe constraints on what can be done in their graph-cut step. Graphs of the sample points have also been used by the constraint programming approaches for phase identification. Here the Delaunay triangulation procedure is used to embed the samples into a connectivity graph, which is used to define the shift continuity and connectivity constraints used by the constraint programming solver.[7,21,28,29]

Different to Kusne's spectral clustering with the graph-cut approach, here we use ideas in image segmentation algorithms in computer vision to solve the phase diagram mapping problem. As shown in Figure 1, phase segmentation can be mapped to the image segmentation problem. The problem of segmenting an image into regions is a problem to cluster image pixels to different objects. A fundamental requirement for image segmentation is the continuity of pixels of the partitioned objects, which matches closely to the phase attribution problem. Similarly, phase diagram samples can be seen as pixels in an image. Image segmentation techniques can be classified into threshold based methods, edge detection methods, region based methods, pixel classification/clustering methods, texture-based segmentation, graph-based segmentation algorithm with region growing and model-fitting refinement steps to solve the phase diagram segmentation problem.[30] Among these methods, the graph-based segmentation algorithm is selected for phase segmentation due to its flexibility and power for dealing with continuity requirement.

■ METHODS

Graph-Based Phase Segmentation (GPhase) Algorithm. Our graph-based phase segmentation (GPhase) algorithm combines Delaunay tessellation, automated XRD peak detection and graph-based image segmentation to create connected phase diagrams. Figure 2 provides a flowchart for the data process of GPhase. Composition data is first tessellated by Delaunay tessellation to generate a composition space graph where each sample is a vertex and the neighboring samples are connected by edges. Peak detection is then performed on the structure data, here in the form of 1D XRD data. The dissimilarity measure, determined by the Euclidean metric between XRD peak data for two samples, serves as the weight of neighboring samples in the composition space graph. The weights between neighboring samples are then fed into the efficient graph-based segmentation algorithm[32] to determine a phase diagram.

Composition Space Graph Generation. The first step of phase segmentation is to generate a composition space graph by composition data. For ternary materials system, the composition data are (x, y, z) composition percentages for three elements. Our goal is to extract the composition neighboring relationships of the samples, which corresponds to the planar geometric neighboring relationship of the pixels in an image. Here Delaunay tessellation, which generates well-connected neighboring samples was applied.[33] Given a set of points P, the Delaunay triangulation of P is a particular triangulation build on P, in which the circum-circle of each simplicial cell in the triangulation does not contain any input point $p \in P$. Delaunay triangulations is implemented by maximizing the minimum angle of all the angles of the triangles in the triangulation. Delaunay triangulations can, however, attempt to connect samples with too dissimilar compositions. Therefore, a user-defined maximum angle was set as the threshold to avoid

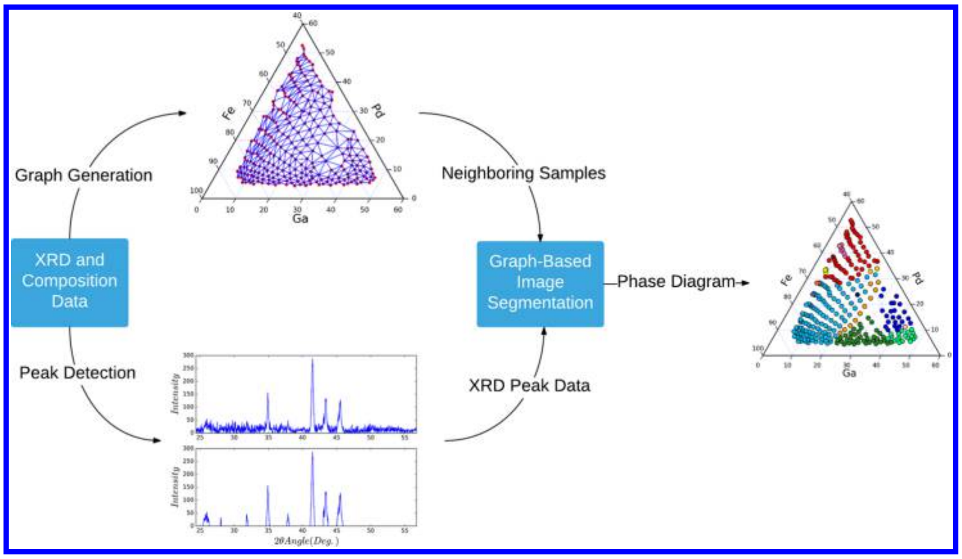


Figure 2. Flowchart of GPhase algorithm. Composition and XRD are collected from the composition spread combinatorial library. The composition data is tessellated to generate a composition space graph. The XRD data is then processed through a wavelet transform peak detection algorithm. A graph-based image segmentation algorithm is then used to do phase diagram determination.

