Pattern Decomposition and Basis Phase Recognition of Inorganic Materials

Graham Antoszewski
ganto@math.umd.edu

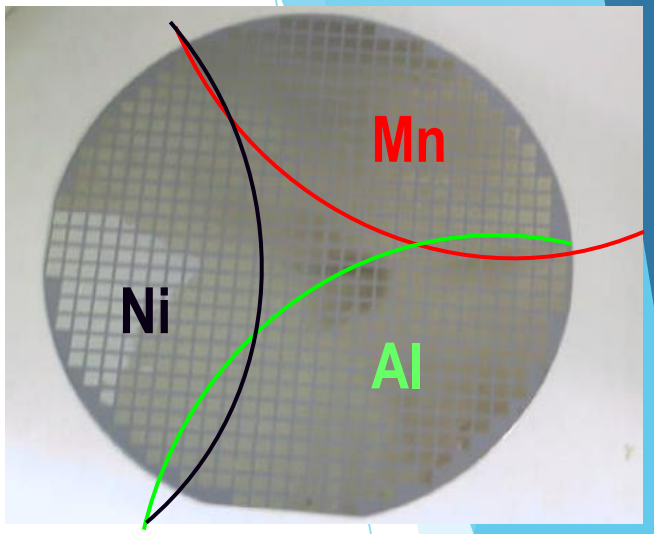
Advisor: Dr. Hector Corrada-Bravo
Center for Bioinformatics and Computational Biology
University of Maryland, Department of Computer Science
hcorrada@umiacs.umd.edu

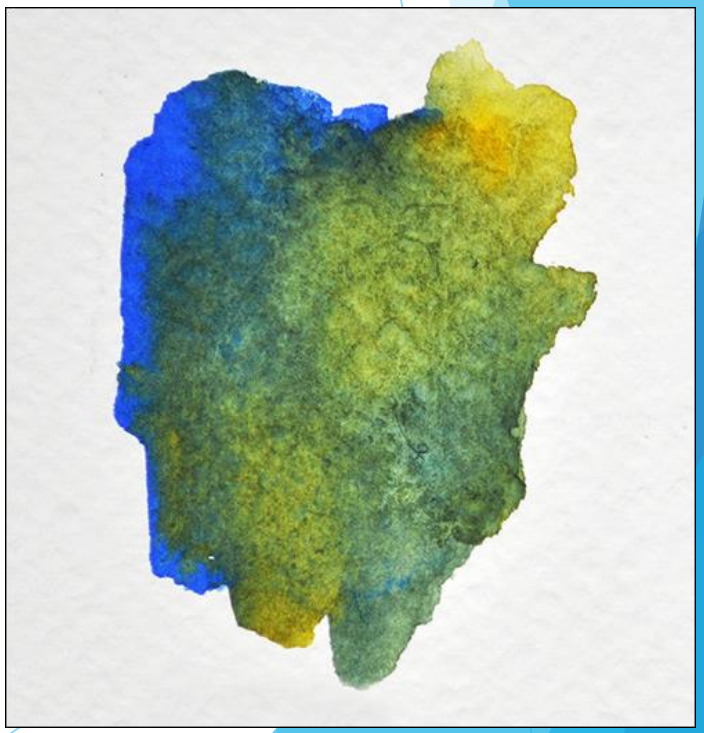
May 11, 2017

Overview of Pattern Decomposition and Phase Recognition

- Mixtures of 3 metals – ternary metal alloy
  - Non-uniform chemical composition
  - Unique structure → Unique chemical properties

- Pattern Decomposition
  - N data points
  - Expressed as linear combination *k* basis vectors (phases)
  - Phases tell us chemical properties

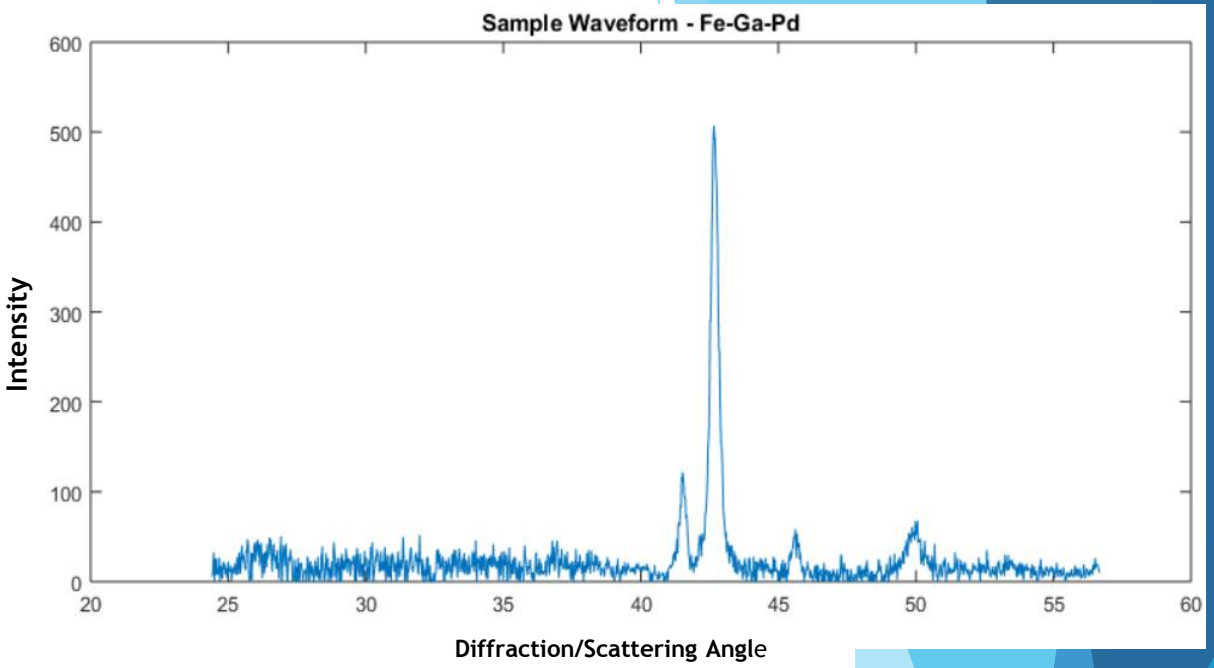


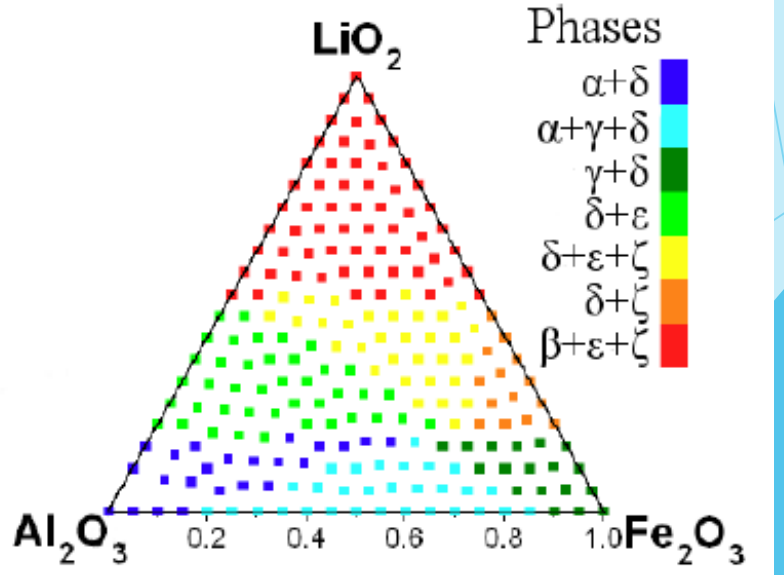


Top: Takeuchi I. (2016) MRS Meeting; Bottom: http://emptyeasel.com/2014/03/10/3-essential-color-mixing-techniques-for-watercolor-painters/

