ACS Combinatorial Science

Research Article

pubs.acs.org/acscombsci

Generating Information-Rich High-Throughput Experimental Materials Genomes using Functional Clustering via Multitree Genetic Programming and Information Theory

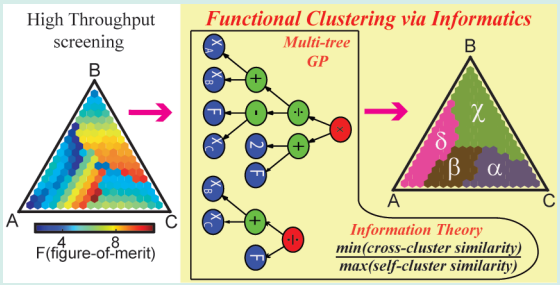
Santosh K. Suram,*[,‡] Joel A. Haber,[‡] Jian Jin,[†] and John M. Gregoire*[,‡]

[‡]Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, California 91125, United States

[†]Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States

Supporting Information

ABSTRACT: High-throughput experimental methodologies are capable of synthesizing, screening and characterizing vast arrays of combinatorial material libraries at a very rapid rate. These methodologies strategically employ tiered screening wherein the number of compositions screened decreases as the complexity, and very often the scientific information obtained from a screening experiment, increases. The algorithm used for down-selection of samples from higher throughput screening experiment to a lower throughput screening experiment is vital in achieving information-rich experimental materials genomes. The fundamental science of material discovery lies in the establishment of composition−structure−property relationships, motivating the development of advanced down-selection algorithms which consider the information value of the selected compositions, as opposed to simply selecting the best performing compositions from a high throughput experiment. Identification of property fields (composition regions with distinct composition-property relationships) in high throughput data enables down-selection algorithms to employ advanced selection strategies, such as the selection of representative compositions from each field or selection of compositions that span the composition space of the highest performing field. Such strategies would greatly enhance the generation of data-driven discoveries. We introduce an informatics-based clustering of composition-property functional relationships using a combination of information theory and multitree genetic programming concepts for identification of property fields in a composition library. We demonstrate our approach using a complex synthetic composition-property map for a 5 at. % step ternary library consisting of four distinct property fields and finally explore the application of this methodology for capturing relationships between composition and catalytic activity for the oxygen evolution reaction for 5429 catalyst compositions in a (Ni−Fe−Co−Ce)O_x library.



KEYWORDS: *materials genomes, high-throughput experimentation, combinatorial science, informatics, down-selection, clustering, functional relationships, multitree genetic programming, information theory*

1. INTRODUCTION

The main pillars in the realization of materials genome-based discovery are experimentation, first-principles computations and materials informatics. Several research efforts have focused on developing theoretical materials genomes for discovery of materials using first-principles calculations.[1−3] Materials informatics methods that efficiently mine data from elemental properties, experimental data and first-principles computations have also been developed and demonstrated as a framework for discovery of new materials.[4,5] Additionally, high-throughput and combinatorial experimentation approaches have led to discovery of new materials for several applications.[6−8] However, there is limited research toward developing a framework for combining informatics methods and high-throughput experimentation strategies to create information-rich experimental materials genomes that accelerate materials discovery and allow efficient integration with large scale computational materials science repositories.[1,9]

High-throughput (HiTp) experimentation typically involves the coarse, rapid measurement of a property of interest for each sample in a material library. Ultimately, the materials of greatest interest are selected for investigation using traditional techniques, which have much lower throughput. To transition between these two extremes, experiments with intermediate throughput and commensurate down-selection rates can be introduced to create a tiered screening scheme. Appropriate down-selection methods are essential to ensure generation of information rich experimental data that lead to knowledge and discovery. While a combinatorial material library may include variation of a number of process parameters such as synthesis temperature or processing parameters,[10,11] we continue this

Received: October 9, 2014
Revised: February 23, 2015
Published: February 23, 2015

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224−233

discussion in the context of composition libraries and note that the discussion is equally applicable to other parameters.

For characterization of composition-property relationships in mission-driven research, the property of interest is typically a performance metric for a target application. The relationship of this property to composition may be governed by any number of chemical and physical attributes, such as phase, crystallinity, microstructure and surface composition etc. While development of HiTp materials characterization[12−14] and related analysis techniques[15,16] is an active field of research, down-selection of samples from a HiTp screen for performing characterization is generally necessary. For a given composition region, a systematic variation in a materials characterization attribute may lead to a corresponding variation in the performance metric. By partitioning a composition space into regions which exhibit systematic trends in performance, samples can be selected for detailed characterization to capture the attribute-property relationships both within and among the composition regions.

Partitioning the composition region can be considered as a composition clustering exercise where the shape, size and location of the cluster in composition space is unknown. Most clustering methods that employ distance and density metrics have limited ability to identify arbitrarily shaped clusters, an issue with ongoing research.[17,18] Whereas, information theory based metrics provide access to higher order statistics[19−21] necessary for clustering/classification in complex data structures. Specifically, decision tree algorithms based on Shannon entropy criterion have been successfully applied as a supervised classification algorithm for unravelling crystal chemistry design rules[22] and discovery of materials.[4] Genetic programming, with its capability to identify arbitrary shaped clusters and perform ergodic optimization, has been successfully applied for supervised classification problems.[23] These desirable aspects result from its inherent concept of evolution of computer programs structured as genetic trees by iteratively performing selection, crossover and mutation operations. While other evolutionary techniques such as genetic algorithms[24] and particle-swarm optimization[25] have also been used for clustering data, they use cluster variance-based fitness metrics and hence are unable to capture nonhyperspherically shaped clusters.

While the above approaches and several others have been applied to (a) capture the function relating the input and output variables[26,27] and/or (b) cluster data based on input variables[28] and/or c) classify complex data structures in supervised classification;[29] there is limited work focused on clustering based on the (dis)similarity in the relationship between the input and output variables. Boric et al.[30] developed a multitree genetic programming based clustering approach that optimizes membership of each data point in a specified number of clusters by maximizing Cauchy−Schwarz divergence[31] cost function. Using this framework, we demonstrate a method that effectively considers systematic composition-property relationships as a metric of cluster membership.

In this article, we introduce the concepts of multitree genetic programming to a materials discovery application. In our approach, the genetic programming trees represent a function space that maps the compositions and HiTp property measurements to membership in a fixed number of clusters. The clustering is defined over the composition space such that the optimized trees cluster the compositions based on the *functional relationships* between composition and measured property. This method of clustering provides the ability to select representative compositions from each cluster for further investigation and characterization, resulting in information rich experimental materials genomes with respect to composition-characterization attribute-property relationships.

