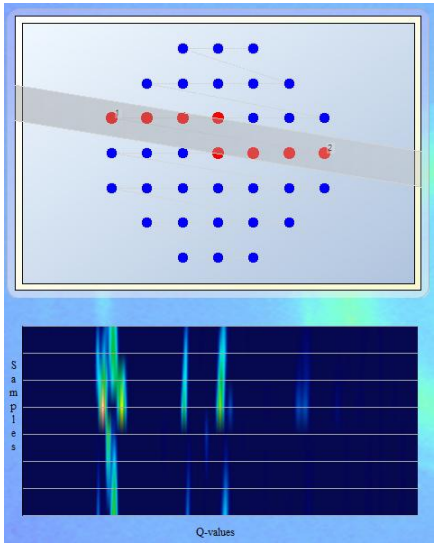
HUMAN COMPUTATION FOR MATERIALS DISCOVERY



Ronan Le Bras	Computer Science
Richard Bernstein	Computer Science
Carla P. Gomes	Computer Science
Bart Selman	Computer Science
R. Bruce van Dover	Materials Science/Physics

December 7, 2012

NIPS Workshop - HCSCS

Motivation

Cornell Fuel Cell Institute

Mission: develop **new materials** for **fuel cells**.

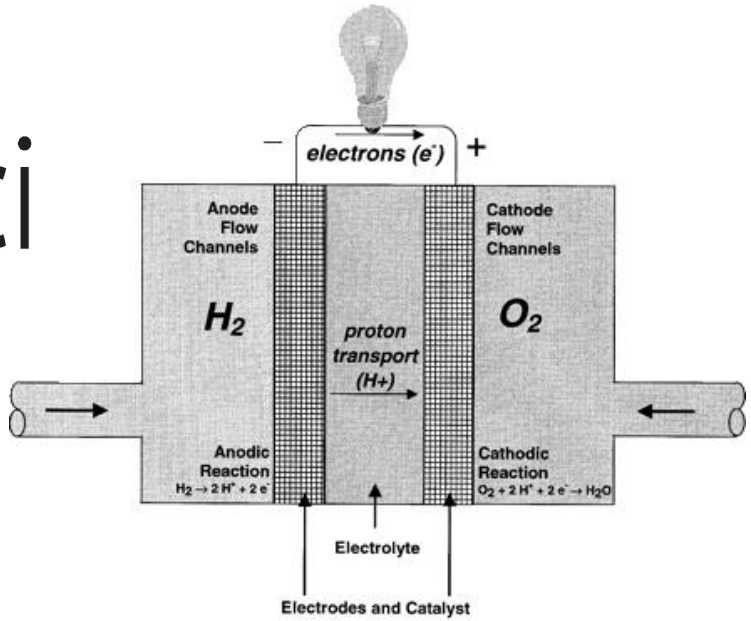


Figure 1. Fuel cell schematic.
Source: Annual Reveiws of Energy and the Environment. http://energy.annualreviews.org/cgi/content/full/24/1/281

An **Electrocatalyst** must:

1) Be electronically **conducting**
2) **Facilitate** both reactions

Platinum is the best known metal to fulfill that role, but:

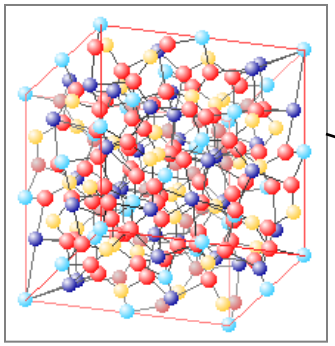
1) The reaction rate is still considered slow (causing **energy loss**)
2) Platinum is fairly **costly**, **intolerant** to fuel **contaminants**, and has a **short lifetime**.

Goal: Find an **intermetallic compound** that is a better catalyst than Pt.

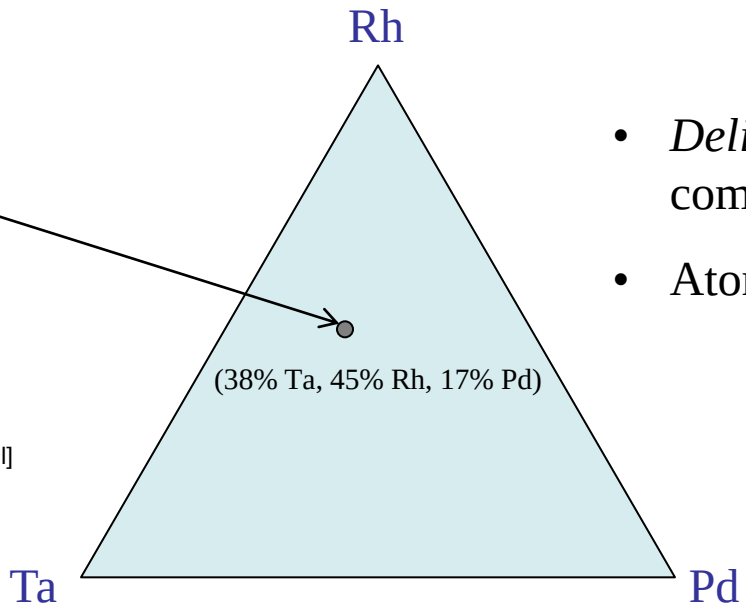
Motivation

Recipe for finding alternatives to Platinum

1) *In a vacuum chamber, place a silicon wafer.*
2) *Add three metals.*
3) *Mix until smooth, using three sputter guns.*
4) *Bake for 2 hours at 650ºC*