X-ray Diffraction Patterns to Basis Phase Diagrams

- Input data – X-ray light diffracted back at certain angles
  - Based on structure of material (basis phases)
- Phase diagrams
  - Same colors → areas of uniform composition → same basis phases
- Physical constraints on our solution
  - Gibbs Phase Rule
  - Connectivity of clusters
  - Peakshifting (error due to alloying process)

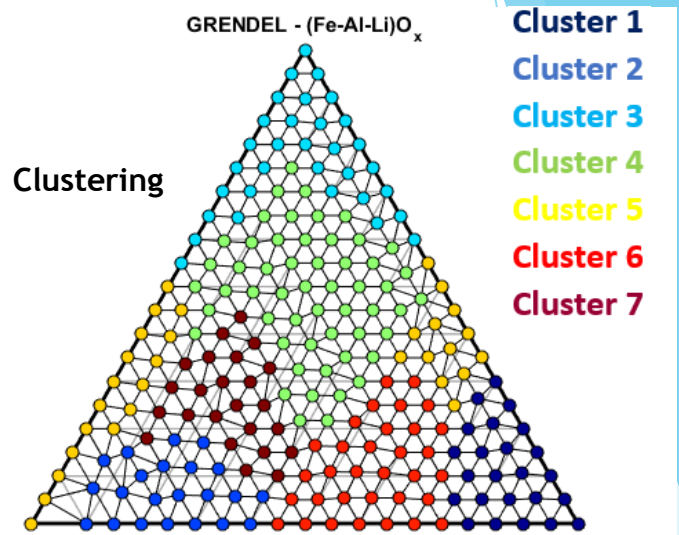


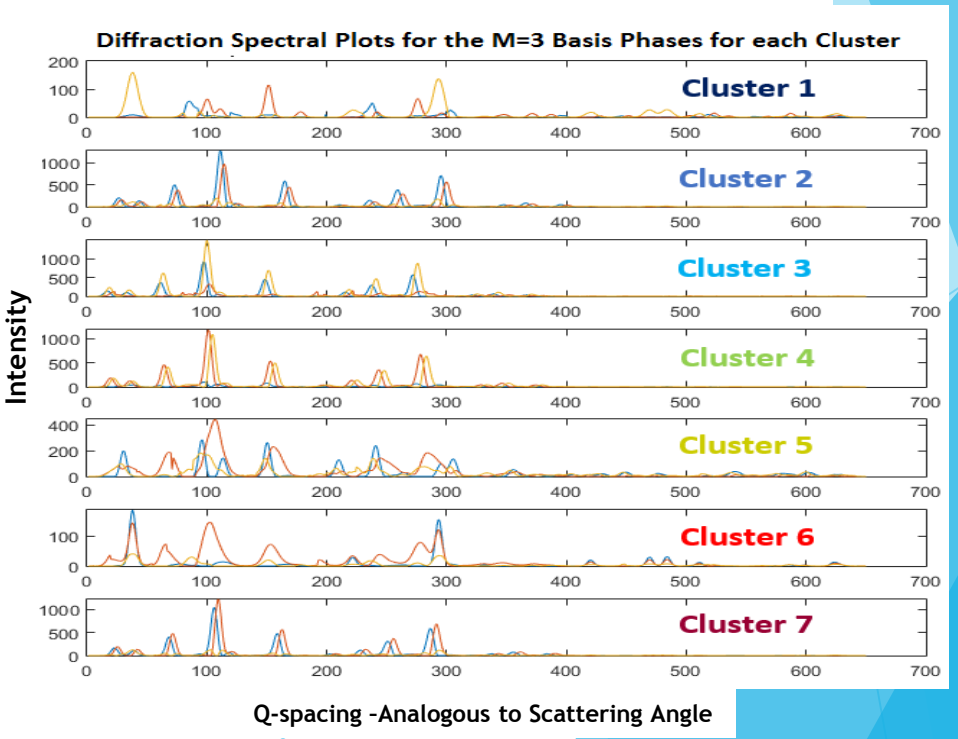


Top: Kusne et al (2015) AAAI 26(44) 44400; Bottom: LeBras et al (2011) AAAI CP'11 508-522

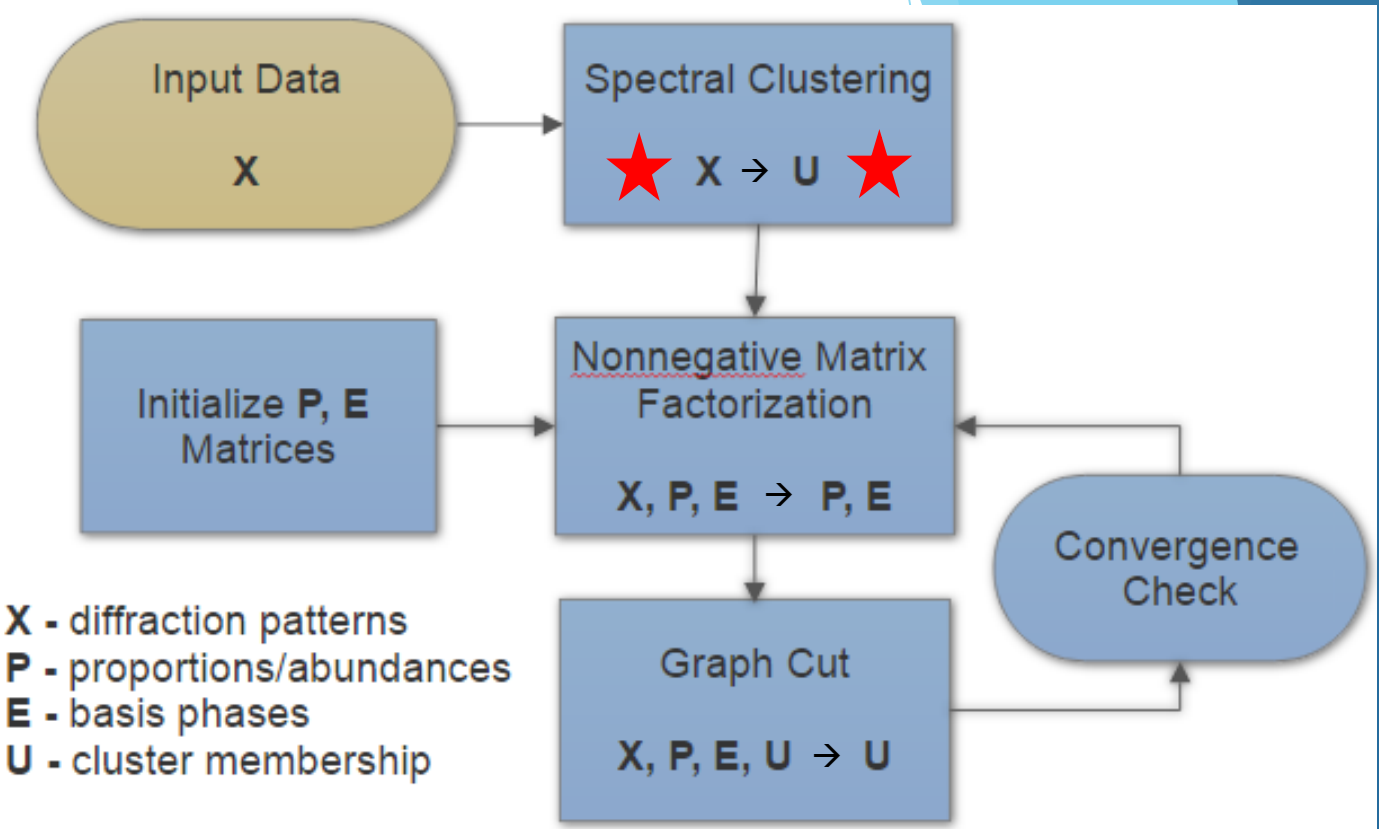
Overall Project Goal

- Develop algorithm to:
  - Obey physical constraints
  - Output clusters, phase diagrams
  - Identify basis phases
- Extend GRENDEL
  - (Graph-based **End**member **E**xtraction and **L**abeling)
  - Develop methods/algorithms to make algorithm results more physically realistic
  - Constraint programming





