

## ARTICLE OPEN

## Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries

Yuma Iwasaki<sup>1,2</sup>, A. Gilad Kusne<sup>3,4</sup> and Ichiro Takeuchi<sup>4</sup>

Machine learning techniques have proven invaluable to manage the ever growing volume of materials research data produced as developments continue in high-throughput materials simulation, fabrication, and characterization. In particular, machine learning techniques have been demonstrated for their utility in rapidly and automatically identifying potential composition–phase maps from structural data characterization of composition spread libraries, enabling rapid materials fabrication–structure–property analysis and functional materials discovery. A key issue in development of an automated phase-diagram determination method is the choice of dissimilarity measure, or kernel function. The desired measure reduces the impact of confounding structural data issues on analysis performance. The issues include peak height changes and peak shifting due to lattice constant change as a function of composition. In this work, we investigate the choice of dissimilarity measure in X-ray diffraction-based structure analysis and the choice of measure's performance impact on automatic composition–phase map determination. Nine dissimilarity measures are investigated for their impact in analyzing X-ray diffraction patterns for a Fe–Co–Ni ternary alloy composition spread. The cosine, Pearson correlation coefficient, and Jensen–Shannon divergence measures are shown to provide the best performance in the presence of peak height change and peak shifting (due to lattice constant change) when the magnitude of peak shifting is unknown. With prior knowledge of the maximum peak shifting, dynamic time warping in a normalized constrained mode provides the best performance. This work also serves to demonstrate a strategy for rapid analysis of a large number of X-ray diffraction patterns in general beyond data from combinatorial libraries.

*npj Computational Materials* (2017)3:4; doi:10.1038/s41524-017-0006-2

## INTRODUCTION

Composition-spread experiments enable rapid mapping of composition–structure–property relationships and are a powerful tool for materials discovery owing to recent developments in high throughput materials synthesis and rapid characterization techniques.<sup>1–5</sup> Machine learning techniques provide a means to convert the large volume of diverse, complex data collected from materials experiments into actionable knowledge refs 6–18 and have been successfully demonstrated in composition spread experiments. Of particular interest is the use of machine learning techniques to rapidly determine potential composition–phase maps from X-ray diffraction (XRD) data from composition spreads.<sup>19</sup> As the vast number of potential materials of three or more materials components is overwhelming compared to the number of expert-analyzed materials (typically unary or binary), this is primarily an unsupervised learning task. The knowledge gained from unsupervised learning analysis of phase data structure can be used to optimize semi-supervised learning analysis in those cases where phase labeled data exists. These techniques also promote a reduced reliance on expert human input, greater throughput, and an end goal of autonomous materials exploration systems. A variety of machine learning methods have been investigated in the literature, beginning with hierarchical clustering of XRD and Raman spectra data, and continuing with recent investigations in the use of constraint

programming and hyperspectral graph-based techniques for concurrent composition–phase map determination and constituent phase identification.<sup>19</sup>

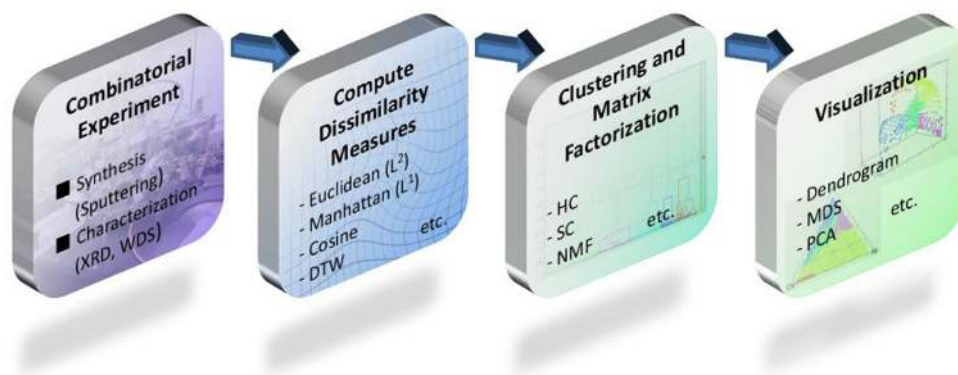
A key challenge for these methods is the selection of an appropriate dissimilarity (or similarity) measure, also known as a kernel function, to quantify the relationship between two pieces of structure data. Once computed, the dissimilarity between the two pieces of structure data is translated into dissimilarity between the structures of the originating samples. Proper selection of dissimilarity measure is essential to achieving quality results from unsupervised learning (and semi-supervised learning) as emphasized by Hastie, *et al.*<sup>20</sup> in their machine learning textbook *Elements of Statistical Learning*, “Specifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm.” This work seeks to provide insight to the experimentalist for dissimilarity measure selection through an efficacy analysis—identifying the impact of dissimilarity measure selection on composition–phase map accuracy and computation time. While particularly beneficial to the high-throughput experiment community where computation time is of concern, the analysis should also benefit the larger materials research community dealing with a large amount of XRD data as well. The structural data of interest for this work is the one-dimensional XRD pattern—a set of intensities, where each intensity value corresponds to a  $2\theta$  value

<sup>1</sup>IoT Devices Research Laboratories, NEC Corporation, Tsukuba 305-8501, Japan; <sup>2</sup>Spin Quantum Rectification Project, ERATO, Japan Science and Technology Agency, Sendai 980-8577, Japan; <sup>3</sup>National Institute of Standards and Technology, Gaithersburg, MD 20899, USA and <sup>4</sup>Department of Materials Science and Engineering, University of Maryland, College park, MD 20742, USA

Correspondence: A. Gilad Kusne (aaron.kusne@nist.gov)

Received: 14 June 2016 Revised: 29 December 2016 Accepted: 30 December 2016

Published online: 03 February 2017



**Fig. 1** Composition spread analysis flow chart: The strategy begins with the **a** synthesis of the composition spread and rapid characterization of composition and structure, followed by **b** computing dissimilarity measures, **c** composition-phase map determination and constituent phase identification via machine learning, and **d** visualization of analysis results via the use of a composition diagram and data dimension reduction techniques applied to the structural data

or  $q$ -spacing. The diffraction data is typically collected as a two-dimensional image of intensities as a function of the instrument parameters  $2\theta$  and  $\chi$ , integrated over the  $\chi$  angle to generate the one-dimensional diffraction pattern, and the  $2\theta$  values are translated to  $q$ -spacing values using the Bragg Law.<sup>21</sup>

A number of confounding factors can impact the performance of the dissimilarity measure. Peak shifting can occur in XRD data due to lattice expansion or contraction, so that two samples sharing the same structure may have differing XRD patterns. The texturing of samples is another challenge. Samples that share the same structure but have different texturing can result in differences in peak heights with a significant likelihood that some peaks may reduce in intensity to the point of being lost in measurement noise. Finally, the typical one-dimensional XRD pattern used to identify material structure is a data reduction from the two-dimensional Ewald's sphere.<sup>21</sup> This data reduction along with measurement signal to noise issues can result in a potential loss of distinguishing information so that two different structures have similar diffraction patterns. The optimal choice of measure takes all of these factors into account to accurately determine dissimilarity, typically through increased complexity of the measure itself and increased computational cost. The data analyst must therefore make a cost benefit decision between accuracy and computational cost. For high-throughput combinatorial library data analysis, computational cost—and the resulting computation time, hold a high priority. This is true also of analyzing other large data sets of collected XRD patterns, which may have been collected in an Edisonian sample-by-sample study. When dealing with non-high-throughput studies where sample synthesis can take an appreciable amount of time, the experimentalist may choose to select those XRD dissimilarity measures with the best performance, despite computational cost.