1.1. Traditional Approaches. Figure 1 shows a composition-property map for a 5 at. % step ternary library

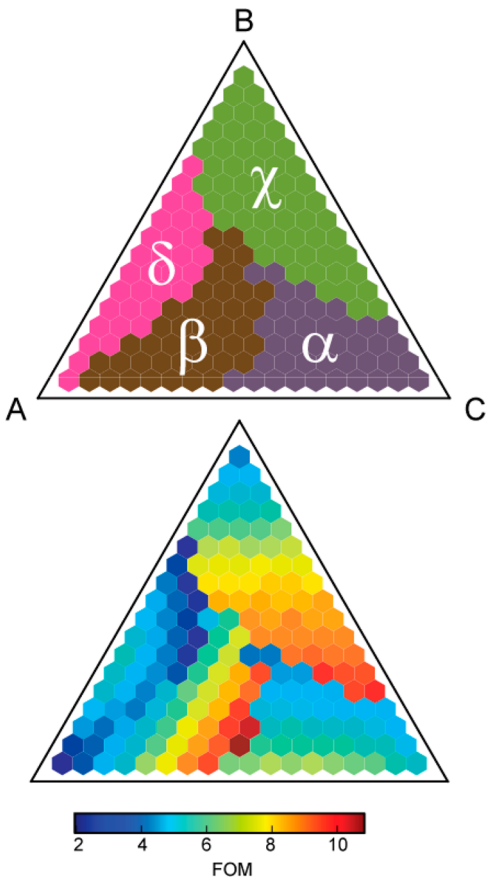


Figure 1. By partitioning a ternary composition space (with 5 at. % step) into 4 property fields (top), a synthetic composition−property plot is obtained by applying distinct polynomial functions to the compositions of each property field (bottom).

consisting of four property fields. The shapes of the property fields are chosen to be nonhyperspherical, as is the norm in HiTp data.[6,32] To generate this synthetic data set, compositions of each property field are mapped to a property value (also called figure-of-merit (FOM)) using distinct polynomial functions. Polynomial functions are chosen since they can approximate common empirical composition−property relationships of polynomial, exponential and logarithmic forms.[33−35] The polynomial functions were chosen to provide FOM variations by less than a factor of 6 to provide a challenge for the clustering algorithm and demonstrate its utility for property measurements with a small dynamic range.

Down-selection of samples from coarse-screening to finer screening procedures is usually performed by choosing the top z percentile of FOM values from the coarse screening technique. Tracking the changes in material characteristics by scanning from compositions with low FOM values to high FOM values is an essential step toward understanding chemistry-property relationships and tailoring chemistries to optimize specific properties. Using a simple percentile cutoff

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224−233

approach defeats this purpose by neglecting all compositions with FOM values that do not exceed the cutoff. Further, the value of z is typically chosen based on throughput matching between successive screening techniques, and we choose values of 5% and 10% for demonstrative purposes, as shown in Figure 2. In both the cases, compositions from property fields α and δ

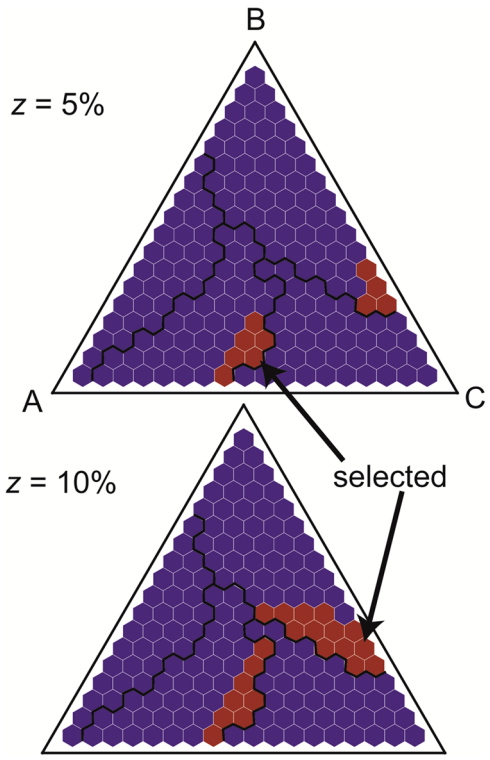


Figure 2. Down-selection of compositions by selecting top z percentile of compositions based on their FOM value (see Figure 1b). The downselected compositions, colored red, are very sensitive to the choice of z, which is usually fixed based on throughput matching of successive experiments. The property field boundaries from Figure 1 are overlaid for comparison.

in the synthetic data set (see Figure 1) would be filtered out from further screening experiments. Additionally, the compositions selected in property fields β and χ are insufficient to capture composition-property trends within these fields because the selected samples occupy only a small composition region of the respective property fields.

Another commonly used approach is to apply k-means algorithm to cluster the data based on FOM values, and the 4 clusters created using the FOM data in Figure 1 are shown in Figure 3 for 2 different inputs for k-means clustering. Using a Euclidian distance metric in FOM space, the clusters are scattered in the composition space, which is unreasonable for a materials science property. Including compositions and performing clustering on the composition−FOM space helps force the connectedness of clusters, and to enact this strategy the composition and FOM vectors were independently rescaled to attain unit standard deviation and provide equal importance to the variations in their values. This approach is successful in forming fairly connected clusters but as shown in Figure 3, the composition clusters fail to represent the property fields in Figure 1.

The above examples elucidate the need for an alternate approach that capture nonhyperspherical clusters and can

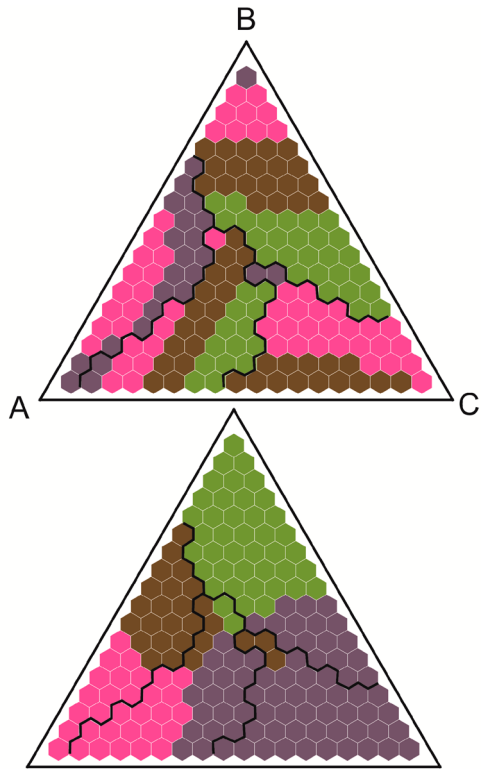


Figure 3. Clustering of the ternary composition library in Figure 1 using a Euclidean distance metric on the FOM space (top) and composition−FOM space (bottom). Clustering using only the FOM yields clusters with compositions scattered over the library, while adding the compositions to the clustering metric yields clusters that are mostly connected in composition space but do not match the original property fields, whose boundaries from Figure 1 are overlaid for comparison.

cluster based on composition−FOM *relationships* instead of FOM values. As discussed earlier, information-theoretic approaches provide access to higher order statistics and satisfy the former requirement. The latter requirement can be met with an appropriate implementation of genetic programming, which has the ability to learn complex data relationships. By combining these approaches, we develop a general framework for identifying property fields that exhibit unique composition−property relationships.

