Pattern Decomposition of Inorganic Materials: Optimizing Computational Algorithm

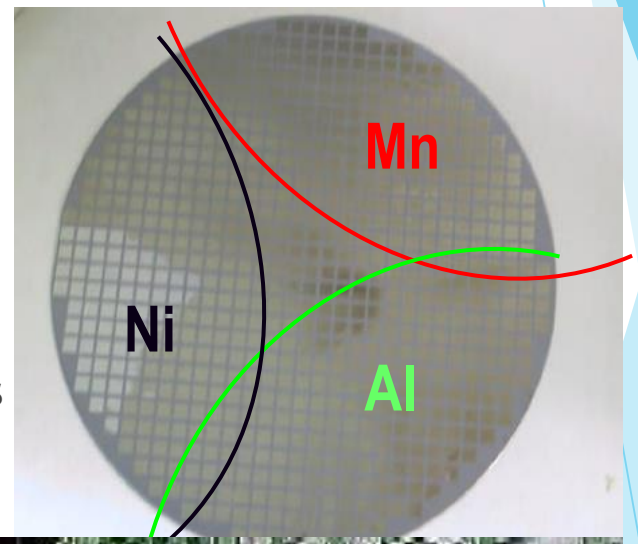
Graham Antoszewski
ganto@math.umd.edu

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

September 27, 2016

Background Information – Materials Sciences

- Inorganic materials – do not contain carbon
- Combinations of metal alloys – ternary systems
  - Unique crystalline structure
  - Depending upon mixture, “phases” are created
  - Phase composition → proportion/combination of given metals
- Crystallographic phases can have certain properties
  - Catalysts
  - Superconductivity



Top: Takeuchi I. (2016) MRS Meeting ; Bottom: courtesy of http://superconductors.org/

Background Information – How Do We See These Phases?

- Given material is sampled using electron probe
  - Use various intensities of light
- Send in x-ray light, which is diffracted back at a certain angle
- Output seen is a continuous waveform
  - Scattering angle
  - Intensity of diffracted light
- Peaks correspond to material detection