Various measures have been used throughout the literature. In this work, we seek to investigate a set of common measures for their relative computational cost in comparing one-dimensional XRD data and their relative accuracy in resulting cluster-based composition-phase map determination. A range of measures have been investigated in the past. Long, *et al.* reported on the use of the Pearson correlation coefficient as part of a software package to perform hierarchical cluster analysis (HCA) for composition-phase map determination.<sup>22</sup> The software package, CombiView, also allows the user to select from a range of measures.<sup>23</sup> Kusne, *et al.*<sup>24</sup> demonstrated the speed and accuracy of using the L1 norm with a mean shift theory clustering method to provide analysis of XRD data as the data was being collected. Baumes, *et al.*<sup>25</sup> investigated the use of the dynamic time warping (DTW) measure to assist in XRD analysis and found it to be resilient to

peak shifting when the range of peak shifting was known. LeBras, *et al.*<sup>26</sup> and Ermon, *et al.*<sup>27</sup> later applied DTW as part of larger constraint programming-based algorithms for concurrent composition-phase map determination and constituent phase identification from XRD data. Another study in the use of high speed methods by Kusne, *et al.*<sup>28</sup> demonstrated the efficacy of the cosine metric in a regularized least square method for concurrent composition-phase map determination and constituent phase identification from both XRD data and Raman spectra.<sup>1</sup>

In this work, we investigate a set of measures for accuracy and computational cost in analyzing XRD data from the Fe–Co–Ni ternary-alloy thin-film composition spread.<sup>29</sup> This work provides the basis for a broader analysis of dissimilarity measures and their efficacies over the wide range of materials systems and their respective structure data artifacts, such as extreme variations in peak heights, variations in number of peaks, and greater number and complexity of phase regions. This work also serves to demonstrate a system for rapid analysis of combinatorial spreads (Fig. 1). The system begins with the synthesis of the thin film composition spread using a combinatorial co-sputtering technique, followed by rapid characterization of composition using wavelength dispersive spectroscopy and structure via XRD. The dissimilarity is computed between each composition-spread sample and a set of machine learning analysis techniques are used to sort the samples into clusters of similar structure and to identify potential constituent phases. The samples are then visualized in a composition diagram with color-coded cluster labels. Contiguous regions of the composition diagram that share a cluster label correspond to potential phase regions of a composition-phase map. Dimension reduction techniques such as multidimensional data scaling are used to visualize the high dimensional XRD data in two or three dimensions to facilitate user evaluation of the analysis results.

#### Dissimilarity measures

The data of interest for this work is the one-dimensional XRD pattern from a Fe–Co–Ni composition spread. Each diffraction pattern is described by a set of intensities with a 1-to-1 correspondence to a set of  $2\theta$  values. The diffraction patterns are measured over the same set of  $2\theta$ , here indexed with  $i \in \{1 \dots N\}$ . There were 1125 diffraction patterns uniformly covering the entire Fe–Co–Ni ternary composition-phase map taken in the  $2\theta$  range of  $42.6^\circ$  to  $47.0^\circ$  (with  $\lambda = 0.15418$  nm for Cu Ka).<sup>29</sup> To demonstrate the robustness of the overall analysis protocol developed here, we have also applied it to another ternary system, namely, Fe–Ga–Pd composition spread, and the results are presented in the [supplementary section](#)<sup>22</sup>.  $2\theta$  is used here rather than  $q$ -spacing due to

the simplicity of having equal sampling spacing of  $\Delta(2\theta)$  between each intensity measurement, allowing the use of the index  $i$  as the dependent variable. The results of this work should generalize to an equivalent set of  $q$ -spacing values with uniform spacing and the appropriately interpolated XRD intensity values. For two diffraction patterns  $\mathbf{s}$  and  $\mathbf{t}$ , the dissimilarity measure is defined as  $D(\mathbf{s}, \mathbf{t})$ . For  $D(\mathbf{s}, \mathbf{t}) = 0$  the two diffraction patterns are assumed to be identical and the corresponding samples are assumed to share the same structure. Larger values of the dissimilarity measure imply greater dissimilarity between the samples' structures.

The set of dissimilarity measures investigated in this work fall into a group of categories (with some falling into more than one): the  $L_1$  norm ('Manhattan'),  $L_2$  norm ('Euclidean'), and the cosine metric are geometry-based measures; Pearson correlation coefficient and Spearman rank correlation coefficient are statistics-based dissimilarity measures; DTW and the earth mover's distance (EMD), also known as the "Wasserstein distance", are measures specifically developed for feature preservation, i.e. resilience to peak shifting; and the Jensen–Shannon divergence (JSD) is based in information theory. These measures, except for DTW and EMD, satisfy the requirements of a metric. DTW is not a metric as it does not satisfy the triangle equality. EMD is a metric under two conditions –the ground distance used must be a metric and the areas under the two diffraction patterns being compared must be equal. While the ground distance used here is the  $L_1$  norm metric, the areas under any two diffraction patterns are not necessarily equal, and EMD is therefore not assumed to satisfy the requirements of a metric. This work introduces the Normalized and Constrained Dynamic Time Warping measure (NC-DTW) as a novel means for computing XRD dissimilarity, and which like the DTW measure, does not satisfy the triangle inequality.

$L_1$  &  $L_2$  norms. The  $L_1$  and  $L_2$  norms are special cases of the  $p$ -norm, given by:

$$D_{p\text{-norm}}(\mathbf{s}, \mathbf{t}) = \left( \sum_{i=1}^N (s_i - t_i)^p \right)^{1/p}, \quad (1)$$

where  $p=2$  is the  $L_2$  norm, commonly known as the Euclidean distance, and  $p=1$  is the  $L_1$  norm, also known as the Manhattan, taxi-cab, or city block distance.