2. ALGORITHM AND DISCUSSION

2.1. Information−Theoretic Approach. Our objective is to cluster the composition space in a ternary library based on composition−FOM relationships. Alternately, we seek to cluster the composition space such that the similarity of composition−FOM relationships among different clusters is minimized while similarity of composition−FOM relationships within a given cluster is maximized. Using an information-theoretic approach, our objective can be stated as minimizing cross "between cluster" information potential while maximizing self "within cluster" information potential. An attractive metric that minimizes cross information potential and maximizes self-information potential for a two class system is the Cauchy−Schwarz divergence[31,36] and is expressed as

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224−233

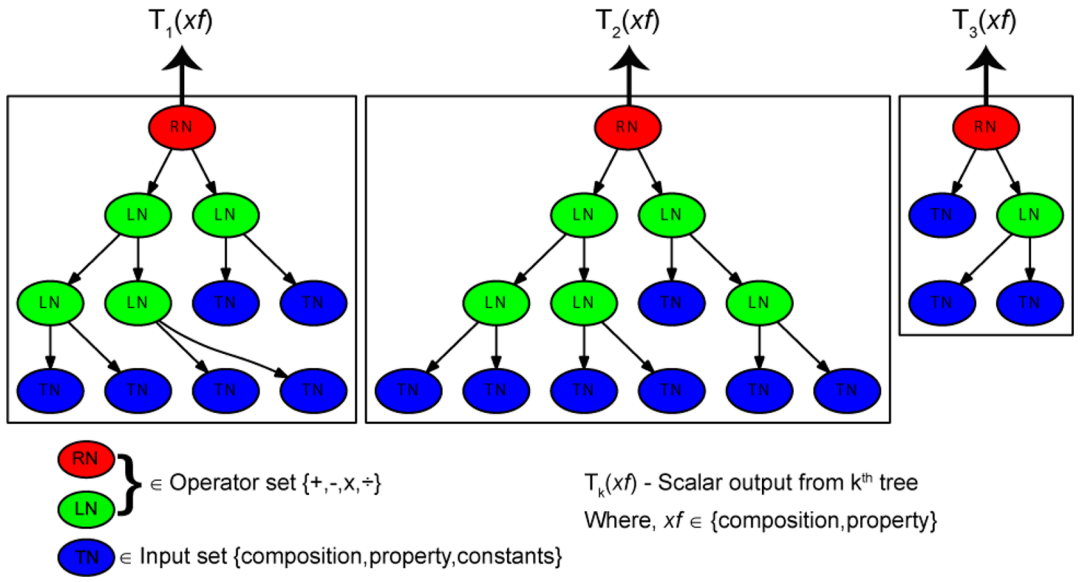


Figure 4. A schematic of a multitree chromosome in an MT-GP approach for 3 clusters and maximum depth 3. Abbreviations used: TN = terminal node, LN = leaf node, RN = root node.

$$D_{cs}(p_1, p_2) = -\ln \frac{\int p_1(x)p_2(x)\,dx}{\sqrt{\int p_1^2(x)\,dx \int p_2^2(x)\,dx}} \quad (1)$$

where $p_k(x)$ is the probability distribution of x in class C_k and x is the (multidimensional) composition coordinate.

In case of discrete data; probability distribution functions can be estimated using a Parzen window[31] with a Gaussian kernel as

$$p(x) = \frac{1}{n}\sum_{i=1}^{n} G(x - x_i, \sigma^2), \text{ where } G(x - x_i, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|(x - x_i)\|^2}{2\sigma^2}\right) \quad (2)$$

The kernel width, σ, is an apriori specified parameter; n is number of observations; d is the dimension of the data set.

Using eq 2, Jenssen et al.[20] show that the divergence function of eq 1 can be estimated as

$$D_{cs}(p_1, p_2) \approx -\ln \frac{\sum_{x_i \in C_1}\sum_{x_j \in C_2} G_{ij,2\sigma^2}}{\sqrt{\sum_{x_i, x_{i'} \in C_1} G_{ii',2\sigma^2} \sum_{x_j, x_{j'} \in C_2} G_{jj',2\sigma^2}}} \quad (3)$$

where $G_{ij,\sigma^2} \equiv G(x_i - x_j, \sigma^2)$.

From Figure 1, every composition in the ternary library belongs to exactly one property field. This relationship can be imposed onto clusters using a membership value (${}^i m_k$) for data point i in cluster k as

$${}^i m_{k'} = 1 \text{ for } k' = k \text{ and } {}^i m_{k'} = 0 \text{ for } k' \neq k \quad (4)$$

and ${}^i\mathbf{m}$ is defined as the vector of membership values for a data point i over the set of clusters. Using these membership notations and extending eq 3 to a c-cluster problem ($c \geq 2$), Boric et al.[30] express the Cauchy–Schwarz divergence function as

$$D_{cs}(p_1, p_2, \ldots p_c) \approx -\ln \frac{\frac{1}{2}\sum_{i,j=1}^{n}(1 - {}^i\mathbf{m}^{T\,j}\mathbf{m})G_{ij,2\sigma^2}}{\sqrt{\prod_{k=1}^{c}\sum_{i,j=1}^{n} {}^i m_k {}^j m_k G_{ij,2\sigma^2}}} \quad (5)$$

However, in this objective function, the denominator scales as a power of the number of clusters (c) whereas the numerator varies comparatively very slowly with c [see Supporting Information Figure 1]. Thus, as c increases, the denominator which quantifies self-information dominates the objective function, decreasing the importance of cross-cluster dissimilarity.

Therefore, we introduce a modified form of Cauchy–Schwarz divergence function such that the numerator and denominator remain invariant to the number of clusters:

$$D_{cs}(p_1, p_2, \ldots p_c) \approx -\ln \frac{\sqrt{\left(\sum_{i,j=1}^{n}(1 - {}^i\mathbf{m}^{T\,j}\mathbf{m})G_{ij,2\sigma^2}\right)\left(\frac{c}{c-1}\right)}}{c\left(\prod_{k=1}^{c}\sum_{i,j=1}^{n} {}^i m_k {}^j m_k G_{ij,2\sigma^2}\right)^{1/2c}} \quad (6)$$

This modified Cauchy–Schwarz divergence function is zero for random clusters, negative for correlated clusters and positive for divergent clusters (see Supporting Information Figure 2 for details of verification of our cost function using a set of random membership values).