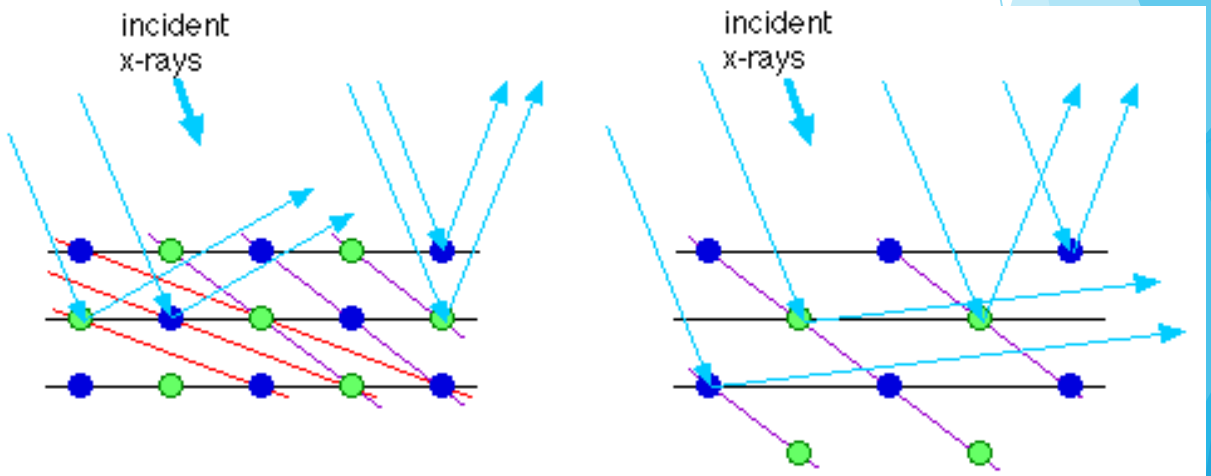
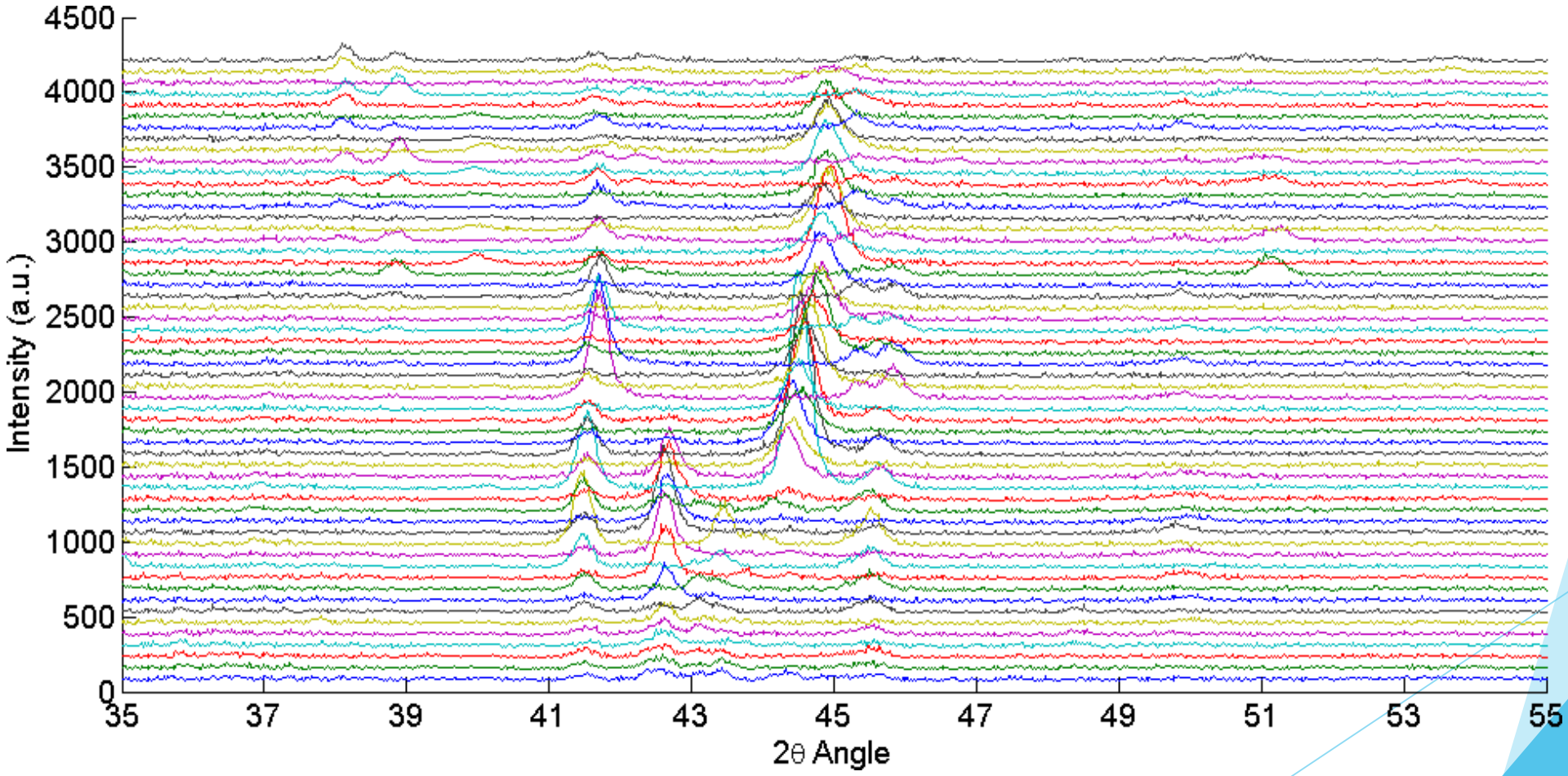


Figure courtesy of