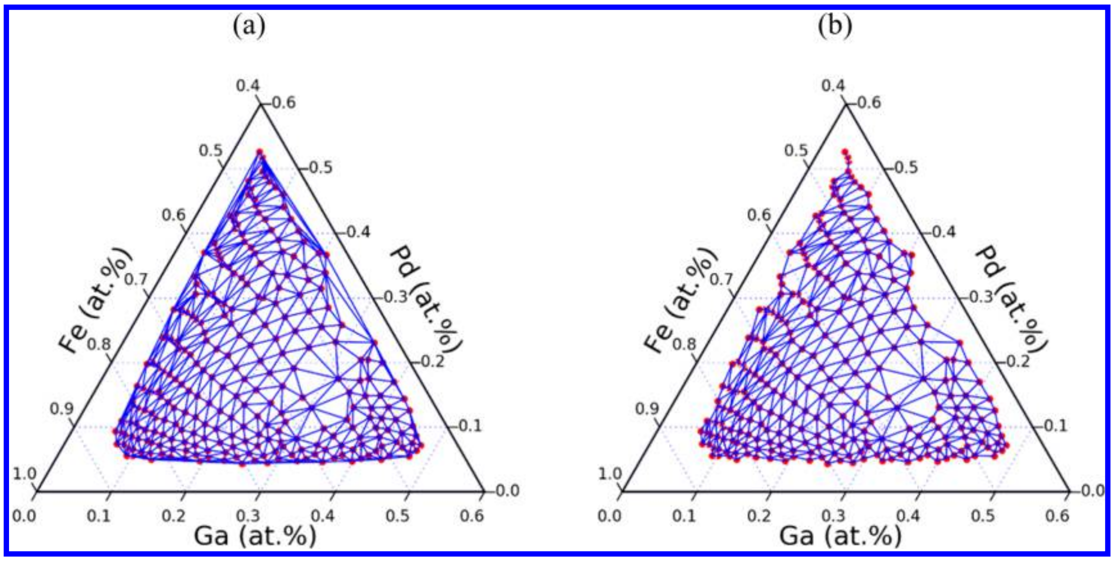


Figure 3. Composition graphs generated with (a) basic tessellated composition space graph generation approach and (b) modified tessellated composition space graph generation approach.

overlong triangles. Figure 3 shows the tessellated Fe−Ga−Pd composition space graph without (Figure 3a) and with (Figure 3b) the maximum angle constraint.

Peak Detection. A peak detection algorithm (peakfindsb) previously reported by O'Haver[34] was used to convert each spectrum vector into peak vectors containing the peak position, shape, and magnitude. Briefly, peak detection in the presence of instrumental noise is performed by first smoothing the slope of the data and then a sliding window is used to detect when the slope transitions from positive to negative. False peaks are removed from the list by ensuring that both the slope and intensity of the peak exceed some threshold values. To prevent smoothing from distorting peak positions, the final fit of the Gaussian peaks is performed on the original data set. From our experiments, it is found that this peak detection algorithm is effective in detecting peaks related to phase mapping.

The main purpose of peak detection is to remove the background signals before calculating the dissimilarity between two XRD samples (Figure 2). For each sample, after all its peaks are detected, the intensity values of the 2θ sampling points that are NOT covered by any peak region (the range centered by the peak position with the Gaussian width on both sides) will be set to zero. The resulting filtered data of the XRD samples will then be used for calculating sample dissimilarities.

Measures for Comparing XRD Patterns between Samples. A dissimilarity measure is used to quantify the difference of XRD patterns between two samples and will be used as the weights in composition space graph. If two compositionally neighboring samples belong to the same phase region, their dissimilarity measure should be small compared to the phase boundary samples. There are several ways to calculate pairwise dissimilarity. Here, it is defined as the Euclidean distance of the 1D XRD peak patterns of two neighboring samples in the composition space graph. If $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$ are 1D XRD filtered intensity vectors for sample p and q respectivley. The dissimilarity is represented as

$$\begin{aligned} D_{ij} &= d(p, q) \\ &= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \end{aligned}$$

Where D_{ij} is the Euclidean metric of dissimilarity between the XRD peak data for samples p and q; n is the number of (2θ) sampling points of the samples.

Graph-Based Segmentation. The phase diagram determination is equivalent to finding all the edges with significant distances and to disconnect the graph so that multiple connected regions will be obtained with each corresponding to one phase region. This can be mapped to the image segmentation problem, which can be solved by efficient graph-based image segmentation algorithm as proposed by Felzenszwalb.[32] This algorithm is closely related to Kruskal's algorithm for constructing a minimum spanning tree of a graph. Instead of segmenting image pixels in image segmentation, XRD patterns will be segmented in our phase segmentation algorithm based on the idea of minimum spanning tree.

Formally, we define $G = (V, E)$ as an undirected graph, of which the vertices are $v_i \in V$, the set of samples to be segmented, and the edges $(v_i, v_j) \in E$ are the pairs of neighboring vertices, depending on the neighboring relationship in the composition space graph. The corresponding weight $w((v_i, v_j))$ for each edge $(v_i, v_j) \in E$ is a non-negative measure of the dissimilarity between neighboring vertices v_i and v_j, which is the Euclidean distance of the 1D XRD peak patterns of two samples in the composition space graph. A phase segmentation S is defined as a partition of V into phase regions. The weight (dissimilarities) of the edges in the same region are expected to be small, i.e., relatively low weights between vertices inside the region and the ones between different regions are expected to be more dissimilar, i.e., relatively higher weights for the edges connecting two different regions.

