University of Maryland, College Park

AMSC 663/664

# Pattern Decomposition of Inorganic Materials: Optimizing Computational Algorithm

*Author:*
Graham Antoszewski
ganto@math.umd.edu

*Supervisor:*
Dr. Hector Corrada-Bravo
Center for Bioinformatics and Computational Biology
University of Maryland, Department of Computer Science
hcorrada@umiacs.umd.edu

October 6, 2016

**Abstract**

Phase pattern decomposition of inorganic materials' crystalline structure is extremely important for the unearthing of new properties such as superconductivity. Previously, this process had meticulously been done by hand, so computer algorithms have been developed to try and uncover these phases. They, however, have yet to combine efficiency and accuracy together. The goal of this project is to do just that by extending the Graph-based endmember extraction and labeling algorithm (GRENDEL). Phase one will be to incorporate physical constraints and prior knowledge to increase the accuracy of our phase composition results, and phase two will be to utilize active learning to minimize the number of sample points needed to analyze a given material to increase efficiency.

# 1 Background Information

Inorganic materials are compounds or mixtures of elements which do not contain any carbon. Of particular interest are combinations of metal alloys called ternary systems, where three different metallic compounds are heated up and combined into one. Because of the heating and cooling process, the crystalline structure of each individual metal has been altered, similar to how an ice cube that is melted and refrozen will not be identical to the initial configuration. This means the phase of the metal has changed, as the phase is defined as a region within a material or compound where the crystal structure and composition is uniform [1]. This means these phases have distinct properties, such as density and index of refraction. Within different areas of the ternary alloy, there can be different phases of each metal as well due to how the atoms restructured and the proportions of each compound at the given point. Each point of this material is made of a different composition of the three input metals, meaning there can be three phases present and at different proportions based on the mixing process of the alloy. An example of a typical thin film sample of a ternary system is seen in Figure 1.

A given phase of a metal, as previously stated, has distinct properties, one of these being a unique diffraction pattern. X-ray diffraction is used to probe a given material, sending in beams of electrons at various intensities and observing the outgoing spectra [1]. An X-ray has a wavelength that is approximately the same as the distance between atoms in a crystal lattice, giving it a better chance to hit the atoms within the structure. The light will hit an electron in the metal, absorb energy, and bounce back at a given angle. Note that this energy exchange only happens at certain incident angles, which is dictated by the Bragg equation,

$$2dsin\theta = n\lambda, \tag{1}$$

where $d$ is the distance between atoms in the lattice, $\theta$ is the incoming scattering angle, $\lambda$ is the wavelength of the x-ray, and $n$ is an integer. The absorbed energy is seen as diffraction peaks at the given angles which satisfy equation 1.

But for an unknown phase, we do not know the distance $d$. Thus, both the source of x-ray light and the detector rotate in order to record data over all possible scattering angles $2\theta \in [0°, 90°]$ ($2\theta$ is defined as the angle between the detector and the incident beam rather than the plan of the material). Figure 2 shows an example of an x-ray spectrum for a single

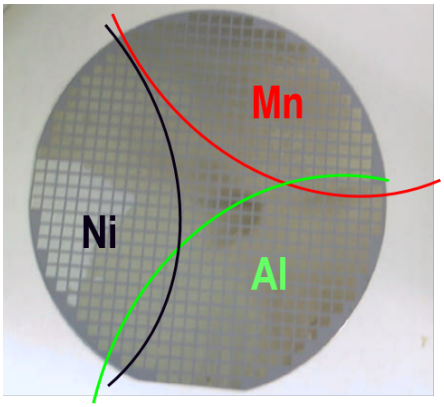


Figure 1: Seen is a thin film of an $Al$-$Mn$-$Ni$ ternary system, with the regions specified by each color being the predominant areas of each of the composite metals. That is, where each metal was initially introduced into the alloy and then mixed. This highlights how the mixing throughout the material is not uniform, as we want to find all possible combinations of constituent phases [2].

sampled point in a material, noting that we probe the data not only over various angles but also a range of light intensities [2]. Each line represents a waveform, with constant lines indicating the light did not interact with the material as the overall energy of the light did not change. Peaks in the diffraction spectrum indicate a detection of a certain phase.

The are three main aspects of a given diffraction peak. The scattering angle $2\theta$ tells us about the metal observed and its particular phase, while the height and width of a given peak can tell us about the given proportion of that phase in the mixture of metals that make up the alloy at the sampled point. One can also notice a shifting of the position of seen peaks over different light intensities, which is a source of error and something that has to be accounted for. Using this data, we can recognize the constituent phases seen in the material along with their respective proportions, and a phase diagram can be made like the one seen in Figure 3 [1]. The alloyed material we wish to sample is usually on a circular thin film, yet we transform the data taken from this shape into a simplex, where each vertex corresponding to the locations of the three initial compounds at the start of the mixing process. Different colors represent areas within the material where similar phase structure is seen, indicating these whole regions will have the same intrinsic chemical properties. Each dot or marker on the simplex corresponds to a probed sample point.