http://physics.bu.edu/py106/notes/Resolution.html

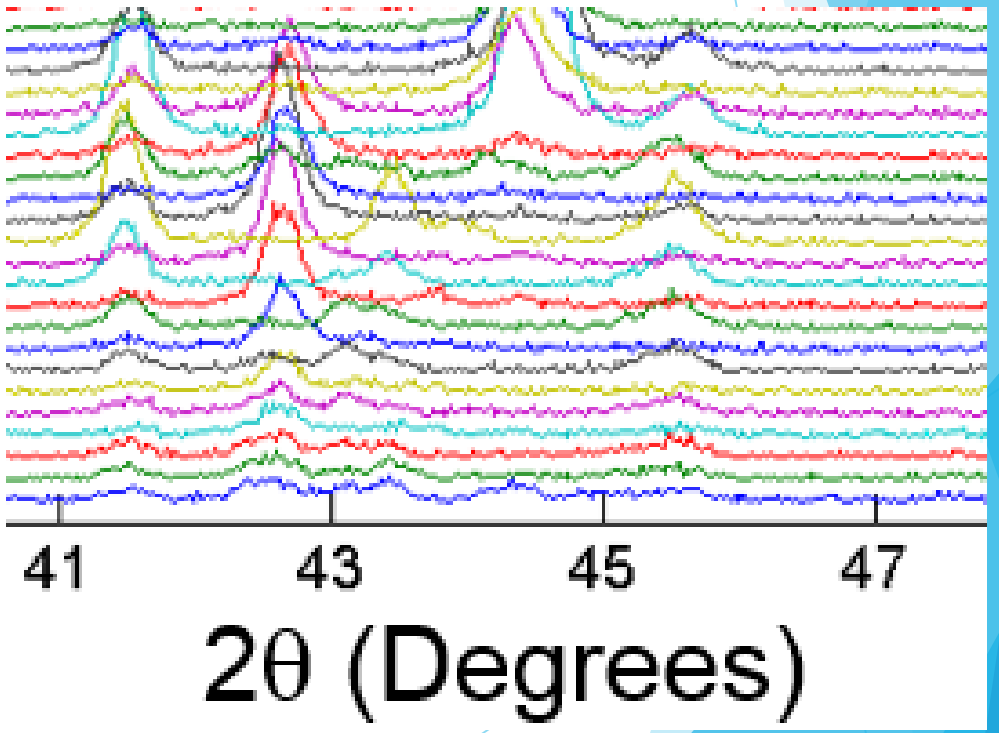
Background Information – X-ray Diffraction