[Source: *Pyrotope*, Sebastien Merkel]



- *Deliberately* inhomogeneous composition on Si wafer
- Atoms are intimately mixed

Motivation

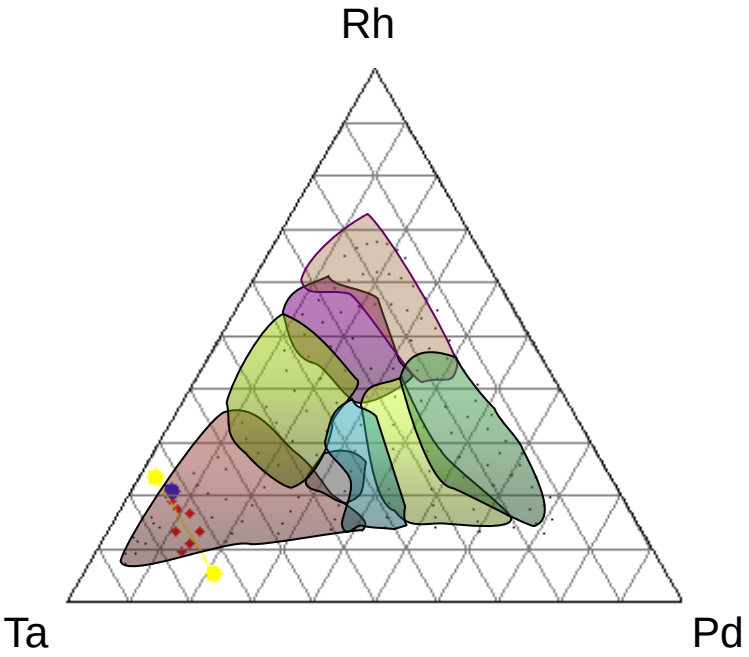


Figure 1: Phase regions of Ta-Rh-Pd

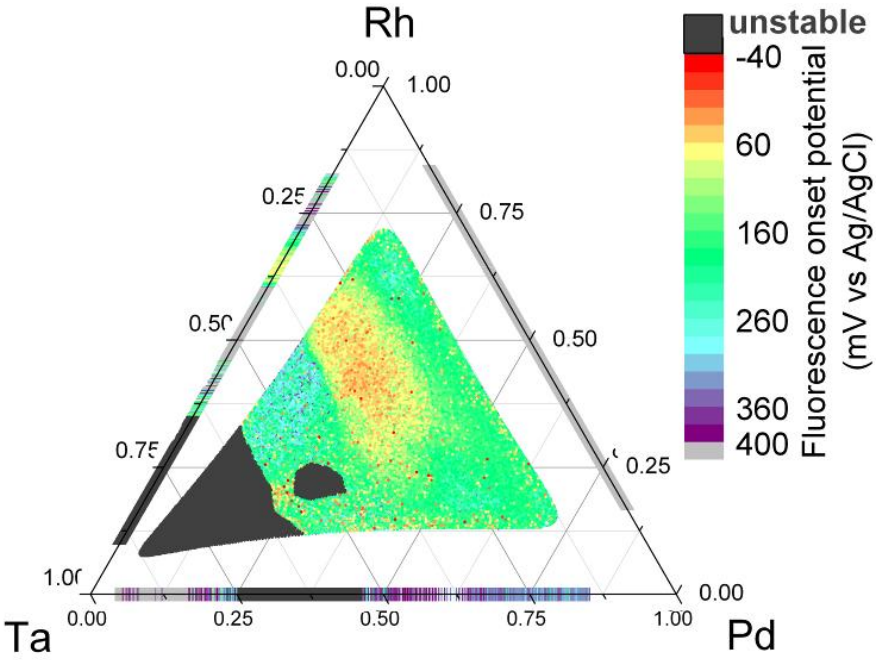
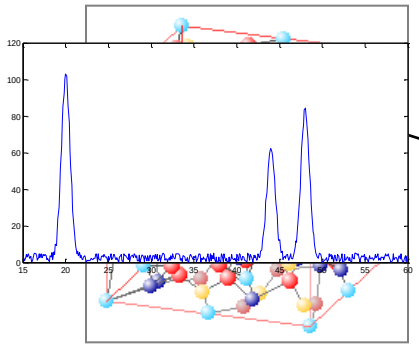


Figure 2: Fluorescence activity of Ta-Rh-Pd

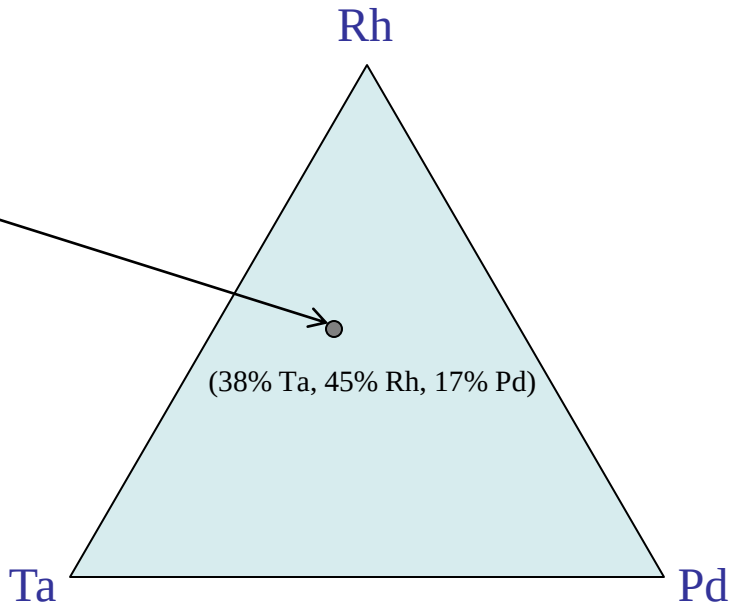
Motivation

Identifying crystal structure using **X-Ray Diffraction** at CHESS

- The XRD pattern **characterizes** the underlying **lattice** at a given point on the silicon wafer.