The intra- and interdifference of regions are defined as follows. The intradifference of a region $C \subseteq V$ is the largest weight in the minimum spanning tree (MST) in this region, that is

$$\mathrm{IntraDiff}(C) = \max_{e \in \mathrm{MST}(C, E_c)} w(e)$$

The interdifference between two regions C_1, $C_2 \subseteq V$ is the minimum weight of all the edges connecting these two regions. That is,

$$\mathrm{InterDiff}(C_1, C_2) = \min_{v_i \in C_1, v_j \in C_2, (v_i, v_j) \in E} w((v_i, v_j))$$

$\mathrm{InterDiff}(C_1, C_2)$ will be defined as $+\infty$ if there is no edge between two regions. Intuitively, two regions are believed to be similar if the interdifferences are small compared to the intradifferences in these two regions to some extent. The comparison is defined as,

$$\begin{aligned} &\mathrm{InterDiff}(C_1, C_2) < \min(\mathrm{IntraDiff}(C_1) + \tau(C_1), \\ &\quad \mathrm{IntraDiff}(C_2) + \tau(C_2)) \end{aligned}$$

where $\tau(C) = k/|C|$, $|C|$ represents the number of vertices in C and k is a threshold parameter for controlling the degree how much the interdifference must be smaller than the intradifference to declare two regions can be merged into one region. The value of $\tau(C)$ for larger regions will be smaller, making them less likely to be merged. The constant k is a key parameter to control how large the regions would be generated by the algorithm. Larger k value tends to create larger regions, as the comparison threshold becomes easier to achieve.

On the basis of this definition of comparison, it has been proved that there exists some segmentation S that is neither too coarse nor too fine for any graph $G = (V, E)$.[32] In addition, the segmentation S produced by the following algorithm is not too fine nor too coarse.[32]

The procedure of phase segmentation is as follows: Construct an undirected graph $G = (V, E)$, with n vertices and m edges, where vertices $v_i \in V$ are the samples to be segmented, edges $(v_i, v_j) \in E$ are vertices pairs, of which the neighboring relationship comes from the composition space graph and the values depend on the Euclidean distance between XRD peak patterns. The output is a phase segmentation S.

1. Sort E by nondecreasing edge weights.
2. Start with a segmentation S_0, where each vertex is in its own region.
3. Construct S_q given S_{q-1} as follows. Let V_i and V_j represent the vertices connecting qth edge in the ordering. If V_i and V_j are in disjoint regions C_1 and C_2 of S_{q-1} and weight (V_i, V_j) is small compared with the internal difference of both those regions in threshold function τ, that is,

   $$\begin{aligned} &w((v_i, v_j)) < \mathrm{minimum}(\mathrm{IntraDiff}(C_1) + \tau(C_1), \\ &\quad \mathrm{IntDiff}(C_2) + \tau(C_2)) \end{aligned}$$

   then merge the two regions; otherwise, do nothing.
4. Repeat step 3 for $q = 1, ..., m$.
5. Return $S = S_m$.

Performance of Evaluation. To evaluate the result of our GPhase algorithm, the standard performance measures of pattern clustering including precision, recall, accuracy, and MCC (Matthews Correlation Coefficient) will be calculated. There are several approaches to evaluate clustering algorithm performances. One effective approach is to view clustering as a series of decisions, one for each of the $N(N - 1)/2$ pairs of sample in the collection.[35] We want to assign two samples to the same phase (cluster) if and only if they are similar. A true positive decision assigns two similar samples to the same phase, a true negative decision assigns two dissimilar sample to different phases. More formally,

- A true positive (TP) test result is one that detects the condition when the condition is present. In phase attribution, it is a sample pair that is correctly predicted by the algorithm as belonging to the same phase.
- A true negative (TN) test result is one that does not detect the condition when the condition is absent. In phase attribution, it is a sample pair that is correctly predicted by the algorithm as not belonging to a common phase.
- A false positive (FP) test result is one that detects the condition when the condition is absent. In phase attribution, it is a sample pair that is incorrectly predicted by the algorithm as belonging to a common phase.
- A false negative (FN) test result is one that does not detect the condition when the condition is present. In phase attribution, it is a sample pair that is incorrectly predicted by the algorithm as not belonging to a common phase.