Takeuchi I. (2016) MRS Meeting

Background Information – Analyzing Diffraction Patterns

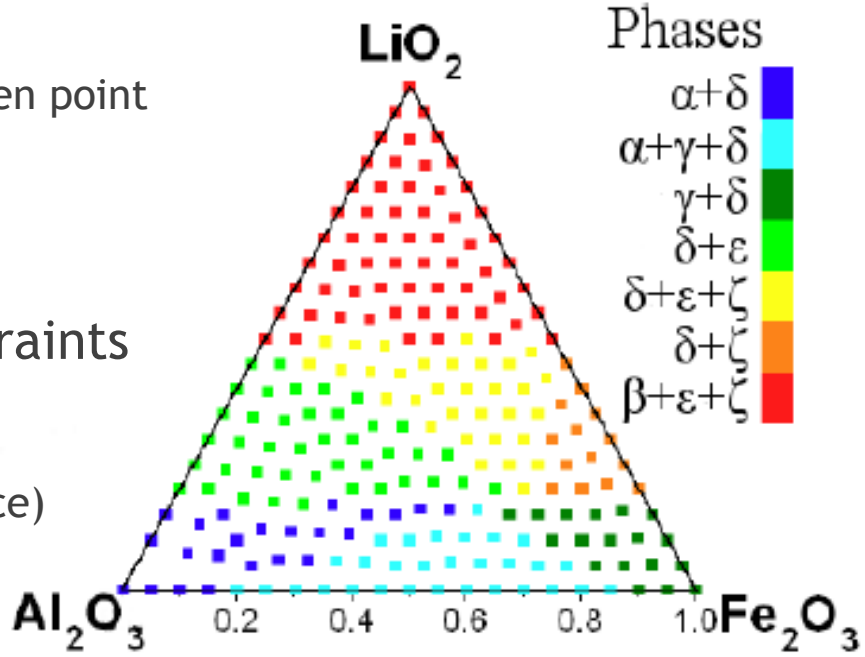
- Three main aspects of a given diffraction peak
  - Scattering angle (2θ)
  - Height of peak (amplitude)
  - Width of peak
- Certain phases of materials previously known
  - Based on 2θ
  - Height, width used as well to determine proportions of constituent phases at each sample point
  - Must be careful of shifts in the peaks with intensity



Takeuchi I. (2016) MRS Meeting

Background Information – Phase diagrams

- After probing all sample points of a material, a simplex can be created
  - Illustration of phase composition at a given point
  - Colors → similar phase structure
- Results must uphold to certain constraints
  - Gibbs phase rule
  - Connectivity (continuity of phases in space)
  - Other laws of physics



LeBras et al (2011) AAAI CP'11 508-522

Background Information – Scientific Computation Component

- Previously, phase diagrams done by hand
- White House Materials Genome Initiative
  - Use libraries of composition/structure data, output phase structure of materials
  - Develop algorithm to do so

- Algorithm must:
  - Obey physical constraints
  - Identify phases accurately
  - Identify regions/clusters of similar phase composition within material
  - Be efficient – short run times so more materials can be analyzed

Background Information – Pattern Decomposition Problem

- Given a system where you observe patterns of a numerical variable at N sample points
- Assume patterns are described by a combination of K basis patterns
  - You wish to uncover these K basis patterns from the N samples

- Example : Cocktail party problem
- Must also adhere to constraints
  - Cocktail problem – size of room, number of people