## 2 Project Objective

Previously, these phase diagrams were done by hand, eying the proportions of the constituent phase composition. This process took so long that a library of materials has already been created which have yet to be analyzed. Thus, the White House Materials Genome Initiative was started in order to encourage an algorithm to take in this structure and composition data as an input and output the desired phase diagrams and phase composition data. This algorithm must obey the laws of physics while also accurately identifying the individual

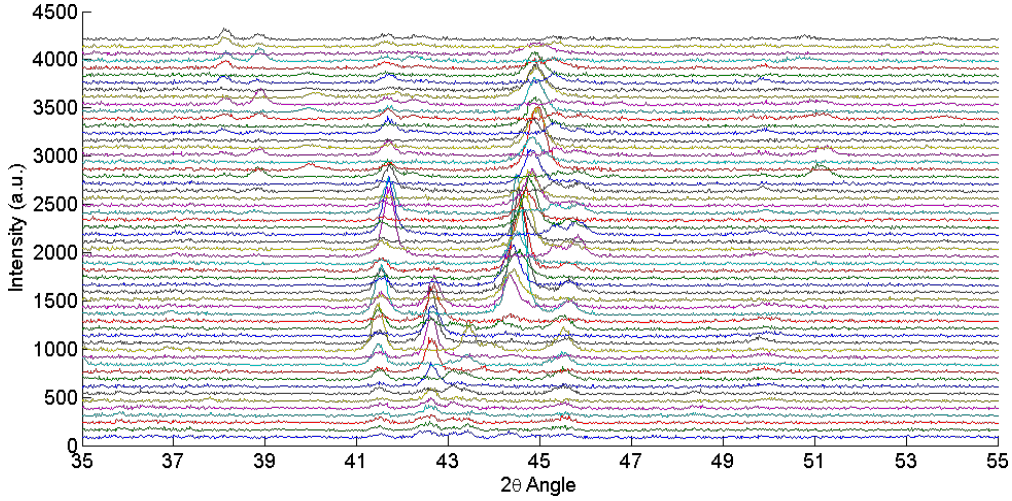


Figure 2: A sample x-ray spectrum, with the x-axis being $2\theta$ = the scattering angle observed and the y-axis being the intensity of light detected. Peaks on this plot represent material detection corresponding to given phases of our metals [2].

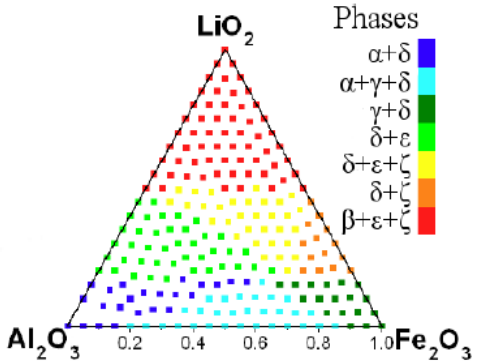


Figure 3: An example of a phase diagram, represented as a simplex. Each vertex corresponds to one of the original compounds in the alloy, colors correspond to similar phase structure between those points, and the Greek symbols in the legend represent the different phases seen in the material [1].

phases, regions or clusters of similar phase composition, and do all of this in an efficient manner so more materials can be evaluated [3].

Current attempts at an algorithm focus on pattern decomposition. Given a set of patterns at $N$ points of a given system, it is assumed these can be described as a combination of $K$ basis patterns. We seek to resolve these basis patterns, which in this case are the $K$ constituent phases that contribute to the diffraction patterns seen in the material. There are two main steps, the first being spectral clustering. Here a similarity matrix is constructed to group points in the material with analogous diffraction patterns, with each group of points being called a *cluster*. This splits up our entire dataset into smaller subproblems, allowing the algorithm to run more efficiently. Second, nonnegative matrix factorization is used to identify the constituent phases and their proportions within each cluster [1]. These steps will be explained in detail in Section 3.

One example of such an algorithm is Graph-based endmember extraction and labeling (GRENDEL). *Endmember* is another word for the constituent phase which make up the phase diffraction pattern of a given point or region within the material. Previous research with GRENDEL [3] uses the term endmembers rather than constituent phases, and seeing as we will be working with GRENDEL collaborations we will keep the same convention for this project. This method seeks to minimize an objective function during the pattern decomposition process, which looks at how well our estimated phase proportions match up with the raw diffraction patterns both within the clusters and over the entire material. GRENDEL runs very fast, with computation times under a minute, but fails to properly take into account physical constraints which leads to inaccuracy [3]. Another attempt at an algorithm, Alternating Mixed Integer Quadratic Optimization, uses mixed integer quadratic problems to minimize the error. Essentially, this boils down to minimizing norms regarding residuals between the original structure and our hypothesized phase structure, yet this method uses prior knowledge to add in physical constraints. It recognizes certain pairs of points and phases that Must-Link and Cannot Link together, which leads to extremely accurate results [4]. AMIQO runs on the order of days, however, so we wish to incorporate the accuracy of AMIQO with the efficiency of GRENDEL.

Furthermore, it takes about a half hour to obtain a full x-ray diffraction spectrum for each point in a material, meaning it can take over a week to go through an entire material. Another goal then must be to find a way to have our algorithm use a minimal number of data points to evaluate an entire material, as an algorithm that runs quickly does no good if the input data cannot be collected quickly too. This requires a program that, rather than taking entire material's spectrum as input, iteratively takes in individual sample points and suggests the most informative point in the material to be the next to be probed.

In summary, current approaches at an algorithm either are computationally inefficient or they yield physically erroneous results. In addition, the sampling process of retrieving new x-ray diffraction data takes too long so even if our program runs quickly, we still cannot do rapid analysis of new materials. Our project objective is to address all three of these issues in one algorithm to combine speed, accuracy, and an optimized sampling process. To do so, will be working to extend the GRENDEL algorithm. GRENDEL begins with a spectral clustering step in order to create initial cluster assignments for all of our sample points within our given material dataset. Then, the objective function, which contains a nonnegative matrix factorization term and a Graph Cut term, is minimized through an iterative process.