*Cosine metric.* The cosine metric gives the cosine of the angle between two vectors, thus measuring only the vector orientation difference, ignoring differences in vector magnitude.

$$D_{\text{cosine}}(\mathbf{s}, \mathbf{t}) = 1 - \frac{\sum_{i=1}^N (s_i \cdot t_i)}{(\sum_{i=1}^N s_i^2)^{\frac{1}{2}} (\sum_{i=1}^N t_i^2)^{\frac{1}{2}}} \quad (2)$$

*Pearson product-moment correlation.* The Pearson product-moment correlation is a measure of linear correlation between two spectrums.

$$D_{\text{Pearson}}(\mathbf{s}, \mathbf{t}) = 1 - \frac{\sum_{i=1}^N (s_i - \bar{s})(t_i - \bar{t})}{(\sum_{i=1}^N (s_i - \bar{s})^2)^{\frac{1}{2}} (\sum_{i=1}^N (t_i - \bar{t})^2)^{\frac{1}{2}}}, \quad (3)$$

where  $\bar{s}$  and  $\bar{t}$  indicate the average values of  $\mathbf{s}_i$  and  $\mathbf{t}_i$ , respectively. The Pearson product-moment correlation is related to the cosine metric by replacing values of  $\mathbf{s}_i$  and  $\mathbf{t}_i$  with  $\mathbf{s}_i - \bar{s}$  and  $\mathbf{t}_i - \bar{t}$ , respectively.

*Spearman rank correlation coefficient.* Each component  $s_i$  is ranked by decreasing value, converting  $s_i$  and  $t_i$  into their ranks  $S_i$  and  $T_i$ , respectively. For example, if  $S_i$  is the second largest value from the set of all  $s_i$ ,  $S_i=2$ . The Spearman

measure is computed using:

$$D_{\text{Spearman}}(\mathbf{s}, \mathbf{t}) = \frac{6 \sum_{i=1}^n d_i}{n^3 - n}, \quad (4)$$

where  $d_i = |S_i - T_i|$  is the difference in rank for  $s_i$  and  $t_i$ .

*Jensen–Shannon divergence (JSD).* The JSD is a symmetric version of the Kullback–Leibler divergence  $D_{\text{KL}}$ , which measures the information lost when using  $\mathbf{s}$  to represent  $\mathbf{t}$ .

$$D_{\text{JSD}}(\mathbf{s}, \mathbf{t}) = \frac{1}{2} D_{\text{KL}}(\mathbf{s}, \mathbf{t}) + \frac{1}{2} D_{\text{KL}}(\mathbf{t}, \mathbf{s}) \quad (5)$$

$$D_{\text{KL}}(\mathbf{a}, \mathbf{b}) = \sum_i a_i \log \frac{a_i}{b_i}$$

*Earth mover's distance (EMD).* The EMD is computed by solving a transport problem. The intensity values of the diffraction patterns being compared are interpreted as mass, with each unit of intensity equal to a unit of mass. The measure is computed as the minimum total work required to deform one diffraction pattern, by moving mass, into the second diffraction pattern.<sup>30</sup> The EMD has a computational complexity of  $O(n^2)$  and is therefore more computationally expensive than the previously mentioned dissimilarity measures, where  $D_{p\text{-norm}}$ ,  $D_{\text{cosine}}$ ,  $D_{\text{Pearson}}$ ,  $D_{\text{Spearman}}$  and  $D_{\text{JSD}}$  have  $O(n)$  computational complexity. A fast-EMD algorithm with  $O(n)$  complexity is used here to reduce computation time.<sup>30</sup>

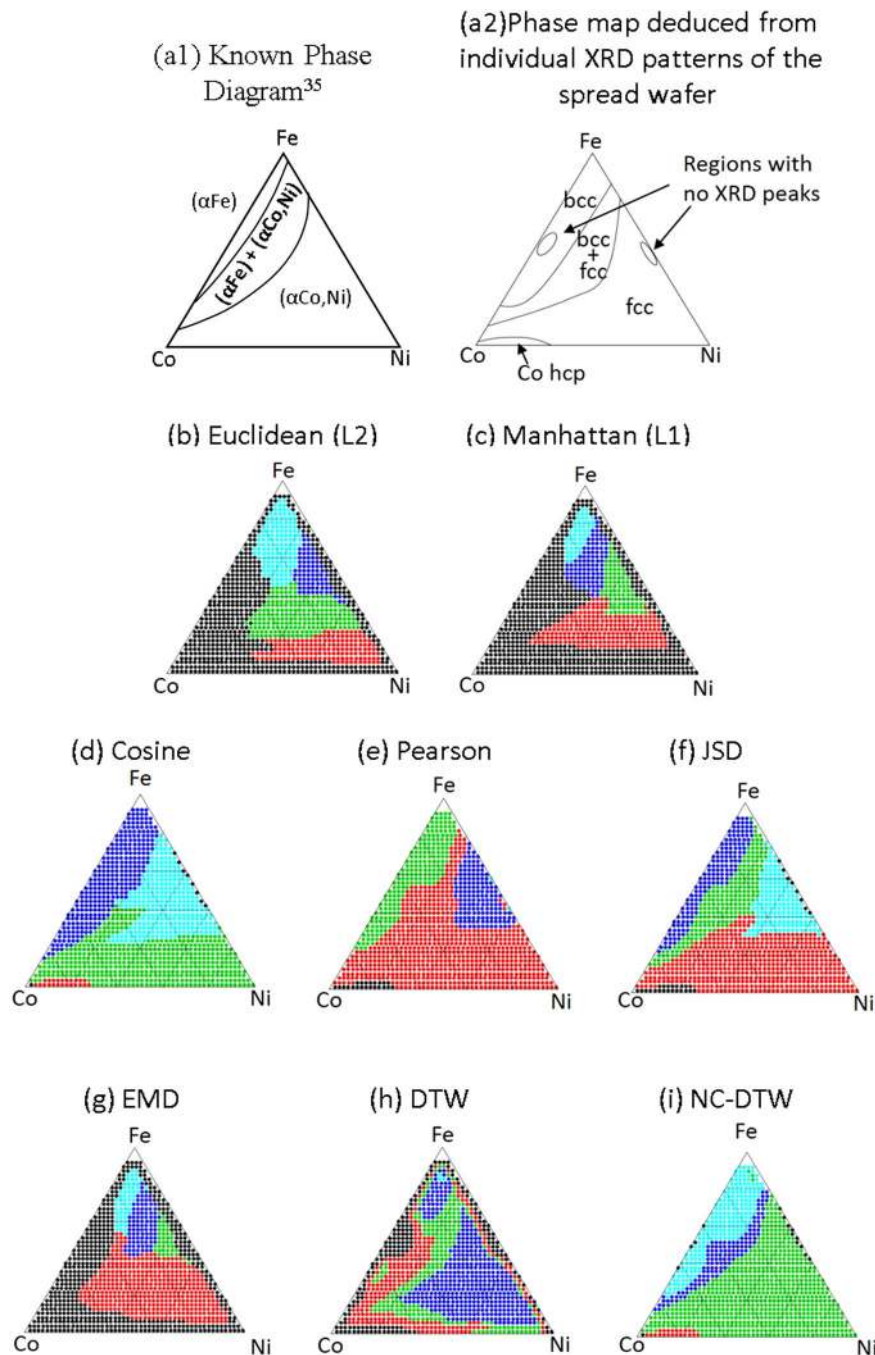
*Dynamic time warping (DTW).* DTW measures the minimum non-linear warping required to map one numeric array to another. DTW also has a computational complexity of  $O(n^2)$ . In this study, a fast-DTW algorithm with  $O(n)$  was employed to shorten the calculation time.<sup>31</sup>