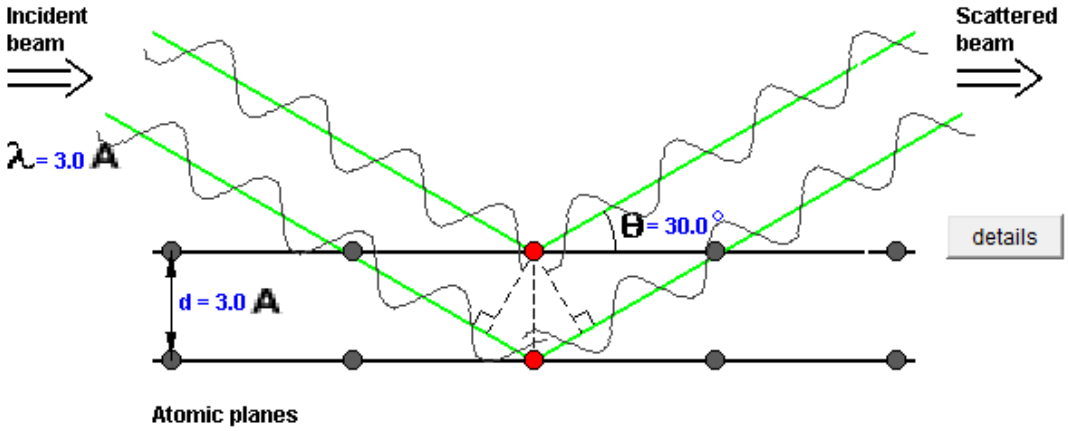


[Source: *Pyrotope*, Sebastien Merkel]

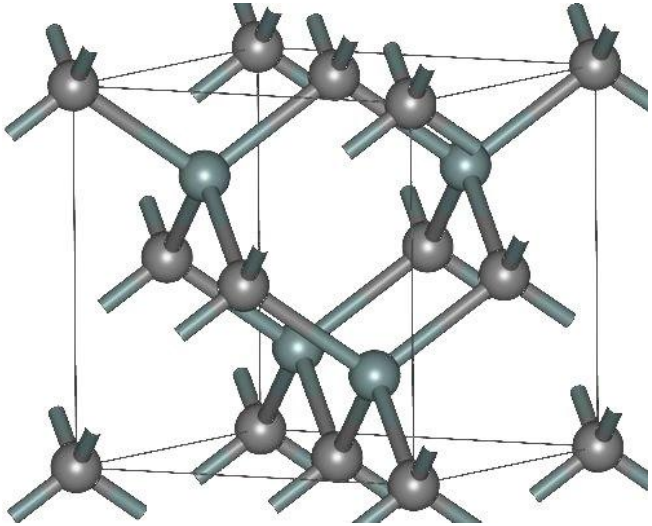


Motivation

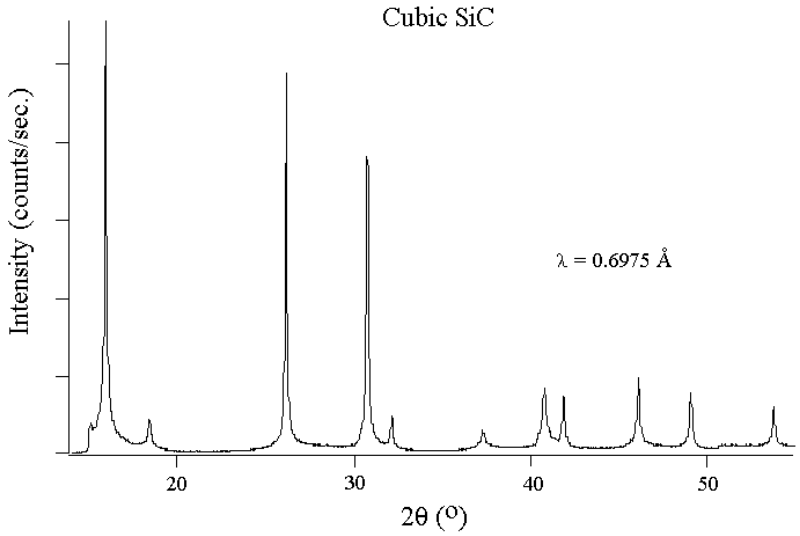
Bragg's law:



From

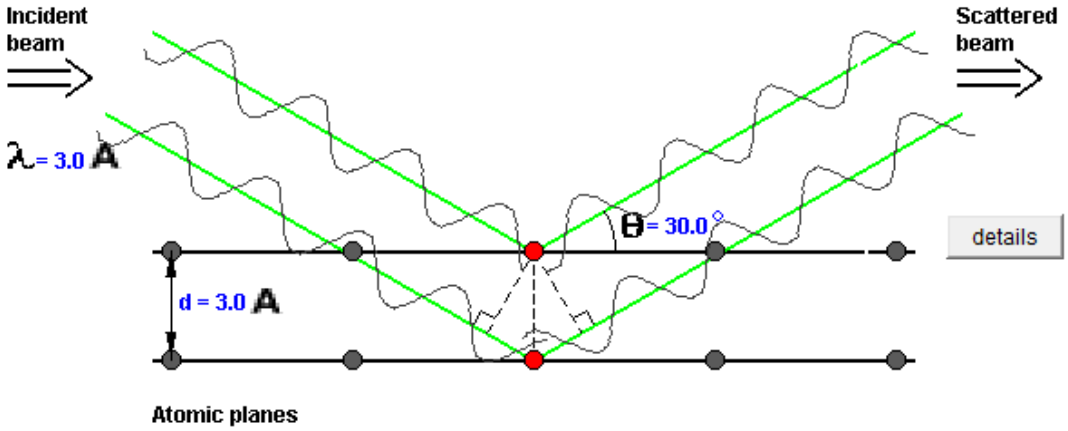


to

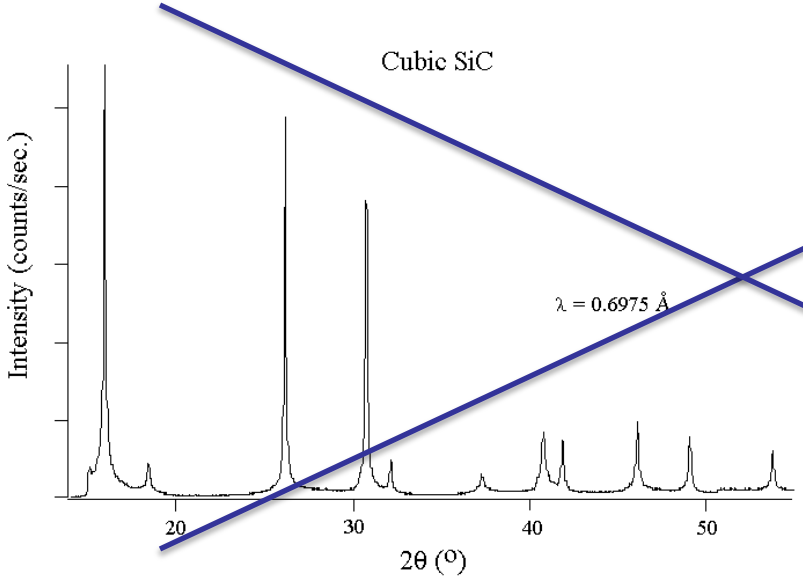


Motivation

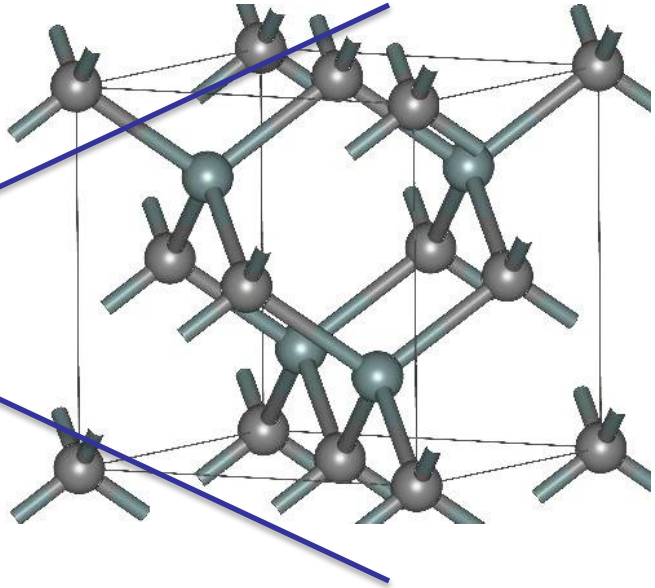
Bragg's law:



From



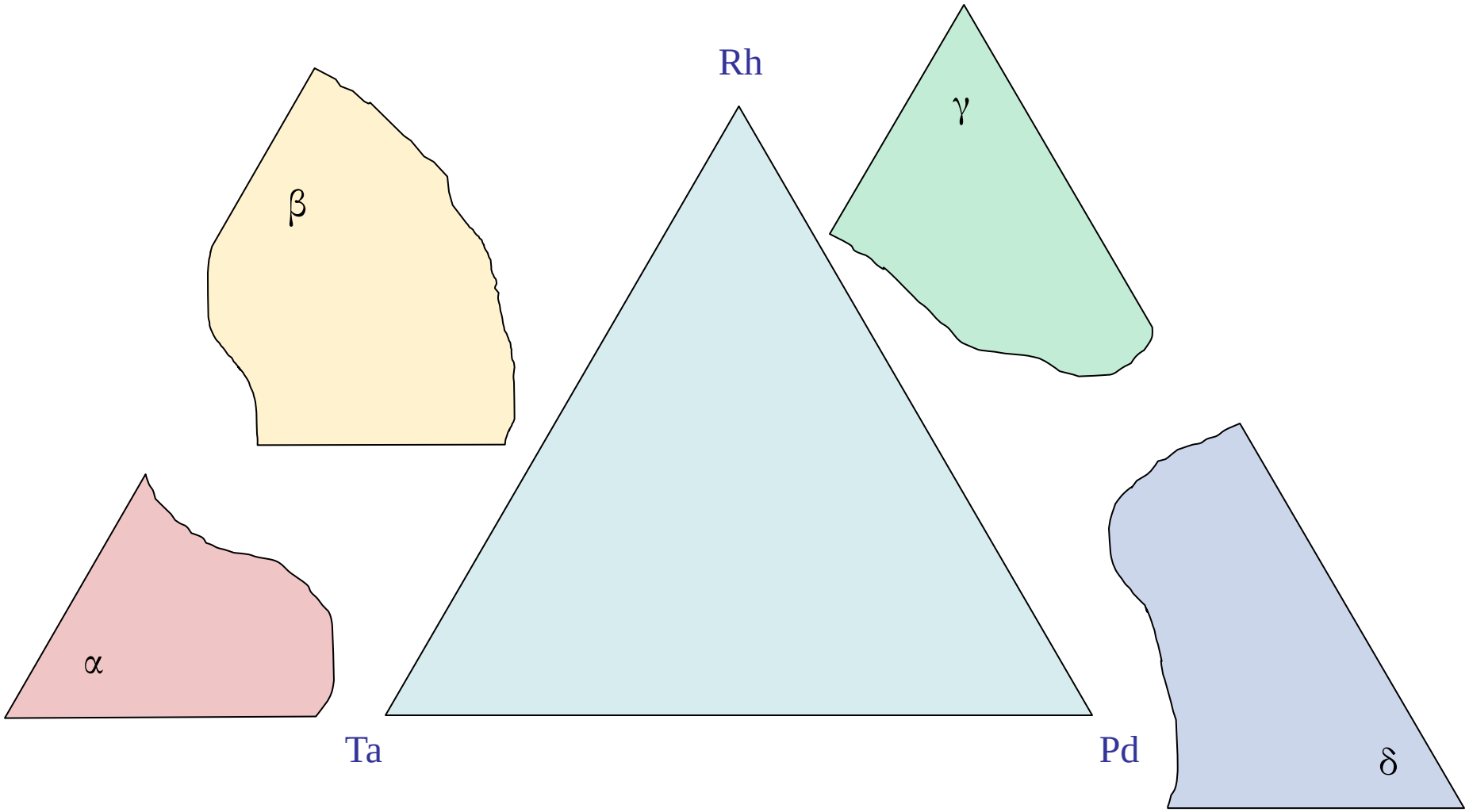
to



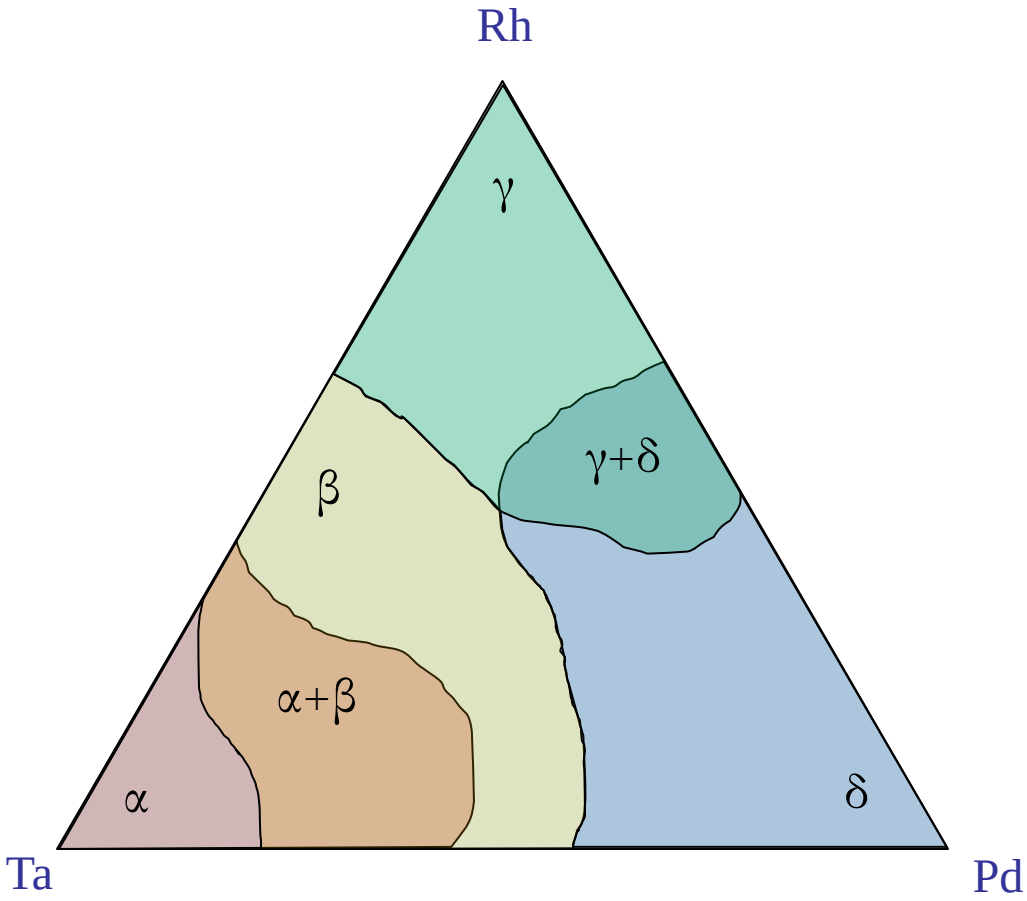
Outline

- ✓ Motivation
- Physical Characteristics
- Problem Definition
- Satisfiability Modulo Theory Approach
- UDiscoverIt
- Empirical Validation
- Conclusions and Future Work