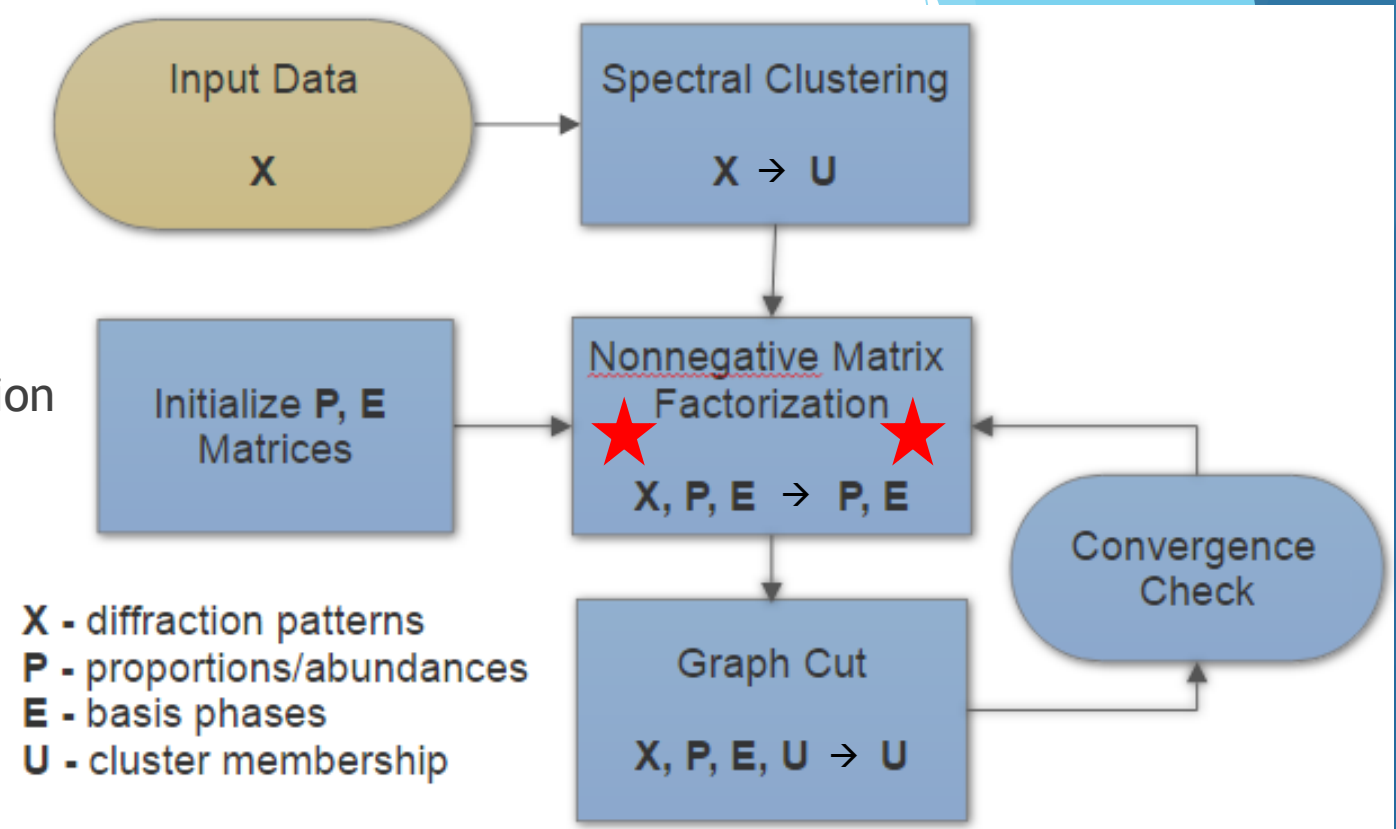
Original GRENDEL Algorithm



- Step 1: Spectral Clustering
  - Diffraction "pattern" → diffraction "spectrum"
  - **X** – input sample data
  - Similarity metric to group data points
    - Cosine Distance, 1 – $\cos(X_i, X_j)$
    - For two sample point diffraction patterns, X_i, X_j : $\cos(X_i, X_j) = \frac{X_i \bullet X_j}{||X_i||_{L_2}||X_j||_{L_2}}$
- Output: initial clustering **U**
  - $U_{i,k} = 1$ → sample point *i* belongs to cluster *k*