*Normalized and constrained dynamic time warping (NC-DTW).* This work introduces the NC-DTW measure as a means for computing XRD dissimilarity. For NC-DTW, each diffraction pattern is normalized by its maximum value. Additionally, the range of potential warping paths from one array to another is limited by setting a window size. The window size defines the maximum number of indices each intensity value can be warped by, e.g. for a window size of  $r$ ,  $h_i$  can only be warped to a value in the set  $[h_{i-r}, \dots, h_{i+r}]$ .<sup>32</sup> Specifically, if the window size is set to 0.5 degrees, two diffraction patterns that share a peak which is separated by less than 0.5 degrees will be identified as similar, while two peaks separated by more than the window size will be considered dissimilar.

## RESULTS

Performance: composition-phase map determination accuracy

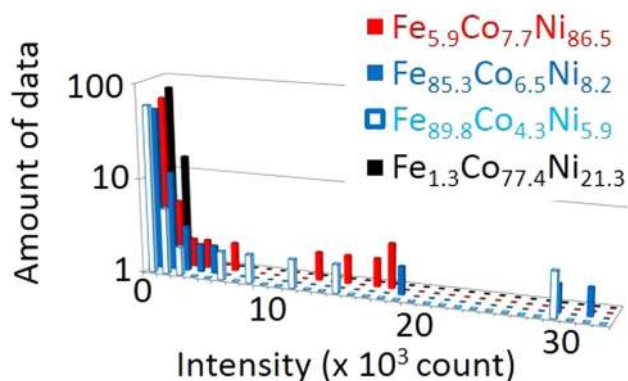
A dissimilarity matrix was computed for the Fe–Co–Ni XRD patterns for each measure. The dissimilarity matrices were then used to sort samples into groups of similar structure using agglomerative HCA and  $k$ -medoids clustering—both methods that rely only on the dissimilarity matrix rather than the original XRD data. For HCA, cluster-cluster dissimilarity was computed using ward's, average, centroid, complete, Mcquitty and median linkage methods. A discussion of the clustering methods and their respective efficacies can be found in ref. 20. The HCA and  $k$ -medoids clustering results were then compared to cluster labels defined by expert analysis<sup>29</sup> using multi-class  $F$ -measure to compute accuracy. The number of clusters was varied between 2 and 10 to observe the impact of cluster number choice on accuracy and to ensure that data effects such as outliers do not skew the overall analysis of measure performance. For visual analysis, the samples were color coded for cluster membership and plotted as a function of composition. The resulting diagram



**Fig. 2** Result of the hierarchical cluster (HC) analysis in the Fe–Co–Ni ternary alloy system with different dissimilarity measures. **a1** Known phase diagram (ref. 35). **a2** Phase map deduced from individual XRD patterns of spread wafer (ref. 29). **b–i** Result of the HC analysis with different measures: **b** Euclidean metric. **c** Manhattan metric. **d** Cosine metric. **e** Pearson metric. **f** JSD. **g** EMD. **h** DTW. **i** NC-DTW

provides a potential composition-phase map with cohesive regions of shared cluster label corresponding to potential phase regions. The phase map deduced from individual XRD patterns of the spread wafer is shown in Fig. 2a2 with the equilibrium phase diagram shown in Fig. 2a1 for reference. The automated composition-phase maps for each measure are shown in Fig. 2b–i for HCA using average linkage method. The number of cluster was set to 5 because the known composition-phase map has five regions with bcc, fcc, hcp, bcc + fcc and no XRD peak detected region.

The Spearman metric performed poorly for all the clustering techniques and number of clusters tested, sorting the majority of data points into one cluster and the rest of the clusters containing one or two points from the edge of the composition diagram, and is excluded from the figure. It was found that the Spearman metric's poor performance is due to the statistical distribution of XRD intensities. Figure 3 shows the data distribution (histogram) with respect to the diffraction intensity for the four XRD pattern used in the later qualitative metric multidimensional data scaling (MMDS) analysis (Section 2.1.1). The majority of intensity values



**Fig. 3** Data distribution (histogram) with respect to the diffraction intensity for four selected XRD patterns

occur at low values, and are typical of noise with no diffraction peaks. The Spearman metric calculates ranks primarily with noise, leading to improper dissimilarity measures and poor clustering, and was therefore found to be an inadequate metric for XRD analysis.

$F$ -measure is an accuracy measure that combines two properties, precision and recall. Precision is defined as the number of samples properly identified as belonging to a phase region, or true positive (TP) count, divided by the sum of the number of TP and the number of samples incorrectly labeled as belonging to the phase region, or false positive. Similarly, recall is defined as the ratio of TP to the sum of TP and the samples incorrectly labeled as not belonging to the cluster, or false negative. The  $F$ -measure is defined as the harmonic mean of precision and recall, with larger values corresponding to higher accuracy. To compute the  $F$ -measure, each sample is labeled by its phase region, given by the expert derived composition-phase map. The set of samples in each expert labeled cluster  $A_m$  are indexed by cluster label  $m$  and the set of samples in the computed clusters  $C_n$  are indexed by cluster label  $n$ . Associations between true cluster labels and automated cluster labels are permuted through permutations of  $n$  and the maximum  $F$ -measure is recorded:

$$F = \sum_{m=1}^K \frac{|A_m|}{N} \max_n \frac{2R_{mn}P_{mn}}{R_{mn} + P_{mn}} \quad (6)$$

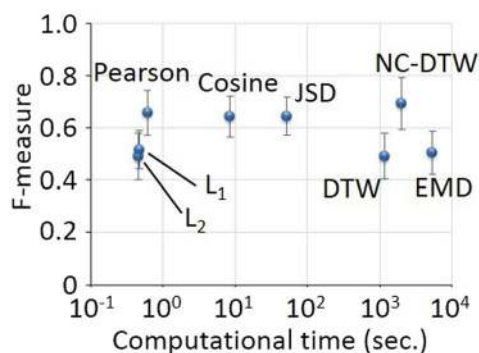
$$P_{mn} = \frac{TP}{TP + FP} = \frac{|A_m \cap C_n|}{|C_n|}$$

$$R_{mn} = \frac{TP}{TP + FN} = \frac{|A_m \cap C_n|}{|A_m|}$$

where  $|A_m|$  and  $|C_n|$  are the number of samples in clusters  $A_m$  and  $C_n$ , respectively.  $K$  is the total number of clusters and  $N$  are the number of XRD patterns, i.e.,  $K=5$  and  $N=1125$ .