Once convergence is attained, GRENDEL will output cluster assignments for each of the sample points within the material as well as a set of constituent phases/endmembers for each cluster [3]. From this output, phase diagrams outlining regions of similar phase structure (clusters) and constituent phase compositions of each cluster can be generated, with an example of a phase diagram seen in Figure 3.

The two new components which we will add to GRENDEL is constraint programming and active learning. Constraints can be added to the objective function, which can be laws of physics or properties of the material that we know prior to computation, to make our solution more accurate and physically realistic. The objective function already has certain constraints, and the challenge of this portion of the project will be uncovering the most important constraints to use, as adding constraints tends to increase the run time of objective function step of GRENDEL.

Active learning pertains to optimization of the sampling process. Rather than using all sample points of a given material, we now realize that our algorithm will be running in conjunction with the sampling process, meaning the input spectral data will be read in iteratively point by point. We will take a small subset of sample points, run GRENDEL to output the current guess of clusters and phases/endmembers, then generalize this analysis to create a predictive phase diagram for the whole material. To estimate the endmember composition of unknown regions of the material where we have not probed, we do a distance-weighted average from the results at the data points we do have. The next point to sample is then chosen by determining the area of highest cluster uncertainty, coinciding with this region's estimated endmember composition conflicting with the cluster compositions output by GRENDEL. This will typically fall at cluster boundaries, as the endmembers of neighboring clusters will differ and overlap to create error [5]. The goal is to probe these areas of uncertainty and skip probing areas that may be superfluous, such as a middle of a large cluster whose endmember composition is already known, to reduce the number of points we need to achieve minimization of the objective function. Further explanation of these new features of GRENDEL are explained in Sections 4.1 and 4.2. Before this, however, the current GRENDEL algorithm must be explained in detail.

# 3 Algorithm - GRENDEL

Figure 4 illustrates the flow of the current implementation of GRENDEL. As an input, both structure and composition data from the Inorganic Crystal Structure Database and other material libraries can be utilized. If $X$ is the input structure and composition data for the whole material of $N$ sample points, we will look at each individual point $X_i$. $X_i$ has dimensions $D \times R$, where $D$ is the number of scattering angles observed. For example, we know $2\theta \in [0, 90]$, so if data is taken with resolution of $0.1^\circ$, then $D = 900$. $R$ is the number of input intensity values used during probing, typically around 50. A given element $X_{i,jr}$ is itself a scattering intensity value seen at the given $j^{th}$ scattering angle by the detector for the $r^{th}$ input probe intensity. To graphically see where each of the $N$ sample points are in terms of our phase diagram, each marker seen on the simplex of Figure 3 is a sample point. Structural data of the material is used in Section 3.2 to place a given sample point $i$ in the correct position on the simplex, whereas the spectral data $X_i$ is used to determine the cluster

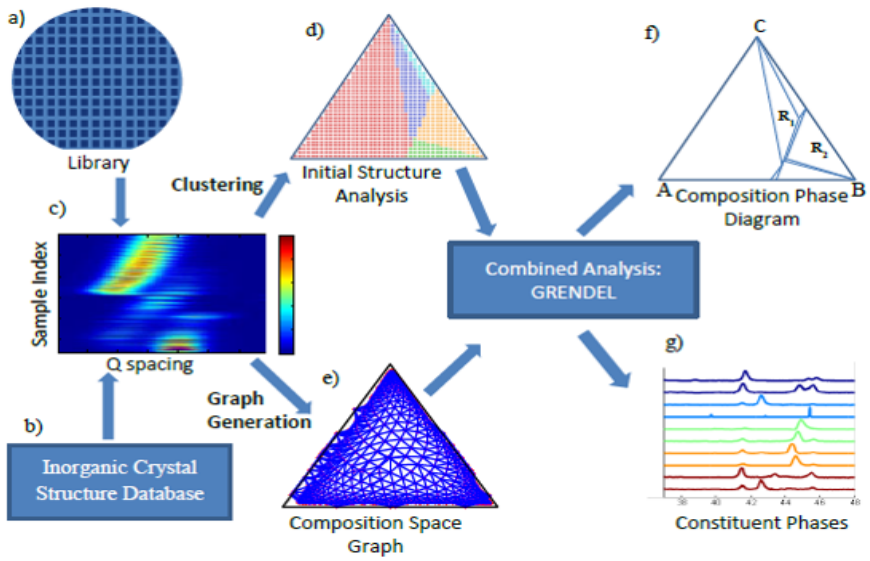


Figure 4: A flow chart of the GRENDEL algorithm. Initial structure and composition data is input, and spectral clustering is done to find areas of similar diffraction spectra. Then, an iterative process of the Graph Cut algorithm and NMF (seen here in the 'Combined Analysis - GRENDEL' box of the flow chart) is implemented in order minimize the objective function which resolve cluster boundaries and the constituent phases within each cluster. Final output plots are the phase diagram and constituent phase plot [3].

assignment and phase composition of each cluster, seen as the different colors for each of the markers [3].