Implementing eq 6 as the objective function in an optimization algorithm is facilitated by defining a continuous membership function, because the binary membership defined in eq 4 does not provide a continuous Cauchy–Schwarz divergence function with respect to alterations in membership of a given sample in a given cluster. In addition, to accurately cluster property fields the membership values should be based on the composition–FOM relationships. Thus, we introduce continuous membership values in the range [0, 1] by defining a membership function $m_k(xf)$ for each cluster such that ${}^i m_k = m_k(xf_i)$ where "i" represents the ith sample in the composition library. It is important to note that the probability distribution functions for Parzen window estimation are defined on the composition space, whereas the membership functions are

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224–233

defined on a combined composition and FOM space, with coordinate represented as *xf*. The inclusion of FOM in the parameter space enables the membership functions to represent composition−FOM relationships. Additionally, by constraining the membership values to sum to one, they can be regarded as a set of posterior probabilities:

$$m_k(xf) = P(C_k|xf), \sum_{k=1}^{c} m_k(xf) = 1 \quad (7)$$

2.2. Genetic Programming Based Clustering. Genetic trees are computer programs capable of learning complex relationships present in the data. In a *c*-class data set, there are *c* functional relationships between composition and FOM that need to be learned or distinguished from each other. Thus, we utilize a multitree genetic programming (MT-GP) framework developed by Muni et al.[37] and Boric et al.,[30] such that each tree learns the functional relationship between composition and FOM for one of the classes in the data. In this representation, each tree (T_k) is defined on the composition−FOM space where the scalar T_k (*xf*) is used to generate membership values, m_k (*xf*), as described below. The problem thus reduces to optimal identification of composition−FOM relationships by MT-GP such that the resulting membership values maximize the Cauchy−Schwarz divergence function (eq 6).

Our MT-GP algorithm was built on top of the framework provided by an open source evolutionary algorithms module, pyevolve.[38] The algorithm is based upon the construct illustrated in Figure 4, where each cluster is represented by a hierarchical tree of root, leaf and terminal nodes in the MT-GP chromosome. The leaf nodes and the root nodes are chosen from the set of operators {+, −, ×, ÷}. The terminal nodes are numerical and the domain includes the composition, FOM parameter space and random integer constants in [0, 10]. For the tree representing a cluster *k*, a sequence of operators comprising a nested algebraic function defined on *xf* terminate with numeric values T_k(*xf*).

Initialization. Chromosomes for genetic programming are built by populating nodes with operators from the operator set or data from the terminal set using the mechanism described below. If the node is not a root node and node-depth is less than a user-defined maximum depth, its value or operator is randomly selected from the operator set and terminal node set. For a root node, the operator is randomly selected from the operator set. For a node whose node depth is equal to the maximum depth, its data is randomly selected from the set of terminal nodes. During the selection of terminal nodes, the entire constant set enters the initial selection as a single parameter and if the constant set is selected then selection of the constant value is made randomly. Using the above strategy, 96 MT-GP chromosomes with four trees each were initialized.

Maximum Depth. The maximum depth of a genetic tree dictates the complexity of function that it is capable of learning. A small value of maximum depth may lead to inaccurate capture of the complexity in the data. A large value of maximum depth may result in very slow convergence of the MT-GP. Thus, this parameter needs to be carefully selected based on the properties of the data set. As an example choice of functional complexity, Xiajing et al.[35] used a third degree polynomial to relate chemical activity to composition for fuel cell materials. In general, third degree polynomials are sufficient to capture a function between composition and FOM, and we observe that a maximum depth of 4 is sufficient to capture a third order polynomial mapping between composition and FOM using the root, leaf and terminal node architecture described above.

Selection. At the beginning of every nonzeroth generation of genetic programming, every individual is populated by selecting the best chromosome out of tp randomly selected chromosomes from the previous generation. Where tp, the tournament pool size, is a user defined parameter that defines selection pressure (here, tp = 2). A large value of tp results in premature convergence. Thus, a relatively small value of tp compared to the number of chromosomes is recommended to exploit the exploratory capabilities of genetic programs. Additionally, the best five individuals from the previous generation are always selected for the next generation as part of an elitism retention strategy. The new population then undergoes crossover and mutation operations which allow tailoring of the genetic trees to learn the composition−FOM relationships.

Crossover. Crossover in a MT-GP approach differs from crossover in traditional genetic programming since a crossover between any two selected parent chromosomes with "*c*" trees can occur using ${}^{c}C_2$ pairs of parents because the *k*th tree in chromosome "*i*" does not have to crossover with the *k*th tree in chromosome "*j*" given that they may not be attempting to learn the same composition-property relationship. Thus, pairs of multitree chromosomes are selected as parents for crossover with a probability p_{cross} (here, set to 1). For each pair of multitree chromosomes selected as parents, pairs of trees are randomly selected with one tree from each of the parent chromosomes contributing to the pair such that every tree in the parent chromosomes is present in exactly one pair. A crossover probability of 0.75 is usually used as a balance between exploratory and exploitative capabilities of traditional genetic programs. In the case of MT-GP, rapid convergence to a robust solution is facilitated by using crossover probably low enough to avoid complete crossover of all tree-pairs and high enough to yield frequent crossover events. To achieve this balance, we parametrize the crossover probability for each pair of trees using a base probability ($p_{treecross}$) and probability multiplier (p_{cm}) such that the probability for crossover of the *k*th randomly selected pair of trees for a given pair of parent chromosomes is $p_{treecross} \times (p_{cm})^{k-1}$. Values of $p_{treecross}$ in the range 0.6−0.8 and p_{cm} in the range 0.8−1.0 were found to be reasonable estimates, although further research is required to identify optimal values of these parameters using various case studies. In this manuscript we use $p_{treecross}$ =0.7 and p_{cm} =0.9.

Mutation and Termination. Generally, probability of mutation is defined uniformly for all the chromosomes in a genetic program. However, this provides the same mutation frequency for simple and complex trees. Learning a complex composition−FOM function using a complex tree representation requires more exploration compared to that required for simple functions/trees. Thus, we consider probability of mutation (p_{mut}) for every branch. This allows larger trees to undergo a greater frequency of mutation than less complex trees. Here, we choose p_{mut} = 0.001 and we notice that this results in an average mutation rate per tree of approximately 2.5% for trees when maximum depth is 4. The MT-GP algorithm was terminated if a change in D_{cs} less than 0.02 was observed over 200 iterations. Typically, convergence was achieved in less than 2000 generations using 96 chromosomes. Using 24 cores within a computing node on Edison (https://www.nersc.gov/users/computational-systems/edison/configuration/) in a shared memory model, typical calculation times for each generation were 0.3 s for the synthetic data set

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224−233

with 231 samples and 0.7 s for the experimental data set with 5429 samples. The algorithm computation time scales sublinearly with number of samples. The evaluation of the objective function is the rate limiting step for large number of samples, and the crossover and mutation processes, which are dependent only on the number of chromosomes, become rate limiting for small number of samples. We observe a similar computing time for near 100% CPU usage on a 4 processor (8 threads) Intel Core i7-3770 CPU.