Table 1. Phase Segmentation Results for All Datasets with Sample Number, Phase Number, Cluster Number, Accuracy, Recall, Precision, and MCC

data set	sample no.	phase	clusters	precision	recall	accuracy	MCC	k
Fe−Ga−Pd	278	6	15	0.934082	0.807942	0.930109	0.823406	1.5
Ni−Mn−Al	535	unlabeled	21	N/A	N/A	N/A	N/A	1.5

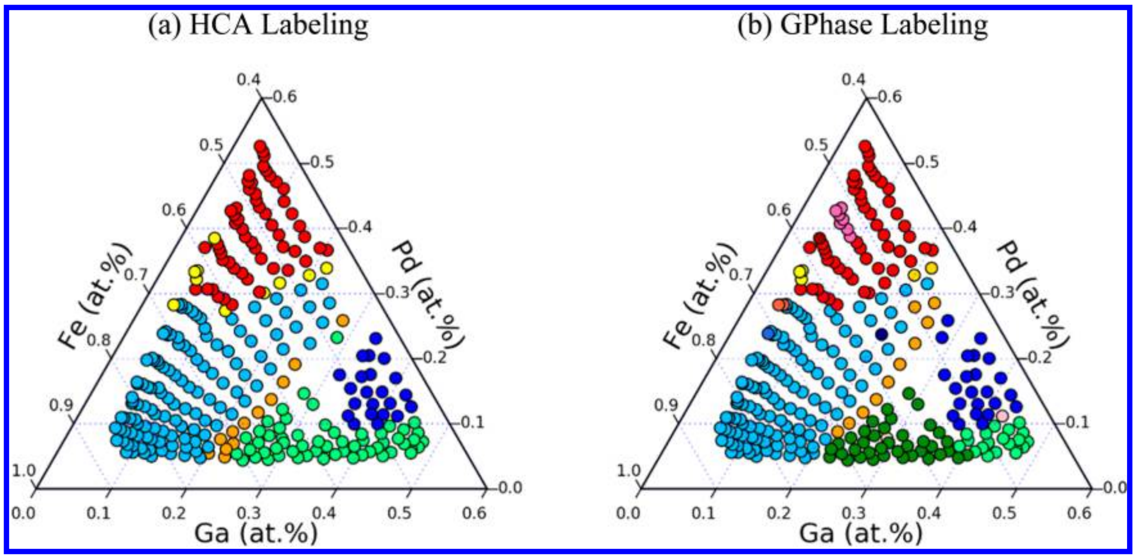


Figure 4. Phase diagram results generated from the Fe−Ga−Pd data set compared with the HCA labeling phase diagram: (a) HCA labeling. (b) GPhase labeling.

For our phase segmentation problem, for each sample against each phase, we will compare its phase attribution to the true phase and classify it into TP/TN/FP/FN. And then precision, recall, accuracy, and MCC are calculated as

- Precision = TP/(TP + FP).
- Recall = TP/(TP + FN).
- Accuracy = (TP + TN)/(P + N).
- $\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$

Precision (also called positive predictive value) is the fraction of sample pairs predicted as same-phase are truly belonging to the same phase, while recall (also known as sensitivity) is the fraction true same-phase sample pairs that are predicted out of all possible such pairs. All these four measures are defined based on the four basic test statistics defined above. These external criteria for evaluating phase prediction performance make it possible to automatically and objectively compare different phase attribution algorithms without manual inspection and matching of the predicted phases against the labeled phases. Such matchings are sometimes very ambiguous when the number of predicted phases is different from the labeled phases.

■ RESULTS

GPhase was run on two open combinatorial diffraction data sets from the Fe−Ga−Pd and Ni−Mn−Al ternary material systems. The details of the synthesis and characterization of the combinatorial libraries can be found elsewhere.[20,36] Briefly, both samples were prepared by sputtering elemental or in the case of Ga, solid solution targets onto 3 in. Si wafers and then vacuumed annealed at 650 °C for 2 h. Diffraction for both wafers was taken using a Bruker D8 with GADDS detector with a Cu Kα source focused to a nominal spot size of 0.5 mm. The 2D diffraction pattern was χ integrated to produce a 1D intensity versus 2θ diffractograms for each position scanned on the wafer. The Fe−Ga−Pd and Ni−Mn−Al data sets contained 278 and 535 individual diffractograms, respectively.

Both sets are available for free online as sample data sets provided with the Combiview, a software for visualization and clustering analysis of XRD data from combinatorial thin film libraries,[37] including the position of each point on the combinatorial wafer, the composition at each point, and the associated diffraction pattern for each point. The Fe−Ga−Pd data has been previously labeled by HCA[20] and has been extensively used to validate new machine learning techniques.[24] Conversely, the Ni−Mn−Al data set has not been previously labeled in detail and thus makes for a good check against overfitting of the parameters used during the peak-finding and graph-based segmentation.

Table 1 shows the performance of our GPhase algorithm for the Fe−Ga−Pd data set for what was determined to be the optimal number of clusters (15). The HCA labeled set contained 6 different phases, and here, 15 clusters were needed accurately reflect the HCA labeled data set. The precision was 0.934, indicating that there was a high probability that if a phase was attributed to a particular region then it was correctly assigned. The recall was 0.807, which indicates a high probability that a large portion of a region containing a specific phase will be identified correctly. The MCC score of 0.823 indicates that accuracy of the predictions was largely not impacted by the highly biased Fe−Ga−Pd data set, in which one phase region contains much more samples than other regions. In general, peak intensity still affected the overall predictive power of this algorithm, likely due to low intensity peaks being missed by the automated peak finding/fitting implementation.