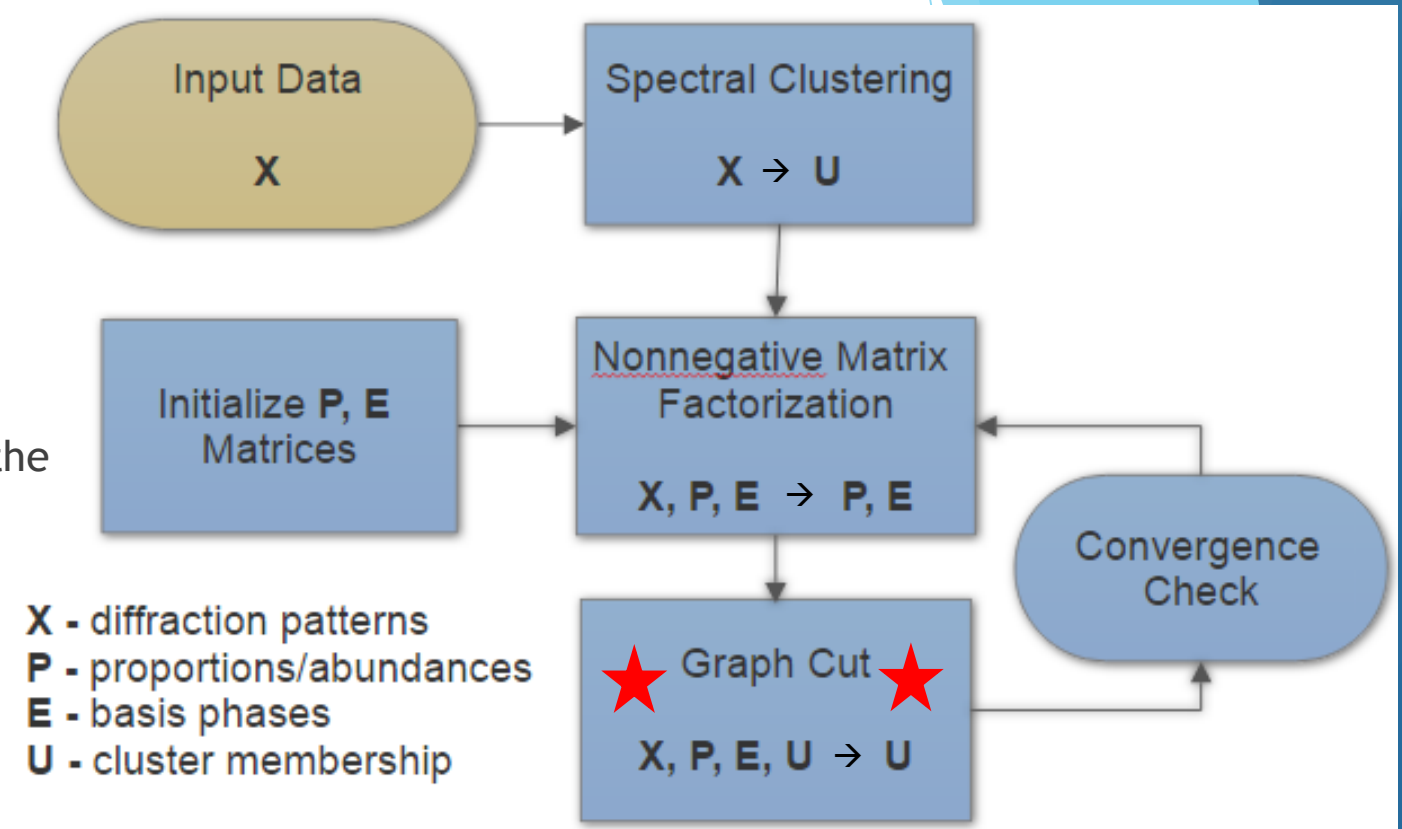
Original GRENDEL Algorithm

- Step 2: Nonnegative Matrix Factorization (NMF)
  - **X** is approximately **P*E**
    - Linear combination of basis phases
  - Find **P, E** by minimizing objective function
  - Ex: Least Squares Error
    - $J_{LS}(X, P, E) = \frac{1}{2}||X - PE||^2_{L_2}$
  - Set derivative of **J** w.r.t. **P, E** equal to zero to create update rules for each matrix
  - Done within each cluster



Original GRENDEL Algorithm

- Step 3: Graph Cut
  - Tries to minimize "cost" function over the entire material to update **U**
- Cost = Data Cost + Smoothness Cost
- Given data point *j* in cluster *i:*
  - Data Cost: $\frac{3}{4}\delta_{\cos}(X_j, \bar{X}_i) + \frac{1}{4}\frac{||X_j - p_{ij}E_i||_{L_2}}{\sum_i ||X_j - p_{ij}E_i||_{L_2}}$
  - Smoothness Cost: 0 if neighboring data points in same cluster, 1 otherwise
  - Balances similarity metrics (Data Cost) with smoothness/connectivity of clusters (Smoothness Cost)
- Convergence check → end program if change between iterations of **P, E, U** is below threshold

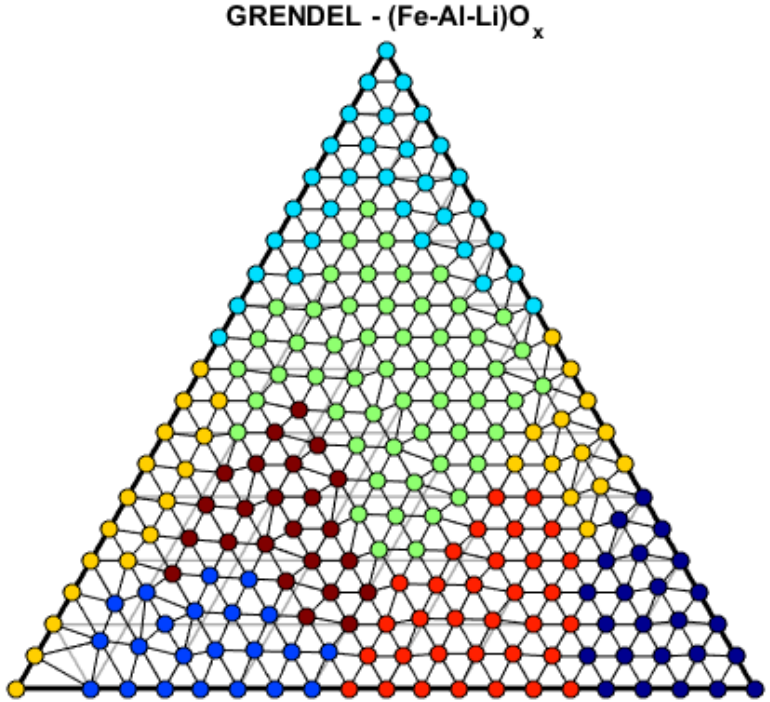


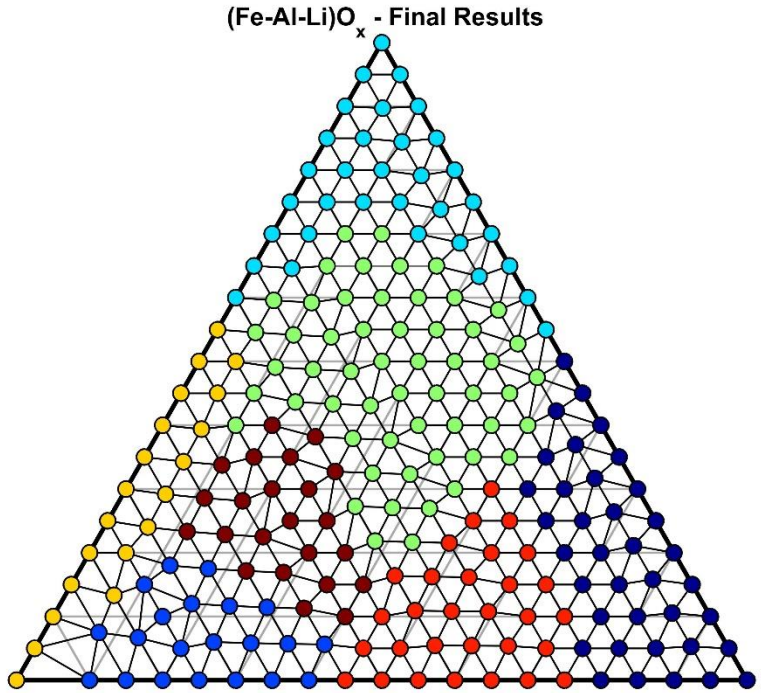
Summary of Project

- Written in MATLAB 2017a
- Data sets:
  - Synthetic diffraction data ((Fe-Al-Li)O_x from Gregoire et al.)
  - Synthetic spectral data from ShiftNMF (Morup M. and Madsen K. H.)
  - Inorganic Crystal Structure Database (Fe-Ga-Pd, from Kusne et al.)
- Last Semester: Cannot Link (connectivity)
- This Semester (previous): ShiftNMF (peakshifting)
- Final Step: Implementing ShiftNMF within existing GRENDEL code

Cannot Link (Review)