## 3.1 Spectral Clustering

Spectral clustering seeks to separate the material data into regions or clusters of similar structure, thus allowing the proceeding steps to be run on smaller subsets to speed up computation. These clusters will also be areas of analogous chemical properties. First, a similarity measure is used to compare how close the diffraction patterns are between two given sample points [6]. This metric is the cosine distance between 2 sample point matrices $X_i$ and $X_j$, given by

$$\delta_{\cos}(X_i, X_j) = 1 - \cos(X_i, X_j), \tag{2}$$

where $\cos(X_i, X_j)$ is the cosine similarity between the two matrices, defined by

$$\cos(X_i, X_j) = \frac{X_i \cdot X_j}{||X_i||\ ||X_j||}. \tag{3}$$

Here, $\cdot$ is the dot product and $||\cdot||$ is the L2 norm, and note that we must vectorize our given matrices in order to do this. Thus, the cosine distance between our points $X_i$ and $X_j$ will be near zero if the diffraction patterns match well and near 2 if they are contradictory [3]. Then, we create a similarity matrix from these cosine distances,

$$S_{ij} = e^{\frac{\delta_{\cos}(X_i, X_j)}{2\sigma^2}}, \tag{4}$$

where $\sigma$ is the spectral clustering bandwidth parameter specific to the given material we are observing. A goal of this project is to optimize this parameter for any given material.

With this, we utilize the K-means function within MATLAB to identify individual clusters of points with similar structures. Using K-means, it is assumed that there are $K$ clusters present. An initial assignment of clusters is made, each with a cluster *centroid*. Since we use the similarity matrix in Equation 4, the mean similarity of each group $\bar{S}_k$ is found by minimizing the total cluster variance between the similarity data of each point $S_i$ in the $k^{th}$ cluster $C_k$ and the desired mean similarity,

$$\min_{C,\{\bar{S}_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||S_i - \bar{S}_k||^2, \tag{5}$$

where $N_k$ is the number of sample points contained in the $k^{th}$ cluster and $C$ is the given cluster assignments of the $i$ data points. Now that we have the mean similarities of each cluster $\bar{S}_k,\ k \in [1, K]$, the data points are then reassigned to clusters where their similarity metrics most resemble the given mean cluster similarity:

$$C(i) = \underset{1 \le k \le K}{\arg\min} ||S_i - \bar{S}_k||^2. \tag{6}$$

This process is repeated until the cluster assignments no longer change, yielding cluster assignments for each data point $X_i,\ i \in \{1, N\}$ [6]. Furthermore, we can also find the mean spectral composition $\bar{X}_k$ of each cluster now by averaging the spectral data of all points within the $k^{th}$ cluster.

## 3.2 Creating the Simplex

The Composition Space graph seen in part (b) of the flow chart in Figure 4 is created using the MATLAB Delaunay tesselation function in MATLAB, then only including edge connections to nearest neighbors of each point. This transforms the circular thin film shape of our structure data into a simplex via triangulation. The vertices of the simplex correspond to the initial metal compounds used to make the ternary system, meaning points closer to these vertices implies the primary component in the mixture at this point will be this particular compound [3].

## 3.3 Objective Function

The main portion of GRENDEL is minimizing the objective function through an iterative process. The objective function is defined as follows:

$$J(E, P, U) = \sum_{k=1}^{K} \Big( \sum_{i=1}^{N} u_{ki}(X_i - E_k p_{ki})^T (X_i - E_k p_{ki}) + \alpha \sum_{h=1}^{M-1} \sum_{l=h+1}^{M} (e_{kh} - e_{kl})^T (e_{kh} - e_{kl}) \Big) \tag{7}$$

Here, $X_i$ is the spectral data for the $i^{th}$ sample point in the material, $K$ is the number of estimated clusters, $N$ is the number of sample points, and $u_{ki}$ is an element of the binary

matrix $U$ that is 1 if $i^{th}$ point belongs to cluster number $k$ and 0 otherwise. In addition, $M$ is the number of endmembers (another word for constituent phases) in a given cluster, $E_k$ is a $D \times M$ matrix where the columns are the individual endmembers that make up the set of endmembers of the $k^{th}$ cluster. That is, the $h^{th}$ column of $E_k$, symbolized as $e_{kh}$, is the diffraction spectra of the $h^{th}$ endmember of the $k^{th}$ cluster. Moreover, $p_{ki}$ is a $M \times R$ matrix of the proportion values for each endmember at each of the $R$ probe intensities used for the $i^{th}$ sample point, and $\alpha$ is a parameter set to 0.0001 to balance the importance of each of the summations [7].

The first term corresponds to nonnegative matrix factorization (NMF) while the second term is analagous to the Graph Cut algorithm, which will be explained further in Sections 3.4 and 3.5, respectively. By minimizing the objective function, the first summation ensures our data points are represented by the correct cluster and its respective phase proportions, and the second summation is a volume constraint on the endmember vectors themselves as well as to make sure the endmember proportions of a cluster match up with the original spectral data as much as possible. By volume constraint, we mean the endmember set of a given cluster should be defined so that clusters are closed and fully connected regions [3].

The GRENDEL algorithm will run until the objective function is minimized and converges to stable values of $E_k$, $p_{ki}$, and $u_{ki}$. After creating initial guesses for each of this matrices, a three-step iterative process is run until convergence. First we solve $\partial J/\partial E_k = 0$ to update our guess for the endmember matrix, yielding the equation

$$E_k = \Big[\Big(\sum_i u_{ki} p_{ki} p_{ki}^T + 2\alpha(M I_{M\times M} - 1_{M\times M})\Big)^{-1}\Big(\sum_i u_{ki} p_{ki} X_i^T\Big)\Big]^T, \tag{8}$$

where $I$ and 1 are the $M \times M$ identity and ones matrices, respectively. We assume endmembers must be positive in order to resemble a physical diffraction pattern, so if an element of $E_k$ is negative, that value is set to zero and the matrix is recomputed via Equation 8.