The HCA labeled and GPhase predicted Fe−Ga−Pd phase diagrams are shown in Figure 4. It can be seen that overall, GPhase performs well across the entire phase diagram. It correctly predicts the extent of the α-Fe phase region, while

DOI: 10.1021/acscombsci.6b00121
ACS Comb. Sci. 2017, 19, 137−144

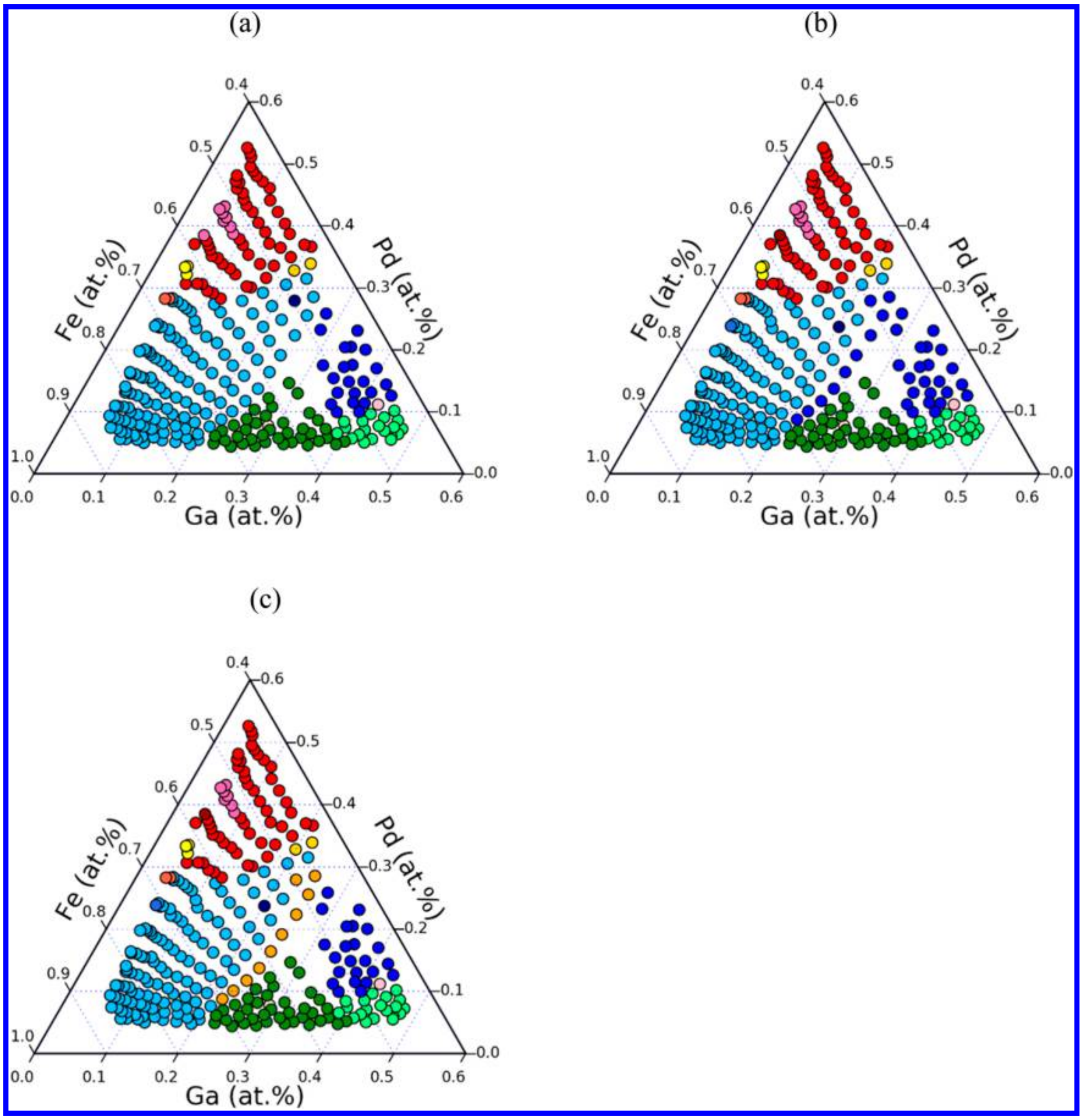


Figure 5. Fe−Ga−Pd clustering analysis from GPhase resulting in (a) 11, (b) 14, and (c) 15 clusters.

also correctly identifying the position of the α-Fe/FCC Fe−Pd boundary. It also successfully labels the region containing two unknown phases[19,24] between 30 and 55 at % Fe. Also, along the Fe−Ga binary edge it identifies an unknown phase for Ga concentrations exceeding 40 at %. This prediction is more inline with predictions from non-negative matrix factorization (NMF),[18] which was previously shown to produce phase labels that differ from HCA in the Fe−Ga−Pd system.