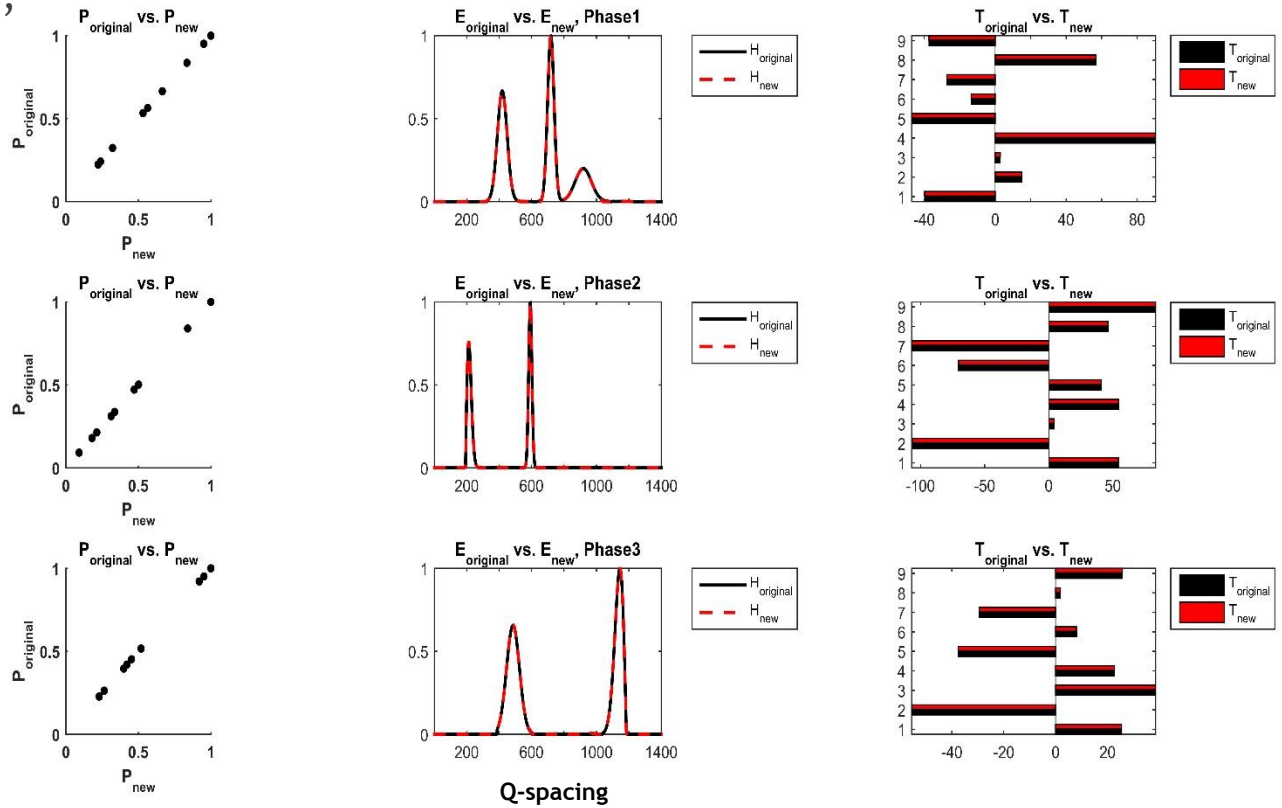
- Used cosine distance as *dissimilarity* metric, creates array of the *p%* most dissimilar pairs of data points (CL)
- Algorithm Overview: After Graph Cut step, makes sure CL pairs are not put in same cluster





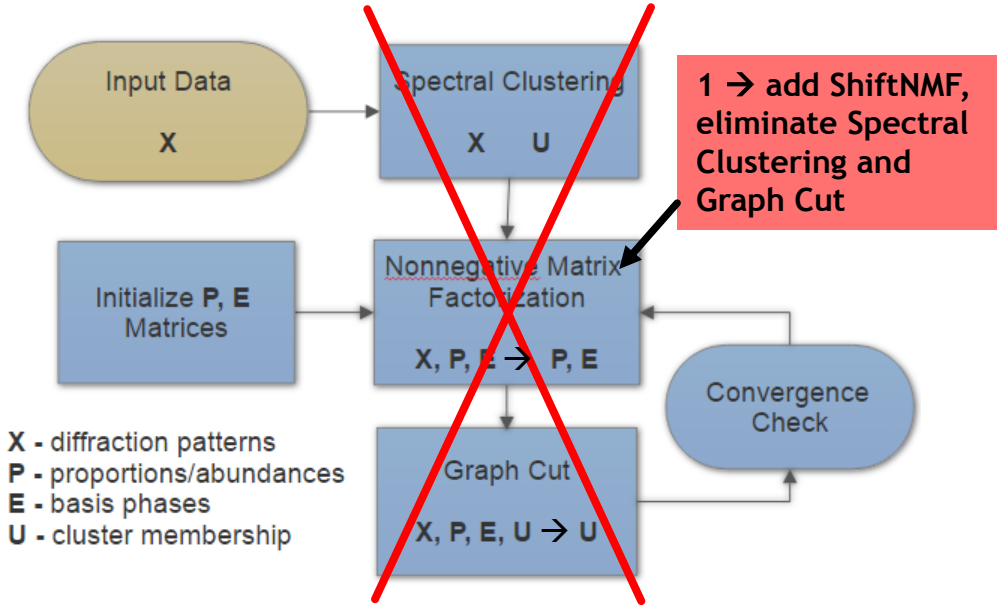
ShiftNMF (Review)

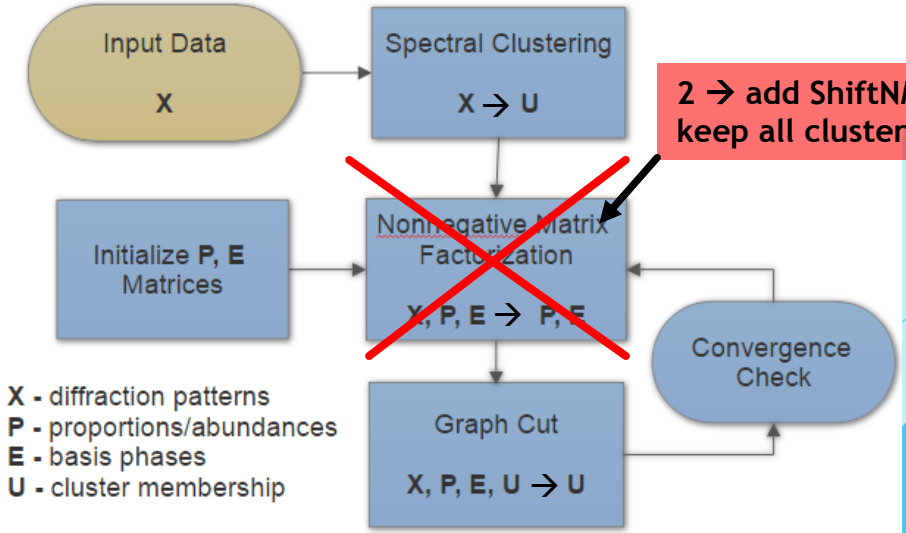
- Algorithm Overview: alter NMF to detect peakshifting within basis phases
- New Objective function: $J_{LS}(X,P,E,T) = \frac{1}{2}||X - PE||_{L_2}^2 = \frac{1}{2M}||X_f - (P_f \bullet \exp(i\omega T))E_f||_{L_2}^2$
  - **T** - matrix of 'peakshifting' delays/values, applied to **P, E** in Fourier Space
  - **P, E** update rules – set derivative of **J** w.r.t. **P, E** equal to zero, respectively
    - Utilizes ratio of negative/positive parts of each gradient
  - **T** update – Newton-Raphson method
  - Cross-correlation step to escape local minima of **J**



Implementing ShiftNMF within GRENDEL

- Two options:
  1. Adding in ShiftNMF outside of clustering (spectral, Graph Cut)
  2. Using ShiftNMF with clustering steps





Experimental Statistics (Random Initial Conditions)

$$R^2 = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}}; \quad \mathrm{SST} = ||X||_{L_2}; \quad \mathrm{SSE} = \frac{1}{2M}||X_f - (P_f \bullet \exp(i\omega T))E_f||_{L_2}^2$$