Background Information – GRENDEL algorithm

- Developed by collaborators at Maryland and NIST
- Given N sample points of a given inorganic material
- 1. Spectral clustering – group similar points together, create similarity matrix
- 2. Graph Cut algorithm – adds in connectivity constraint between clusters
- 3. Nonnegative Matrix Factorization – determines number of constituent phases that compose each cluster
- Graph Cut and NMF guided by objective function, minimize error between:
  - Original structure of material (diffraction spectra) to our phase proportions
  - Volume constraint – proportions of phases realistic

Background Information – AMIQO

- AMIQO – Mixed Integer Quadratic Problem
  - n points, m intensity values, k basis patterns/phases
  - A = original input data (m x n)
  - W = presence of a given phase (binary values, m x k)
  - H = proportion corresponding to given phase at each point (k x n)
  - Must-Link and Cannot-Link pairs of points (clustering)
  - Prior knowledge
- Still uses NMF, spectral clustering steps in the iterative process

$$\min_{W,H,b} \quad ||A - WH||_2$$

$$\text{s.t.} \quad W, H \geq 0, \sum_j h_{i,j} = 1, \sum_i b_{i,j} \leq S$$

$$b_{i,j} \geq h_{i,j},\ b_{i,j} \in \{0,1\}\ i \in [1,k], j \in [1,n]$$

$$b_{i,i_s} = b_{i,j_s}\ i \in [1,k], (i_s, j_s) \in ML$$

$$b_{i,i_s} + b_{i,j_s} \leq 1\ i \in [1,k], (i_s, j_s) \in CL$$

Background Information – Issues with Algorithms

- GRENDEL – good run time (< 1 min), efficient, but lack of physical constraints (connectivity)
- AMIQO – upheld constraints, yet took too long (days) to run
- Sampling time of the material takes 30 minutes per point
  - Whole material → Potentially over a week
  - This runs independent of program
  - Need to reduce number of sample points probed
- Want a program to combine speed, accuracy, and use the minimum amount of sample points

Project Goal – Extending GRENDEL

- Increase accuracy of pattern decomposition algorithm by incorporating constraints
  - Laws of physics
  - Prior knowledge of material
  - Affects cluster analysis and overall phase composition
- Decrease time needed to probe given material in the lab
  - Minimize data points needed to resolve constituent phases (endmembers)

Approach (Part 1) – Constraint Programming

- Add laws of physics into objective function
- Incorporate new constraints based on prior knowledge
- Cannot Link, Must-Link pairs of points like in AMIQO

$$\min_{W,H,b} \quad ||A - WH||_2$$

$$\text{s.t.} \quad W, H \geq 0, \sum_j h_{i,j} = 1, \sum_i b_{i,j} \leq S$$

$$b_{i,j} \geq h_{i,j},\ b_{i,j} \in \{0,1\}\ i \in [1,k], j \in [1,n]$$

$$b_{i,i_s} = b_{i,j_s}\ i \in [1,k], (i_s, j_s) \in ML$$

$$b_{i,i_s} + b_{i,j_s} \leq 1\ i \in [1,k], (i_s, j_s) \in CL$$

Approach (Part 2)– Active Learning

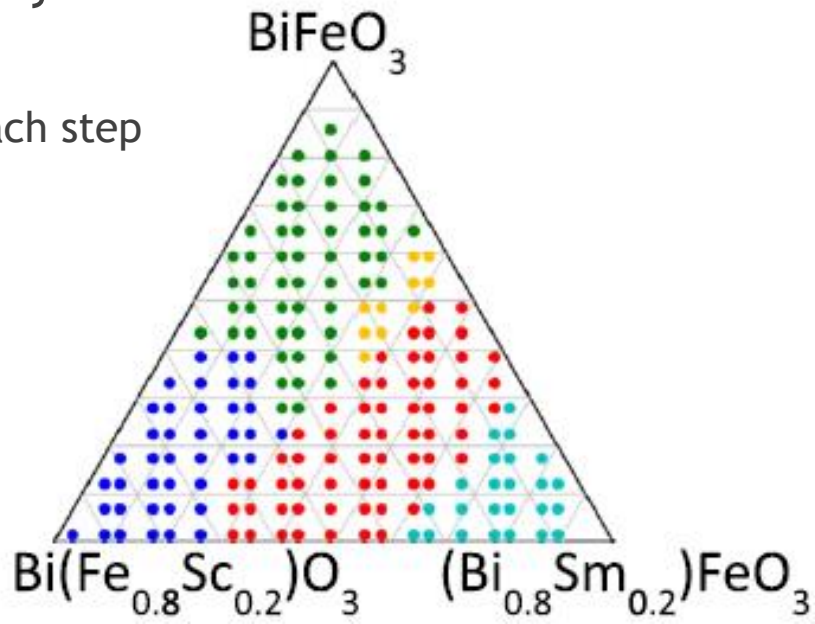
- Using previous data, suggest next informative point to sample
- Hierarchical sampling
  - Look at each point, assess similarity within given cluster
  - Determine area with the lowest similarity to cluster
  - Sample this spot, reassess clustering
- Pinpoint most important areas to probe (cluster boundaries)
  - Goal – reach under desired threshold of accuracy in less iterations (less sample points)