Second, we try to minimize Equation 7 for $p_{ki}$. Our proportions of endmembers in cluster $k$ must sum to 1 for each sample point $X_i$ in the cluster in order to be physically realistic, $\sum_{h=1}^{M} p_{kih} = 1$. To ensure this, we use a Lagrange multiplier $\lambda_k$. Proportions must be nonnegative as well, so our update of $p_{ki}$ becomes

$$p_{ki} = \max\Big((E_k^T E_k)^{-1}(E_k^T X_i - \frac{\lambda_k}{2} 1_{M\times 1}), 0\Big) \tag{9}$$

with

$$\lambda_k = 2\frac{1_{1\times M}(E_k^T E_k)^{-1} E_k^T X_i - 1}{1_{1\times M}(E_k^T E_k)^{-1} 1_{M\times 1}} \tag{10}$$

If a particular proportion value is chosen to be 0 because the first term in Equation 9 is negative, then the other proportions for the $i^{th}$ sample point must be normalized in order to have them sum to one.

Third, the cluster membership value $u_{ki}$ utilizes the Lagrange multiplier like for $p_{ki}$ in order to have these values sum to one for a point $X_i$, $\sum_{k=1}^{K} u_{ki} = 1$. In other words, we want a given point to be assigned to only one cluster. The update equation is

$$u_{ki} = \frac{\frac{1}{(X_i - E_k p_{ki})^T (X_i - E_k p_{ki})}}{\sum_{k=1}^{K}\Big(\frac{1}{(X_i - E_k p_{ki})^T (X_i - E_k p_{ki}) p_{ki})}\Big)}. \tag{11}$$

Note that this will yield results ranging between 0 and 1, or *soft* cluster membership where a sample point $X_i$ can be shared between clusters. Our version of GRENDEL assumes *hard* membership, or that a given point can only be contained in one cluster, so we set the maximum value of $u_{ki}, k \in [1, K]$ to be 1 and set all other $u_{ki}$ to be 0.

These three updates are repeated until convergence is reached for each variable, meaning we have minimized our objective function 7 [7]. Further explanations behind the methodology of using NMF and Graph Cut algorithms will be explained in the next two sections.

## 3.4 Nonnegative Matrix Factorization (NMF)

The NMF part of the objective function is seen as the first summation in Equation 7. It is given that our input spectral data for sample point $i$, $X_i$, of dimension $D \times R$, is assigned via spectral clustering to a given cluster $C_k$. We assume that this $X_i$ can be represented as proportions of the constituent phases/endmembers that make up its assigned cluster endmember set, $E_k$. That is, represented as linear combinations of the basis phase patterns we wish to find:

$$X_i \approx E_k \; p_{ki} \tag{12}$$

Here, all matrices are defined as in the objective function 7. We assume that the diffraction spectrum follows a Poisson distribution in terms of peaks, so we look to maximize the log-likelihood assuming $E_k p_{ki}$ represents the mean of the distribution:

$$L(E, P) = \sum_{j=1}^{D} \sum_{r=1}^{R} \left( X_{i,jr} \log(E_{k,j} p_{ki,r}) - E_{k,j} p_{ki,r} \right) \tag{13}$$

Explanations of the given indices $j$ and $r$ are described in the first paragraph of Section 3. This makes $X_{i,jr}$ a single value, $E_{k,j}$ a row vector and $p_{ki,r}$ a column vector. This methodology is justified by our assumption of nonnegative data, meaning we anticipate only spikes in intensity with the diffraction pattern. Physically, this is exactly what happens, as our incoming light can only absorb energy according to Equation 1. Rather than compute the maximum of $L(E, P)$, we can use an iterative algorithm which will converge each element of $E_k$ and $p_{ki}$ to achieve this maximum. A reminder that index $j$ refers to the dimension $D$, $h$ refers to $M$, and $r$ to $R$:

$$E_{k,jh} \leftarrow E_{k,jh} \frac{\sum_{r=1}^{R} p_{ki,hr} X_{i,jr} / (E_{k,j} p_{ki,r})}{\sum_{r=1}^{R} p_{ki,hr}} \tag{14a}$$

$$p_{ki,hr} \leftarrow p_{ki,hr} \frac{\sum_{j=1}^{D} E_{k,jh} X_{i,jr} / (E_{k,j} p_{ki,r})}{\sum_{j=1}^{D} E_{k,jh}} \tag{14b}$$

When each and every element of $E_k$ and $p_{ki}$ no longer change between iterations, we have found the maximum of the log-likelihood function and thus the representative endmembers, or constituent phases, and their respective proportions for each given sample point $X_i$ [6]. Rather than use this formulation, GRENDEL utilizes NMF as the first and second updates of the objective function to identify $E_k$ and $p_{ki}$, shown by Equations 8 and 9, respectively. We do this because it is assumed that the endmember set $E_k$ for the $k^{th}$ cluster should be

the same for all points contained in the cluster, as we expect uniform composition within these regions of the material. Update Equations 14, however, only concern themselves with individual points $X_i$, making it possible to incorrectly have $E_k$ converge to two different solutions for two different sample spectra $X_{i1}$ and $X_{i2}$ within a cluster. Therefore, it is more accurate to update $E_k$ by summing over all sample points in the cluster like in our objective function.