Implementation	Input Data	Max R^2	Mean R^2	T-statistic (w.r.t. GRENDEL stats)	P-value
Original GRENDEL	Original	0.9512	0.9493	N/A	N/A
	Zero-padded	0.9508	0.9481		
ShiftNMF without clustering	Original	0.9764	0.9607	6.308	3.91E-7
	Zero-padded	0.9765	0.9583	5.537	2.27E-6
ShiftNMF with clustering	Original	0.9893	0.9872	77.236	3.71E-45
	Zero-padded	0.9894	0.9873	53.539	4.03E-35

Experimental Statistics (nnmf() Initial Conditions)

Implementation	Max R^2	Mean R^2	T-statistic (w.r.t. GRENDEL stats)	P-value
Original GRENDEL	0.9545	0.9450	N/A	N/A
ShiftNMF without clustering	0.9771	0.9721	30.385	9.14E-28
ShiftNMF with clustering	0.9888	0.9867	97.088	6.70E-66

- (Fe-Al-Li)O_x synthetic data
- nnmf() – MATLAB nonnegative matrix factorization function
- Both ShiftNMF strategies yield better results

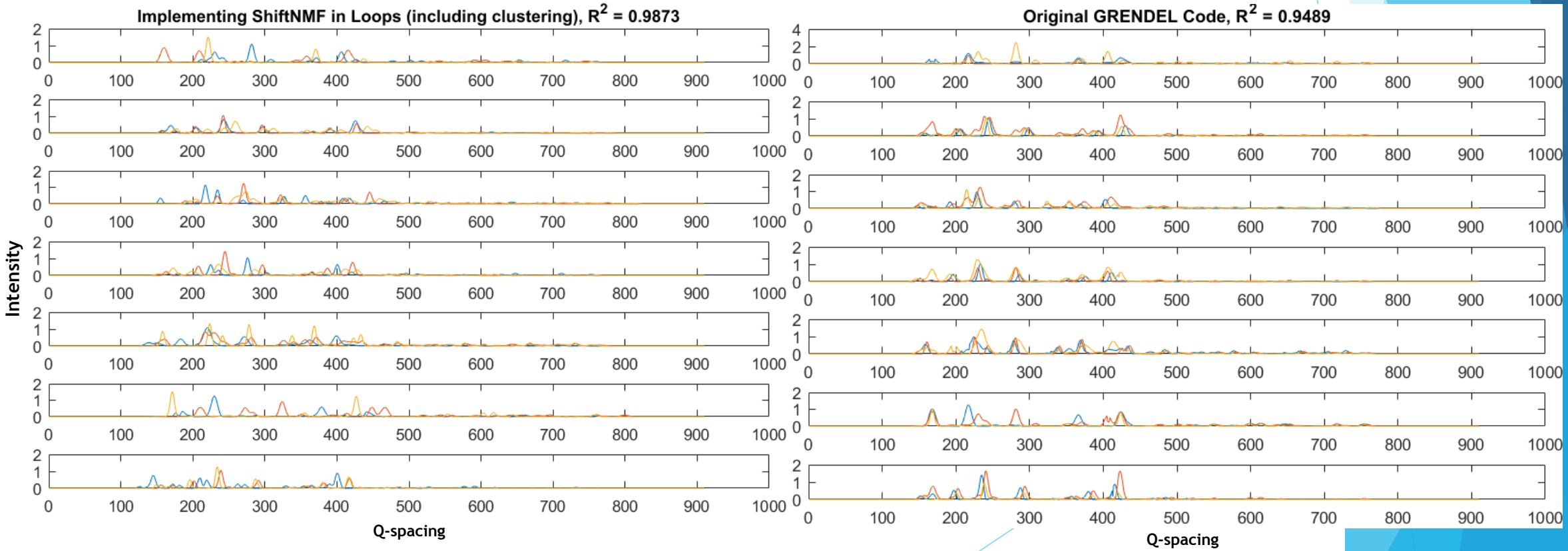
Experimental Statistics – Fe-Ga-Pd data set

Implementation	Max R^2	Mean R^2	T-statistic (w.r.t. GRENDEL stats)	P-value
Original GRENDEL	0.8871	0.8840	N/A	N/A
ShiftNMF without clustering	0.9155	0.9138	23.491	9.45E-11
ShiftNMF with clustering	0.9065	0.9055	17.450	6.89E-09

- Inorganic Crystal Structure Database (real sample, true values unknown)
- Zero-padded the input data **X**
- More data points → Makes clustering attempts more inaccurate!
- What could be the issue?

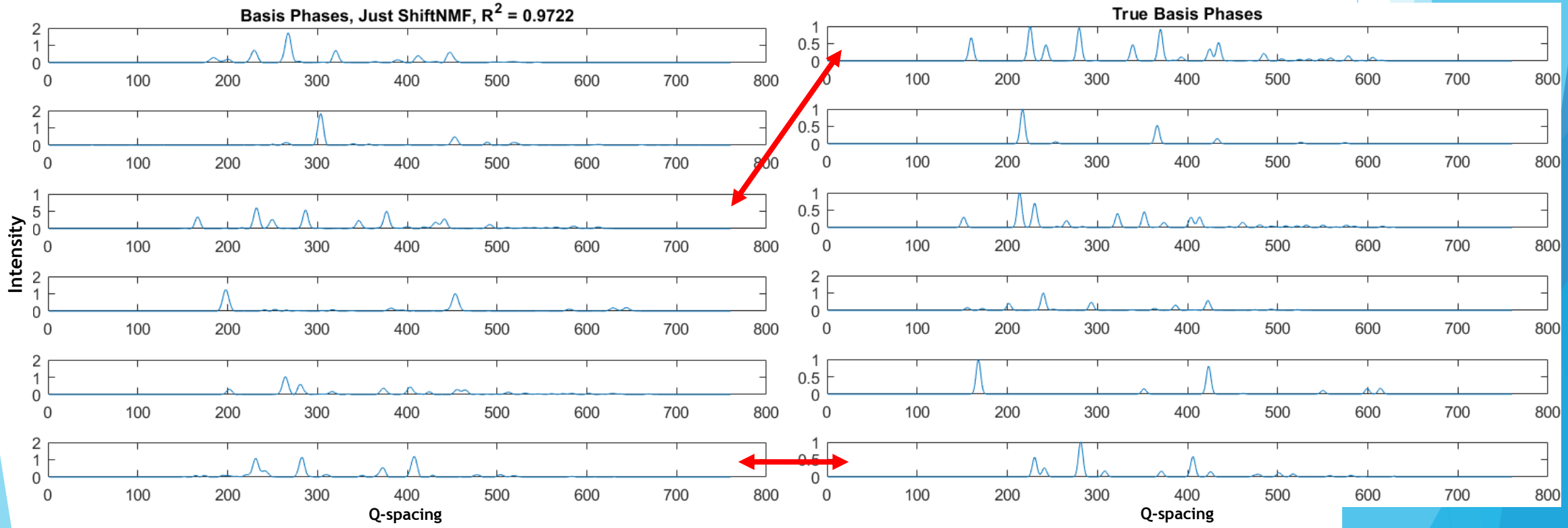
Adding ShiftNMF in with Clustering Creates Error

- Clustering is performed without peakshifting delays **T**
- Result: Anywhere from 15-18 out of 21 distinct phases (not 6 as it should be)