Implementation

- Language - MATLAB R2015a
  - Potential collaboration with C++
- Hardware - personal computer
  - ASUS, 8 GB RAM
- Data sets – Inorganic Crystal Structure Database
  - spectral and structural data from previous research efforts

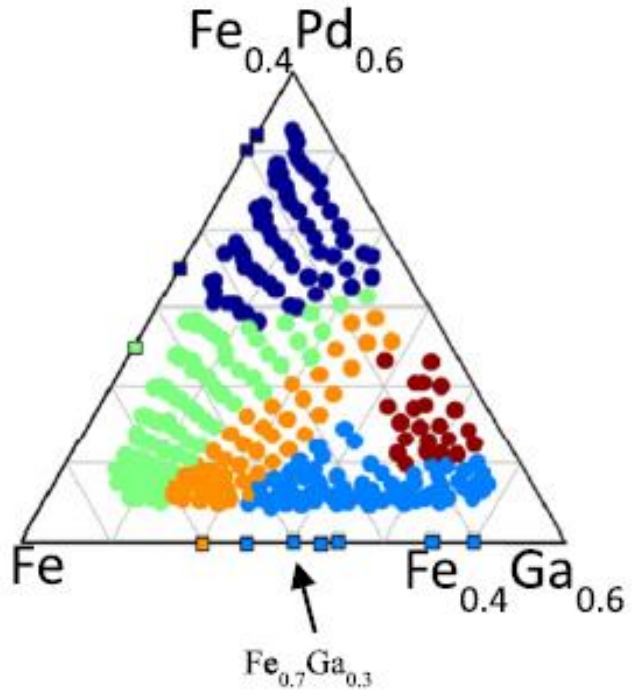
Validation Methods/Test Problems

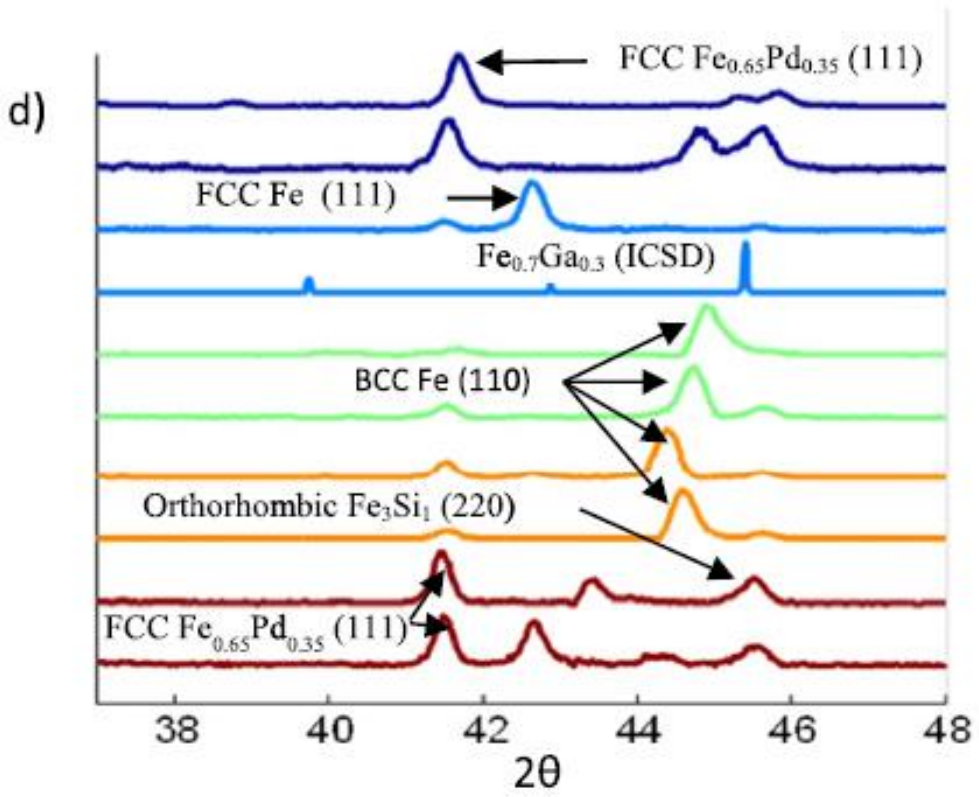
- Phase decomposition of data sets already done by hand
  - Compare our results to these diagrams at each step
- Use previous GRENDEL results
  - Validates increased accuracy, efficiency
- Test problems - Fe-Ga-Pd and $(Bi,Sm)(Sc,Fe)O_3$ thin films