Figure 4 shows the mean and standard deviation (indicated by error bars) of  $F$ -measure accuracies, over both clustering method and number of clusters, computed for each dissimilarity measure as a function of computation time. The  $F$ -measure accuracy for each measure and clustering method with the number of clusters varied from 2 to 10 are shown in the Supplemental Figure S3.

Figure 4 shows that the cosine, Pearson and JSD measures all perform better than the Euclidean, Manhattan, DTW, and EMD measures in  $F$ -measure accuracy. This result is consistent with intuition from Fig. 2a–i. The cosine, Pearson and JSD measures also come at a significantly reduced computational cost than the DTW and EMD measures. Of interest is the significantly greater computational cost of the cosine metric compared to the Pearson metric, as these only differ in a simple computation, the subtraction of variable mean in the Pearson metric. The difference



**Fig. 4** The computational cost and  $F$ -measure for each measure.  $F$ -measure statistics were computed over the various clustering techniques and over the range of cluster number 2–10

in computational time was found to be due to the use of different computation packages as discussed in the Supplemental.

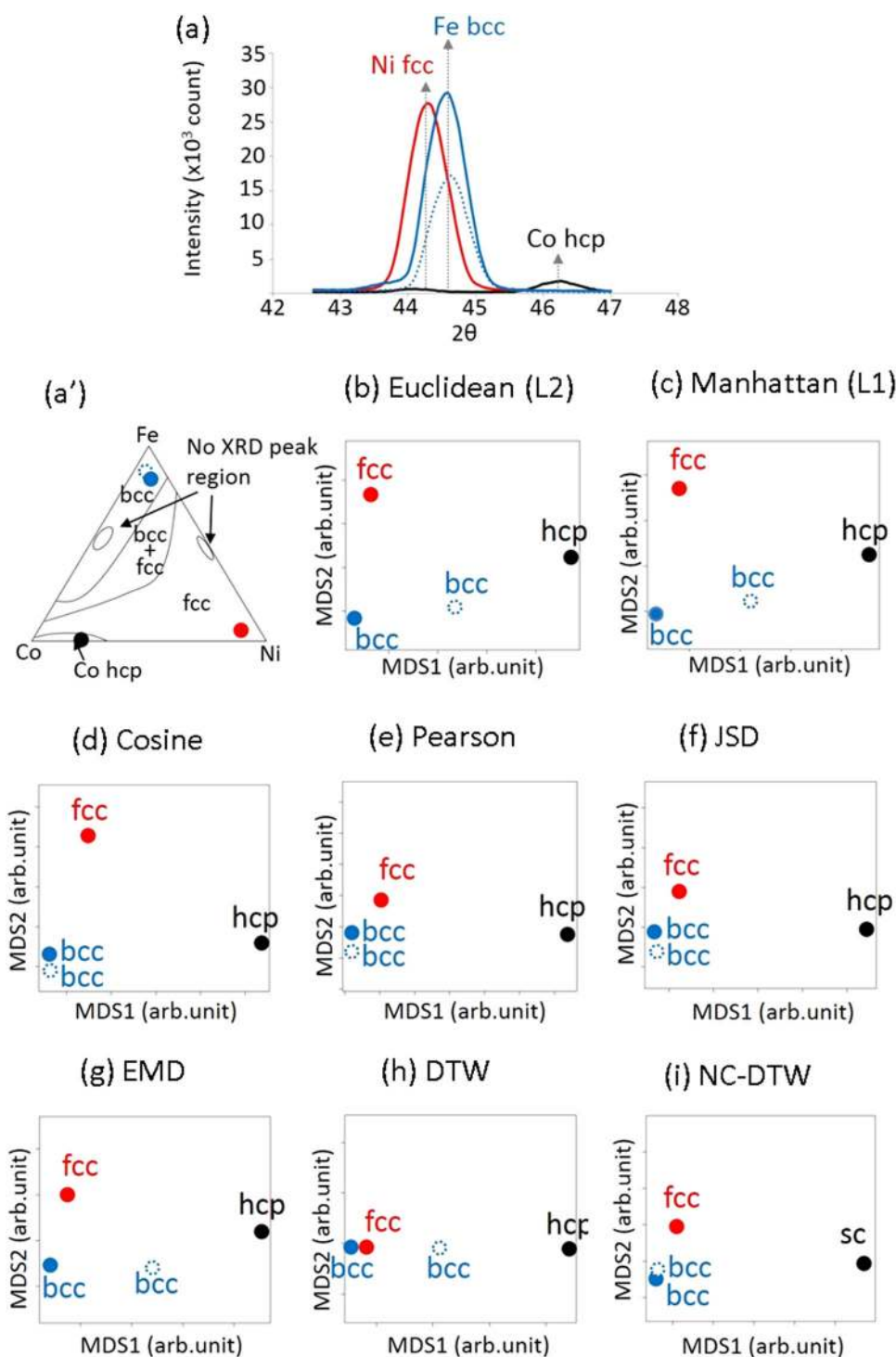
This work introduces the use of the NC-DTW measure, which utilizes prior knowledge of the maximum peak shift magnitude to improve peak-mapping and dissimilarity evaluation. Here the NC-DTW peak shift window is set to 0.6 degrees, permitting peak matching across  $\pm 0.6$  degrees, as the magnitude of peak shift for the bcc structure is approximately 0.6 degrees. As seen in Fig. 5, use of the NC-DTW measure provides the greatest accuracy, while still at a significant computational cost (2003.41 s). Also the clustering result shown in Fig. 2i is quite similar to the correct composition-phase map shown in Figure 2a. The *light blue*, *red*, *green* and *blue* regions in Fig. 2i represent bcc, hcp, fcc and bcc + fcc phases, respectively.

To gain further insight into the relative performance of the various measures, a qualitative analysis using MMDs was performed.

*Qualitative analysis: dimension reduction with metric multidimensional data scaling (MMDs).* To gain a qualitative understanding of the clustering performance of each measure, MMDs was used to map the measure space (defined by the minimum number of dimensions required to preserve the dissimilarities between each set of samples) to the more easily visualized two or three-dimensional space, while minimizing the loss of inter-sample dissimilarities.<sup>22</sup> The two or three-dimensional projection allows for a more easily visualization of cluster results to identify the impact of measure choice on cluster performance.