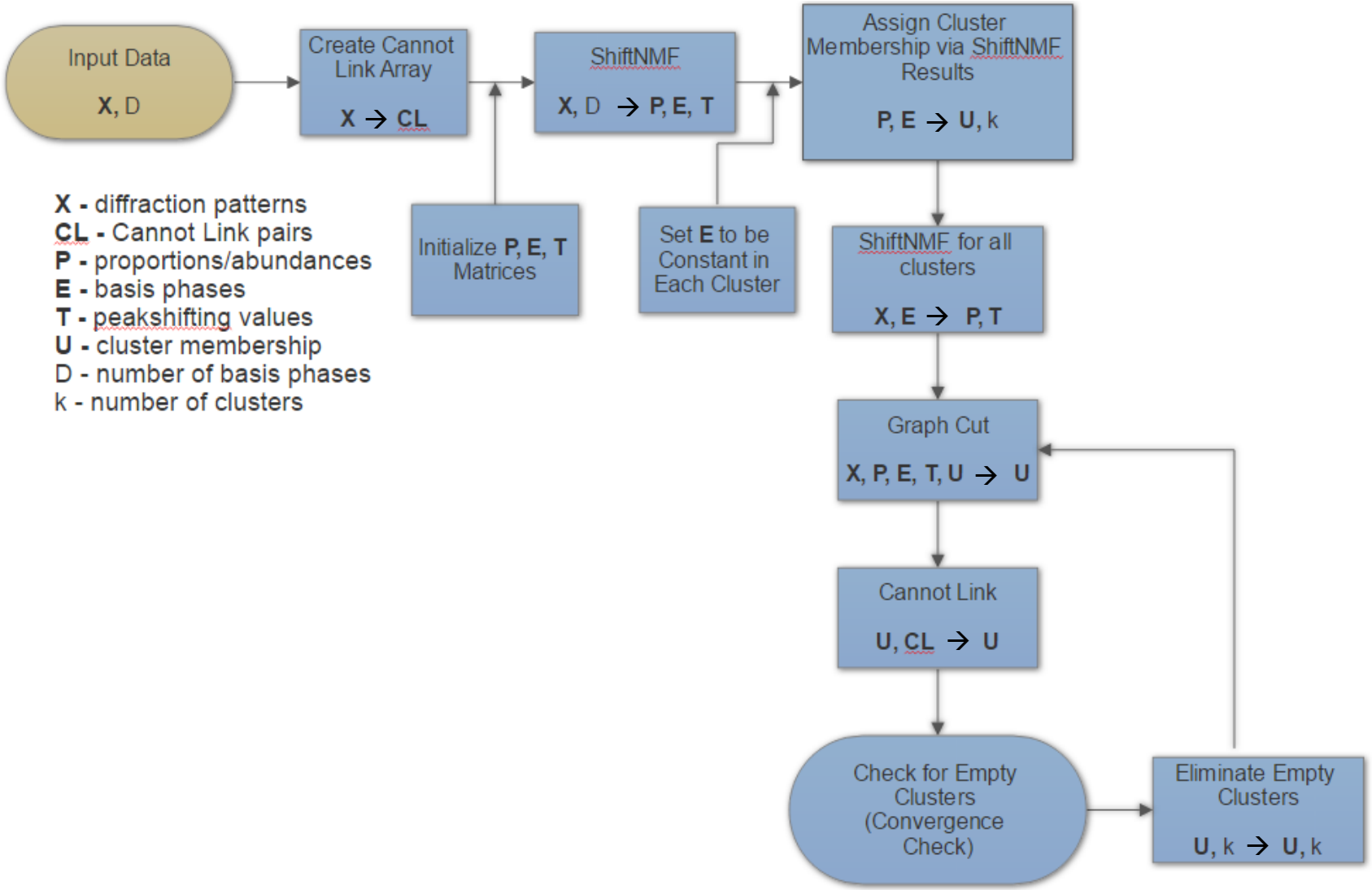
Comparing Basis Phases of Adding in ShiftNMF without Cluster to True Values

- See agreement with certain basis phases, even without adhering to Gibbs Phase Rule in ShiftNMF
- Have the same about of basis phases (6) we expect



New Algorithm - ShiftGRENDEL

- Attempt to formulate an algorithm that ensures correct number of basis phases, incorporates peakshifting delays **T** into clustering:

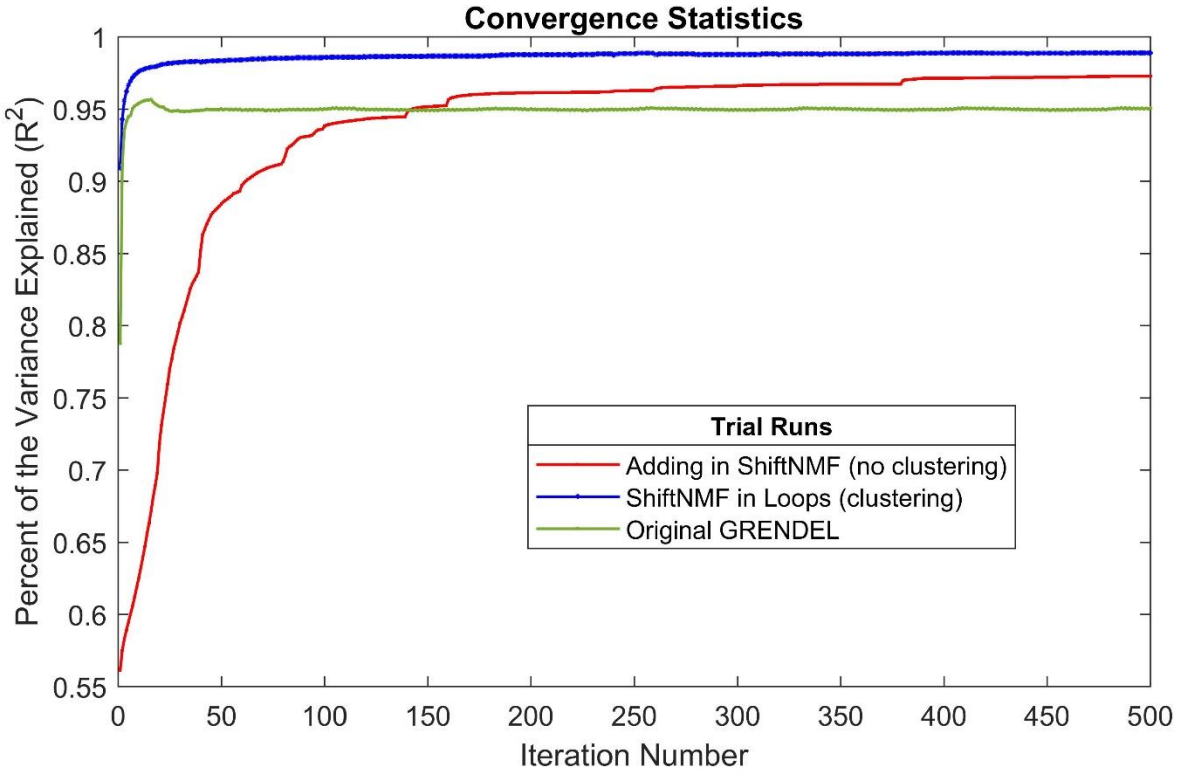


Testing ShiftGRENDEL

- Tested on (Fe-Al-Li)O_x synthetic data
- Original GRENDEL:
  - Mean R^2 → 0.9484
  - Max R^2 → 0.508
- ShiftGRENDEL
  - Mean R^2 → 0.9752
  - Max R^2 → 0.9609
  - T-statistic → 6.779
  - P-value → 4.90E-8

Timing Data, Speeding Up ShiftGRENDEL

- Average run-times (500 iterations)
  - Original GRENDEL – 30.1 (seconds)
  - ShiftNMF without clustering – 174.8
  - ShiftNMF with clustering – 147.7
  - ShiftGRENDEL – **1281.3**
- How to speed up:
  - Running ShiftNMF in the separate clusters parallelizable
  - Figure: Solution reached in < 500 iterations, do not need to run algorithm that long