2.3. Calculating Membership. Boric et al.[30] related the output of the trees $T_k(xf)$ to membership values $m_k(xf)$ using a Sigmoid transformation followed by normalization:

$$T'_k(xf) = \frac{1}{1 + e^{-T_k(xf)}} \text{ and } m_k(xf) = \frac{T'_k(xf)}{\sum_{k'=1}^{c} T'_{k'}(xf)} \quad (8)$$

Since the output of each tree represents a distinct function in composition and FOM, different trees result in outputs of varying magnitudes. This could result in membership values that are skewed toward a particular function. To avoid this, we obtain relative memberships within each class by first normalizing the output of the trees $T_k(xf)$ with respect to the minimum and maximum values of $T_k(xf)$ and then normalizing the relative memberships such that $m_k(xf)$ represent posterior probabilities (eq 9).

$$T'_k(xf) = \frac{T_k(xf) - T_k^{\min}(xf)}{T_k^{\max}(xf) - T_k^{\min}(xf)} \text{ and } m_k(xf) = \frac{T'_k(xf)}{\sum_{k'=1}^{c} T'_{k'}(xf)} \quad (9)$$

The most representative class label set $\hat{k}(x)$ is computed using

$$\hat{k}(x) = \underset{k}{\operatorname{argmax}}(m_k(xf)) \quad ([10])$$

Note that the composition vectors and the FOM vectors are converted to unit standard deviation prior to the MT-GP analysis. Unit standard deviation ensures that variations in each feature vector are given equal importance.

3. RESULTS AND DISCUSSION

3.1. Synthetic Data Set. Figure 5 shows the optimal membership set obtained after clustering the data set shown in Figure 1 assuming the presence of four clusters. Here we choose a Gaussian Parzen window size $\sigma = 0.17$ at. %; based on the ability to obtain crisp, compositionally connected and robust membership values. Figure 5 also shows the clustering of compositions based on their maximum membership class ($\hat{k}(x)$). Since the number of property fields in the synthetic data set and the number of clusters used in the MT-GP algorithm are the same, the association of a synthetic property field and calculated cluster is easily made by evaluating the maximum intersection of the composition points. The clusters in Figure 5 are colored corresponding to the association of property fields in Figure 1, and comparison between these composition maps reveals 14 misclassified samples, approximately 8% of the data points. The misclassified samples lie on the boundaries between different property fields, where the continuous membership parameters show partial membership in each of the neighboring fields. That is, the MT-GP algorithm produces the correct property fields with the boundaries blurred by 1 or 2 composition intervals.

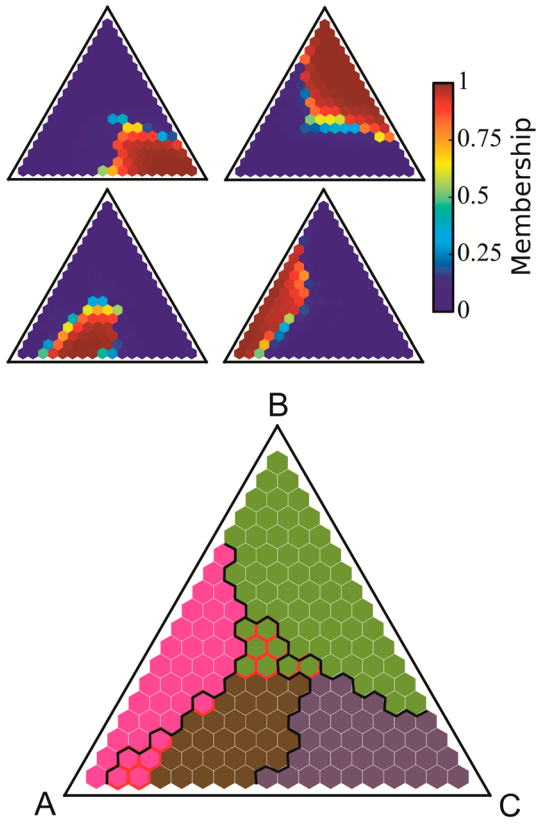


Figure 5. (top) Maps of the membership of each composition in the four optimized MT-GP trees. (bottom) The four clusters obtained by taking the maximum membership for each composition with the property field boundaries from Figure 1 overlaid for comparison. The 14 misclassified compositions are marked by red borders.

In the Supporting Information, we show that the optimal clustering obtained using Sigmoid function based transformation suggested by Boric et al.[30] has poor agreement to the synthetic property fields. We also highlight the loss of information during this transformation to elucidate the need for linear scaling based transformation (eq 9) for the MT-GP approach to accurately capture property fields.

3.2. Experimental Data Set. To demonstrate structure−property relationship clustering on experimental data, we use a combinatorial electrochemistry data set from the recent discovery of a family of electrocatalysts for the oxygen evolution reaction.[39] The metal oxide composition library covers all possibly mixtures of Ni, Fe, Co and Ce with 3.33 at. % composition steps, corresponding to 5456 samples. A quintessential FOM describing the performance of electrocatalysts for solar fuels applications is the overpotential (η) required to deliver a geometric current density of 10 mA cm^{-2}, where lower values correspond to higher performance. This FOM is mapped over the pseudoquaternary composition space Figure 6, where the (Ni−Fe−Co−Ce)O_x composition space is mapped not as a tetrahedron but instead as a series of Ni−Fe−Co pseudoternary triangles with increasing Ce concentration. With approximately 0.5% of missing data, the 5429 FOM values and corresponding 4-component compositions are used as the source data set for the MT-GP algorithm with 4 trees, each with maximum depth 4, and $\sigma = 0.17$. While the precise number of clusters, and hence number of MT-GP trees, required to capture all the features in the data set are unknown, we choose

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224−233

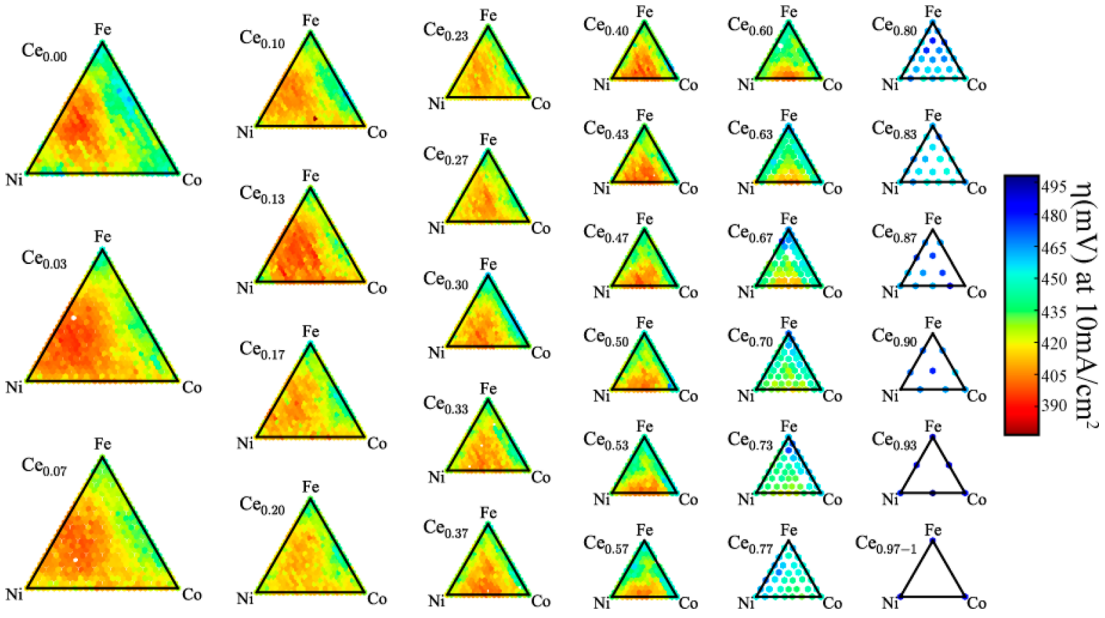


Figure 6. Overpotential (η) for oxygen evolution reaction at 10 mA cm^{-2} current density for 5429 catalyst compositions on a 3.33 at. % step (Ni−Fe−Co−Ce)O_x quaternary library.