Kusne et al (2015) AAAI 26(44) 444002

Test Problems





Kusne et al (2015) AAAI 26(44) 444002

Expected Results

- Run time ~ 1 minute
- Agreement with handmade analysis
  - > 80% for low number of constraints, approach 100% as more are added
- Active learning – significant decrease in sample points needed
  - Keep up efficiency, accuracy
  - Full sample analyzed in 1-2 days

Concluding Remarks – Why Are We Doing This Again?

- Pattern decomposition – unearthing new properties of inorganic materials
- Application advancements outpacing the materials to do it
- Want rapid analysis of these resources – computer algorithm
- High accuracy, high efficiency program – discover new properties quicker

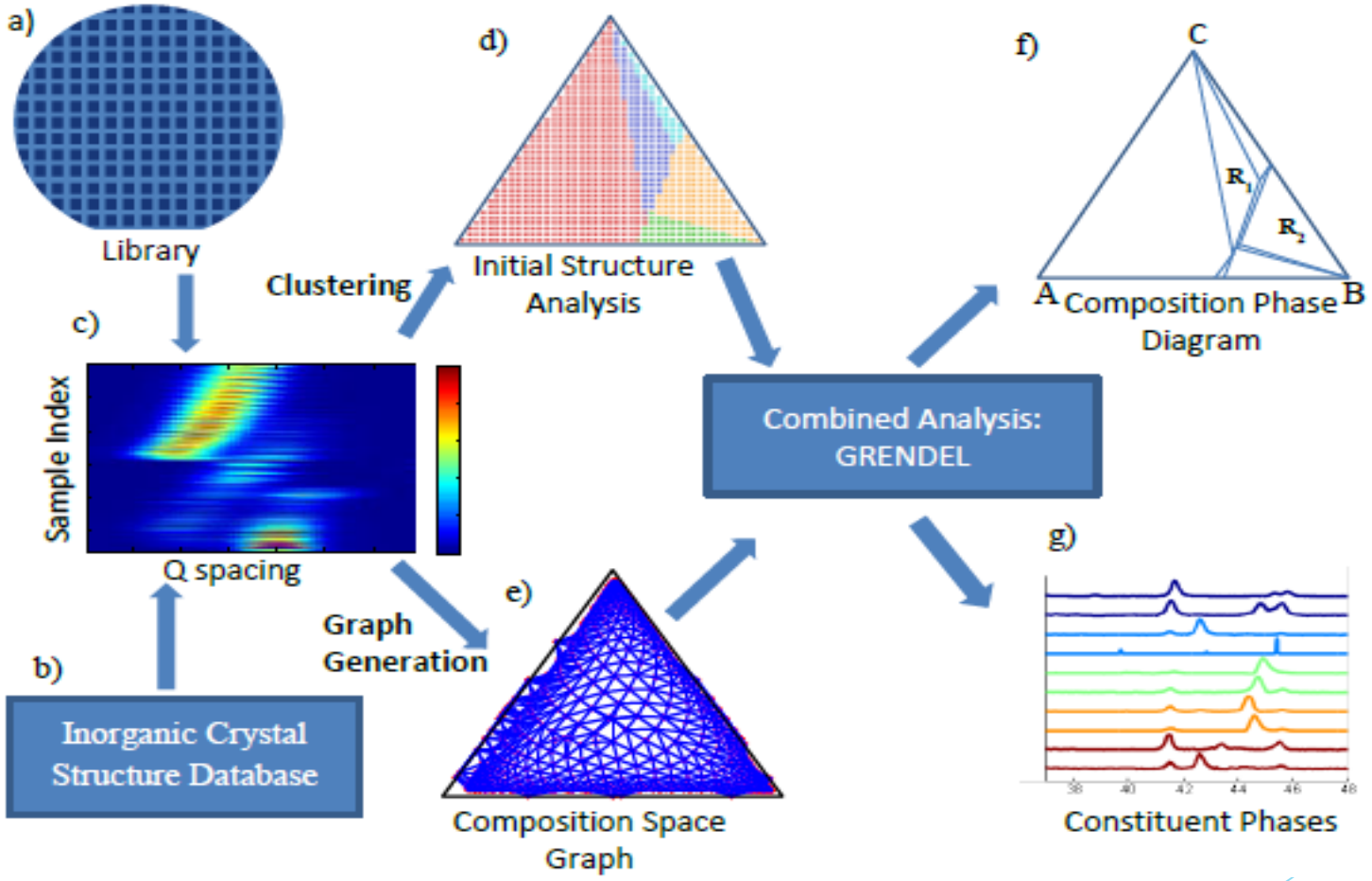
Timeline/Milestones

- Fully understand, replicate previous code/results – mid/late October
- Phase 1 – Constraint Programming
  - Add constraints/prior knowledge, increase accuracy of results for one sample material – mid November
  - Generalize constraints, increase accuracy for all data sets given – early/mid December
- Phase 2 – Active Learning
  - Have algorithm to predict next best point to sample – early/mid February
  - Optimize the sampling algorithm for one material – early/mid March
  - Optimize algorithm for all material data given – mid/late April

Deliverables

- Final code/algorithm
- Results for given materials
  - Phase diagrams
  - Spectral graphs
  - Constituent phase compositions
- Mid-year report and presentation
- End of the year report and presentation

Scientific Computation Algorithm – GRENDEL



Kusne et al (2015) AAAI 26(44) 444002

Scientific Computation Algorithm – Spectral Clustering

- Takes in diffraction data, creates a similarity matrix
  - i,j – sample points
  - $d_{i,j}$ – cosine distance (1 – cosine of difference in scattering angles)
  - σ – spectral clustering bandwidth parameter

$$S_{ij} = e^{\frac{d_{ij}}{2\sigma^2}}$$

$$\mathbf{L} = \mathbf{G} - \mathbf{W}$$

- Creates set of edgeweights **W** according to **S**
- **G** → diagonal matrix, entries are sums of corresponding rows of **W**
- Find smallest eigenvalues, corresponding eigenvectors of Laplacian **L**
- use MATLAB k-means function to assign points to clusters

Scientific Computing Algorithm – Graph Cut

- General “cost” equation

$$V = \lambda_d \sum_i V^i(L_i) + \lambda_s \sum_{i,j \in N} V^{i,j}(L_i, L_j)$$

- Data cost matrix

$$V^j(L_j = i) = \frac{3}{4}\partial_{cos}(\boldsymbol{x}_j, \bar{\boldsymbol{x}}_i) + \frac{1}{4}\frac{\|\boldsymbol{x}_j - \boldsymbol{E}_i\boldsymbol{p}_{ij}\|_2}{\sum_i \|\boldsymbol{x}_j - \boldsymbol{E}_i\boldsymbol{p}_{ij}\|_2}$$

- Smoothness cost – 0 if cluster labels match, 1 otherwise

- Minimize V, noting we sum over all sample points

Scientific Computation Algorithm – Nonnegative Matrix Factorization

- Assume our spectral input data can be represented by proportions of constituent phases

$$\mathbf{X} \approx \mathbf{WH}$$

- Similar to the AMIQO minimizing function
- Found by maximizing $L(\mathbf{W}, \mathbf{H}) = \sum_{i=1}^{N} \sum_{j=1}^{p} [x_{ij} \log(\mathbf{WH})_{ij} - (\mathbf{WH})_{ij}].$
- Solution can be found iteratively using

$$w_{ik} \leftarrow w_{ik} \frac{\sum_{j=1}^{p} h_{kj} x_{ij} / (\mathbf{WH})_{ij}}{\sum_{j=1}^{p} h_{kj}}$$

$$h_{kj} \leftarrow h_{kj} \frac{\sum_{i=1}^{N} w_{ik} x_{ij} / (\mathbf{WH})_{ij}}{\sum_{i=1}^{N} w_{ik}}$$