To explore the impact of the cluster number on the accuracy of the predicted phase diagram, a series of predictions were performed where the number of clusters was varied between 12, 14, and 15 (Figure 5). The number of clusters from these predictions were controlled by changing the "*k*" values in the threshold function to 1.7 for 12 clusters, 1.6 for 14 clusters, and 1.5 for 15 clusters. The results were then compared to results from HCA and NMF. Figure 5a shows that when 12 clusters are used the majority of the phase diagram is described by 6 clusters, with the other 6 clusters containing fewer than 5 points each. Note that the Ga-rich portion of the phase diagram has already deviated from the HCA predictions. This attribution of a distinct phase is supported by results from NMF and is likely resulting in an artificial suppression of the recall and MCC values reported above. Upon increasing the number of clusters to 14 in Figure 5b, the memberships of the clusters change slightly, particularly for compositions containing between 20 and 30 at % Ga and 8 and 30 at % Pd, but no new substantial clusters emerge. It should be noted that the value of *k* not only influences the number of clusters that can form, but can also impact cluster membership, particularly at the cluster edges, as it is used to calculate the internal differences. Once the number of clusters is increased to 15 that set of compositions forms its own cluster, closely matching the behavior observed in HCA studies.

Overall, GPhase performs as well as HCA in determining the individual phase regions, while preserving information about compositional linkages. Our GPhase algorithm always creates connected phase regions by design without any **repairing** step as needed in other approaches. The latter point is important when exploring new materials systems, as in our experience, depending upon the distance metric and linkage type chosen when performing HCA, it is common to observe compositionally disjoint clusters. Further, the peak detection and fitting act to remove noise and the background and increase the contrast between diffraction patterns with low signal-to-noise ratios. This enables the detection of the phase region in the Ga-rich portion of the phase diagram.

After validating that GPhase could reproduce HCA labeling of the Fe−Ga−Pd ternary system, it was applied to the phase mapping of a previously unlabeled Ni−Mn−Al combinatorial data set. Figure 6 presents the phase diagram generated when a k value of 1.5 is used, which here corresponds to 21 distinct clusters. To date, no expert labeled phase mapping of this sample has been published in the literature so only qualitative comparisons can be made to previously reported phase diagrams and functional properties. For instance, a previous study of the mechanical properties of Ni−Mn−Al thin film samples measured via nanoindentation provides confirmation of the relative positions of the phase regions. The large blue colored region in the center of the predicted phase diagram can be correlated to the β−Ni−Mn−Al solid solution phase in the equilibrium phase diagram. In particular, the regions identified as being within the austenitic and martensitic regions of the phase diagram, is captured by the current phase mapping. Meanwhile, the Al-rich portion of the phase diagram is labeled as being structurally distinct from β−Ni−Mn−Al, this agrees qualitatively with previous measurements of hardness via nanoindentation measurements on the same sample, where a marked decrease in hardness is associated with the Al-rich corner of the mapped ternary. Interestingly, it is clear from the diffraction patterns that the Ni-corner of the phase diagram exhibit a substantial interface reaction with the Si substrate, forming Ni-silicides. Here, GPhase is clearly biased by a combination of the intensity of the most intense peak at 43.98° 2θ and the slow variation in intensity of the silicides as the Mn concentration is increased. More detailed analysis is currently being performed to further refine the phase mapping of this Ni−Mn−Al sample.

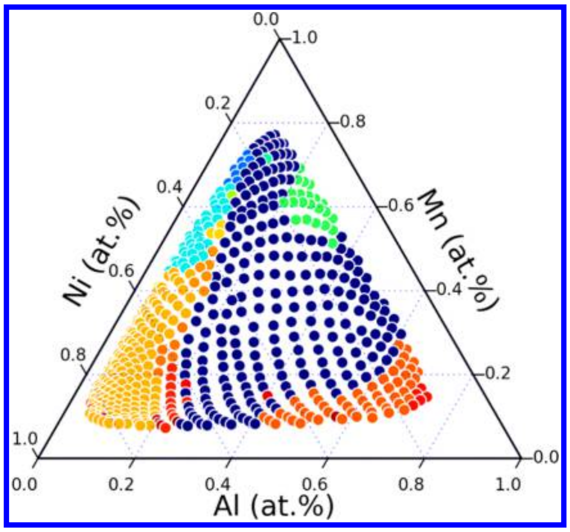


Figure 6. Phase diagram results for the unlabeled Ni−Mn−Al data set by GPhase algorithm.

Another noticeable advantage of GPhase algorithm compared to previous approaches is that our algorithm's running time is quite reasonable. With about 300 samples each with 1600 sampling points, the running time of GPhase is about 5−8 min running on a Desktop with 3.8 GHz CPU. The majority of the running time is used for the peak detection, which may be reduced further using less time-consuming peak detection procedure.