The four XRD patterns were selected with different compositions as shown in Fig. 5a. A *solid-red* and a *solid-black* pattern represent Ni fcc and Co hcp structures, respectively. Though a *solid-blue* and *dot-blue* patterns show different peak intensity, both of them show the Fe bcc structure. The compositions of the *solid-red*, *solid-blue*, *dot-blue* and *solid-red* patterns are samples whose compositions are  $\text{Fe}_{5.9}\text{Co}_{7.7}\text{Ni}_{86.5}$ ,  $\text{Fe}_{85.3}\text{Co}_{6.5}\text{Ni}_{8.2}$ ,  $\text{Fe}_{89.8}\text{Co}_{4.3}\text{Ni}_{5.9}$ , and  $\text{Fe}_{1.3}\text{Co}_{77.4}\text{Ni}_{21.3}$ , respectively, as shown in Fig. 5a'.

Figure 5b–i is the 2D MDS results for four selected patterns on the Euclidean, Manhattan, cosine, Pearson, JSD, EMD, DTW, and NC-DTW measures, respectively. The better dissimilarity measures result in the *solid-blue* and *dotted-blue* patterns appearing closer due to similar structural phase (bcc), while the distance between the different structure phases should be large. On the 2D MMDs mapping in the Euclidean, Manhattan, EMD, and DTW the distance between the *solid-blue* and *dotted-blue* is large in spite of the fact that they are from the same structural phase because these measures are strongly affected by the peak intensity difference. On the other hand, the cosine, Pearson, and JSD dissimilarities allow us to locate same structural phases closer in the 2D MMDs mapping since they mainly judge the peak shape, not peak

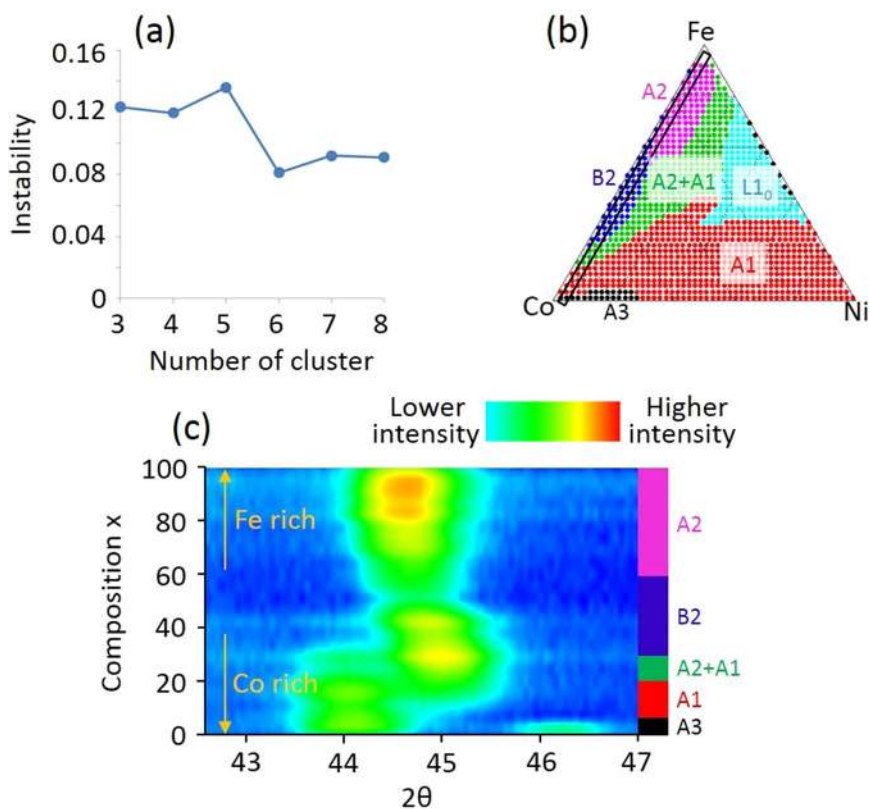


**Fig. 5** Result of the MMDS analysis. **a** Four selected XRD patterns. The *solid-blue* and *dotted-blue* are from the bcc phase. The *solid-red* and *solid-black* are fcc and hcp structural phase, respectively. **a'** Composition mapping for four selected XRD patterns. **b–i** Result of the 2D MMDS analysis with different metric. **b** Euclidean metric. **c** Manhattan metric. **d** Cosine metric. **e** Pearson metric. **f** Jensen–Shannon Divergence. **g** EMD. **h** DTW **i** NC-DTW

intensity differences. The cosine, Pearson and JSD metrics permit us to separate the different structural phases and cluster the same structural phases in the defined space.

When the structural phases are roughly divided into only fcc, bcc and hcp phases, and we expect to obtain the “correct” cluster result as shown in Fig. 2a (known composition-phase map), A2 and B2 (A1 and L1<sub>0</sub>) have to be recognized as the same structural

phase in metric space because both of A2 and B2 (A1 and L1<sub>0</sub>) are based on the bcc (fcc) structure. In order to end up with the “correct” clustering result, the peak shift information due to the difference in the lattice constant need to be accounted for, and the diffraction peaks from A2 and B2 (A1 and L1<sub>0</sub>) need to be identified as arising from the same/equivalent reflections. The EMD and DTW were expected to take care of the peak shift issue



**Fig. 6** Automated determination of the cluster number using the bootstrap method. **a** Result of cluster instability by the bootstrap method. **b** Result of cluster analysis with six clusters determined by bootstrap method. **c** Composition spread for XRD pattern in the black rectangle ( $\text{Fe}_x\text{Co}_{100-x}\text{Ni}_{2.5\pm 1.0}$ ) in Fig. 5b

because of their robustness against such peak shifting. However, the EMD and DTW led to poor cluster results as shown in Fig. 2g,h due to the following two reasons. First, EMD and DTW are strongly affected by the peak intensity difference, and accordingly, the distance between the *solid-blue* and *dotted-blue* in the MMDS space (Figs. 5g,h) is large despite the fact that they are from the same structural phase. Second, misidentification of diffraction peaks may occur. For example, the DTW has “mixed-up” the fcc and bcc peaks since the distance between *solid-red* (fcc) and *solid-blue* (bcc) is very small in Fig. 5h due to the measures’ robustness against peak location differences.

The NC-DTW measure introduced in this work resolves these issues. Compare to the MMDS mapping for DTW in Fig. 5h, the distance between *solid-blue* and *dotted-blue* in Fig. 5i is small due to the normalization. This indicates that the normalization decreases the influence of peak intensity differences on cluster results. Moreover, the peak shift window constraint allows us to increase the distance between fcc and bcc by defining an allowable peak shift range.