Scientific Computation Algorithm – Objective Function

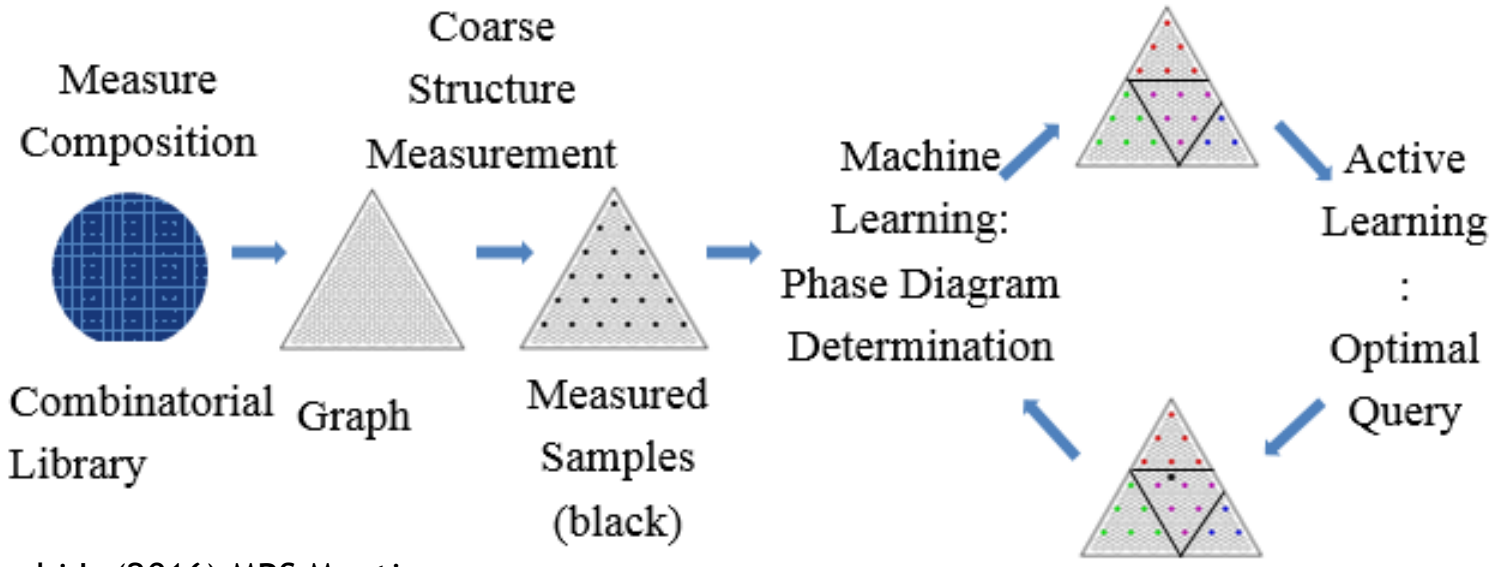
$$J(\boldsymbol{E}, \boldsymbol{P}, \boldsymbol{U}) = \sum_{i=1}^{C}\left(\sum_{j=1}^{N} u_{ij} \left(\boldsymbol{x}_j - \boldsymbol{E}_i \boldsymbol{p}_{ij} \right)^{\mathrm{T}} \left(\boldsymbol{x}_j - \boldsymbol{E}_i \boldsymbol{p}_{ij} \right) + \alpha \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} (\boldsymbol{e}_{ik} - \boldsymbol{e}_{il})^{\mathrm{T}} (\boldsymbol{e}_{ik} - \boldsymbol{e}_{il}) \right).$$

- **E** – endmembers (constituent phases) within a cluster
- **P** – endmember proportions
- **U** – cluster membership (binary)

- Once minimized, we arrive at our final phase composition and phase diagram

Scientific Computation Algorithm – How to Incorporate Active Learning?

- Read in sample points one by one
- Extend the given clustering to unknown areas of material
- Choose next sample point to be one with highest uncertainty/error
- Utilize objective function to choose this



Takeuchi I. (2016) MRS Meeting

Bibliography

- LeBras R., Damoulas T., Gregoire J.M., Sabharwal A., Gomes C.P., and van Dover R.B., 2011. Constraint reasoning and kernel clustering for pattern decomposition with scaling. AAAI. CP'11: pp.508-522.
- Kusne A.G., Keller D., Anderson A., Zaban A., and Takeuchi I., 2015. High-throughput determination of structural phase diagram and constituent phases using GRENDEL. Nanotechnology. 26(44): pp. 444002.
- Ermon S., LeBras R., Suram S., Gregoire J.M., Gomes C.P., Selman B., and van Dover R.B., 2015. Pattern decomposition with complex combinatorial constraints: application to materials discovery. AAAI Conference on Artificial Intelligence. Available at http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10020
- Hastie T., Tibshirani R., and Friedman J., 2013. *The Elements of Statistical Learning - Data Mining, Interference, and Prediction*. ed. 2 (Berlin: Springer).
- Settles B., 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning #18 (Morgan & Claypool).
- Kan D., Suchoski R. Fujino S., Takeuchi I., 2009. Combinatorial investigation of structural and ferroelectric properties of A- and B- site co-doped BiFeO3 thin films. Integrated Ferroelectrics. 111: pp. 116-124.
- Takeuchi I., 2016. Data Driven Approaches to Combinatorial Materials Science. Materials Research Society Spring Meeting (presentation).