CONCLUSIONS

Here we demonstrated the application of a graph-based phase segmentation algorithm, GPhase, for unsupervised phase attribution that accounts for compositional similarity in cluster attribution. We showed that when comparing to HCA labels we were able to obtain precision 0.934, recall 0.808 and a MCC of 0.823. Interestingly, we found that our method of preprocessing the data substantially improved our ability to discern the presence of an Al-rich phase in the phase diagram that is not observed via HCA using Pearson's weighting and a centroid linkage. GPhase was then applied to the Ni−Mn−Al combinatorial structure data set to create a tentative phase mapping. The obtained mapping was in qualitative agreement with previously reported mechanical properties, as measured through nanoindentation. Compared to previous approaches for phase mapping, GPhase is able to handle the disjoint region issue (the connectivity issue) and the shift continuity constraint of the identified phase maps. Graph partition based GPhase can also lead to more meaningful phase regions based on the boundary properties between neighboring phase regions. Interested readers can obtain the algorithm from Dr. Jianjun Hu at (jianjunh@cse.sc.edu) or try the online web service from http://mleg.cse.sc.edu/gphase to be released.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jianjunh@cse.sc.edu.

ORCID

Jianjun Hu: 0000-0002-8725-6660

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank Jonathan Kenneth Bunn and Benjamin Ruiz-Yi for their evaluation and interpretation of the experimental results.

REFERENCES

(1) Ward, C. Materials Genome Initiative for Global Competitiveness. *23rd Advanced Aerospace Materials and Processes (AeroMat) Conference and Exposition*, Asm: 2012.

(2) Arriola, D. J.; Carnahan, E. M.; Hustad, P. D.; Kuhlman, R. L.; Wenzel, T. T. Catalytic production of olefin block copolymers via chain shuttling polymerization. *Science* **2006**, *312* (5774), 714−719.

(3) Barr, G.; Dong, W.; Gilmore, C. J. High-throughput powder diffraction. II. Applications of clustering methods and multivariate data analysis. *J. Appl. Crystallogr.* **2004**, *37* (2), 243−252.

(4) Baumes, L. A.; Collet, P. Examination of genetic programming paradigm for high-throughput experimentation and heterogeneous catalysis. *Comput. Mater. Sci.* **2009**, *45* (1), 27−40.

(5) Berry, M. W.; Browne, M.; Langville, A. N.; Pauca, V. P.; Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis* **2007**, *52* (1), 155−173.

(6) Cunningham, G. J. *Application of cluster analysis to high-throughput multiple data types*; University of Glasgow, 2011.

(7) Ermon, S.; Le Bras, R.; Suram, S. K.; Gregoire, J. M.; Gomes, C. P.; Selman, B.; van Dover, R. B. In *Pattern Decomposition with Complex Combinatorial Constraints: Application to Materials Discovery*; AAAI, 2015; pp 636−643.

(8) Farrusseng, D.; Clerc, F. Diversity management for efficient combinatorial optimization of materials. *Appl. Surf. Sci.* **2007**, *254* (3), 772−776.

(9) Rodemerck, U.; Baerns, M.; Holena, M.; Wolf, D. Application of a genetic algorithm and a neural network for the discovery and

optimization of new solid catalytic materials. *Appl. Surf. Sci.* **2004**, *223* (1), 168−174.

(10) Hunt, W. H., Jr Materials informatics: Growing from the bio world. *JOM* **2006**, *58* (7), 88−88.

(11) Kullmann, O. Theory and Applications of Satisfiability Testing-SAT 2009. *Proceedings of the 12th International Conference, SAT 2009*, Swansea, UK, June 30−July 3, 2009; Springer: 2009; Vol. *5584*.

(12) Peurrung, L.; Ferris, K.; Osman, T. M. The materials informatics workshop: Theory and application. *JOM* **2007**, *59* (3), 50−50.

(13) Holena, M.; Cukic, T.; Rodemerck, U.; Linke, D. Optimization of catalysts using specific, description-based genetic algorithms. *J. Chem. Inf. Model.* **2008**, *48* (2), 274−282.

(14) Suh, C.; Gorrie, C.; Perkins, J.; Graf, P.; Jones, W. Strategy for the maximum extraction of information generated from combinatorial experimentation of Co-doped ZnO thin films. *Acta Mater.* **2011**, *59* (2), 630−639.

(15) Valero, S.; Argente, E.; Botti, V.; Serra, J. M.; Serna, P.; Moliner, M.; Corma, A. DoE framework for catalyst development based on soft computing techniques. *Comput. Chem. Eng.* **2009**, *33* (1), 225−238.

(16) Wolf, D.; Buyevskaya, O.; Baerns, M. An evolutionary approach in the combinatorial selection and optimization of catalytic materials. *Appl. Catal., A* **2000**, *200* (1), 63−77.

(17) West, D. Phase equilibria in iron ternary alloys. *Br. Corros. J.* **1989**, *24* (1), 16−17.