From these qualitative and quantitative analyses, we conclude that the cosine and Pearson metrics should be the measures of choice in cluster analysis of XRD patterns due to their high accuracy and low computational cost. Furthermore, when prior knowledge of the maximum peak shift magnitude is available, the NC-DTW should be selected if adequate time exists for the increased computational cost.

**Impact of measure on parameter selection: number of clusters.** For this study, expert derived labels were available to quantify the accuracy of clustering. However, for the high throughput systems described in Fig. 1, this is rarely the case. Typically, only the structure and composition data is available and the automated clustering results are presented to the expert for verification.

For these cases, a set of factors can impact performance—the choice of measure, the choice of clustering method, and the choice of clustering method parameters. There are numerous methods for quantifying clustering performance when correct class assignment is unknown,<sup>20</sup> though expert analysis is typically taken as the final say. These methods are used for identifying the optimal cluster method parameters, including the number of clusters. For this work, cluster performance is determined by the bootstrap method, which minimize cluster instability.<sup>33</sup> Figure 6a shows the cluster instability calculated by the bootstrap method for the HCA-average linkage method with the JSD metric in the Fe–Co–Ni system. When the number of clusters is equal to six, the cluster analysis is the most stable. Therefore, this result suggests setting the number of cluster to six.

Figure 6b shows the cluster result with six clusters by the HCA-average linkage method with the JSD metric. Comparing to Fig. 2f, the bcc phase was separated into two phases (*blue* and *pink* region in Fig. 6b). This result reflects the difference of lattice constant which causes the peak shift in XRD pattern. The Co atomic radius is slightly smaller than the Fe atomic radius, and as a result, the diffraction peak of bcc structure with Co-rich composition shifts to the higher angle as shown in Fig. 6c, which displays XRD patterns in the *black rectangle* ( $\text{Fe}_x\text{Co}_{100-x}\text{Ni}_{2.5\pm 1.0}$ ) in Fig. 6b.

The result in Fig. 6b can be explained by the Strukturbericht designation, where  $\text{Fe}_{50}\text{Co}_{50}$  and  $\text{Fe}_{50}\text{Ni}_{50}$  can be the B2 and L1<sub>0</sub> structure, respectively. Therefore, the *pink*, *blue*, *red*, *black*, *light blue* and *green* region represent the A2, B2, A1, A3, L1<sub>0</sub> and A1 + A2 structure in Fig. 6b.

#### Existing problems and future work

**Window size in NC-DTW:** In Section 2.1, the window size  $r$  was set to 0.6 degrees. since we knew that the magnitude of the peak shift was about 0.6 degrees. in the Fe–Co–Ni system. However, the

magnitude of the peak shift is usually unknown when the combinatorial data is analyzed for the first time. Therefore, a method which determines the window size automatically is essential when the NC-DTW is used in the cluster analysis. One possible method to decide the window size is a deductive approach by the condensed matter theory. For example, Ab initio calculation might allow us to estimate the magnitude of peak shift on XRD patterns. Another possible method is a recursive approach by statistics. For example, the magnitude of peak shift might be predicted by supervised machine learning when a large amount of data for XRD patterns from other materials has already been stored in a database.

Extension to the other fields: The cluster analysis with the cosine, Pearson, JSD, and NC-DTW are found to be the ideal metrics for combinatorial XRD data. It would be interesting to determine the best metric for Raman spectroscopy (Raman), X-ray magnetic circular dichroism (XMCD) and other physical property data. One must carefully choose the appropriate metric(s) to use taking into account the detailed characteristics of different types of physical property data. For example, the cosine metric which determines the only peak shapes (not peak intensity) is not likely to be suitable for XMCD data because peak intensity information of XMCD is very important. We have previously shown that Pearson is effective as one metric for carrying out clustering of Raman spectroscopy data.<sup>34</sup> We are currently in the process of evaluating different metrics for a variety of data formats from other materials characterization techniques.

## DISCUSSION

We have investigated the effect of the dissimilarity measures on clustering analysis for XRD data. Clustering with the Euclidean, Manhattan, cosine, Pearson, Spearman, JSD, DTW and NC-DTW measures were carried out on XRD data from the Fe-Co-Ni ternary alloy spread. It was found that the cosine, Pearson, JSD and NC-DTW measures are the most suitable for XRD data analysis, with the cosine, Pearson, and JSD measures providing optimal results in the presence of peak height change and NC-DTW providing optimal results when prior knowledge of peak shifting magnitude is available to be incorporated into the analysis. In addition, selecting the cosine measure over DTW was shown to reduce dissimilarity computation time by two orders of magnitude (fastDTW: 1165 s versus cosine metric: 8.52 s). Similar results were obtained for another ternary system, and the results are provided in the [supplementary section](#). Dissimilarity measure selection was shown to provide the translation invariance and scale invariance required to properly handle peak height and peak shift changes. For improved results, further physical constraints such as Gibbs phase rule can be introduced through constraints applied in the clustering algorithm.<sup>19</sup>

We have looked at composition spreads of a number of other materials systems (about 10, all metallic alloys or oxides) using some of the same metrics and the clustering algorithms. Some systems were indeed more complicated than others: while most of the samples were textured, some were polycrystalline, and some contained diffraction peaks from impurity phases. The result on the relative performance of different metrics were found to be the same or very similar to the one described here for the Fe-Co-Ni spread. We have chosen the Fe-Co-Ni system result for the current manuscript as the most representative and most effectively illustrative because it has a known phase diagram, and because the phases on the composition spread have been identified “manually” one by one prior to the current work.<sup>29</sup>

This study provides the basis for a broader analysis of dissimilarity measures and their efficacy across different material systems and their respective data features. In particular, future work will investigate whether the lessons learned in this study may be extended to systems of greater number and complexity of

phase regions by appropriately increasing the composition sampling resolution of the composition spread, to ensure well defined phase regions. Furthermore, the effect of data artifacts associated with XRD measurement such as limited low 2 $\theta$  range sampling or low signal to noise ratio will also be addressed. Additionally, the impact of uncertainty in measurement and data analysis will be analyzed for their propagated effects on the final composition-phase map.

It is important to point out that the challenge of having to analyze a large amount of data in materials science goes beyond screening combinatorial libraries. Mapping spatially resolved properties of samples is increasingly a desirable mode of materials interrogation as various characterization techniques (such as scanning probe microscopy and synchrotron micro diffraction) continue to become powerful with higher and higher spatial resolution. The present result therefore is applicable to any experiment where a large number of XRD patterns is collected.