## 3.5 Graph Cut Algorithm

The second summation of the objective function 7 is GRENDEL's incorporation of the Graph Cut algorithm, and is seen as the third update Equation 11 for $u_{ki}$. To compute the update of these cluster memberships $u_{ki}$, we use a specific MATLAB wrapper available online at *http://www.wisdom.weizmann.ac.il/∼bagon/matlab.html* [3]. A discussion of the Graph Cut cost equation is seen below to explain why we use Graph Cut in our analysis, although we admit further explanation of how this cost equation relates to the second summation of the objective function is necessary.

The general cost $V$ of the cluster labeling of all input spectral data $X_i, i \in [1, N]$, is described as:

$$V = \lambda_d \sum_i V^i(C_i) + \lambda_s \sum_{i,j \in N} V^{i,j}(C_i, C_j), \tag{15}$$

where $V^i(C_i)$ is the data cost for a point $i$, or the cost to assign a cluster label $C$ to $i$, and $V^{i,j}(C_i, C_j)$ is the smoothness cost, or the cost to assign the labels $C_i$ and $C_j$ to the neighboring points $i$ and $j$. Note that $C$ can take the same values as our index $k$, $C, k \in [1, K]$. Here though, $C_i$ is the cluster label value of sample point $i$ and $k$ is still an index of our matrix notation, such as $p_{ki}$. $\lambda_d$ and $\lambda_s$ are data cost and smoothness cost weights, respectively, which are parameters chosen to balance the smoothness cost, which emphasizes connectivity of clusters so they are all closed regions, and data cost, which emphasizes the similarity of points within a given cluster.

The data cost in Equation 15 is given by

$$V^i(C_i = k) = \frac{3}{4}\delta_{\cos}(X_i, \bar{X}_k) + \frac{1}{4}\frac{||X_i - E_k p_{ki}||_2}{\sum_i ||X_i - E_k p_{ki}||_2}, \tag{16}$$

where $\delta_{\cos}(X_i, \bar{X}_k)$ is the cosine distance between diffraction peaks of sample point $i$ and the mean spectra of the currently assigned cluster $C_i = k$, $||\cdot||_2$ is the L2 norm, and $E_k$ and $p_{ki}$ are defined as in Section 3.3. The first terms makes sure that the spectral data (diffraction pattern) of a point $X_i$ matches with the assigned cluster's mean spectra, while the second term makes sure that this cluster's endmembers, or constituent phase composition, correctly represent the sample point's spectral data $X_i$.

The smoothness cost $V^{i,j}(C_i, C_j)$ is 0 if $i$ and $j$, which must be neighboring data points, belong to the same cluster and 1 if they do not. While it is not exactly noticeable in Equation 15, the smoothness cost summation is restricted to only neighboring points $i$ and $j$ rather than summing over all possible pairs of points. This makes sense, for if we want smooth and continuous clusters, we expect most of the adjacent data points of sample point $i$ to also be in the same cluster unless it is on a boundary. Adding these two terms together,

$V$ is minimized and all sample points are reassigned into the clusters based on this minimized result. But, as stated before, this is done using the MATLAB wrapper referenced above [3].

# 4 Approach to Extend GRENDEL

## 4.1 Constraint Programming

First, we will introduce constraint programming within the objective function. Using both the laws of physics and prior knowledge, we will add constraints on certain matrices and variables which will create better agreement between our results and the original spectral data. The constraints must be followed at all iterations of the objective function until convergence. While this may increase computation time, it is important to make sure our final phase composition is accurate. We will look to follow the strategy of the AMIQO algorithm.

One example of a constraint to be added is the Gibbs phase rule. Our material is considered to be in equilibrium or steady-state. That is, it is not undergoing any chemical processes such as melting or evaporation, and the chemical composition is stable. At equilibrium, a compound or element must be in a set crystalline structure, corresponding to a set phase. Thus, within our ternary system there can only be three phases seen at a given point due to the three input compounds. Every point assigned to a given cluster $k$ should also be represented by the same set of endmembers $E_k$, so this means that at most 3 phases can be seen in a given cluster [1]. As $M$ is defined as the number of endmembers seen in a given cluster, this constraint is written as

$$M \leq 3. \tag{17}$$

Another constraint to be added will be Must-Link and Cannot-Link pairs of points. Essentially, if a pair of points $i$ and $j$ have spectra $X_i$ and $X_j$ that are described by the exact same phase/endmember, then we say that they must be linked together in the same cluster no matter what cluster it is. If they do not contain the same phase, then they cannot be linked as their phase composition varies, meaning they must be in different clusters [4]. This is written as

$$u_{ki} \in \{0,1\}, k \in [1,K], i \in [1,N] \tag{18a}$$
$$u_{ki} = u_{kj}, k \in [1,K], (i,j) \in MustLink \tag{18b}$$
$$u_{ki} + u_{kj} \leq 1, k \in [1,K], (i,j) \in CannotLink \tag{18c}$$

Lastly, a constraint already added into GRENDEL is the nonnegativity of endmember waveforms, or columns of $E_k$, and proportions of these endmembers $p_{ki}$. Also, the proportions of the endmembers within a given cluster $k$ should sum up to 1 for each sample point $i$ in cluster $k$ [3]. That is,

$$E_k, p_{ki} \geq 0, \quad \sum_{h=1}^{M} p_{ki,h} = 1 \quad \forall k \in [1,K] \tag{19}$$

Other types of physical constraints will be added and tested in order to figure out the importance of each constraint. At the end of this first phase of research, we hope to have

a set of constraints robust enough to increase the accuracy of our phase compositions and resulting diagrams, but also small enough that the computation time is not significantly increased.