(18) Kusne, A.; Keller, D.; Anderson, A.; Zaban, A.; Takeuchi, I. High-throughput determination of structural phase diagram and constituent phases using GRENDEL. *Nanotechnology* **2015**, *26* (44), 444002.

(19) Long, C.; Bunker, D.; Li, X.; Karen, V.; Takeuchi, I. Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. *Rev. Sci. Instrum.* **2009**, *80* (10), 103902.

(20) Long, C.; Hattrick-Simpers, J.; Murakami, M.; Srivastava, R.; Takeuchi, I.; Karen, V.; Li, X. Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. *Rev. Sci. Instrum.* **2007**, *78* (7), 072217.

(21) LeBras, R.; Damoulas, T.; Gregoire, J. M.; Sabharwal, A.; Gomes, C. P.; Van Dover, R. B., Constraint reasoning and kernel clustering for pattern decomposition with scaling. In *Principles and Practice of Constraint Programming-CP 2011*; Springer: 2011; pp 508−522.

(22) Barr, G.; Dong, W.; Gilmore, C. J. PolySNAP: a computer program for analysing high-throughput powder diffraction data. *J. Appl. Crystallogr.* **2004**, *37* (4), 658−664.

(23) Corma, A.; Díaz-Cabanas, M. J.; Moliner, M.; Martínez, C. Discovery of a new catalytically active and selective zeolite (ITQ-30) by high-throughput synthesis techniques. *J. Catal.* **2006**, *241* (2), 312−318.

(24) Bunn, J. K.; Han, S.; Zhang, Y.; Tong, Y.; Hu, J.; Hattrick-Simpers, J. R. Generalized machine learning technique for automatic phase attribution in time variant high-throughput experimental studies. *J. Mater. Res.* **2015**, *30* (07), 879−889.

(25) Kusne, A. G.; Gao, T.; Mehta, A.; Ke, L.; Nguyen, M. C.; Ho, K.-M.; Antropov, V.; Wang, C.-Z.; Kramer, M. J.; Long, C.; Takeuchi, I. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **2014**, *4*, 6367.

(26) Hattrick-Simpers, J. R.; Gregoire, J. M.; Kusne, A. G. Perspective: Composition−structure−property mapping in high-throughput experiments: Turning data into knowledge. *APL Mater.* **2016**, *4* (5), 053211.

(27) Suram, S. K.; Haber, J. A.; Jin, J.; Gregoire, J. M. Generating Information-Rich High-Throughput Experimental Materials Genomes using Functional Clustering via Multitree Genetic Programming and Information Theory. *ACS Comb. Sci.* **2015**, *17* (4), 224−233.

(28) Ermon, S.; Le Bras, R.; Gomes, C. P.; Selman, B.; Van Dover, R. B. In Smt-aided combinatorial materials discovery. *International Conference on Theory and Applications of Satisfiability Testing*; Springer, 2012; pp 172−185.

(29) LeBras, R.; Bernstein, R.; Gomes, C. P.; Selman, B.; Van Dover, R. B. In *Crowdsourcing Backdoor Identification for Combinatorial Optimization*; IJCAI, 2013; pp 3−9.

(30) Ilea, D. E.; Whelan, P. F. Image segmentation based on the integration of colour−texture descriptors—A review. *Pattern Recognition* **2011**, *44* (10), 2479−2501.

(31) Mohita, N. Soil Groups: 8 Major Soil Groups available in India. http://www.yourarticlelibrary.com/soil/soil-groups-8-major-soil-groups-available-in-india/13902/ (accessed Oct 8, 2016).

(32) Felzenszwalb, P. F.; Huttenlocher, D. P. Efficient graph-based image segmentation. *International Journal of Computer Vision* **2004**, *59* (2), 167−181.

(33) Cignoni, P.; Montani, C.; Scopigno, R. DeWall: A fast divide and conquer Delaunay triangulation algorithm in E d. *Computer-Aided Design* **1998**, *30* (5), 333−341.

(34) O'Haver, T. *A pragmatic introduction to signal processing with applications in scientific measurement*, 2nd ed.; CreateSpace Independent Publishing Platform; 2016.

(35) Manning, C. D.; Raghavan, P.; Schütze, H. *Introduction to information retrieval*; Cambridge University Press: New York, 2008; p xxi.

(36) Famodu, O. O.; Hattrick-Simpers, J.; Aronova, M.; Chang, K.-S.; Murakami, M.; Wuttig, M.; Okazaki, T.; Furuya, Y.; Knauss, L. A; Bendersky, L. A; Biancaniello, F. S; Takeuchi, I. Combinatorial Investigation of Ferromagnetic Shape-Memory Alloys in the Ni-Mn-Al Ternary System using a Composition Spread Technique. *Mater. Trans.* **2004**, *45* (2), 173−177.

(37) Long, C. Combiview, software for visualization and clustering analysis of X-Ray diffraction data from combinatorial thin film libraries. https://sourceforge.net/projects/xrdsuite/ (accessed Sep 5, 2016).