Recap of AMSC 663/664 Work

- Cannot Link algorithm → increased connectivity of clusters
  - "Expert knowledge" constraint – based on observation, not law of physics
- ShiftNMF algorithm → takes peakshifting into account to correctly identify basis phases
  - Physical constraint – based on fundamental concept of physics/chemistry
- ShiftGRENDEL algorithm → alter GRENDEL program by incorporating Cannot Link and ShiftNMF
  - Provides framework to create first physically realistic basis phase recognition of inorganic materials, except …

Unresolved Issues – Summer Work

- Gibbs Phase Rule yet to be incorporated properly
  - LASSO method to implement constraint
- Develop mechanism to stop algorithm when algorithm is seen reaching local minimum
  - Attempt to save time, restart when undesirable result is detected early
- If both of these steps are successfully executed → First ever unsupervised method to identify clustering and basis phases of inorganic materials

Timeline/Milestones (OLD)

- Fully understand, replicate previous code/results – mid/late October
- Phase 1 – Constraint Programming
  - Add connectivity constraints, expert prior knowledge for given samples - November
  - Add constraints for peak shifting - January
  - Potential addition of other physical laws, ~~Mixed Integer Programming~~ - February
- ~~Phase 2 – Active Learning (Time permits)~~
  - ~~Have algorithm to predict next best point to sample – March~~
  - ~~Optimize the sampling algorithm for one material – mid April~~
  - ~~Optimize algorithm for all material data given – late April~~

Timeline/Milestones (Final Revision)

- Fully understand, replicate previous code/results – mid/late October
- Stage 1 – Connectivity Constraint
  - Write Cannot Link algorithm – November
  - Validate and optimize parameters – December
- Stage 2 – Peakshifting Constraint
  - Locate and understand algorithm, ShiftNMF – January
  - Write ShiftNMF algorithm – February
  - Validation – March
- Stage 3 – Optimization of GRENDEL
  - Develop method to integrate ShiftNMF with Graph Cut - April
  - Collect final results, decrease run time of algorithm - May

Deliverables

- Codes:
  - Original GRENDEL (with Cannot Link included)
  - ShiftNMF algorithm (with demo to test ShiftNMF on its own)
  - Algorithm adding ShiftNMF into GRENDEL without clusters
  - Algorithm adding ShiftNMF with clustering
  - ShiftGRENDEL
- All data sets used in testing
- Sample phase diagrams, basis phase spectral plots seen in reports

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).
- Zare A., Gader P., Bchir O., and Frigui H., *Piecewise Convex Multiple-Model Endemember Detection and Spectral Unmixing,* IEEE Transactions on Geoscience and Remote Sensing **51** (2013), no. 5: 2853-2862.
- Boykov Y. and Kologorov V., *An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision*, IEEE Transactions on PAMI, **26** (2004), no. 9: 1124-1137.
- Morup M., Madsen K. H., and Hansen L. K., *Shifted Non-negative Matrix Factorization*, IEEE International Workshop on Machine Learning for Signal Processing, (2007): pp. 139-144.
- Xue Y., Bai J., Le Bras R., Rappazzo B., Bernstein R., Bjork J., Longpre L., Suram S., van Dover R., Gregoire J., and Gomes C., *Phase-Mapper: An AI Platform to Accelerate High Throughput Material Discovery*, CoRR, **1610** (2016).
- Suram S., Xue Y., Bai J., Le Bras R., Rappazzo B., Bernstein R., Bjorck J., Zhou L., van Dover R., Gomes C., and Gregoire J., *Automated Phase Mapping with AgileFD and its Application to Light Absorber Discovery in the V-Mn-Nb Oxide System*, arXiv:1610.02005 (2016).
- Information about White House Genome Initiative courtesy of https://www.whitehouse.gov/mgi

Appendix: Cannot Link Constraint Algorithm

- Analysis of algorithm – NMF updates of **E** and **P** are what violate cluster connectivity requirement
- Algorithm:

  Compute cosine distance between all pairs

  Assign top p% dissimilar pairs to 'Cannot Link' array

  After initial Graph Cut:

    Remove pairs in CL which are initially clustered together

  After all subsequent Graph Cut iterations:

    Loop through all CL pairs:

    If pair in same cluster

      If 1st point changed cluster

        Revert cluster assignment of 1st point to old cluster

      Else

        Revert 2nd point's cluster assignment

      end

    end

Appendix B: ShiftNMF – P, E update rules

P update rule

$$E_{f,T} = E_f \circ \exp(i\omega T)$$

$$\text{grad}_P = \frac{-1}{M}(X_f - P_{f,T}E_f)E_{f,T}^H$$

$$\text{grad}_P^- = \frac{1}{M}X_f E_{f,T}^H$$

$$\text{grad}_P^+ = \frac{1}{M}P_f E_{f,T}E_{f,T}^H$$

$$G^+ = \text{ifft}(\text{grad}_P^+), \quad G^- = \text{ifft}(\text{grad}_P^-)$$

$$P = P \circ \left(\frac{G^-}{G^+}\right)^\alpha$$

Guaranteed convergence for $\alpha = 1$

E update rule

$$P_{f,T} = P_f \circ \exp(i\omega T)$$

$$\text{grad}_E = \frac{-1}{M}P_{f,T}^H(E_f - P_{f,T}E_f)$$

$$\text{grad}_E^- = \frac{1}{M}P_{f,T}^H P_{f,T}E_f$$

$$\text{grad}_E^+ = \frac{1}{M}P_{f,T}X_f$$

$$G^- = \text{ifft}(\text{grad}_E^-), \quad G^+ = \text{ifft}(\text{grad}_E^+)$$

$$E = E \circ \left(\frac{G^-}{G^+}\right)^\alpha$$

If $J_{new} \geq J_{old}$, then reduce α until $J_{new} < J_{old}$

Appendix C: ShiftNMF – T update rule

- Utilizes Newton-Raphson method:
  - $\mathbf{T} = \mathbf{T} - \eta \mathbf{B}^{-1}\mathbf{g}$
  - η – step size parameter
  - **B** – Hessian
  - **g** - gradient

$$P_{f,T} = P_f \circ \exp(i\omega T)$$

$$Q_f = P_{f,T} E_f$$

$$Y_f = X_f - Q_f$$

$$g = \frac{-1}{M} \sum_{\omega} 2\omega \Im[Q_f Y_f^*]$$

$$B = \left\{ \begin{array}{ll} \frac{-2}{M} \sum_{\omega} \omega^2 \Re[Q_f Q_f^*], & \text{for diagonal entries} \\ \frac{-2}{M} \sum_{\omega} \omega^2 \Re[Q_f (Q_f^* + Y_f^*)], & \text{else} \end{array} \right\}$$

$$T = T - \eta B^{-1} g$$

$$\text{If } J_{new} \geq J_{old}, \quad \text{then reduce } \eta \text{ until } J_{new} < J_{old}$$

Appendix D: ShiftNMF – Cross-Correlation Step

- Due to complexity of the objective function, local minima are abundant
- To avoid these, every 20 iterations we run a 'cross-correlation step'
- Done in random permutation order to shake up our **T** matrix

Randomly select d' phase, n' data point

Let $X_{n',f} = \text{fft(X) at n'}, \quad P_{n',f} = \text{fft(P) at n'}$

Let $E_{d',f} = \text{fft(E) at d'}$

$$R_{n',f} = X_{n',f} - \sum_{d \neq d'} P_{n',f} E_{d,f}$$

$$C_{n',f} = R^*_{n',f} E_{d',f}$$

$$t = \arg\max C_{n',f}$$

$$T_{n',d'} = t - (M+1)$$