## 4.2 Active Learning

The probing process employed to create the x-ray diffraction spectra of a given material takes half an hour per sample point. To allow for fast analysis of a large set of unknown materials, one of the major goals of this project, we need to minimize the number of data points necessary to create accurate analysis with our algorithm. Previously, GRENDEL used the spectra of the entire sample as an input, but for this approach we must change the input to iteratively read in single sample points of the material, so our value of $N$ increases by one for each run of GRENDEL. This simulates the algorithm running in conjunction with the diffraction process. We start with a initial number of sample point spectral data, say ten data points, and GRENDEL is run on this small subset of points within the material to initially estimate clusters and cluster endmember composition as described in Section 3.

Then, these findings are utilized to predict the endmember composition over the entire material. To do so, an unknown point in the material is assumed to have an endmember composition that is a distance-weighted average of the phase/endmember compositions of the already-sampled data. At this stage in the project a formal mathematical equation to describe this process has yet to be formulated. This step will depend on the constraint programming described in Section 4.1 to determine if there is any constraint or factor other than Euclidean distance that needs to be incorporated to better our prediction.

After this, we assign a probability or similarity metric, $\rho_{ki}$, that quantifies how well the $k^{th}$ cluster endmember composition represents the estimated values at the unknown sample point $i$. Again, this particular portion of the active learning algorithm must be determined after constraint programming. We could use the cosine distance between the cluster endmember matrix $E_k$ and the predicted endmember matrix at point $i$ akin to spectral clustering, but we may find it is better to work with the variance between the cluster and estimated sample point compositions in order to add in constraints at this stage as well. We then choose the next point to probe to be the area within the material with the lowest probability to be assigned to any unique cluster, or correspondingly the lowest similarity with each of the observed clusters [5].

This will focus the sampling process on cluster boundaries to resolve these edges. At these boundaries, we expect endmember compositions of the adjacent clusters to overlap when we generate our predictive endmember composition described in the previous paragraph. This will yield low probability values since the mixing of both compositions will create error. Consequently, areas in the middle of certain clusters will be excluded from sampling, exactly what we want to happen. Probing these regions are a waste of time, as it is known for a given material that within a cluster, the endmember composition (proportions of constituent phases with determine chemical properties) will be uniform [1]. Figure 5 shows a basic representation of the active learning algorithm which will be implemented. Upon completion, we expect to decrease the number of sample points needed to output the same phase diagrams and constituent phase compositions generated with the full dataset of $N$ sample points, the final outputs of the GRENDEL algorithm.

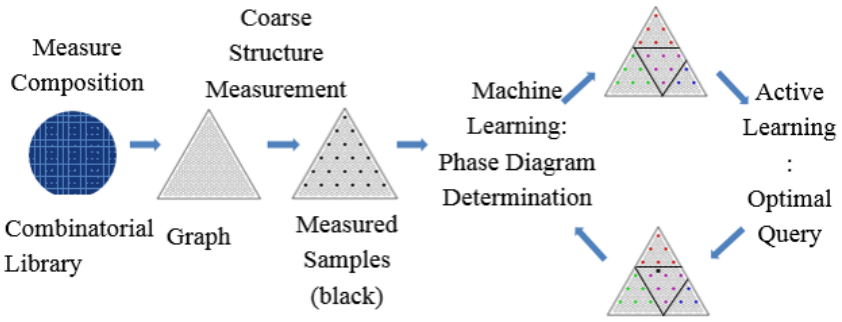


Figure 5: A flow chart of the active learning procedure. Input data is read in, an initial analysis is done, then this cluster and phase data is extended as a continuous distribution through the rest of the material. The next point chosen to sample will have the highest uncertainty in composition, pinpointing cluster boundaries and areas where sampling is more sparse [2].

# 5 Implementation

The algorithm will be written primarily in MATLAB R2015a, potentially with some collaboration in C++ for the constrain programming. This will be run on a personal ASUS laptop with a 2.4 GHz Intel processor and 8 GB of RAM. If needed a high-performance computer may be used, although previous implementations have run well on the current computer specifications.

# 6 Datasets

The datasets to be used to test this algorithm fall into one of two categories. The first group of materials data will be taken from the Inorganic Crystal Structure Database, a large library of material data. The ICSD will give us spectral composition, structural, and phase composition data. These materials have already been analyzed and thus can serve as ground truths used to validate our algorithm's phase composition results. The second group consists of materials not yet analyzed by hand. These will also have structural and spectral data, but the phase compositions have yet to be determined. The second group of datasets will be given to us by the current collaborators of GRENDEL, who have previously analyzed these materials. We will look to compare the output of our new extended GRENDEL algorithm with their past conclusions, particularly looking to make sure our run times are still comparable to their computation times.

# 7 Validation Methods

Certain test data sets have already been analyzed by hand, meaning phase compositions and diagrams of these materials have already been created. Comparing the outputted plots and compositions to these ground truths will verify the overall accuracy of the algorithm.