## METHODS

The set of dissimilarity measures were computed for the Fe-Co-Ni and Fe-Ga-Pd XRD using the software packages listed in the [Supplemental](#). Dissimilarity measure results were then used to cluster the diffraction patterns into phase regions using agglomerative HCA and *k*-medoids clustering. For HCA, cluster-cluster dissimilarity was computed using ward's, average, centroid, complete, Mcquitty and median linkage methods. The clustering results were then compared to phase region labels defined by expert analysis<sup>27</sup>, with the accuracy computed using multi-class *F*-measure. The number of clusters was varied between 2–10 to observe the impact of cluster number choice on accuracy and to ensure that data effects such as outliers do not skew the overall analysis of measure performance. To gain a qualitative understanding of the clustering performance of each dissimilarity measure, MMDS was used to map each measure space (defined by the minimum number of dimensions required to preserve the dissimilarities between each set of samples) to the more easily visualized two-dimensional space, while minimizing the loss of inter-sample dissimilarities.<sup>22</sup> The two-dimensional projections were then used to identify the impact of measure choice on cluster performance.

## ACKNOWLEDGEMENTS

We thank Sean. W. Fackler, Tieren Gao and Jie Yong for valuable discussions. We also thank Young Yoo and X.-D. Xiang for providing the raw data from ref. [29](#). This work was supported by NIST and NEC and partially supported by ONR N000141512222.

## AUTHOR CONTRIBUTIONS

Y.I., A.G.K., and I.T. conceived the idea for the present work. Y.I. and A.G.K. carried out the computation. Y.I., A.G.K., and I.T. wrote the manuscript together.

## COMPETING INTERESTS

The authors declare no competing interests.

## REFERENCES

1. Koinuma, H. & Takeuchi, I. Combinatorial solid-state chemistry of inorganic materials. *Nat. Mater.* **3**, 429–438 (2004).
2. Takeuchi, I. et al. Identification of novel compositions of ferromagnetic shape-memory alloys using composition spreads. *Nat. Mater.* **2**, 180–184 (2003).
3. Takeuchi, I., Dover, R. Bvan & Koinuma, H. Combinatorial synthesis and evaluation of functional inorganic materials using thin-film techniques. *MRS Bull.* **27**, 301–308 (2002).
4. Takeuchi, I. et al. Monolithic multichannel ultraviolet detector arrays and continuous phase evolution in Mg<sub>x</sub>Zn<sub>1-x</sub>O composition spreads. *J. Appl. Phys.* **94**, 7336–7340 (2003).
5. Fukumura, T. et al. Rapid construction of a phase diagram of doped Mott insulators with a composition-spread approach. *Appl. Phys. Lett.* **77**, 3426–3428 (2000).
6. Fischer, C. C., Tibbetts, K. J., Morgan, D. & Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **5**, 641–646 (2006).



7. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 1–6 (2013).
8. Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
9. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
10. Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R. & Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).
11. Montavon, G. et al. Machine learning of molecular electronic properties in chemical compound space. *New. J. Phys.* **15**, 095003 (2013).
12. Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22**, 3762–3767 (2010).
13. Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **13**, 17930–17955 (2011).
14. Balabin, R. M. & Lomakina, E. I. Neural network approach to quantum-chemistry data: accurate prediction of density functional theory energies. *J. Chem. Phys.* **131**, 074104 (2009).
15. Hansen, K. et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory. Comput.* **9**, 3404–3419 (2013).
16. Saad, Y. et al. Data mining for materials: computational experiments with AB compounds. *Phys. Rev. B* **85**, 104104 (2012).
17. d'Avezac, M., Luo, J.-W., Chanier, T. & Zunger, A. Genetic-algorithm discovery of a direct-gap and optically allowed superstructure from indirect-gap Si and Ge semiconductors. *Phys. Rev. Lett.* **108**, 027401 (2012).
18. Mueller, T., Kusne, A. G. & Ramprasad, R. Machine learning in materials science. *Rev. Comput. Chem.* **29**, 186–273 (2016).
19. Hatrick-Simpers, J., Gregoire, J. & Kusne, A. G. Perspective: composition – structure – property mapping in high-throughput experiments: turning data into knowledge. *APL Mater.* **4**, 053211 (2016).
20. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer, 2009).
21. Graef, M. D. & McHenry, M. E. *Structure of materials: an introduction to crystallography, diffraction and symmetry*. (Cambridge University Press, 2012).
22. Long, C. et al. Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. *Rev. Sci. Instrum.* **78**, 072217–072217 (2007).
23. Takeuchi, I. et al. Data management and visualization of x-ray diffraction spectra from thin film ternary composition spreads. *Rev. Sci. Instrum.* **76**, 062223–062223 (2005).
24. Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 1–7 (2014).
25. Baumes, L. A., Moliner, M., Nicoloyannis, N. & Corma, A. A reliable methodology for high throughput identification of a mixture of crystallographic phases from powder X-ray diffraction data. *CrystEngComm* **10**, 1321–1324 (2008).
26. LeBras, R. et al. Constraint reasoning and kernel clustering for pattern decomposition with scaling. In *International Conference on Principles and Practice of Constraint Programming* (ed. Jimmy, L.) 508–522 (Springer, Berlin Heidelberg, 2011).
27. Ermon, S. et al. *Pattern Decomposition with Complex Combinatorial Constraints: Application to Materials Discovery* 636–643 (The AAAI Press, Palo Alto, CA), <http://www.aaai.org/Library/AAAI/aaai15contents.php> (2015).
28. Kusne, A. G., Keller, D., Anderson, A., Zaban, A. & Takeuchi, I. High-throughput determination of structural phase diagram and constituent phases using GRENDEL. *Nanotechnology* **26**, 444002 (2015).
29. Yoo, Y. K. et al. Identification of amorphous phases in the Fe–Ni–Co ternary alloy system using continuous phase diagram material chips. *Intermetallics* **14**, 241–247 (2006).
30. Pele, O. & Werman, M. Fast and robust earth mover's distances. In *IEEE 12th International Conference on Computer Vision* 460–467, doi:10.1109/ICCV.2009.5459199 (2009).
31. Salvador, S. & Chan, P. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* **11**, 561–580 (2007).
32. Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26**, 43–49 (1978).
33. Fang, Y. & Wang, J. Selection of the number of clusters via the bootstrap method. *Comput. Stat. Data Anal.* **56**, 468–477 (2012).
34. Kan, D., Long, C. J., Steinmentz, C., Lofland, S. E. & Takeuchi, I. Combinatorial search of structural transitions: Systematic investigation of morphotropic phase boundaries in chemically substituted BiFeO<sub>3</sub>. *J. Mater. Res.* **27**, 2691–2704 (2012).
35. Raynor, G.V. & Rivlin, V.G. *Phase equilibria in iron ternary alloys - a critical assessment of the experimental literature*. (The Institute of Metals, London, UK, 1988).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017

Supplementary Information accompanies the paper on the *npj Computational Materials* website (doi:10.1038/s41524-017-0006-2).