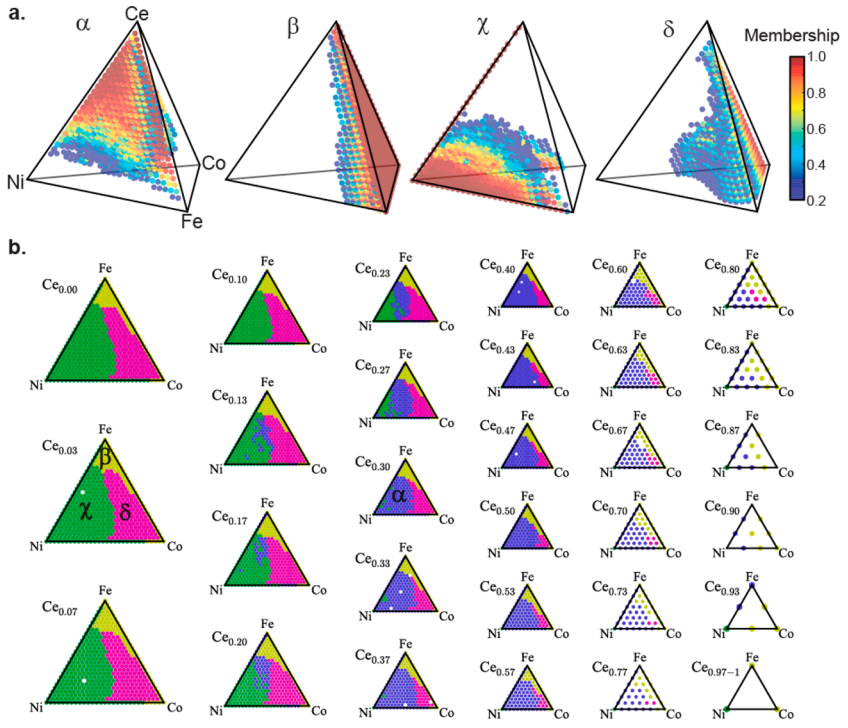


Figure 7. (a) Quaternary plots of membership values in four optimal clusters and (b) mapping of the most representative cluster onto quaternary compositions in a (Ni−Fe−Co−Ce)O_x library.

4 clusters to demonstrate the capability of our algorithm to capture important composition−FOM relationships.

Since division is one of the genetic operators and compositions along ternary faces, binary lines, and unary end points have at least one composition component as zero, all the compositions where shifted by $\Delta = 0.01$ at. % to avoid division by zero. For the synthetic data set described above, binary compositions were excluded to avoid this issue. The membership values for the 4 trees are shown as tetrahedral composition plots in Figure 7a, where only the points with membership in excess of 0.2 are shown and the points are plotted with 70% opacity to facilitate the visualization of the compositional clusters. Using maximum-membership to define representative clusters, the stacked-ternary representation of the 4 clusters is shown in Figure 7b and can be directly compared to Figure 6.

For this experimental data set, as with practically any experimental data set in high-order composition space, there is no known optimal solution for composition clusters. For this data set, 2 unique, highly active catalyst composition regions have been identified and classified through additional electrochemical characterization.[40] The recently discovered catalyst composition region contains little to no Fe and approximately

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224−233

50% Ce (region 1). Traditional mixed-transition-metal oxides with at least approximately 50% Ni comprise the low-Ce region of highly active catalysts (region 2).

Possible metrics for evaluating the clustering result include comparison with other measured properties such as crystallographic phase. As described in ref.,[41] these mixed metal oxides are X-ray amorphous and the greatest understanding of composition-property relationships in this system comes from extensive characterization of composition from Region 1, $Ni_{0.3}Fe_{0.07}Co_{0.2}Ce_{0.43}O_x$. The characterization experiments reveal that the behavior of this catalyst is markedly different from the compositions in region 2 because of the existence of a biphasic nanostructure.

To facilitate the interpretation of the clustering result of Figure 7, we consider the distribution of FOM values that exists within each cluster. The 4 histograms are shown in Figure 8

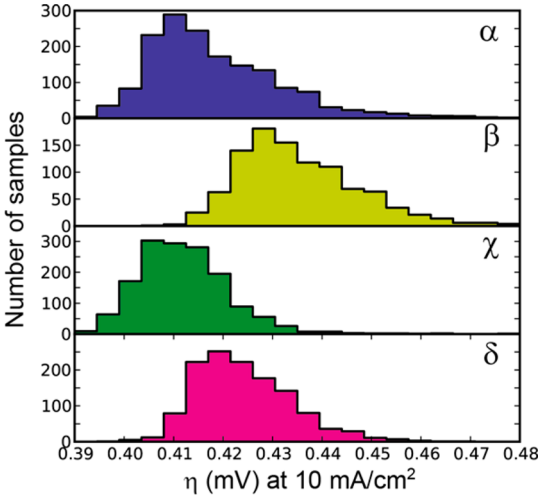


Figure 8. Histograms of the FOM for the 4 optimal clusters.

and clearly demonstrate that the MT-GP does not cluster by FOM value, as each of the 4 clusters contains samples with a wide range of FOM values. The 2 clusters that contain most of the best catalysts (lowest FOM values) are α and χ. The α cluster covers most of the Ni−Co rich compositions with Ce concentration in excess of 20 at. %, in excellent agreement with Region 1 described above. Likewise, the χ cluster covers Region 2 and is comprised of most Ni-rich compositions with Ce concentrations no more than 20 at. %. The identification and separation of these composition regions demonstrates that the MT-GP algorithm identifies the same composition clusters noted in experimental reports. There is no experimental basis for evaluating the 2 additional clusters, although we may draw insight from their compositional coverage. Composition cluster δ contains many Co-rich compositions, and composition cluster β traverses the Fe−Co−Ce ternary face, possibly signifying the unique behavior of Ni-free catalysts. Given the presence of experimental noise and limited dynamic range in the measured FOM, the excellent clustering results suggest that the MT-GP algorithm can be successfully deployed for automated down-selection routines, for example by choosing representative samples from each cluster or choosing samples that span the composition regions of the two high-performance clusters.