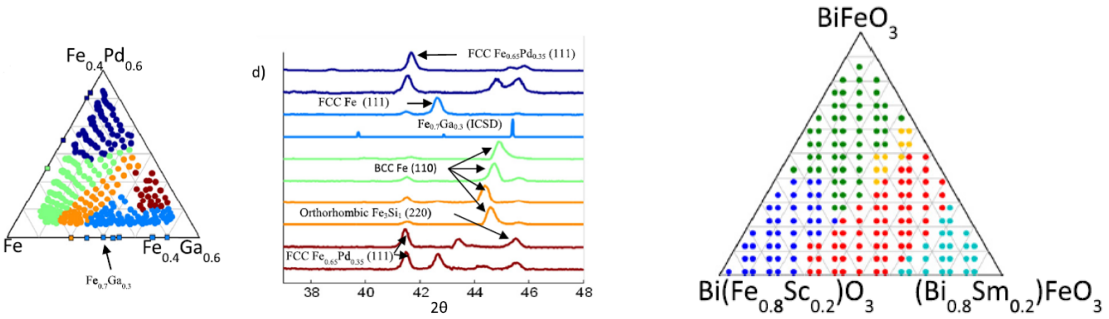


Figure 6: To the left, we have the $Fe$-$Ga$-$Pd$ ternary system. Seen is both the phase diagram simplex and constituent phase plot from previous GRENDEL experiments [3]. To the right is the $(Bi, Sm)(Sc, Fe)O_3$ alloy, where only the phase diagram is given. Different colors in each figure represent the separate clusters within each material [8].

During the constraint programming step, we expect to match these verified phase diagrams, and with our active learning stage we will do the same only with significantly less sample points utilized as inputs. Previous phase pattern decomposition experiments have yielded 80% accuracy with ICSD analysis, so improvement on this is planned [4].

As stated previously, some of our sample datasets will not have phase decomposition data, but results from prior GRENDEL research is available. To validate this material data we will want to compare our output to the aforementioned GRENDEL diagrams, and to more importantly to measure the efficiency of our program. Computation times should be similar to those of GRENDEL's past runs, staying on the order of a minute [3]. Examples of these data already utilized in GRENDEL research are the $Fe$-$Ga$-$Pd$ and $(Bi, Sm)(Sc, Fe)O_3$ ternary systems, seen in Figure 6. Phase diagrams and a constituent phase plot is seen, the latter being a graphical representation of the diffraction spectra for each of the endmembers/phases seen in the material. Each line in this particular plot corresponds to a column of $E_k$ for all $K$ clusters.

$Fe$-$Ga$-$Pd$ is speculated to have permanent magnetic properties, as it has been seen in the $Fe$-$Pd$ binary system. $(Bi, Sm)(Sc, Fe)O_3$ consists of Perovskite Bismuth Ferrite ($BiFeO_3$), Scandium, and Samarium. Rare earth metals are known to bring out ferroelectric properties of $BiFeO_3$, with potential applications in capacitors and switches due it its ability to change its polarity with applied electric fields [8].

# 8 Timeline

The project will be divided into three separate stages:

1. Fully understand GRENDEL, replicate the previous results (*mid/late October*)
2. Constraint Programming
   (a) Add constraints and prior knowledge to increase the accuracy of results for one sample material (*mid November*)

(b) Generalize physical constraints to increase accuracy for all data sets given (*early/mid December*)

3. Active Learning

   (a) Have an algorithm to predict the next best point to sample (*early/mid February*)

   (b) Optimize the sampling algorithm for one material (*early/mid March*)

   (c) Optimize algorithm for all given material data (*mid/late April*)

# 9 Deliverables

I will be delivering my final algorithm, which will compute the constituent phase composition and clustering. This code will also include the active learning component, so the number of sample points needed to generate the desired results will be outputted as well. When available, percent agreement between our output and previously-generated phase decomposition data will be available. Phase diagrams, spectral graphs, and numerical phase composition proportions will be generated for each of the given sample materials. In addition, mid-year and end of the year reports and presentations will be given per requirements of the course.

# References


[1] Lebras R., Damoulas T., Gregoire J.M., Sabharwal A., Gomes C.P., and van Dover R.B., *Constraint reasoning and kernel clustering for pattern decomposition with scaling*, AAAI **CP'11** (2011), 508–522.

[2] Takeuchi I., *Data Driven Approaches to Combinatorial Materials Science*, Materials Research Society Spring Meeting presentation, University of Maryland, College Park, 2016.

[3] Kusne A.G., Keller D., Anderson A., Zaban A., and Takeuchi I., *High-throughput determination of structural phase diagram and constituent phases using GRENDEL*, Nanotechnology **26** (2015), no. 44, 444002.

[4] Ermon S., LeBras R., Suram S., Gregoire J.M., Gomes C.P., Selman B., and van Dover R.B., *Pattern decomposition with complex combinatorial constraints: application to materials discovery*, AAAI Conference on Artifical Intelligence (1972), available at `http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10020`.

[5] Settles B., *Active Learning*, 18th ed., Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan and Claypool, 2012.

[6] Hastie T., Tibishirani R., and Friedman J., *The Elements of Statistical Learning - Data Mining, Interference, and Prediction*, 2nd ed., Springer, 2013.

[7] Zare A., Gader P., Bchir O., and Frigui H., *Piecewise Convex Multiple-Model Endmember Detection and Spectral Unmixing*, IEEE Transactions on Geoscience and Remote Sensing **51** (2013), no. 5, 2853–2862.

[8] Kan D., Suchoski R., Fujino S., and Takeuchi I., *Combinatorial investigation of structural and ferroelectric properties of A- and B- site co-doped BiFeO3 thin films*, Integrated Ferroelectrics **111** (2009), 116-124.