3.3. Parameter Optimization. The free parameters in our clustering approach are the Parzen window size (σ) and the number of clusters (c). From a coarse sensitivity analysis on the synthetic data set, we observe that $\sigma = 0.1$ is less than optimal window size and is likely to result in noisy membership functions. Whereas, $\sigma = 0.3$ misrepresents the features in the data set (see Supporting Information Figure 4). This also indicates that $\sigma = 0.42$ obtained from Silverman's rule of thumb in accordance with the protocol proposed by Jenssen et al.[20] is inapplicable for this data set. While a value of $\sigma = 0.17$ captures the complexities in the synthetic data set, further sensitivity analyses is required to identify an acceptable range for σ to enable automated execution of the MT-GP algorithm.[42] Additionally, automatic detection of the number of clusters is required to easily adapt this algorithm to experimental data sets, where no prior knowledge on the number of clusters is available. The requirement for a priori specification of the number of clusters is a limitation common to many clustering algorithms. However, the modified Cauchy−Schwarz divergence function introduced in eq 6 allows us to quantitatively compare the divergence information obtained for various numbers of clusters, motivating further research for automatic determination of the optimal number of clusters.

Additional future work will incorporate different approaches for parametrizing composition space with consideration of the implications for standardizing the source data and defining the objective function. In the approach described above, the transformations from discrete data to probability distribution functions and the connectedness of clusters employ the Gaussian kernel defined in Euclidean space. By treating composition variables as Euclidean coordinates, the algorithm successfully identified clusters in the composition space, although with apparent artifacts on the edges of the experimental composition space. For example, composition cluster χ in Figure 7 is a compact cluster of Ni-rich compositions with the exception of an outcropping of compositions along the binary Ni−Ce and Ni−Co lines, possibly affected by composition shift using offset parameter Δ. Clustering of these low-order compositions is sensitive to the choice of Δ, and in general, the application of non-Euclidean compositional distance metrics needs to be explored.

While there is a need for further research to develop a nonparametric MT-GP based clustering algorithm that ultimately provides automatic down-selection of compositions for combinatorial experimentation, this article using information theory, a modified Cauchy−Schwarz divergence function and multitree genetic programming establishes a protocol for identifying distinct, complex composition-property fields from combinatorial materials science data. This methodology presents a significant step toward developing information-rich experimental materials genomes.

Summary. High-throughput experiments generally produce a figure of merit for each sample in a material library. Clustering samples by a measured performance metric does not facilitate the selection of samples required for establishing composition-property relationships. We present a new algorithm based on genetic programming and information theory which clusters samples by the functional relationship between composition and figure of merit (FOM). The membership of the samples in a given cluster is represented by a genome of algebraic operations on the source composition and FOM data, where the algebraic operations are indicative of the measured composition−FOM trend in a particular composition region. By implementing this algorithm in a sample down-selection scheme, the information value of the sample subset can be maximized with respect to understanding composition-property

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224−233

relationships, guiding data-driven material discoveries. We demonstrate this approach by clustering composition regions with distinct composition−FOM relationships in a ternary synthetic data set, where the synthetic composition clusters are well-reproduced by the automated algorithm. By applying the genetic program clustering to 5429 measurements of oxygen evolution electrocatalytic activity in the $(Ni-Fe-Co-Ce)O_x$ composition space, the 2 distinct catalyst composition regions from the literature are correctly identified. The successful application of the algorithm to both synthetic and experimental data sets demonstrates its utility for the development of autonomous down-selection schemes for rapid mapping of composition-property relationships.

■ ASSOCIATED CONTENT

Supporting Information

Plots showing properties of the Cauchy−Schwarz divergence function and false color maps illustrating genetic programming trees and clustering algorithm. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: sksuram@caltech.edu. Office: 626-395-2606. Fax: 626-395-1577

*E-mail: gregoire@caltech.edu. Office: 626-395-2613. Fax: 626-395-1577.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is performed by the Joint Center for Artificial Photosynthesis, a DOE Energy Innovation Hub, supported through the Office of Science of the U.S. Department of Energy under Award Number DE-SC000499. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors thank Dr. Misha Z. Pesenson for helpful discussions.

■ REFERENCES

(1) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, No. 011002.

(2) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191−201.

(3) Wu, Y.; Lazic, P.; Hautier, G.; Persson, K.; Ceder, G. First Principles High Throughput Screening of Oxynitrides for Water-Splitting Photocatalysts. *Energy Environ. Sci.* **2013**, *6*, 157.

(4) Balachandran, P. V.; Broderick, S. R.; Rajan, K. Identifying the "Inorganic Gene" for High-Temperature Piezoelectric Perovskites through Statistical Learning. *Proc. R. Soc. London, Ser. A* **2011**, *467*, 2271−2290.

(5) Dey, P.; Bible, J.; Datta, S.; Broderick, S.; Jasinski, J.; Sunkara, M.; Menon, M.; Rajan, K. Informatics-Aided Bandgap Engineering for Solar Materials. *Comput. Mater. Sci.* **2014**, *83*, 185−195.

(6) Green, M. L.; Takeuchi, I.; Hattrick-Simpers, J. R. Applications of High Throughput (combinatorial) Methodologies to Electronic, Magnetic, Optical, and Energy-Related Materials. *J. Appl. Phys.* **2013**, *113*, No. 231101.

(7) Potyrailo, R.; Rajan, K.; Stoewe, K.; Takeuchi, I.; Chisholm, B.; Lam, H. Combinatorial and High-Throughput Screening of Materials Libraries: Review of State of the Art. *ACS Comb. Sci.* **2011**, *13*, 579−633.

(8) Rajan, K. Combinatorial Materials Sciences: Experimental Strategies for Accelerated Knowledge Discovery. *Annu. Rev. Mater. Res.* **2008**, *38*, 299−322.

(9) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O. AFLOWLIB.ORG: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.* **2012**, *58*, 227−235.

(10) Caskey, C. M.; Richards, R. M.; Ginley, D. S.; Zakutayev, A. Thin Film Synthesis and Properties of Copper Nitride, a Metastable Semiconductor. *Mater. Horizons* **2014**, *1*, 424.

(11) Chikyow, T.; Ahmet, P.; Nakajima, K.; Koida, T.; Takakura, M.; Yoshimoto, M.; Koinuma, H. A Combinatorial Approach in Oxide/Semiconductor Interface Research for Future Electronic Devices. *Appl. Surf. Sci.* **2002**, *189*, 284−291.

(12) Gregoire, J. M.; Dale, D.; Kazimirov, A.; DiSalvo, F. J.; van Dover, R. B. High Energy X-Ray Diffraction/x-Ray Fluorescence Spectroscopy for High-Throughput Analysis of Composition Spread Thin Films. *Rev. Sci. Instrum.* **2009**, *80*, 123905.

(13) Kan, D.; Long, C. J.; Steinmetz, C.; Lofland, S. E.; Takeuchi, I. Combinatorial Search of Structural Transitions: Systematic Investigation of Morphotropic Phase Boundaries in Chemically Substituted BiFeO3. *J. Mater. Res.* **2012**, *27*, 2691−2704.

(14) Hattrick-Simpers, J. R.; Hurst, W. S.; Srinivasan, S. S.; Maslar, J. E. Optical Cell for Combinatorial in Situ Raman Spectroscopic Measurements of Hydrogen Storage Materials at High Pressures and Temperatures. *Rev. Sci. Instrum.* **2011**, *82*, No. 033103.

(15) 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.

(16) 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. *Proc. 17th Int. Conf. Princ. Pract. Constraint Program* **2011**, 508−522.

(17) Chaoji, V.; Hasan, M. Al; Salem, S.; Zaki, M. J. SPARCL: Efficient and Effective Shape-Based Clustering. *2008 Eighth IEEE Int. Conf. Data Min.* **2008**, 93−102.

(18) Wan, R.; Wang, L.; Su, X. ASCCN: Arbitrary Shaped Clustering Method with Compatible Nucleoids. *Int. J. Data Warehousing Min.* **2010**, *6*, 1−15.

(19) Gokcay, E.; Principe, J. C. Information Theoretic Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 158−171.

(20) Jenssen, R.; Erdogmus, D.; Hild, K.; Principe, J. C.; Eltoft, T. Optimizing the Cauchy−Schwarz PDF Distance for Information Theoretic, Non-Parametric Clustering. *Int. Work. Energy Minimization Methods Comput. Vis. Pattern Recognit.* **2005**, 34−35.

(21) Jaynes, E. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620−630.

(22) Kong, C. S.; Luo, W.; Arapan, S.; Villars, P.; Iwata, S.; Ahuja, R.; Rajan, K. Information-Theoretic Approach for the Discovery of Design Rules for Crystal Chemistry. *J. Chem. Inf. Model.* **2012**, *52*, 1812−1820.

(23) Muni, D. P.; Pal, N. R.; Das, J. A Novel Approach to Design Classifiers Using Genetic Programming. *IEEE Trans. Evol. Comput.* **2004**, *8*, 183−196.

(24) Bandyopadhyay, S.; Maulik, U. Nonparametric Genetic Clustering: Comparison of Validity Indices. *IEEE Trans. Syst. Man Cybern. Part C (Applications Rev.* **2001**, *31*, 120−125.

(25) Van der Merwe, D. W.; Engelbrecht, A. P. Data Clustering Using Particle Swarm Optimization. *2003 Congr. Evol. Comput.* **2003**, *1*, 215−220.

(26) Broderick, S. R.; Rajan, K. Eigenvalue Decomposition of Spectral Features in Density of States Curves. *EPL* **2011**, *95*, No. 57005.

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224−233

(27) Shi, X.; Luo, J.; N. Njoki, P.; Lin, Y.; Lin, T.-H.; Mott, D.; Lu, S.; Zhong, C.-J. Combinatorial Assessment of the Activity-Composition Correlation for Several Alloy Nanoparticle Catalysts. *Ind. Eng. Chem. Res.* **2008**, *47*, 4675−4682.
(28) Potyrailo, R.; Mirsky, V. M. *Combinatorial Methods for Chemical and Biological Sensors*; Springer Science & Business Media: New York, 2009; p 125.
(29) Li, H.; Liang, Y.; Xu, Q. Support Vector Machines and Its Applications in Chemistry. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 188−198.
(30) Boric, N.; Estévez, P. A. Genetic Programming-Based Clustering Using an Information Theoretic Fitness Measure. *2007 IEEE Congr. Evol. Comput. (CEC 2007)* **2007**, 31−38.
(31) Jenssen, R.; Principe, J. C.; Erdogmus, D.; Eltoft, T. The Cauchy−Schwarz Divergence and Parzen Windowing: Connections to Graph Theory and Mercer Kernels. *J. Franklin Inst.* **2006**, *343*, 614−629.
(32) Gregoire, J. M.; Xiang, C.; Liu, X.; Marcin, M.; Jin, J. Scanning Droplet Cell for High Throughput Electrochemical and Photoelectrochemical Measurements. *Rev. Sci. Instrum.* **2013**, *84*, 024102.
(33) Saunders, N.; Miodownik, A. P. *CALPHAD (Calculation of Phase Diagrams) A Comprehensive Guide*; Elsevier: New York, New York, USA, 1998; pp 91−129.
(34) Srinivasan, S.; Rajan, K. Property Phase Diagrams for Compound Semiconductors through Data Mining. *Materials (Basel)* **2013**, *6*, 279−290.
(35) Shi, X.; Luo, J.; N. Njoki, P.; Lin, Y.; Lin, T.-H.; Mott, D.; Lu, S.; Zhong, C.-J. Combinatorial Assessment of the Activity-Composition Correlation for Several Alloy Nanoparticle Catalysts. *Ind. Eng. Chem. Res.* **2008**, *47*, 4675−4682.
(36) Principe, J.; Xu, D.; Fisher, J. Information Theoretic Learning. In *Unsupervised Adaptive Filtering*; John Wiley and Sons: New York, 2000; Vol. *1*, pp 265−319.
(37) Muni, D. P.; Pal, N. R.; Das, J. A Novel Approach to Design Classifiers Using Genetic Programming. *IEEE Trans. Evol. Comput.* **2004**, *8*, 183−196.
(38) Perone, C. S. Pyevolve: A Python Open-Source Framework for Genetic Algorithms. *ACM SIGEVOlution* **2009**, *4*, 12−20.
(39) Haber, J. A.; Cai, Y.; Jung, S.; Xiang, C.; Mitrovic, S.; Jin, J.; Bell, A. T.; Gregoire, J. M. Discovering Ce-Rich Oxygen Evolution Catalysts, from High Throughput Screening to Water Electrolysis. *Energy Environ. Sci.* **2014**, *7*, 682.
(40) Haber, J. A.; Xiang, C.; Guevarra, D.; Jung, S.; Jin, J.; Gregoire, J. M. High-Throughput Mapping of the Electrochemical Properties of (Ni-Fe-Co-Ce)Ox Oxygen-Evolution Catalysts. *ChemElectroChem.* **2013**, *0000*, 1−5.
(41) Haber, J. A.; Anzenburg, E.; Yano, J.; Kisielowski, C.; Gregoire, J. M. Multi-Phase Nanostructure of a Quinary Metal Oxide Electrocatalyst Reveals a New Direction for OER Electrocatalyst Design. *Adv. Energy Mater.* **2015**, DOI: 10.1002/aenm.201402307.
(42) Jenssen, R.; Principe, J. C.; Eltoft, T. Cauchy−Schwartz Pdf Divergence Measure for Non-Parametric Clustering. Presented at the IEEE Norway Section Interanational Symposium on Signal Processing, Bergen, Norway, 2003.

DOI: 10.1021/co5001579
ACS Comb. Sci. 2015, 17, 224−233