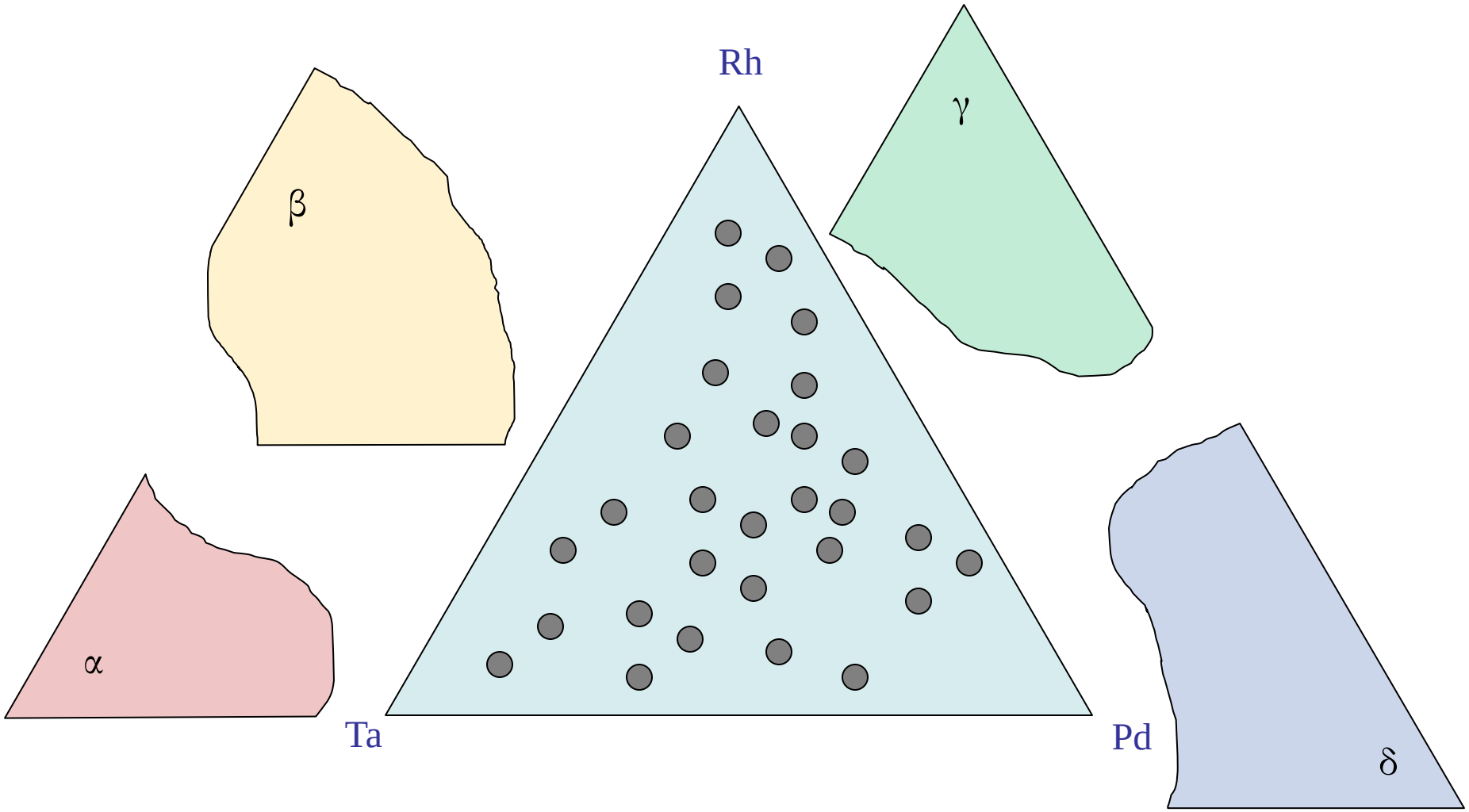
Physical Characteristics



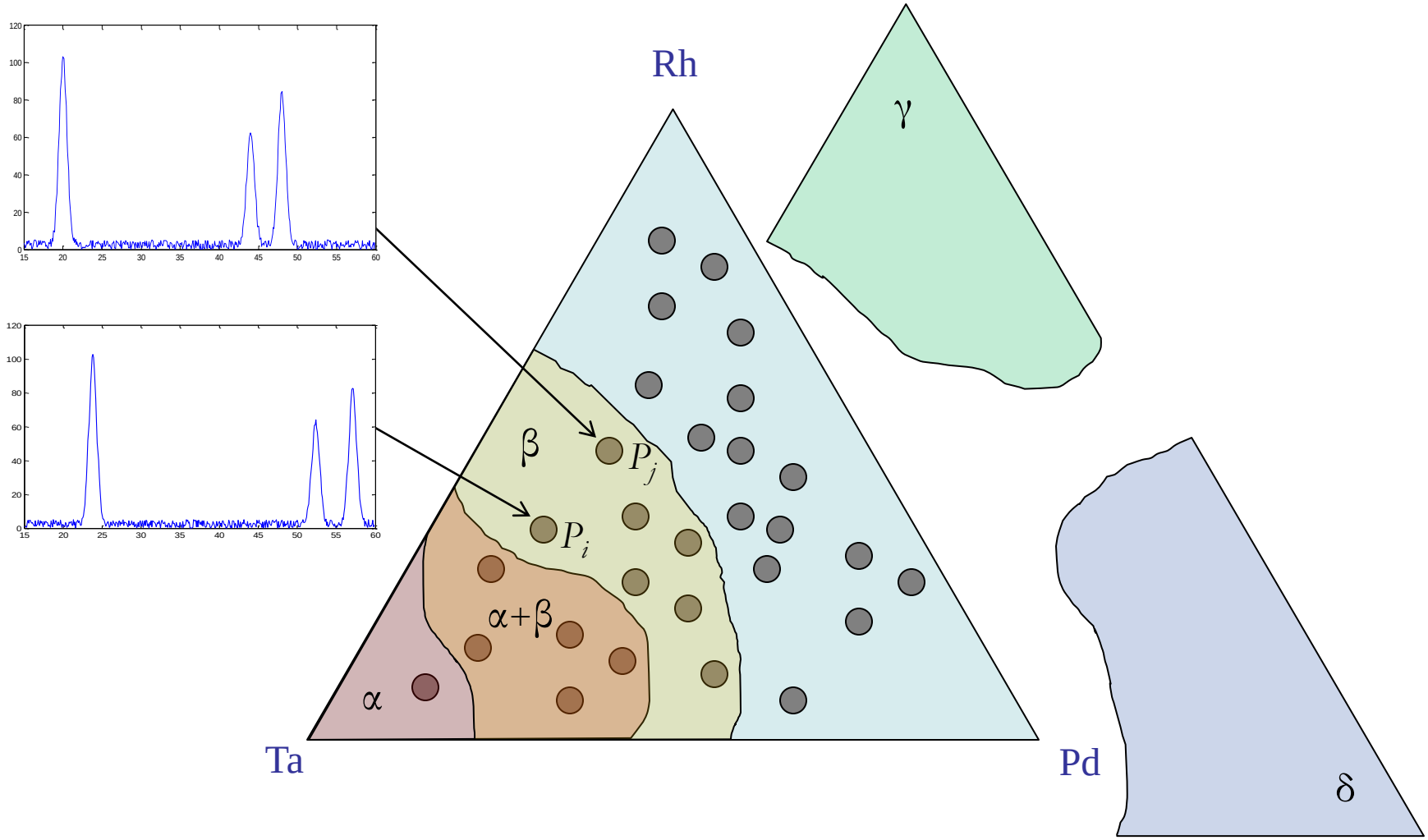
Physical Characteristics



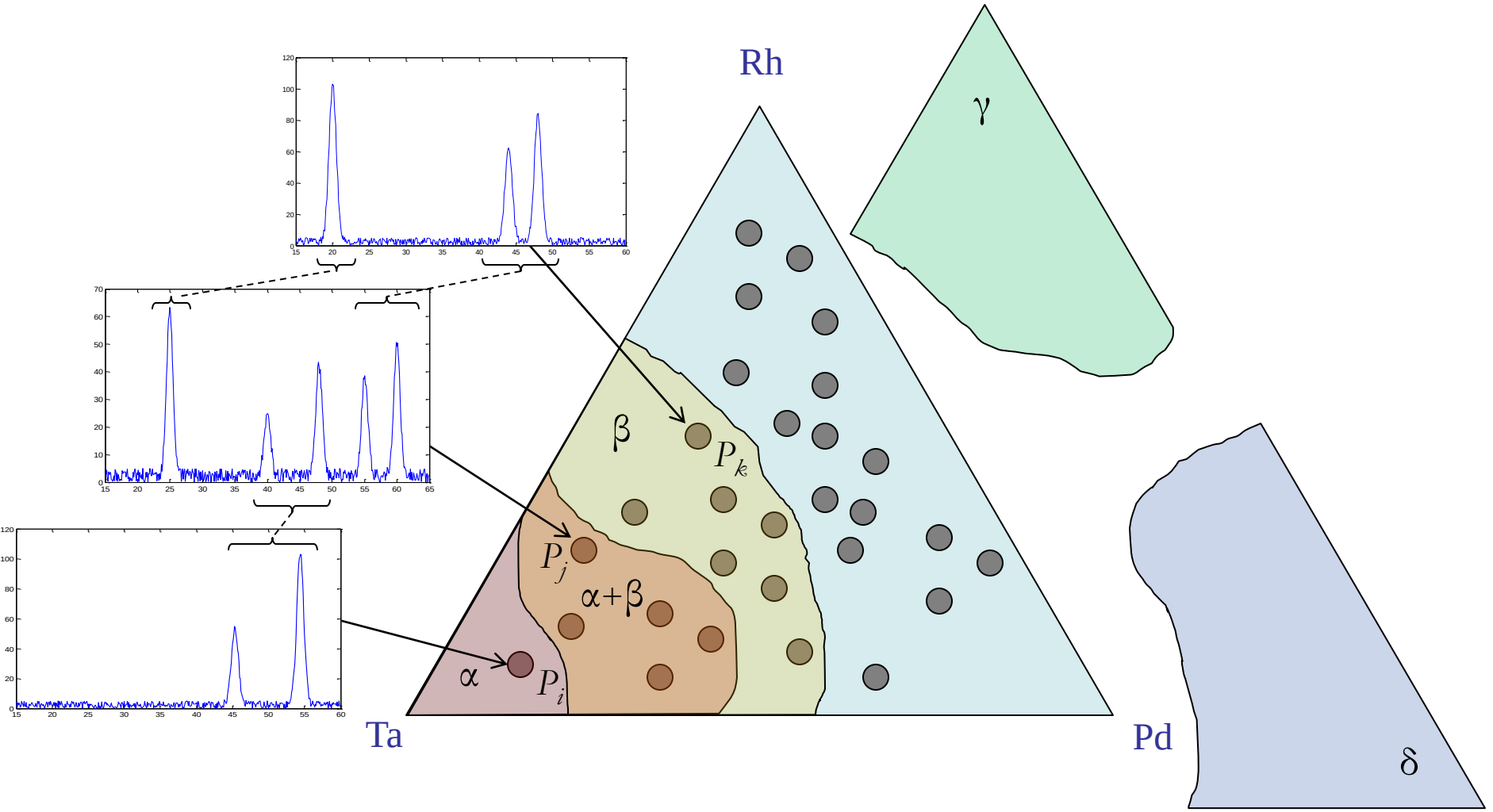
Physical Characteristics



Physical Characteristics

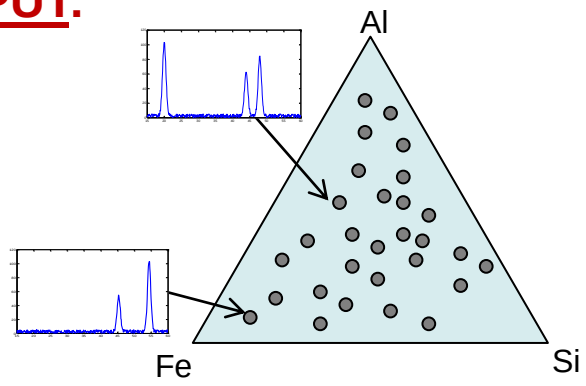


Physical Characteristics



Physical Characteristics

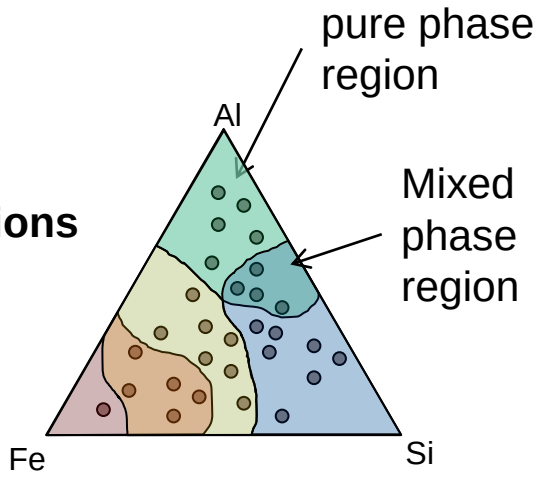
INPUT:



OUTPUT:

***m* phase regions**
- ***k* pure regions**
- ***m-k* mixed regions**

XRD pattern characterizing pure phases



Additional Physical characteristics:

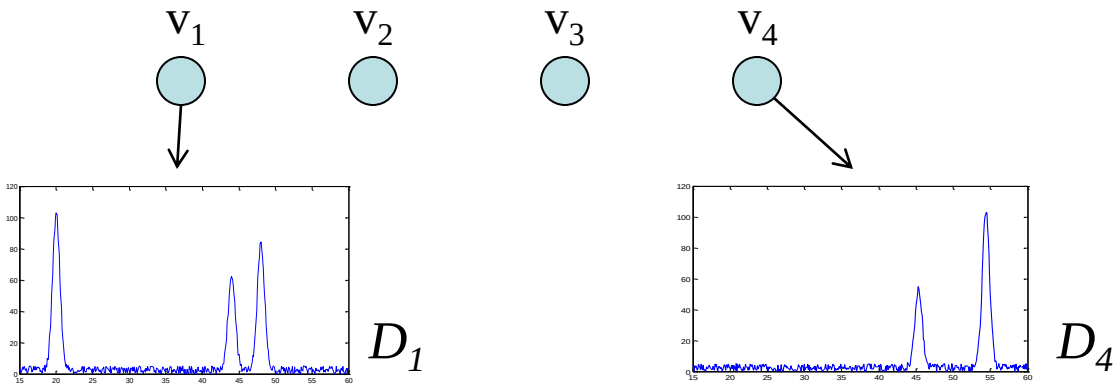
- Phase Connectivity
- Mixtures of ≤ 3 pure phases
- Peaks shift by ≤ 15% within a region
  - Continuous and Monotonic
- Noisy detection of peaks

Outline

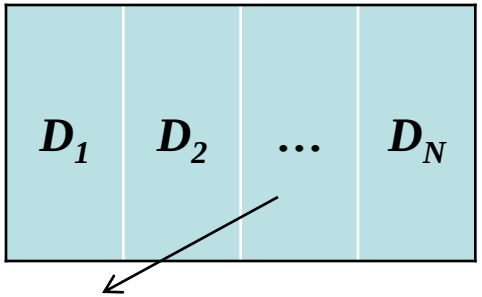
- ✓ Motivation
- ✓ Physical Characteristics
- Problem Definition
- Satisfiability Modulo Theory Approach
- UDiscoverIt
- Empirical Validation
- Conclusions and Future Work

Problem Definition

- **Input:**
  - A list of points on the silicon wafer



  - A real vector $\boldsymbol{D_i}$ per vertex v_i (diffraction patterns)
  - K = user specified number of pure phases
- **Goal**: a basis of K vectors for



$$\boldsymbol{D_i} = a_{i1}\boldsymbol{B_1} + \ldots + a_{iK}\boldsymbol{B_K}$$

Problem Definition

- There is **experimental noise**

Minimize norm instead

$$\boldsymbol{D_i} = a_{i1}\boldsymbol{B_1} + \ldots + a_{iK}\boldsymbol{B_K} \quad \Rightarrow \quad \boldsymbol{min}\ \|\boldsymbol{D_i} - (a_{i1}\boldsymbol{B_1} + \ldots + a_{iK}\boldsymbol{B_K})\|$$

- **Non-negative** basis vectors and coefficients

$$\boldsymbol{B_i} \geq \boldsymbol{0}\ ,\ a_{ij} \geq 0$$

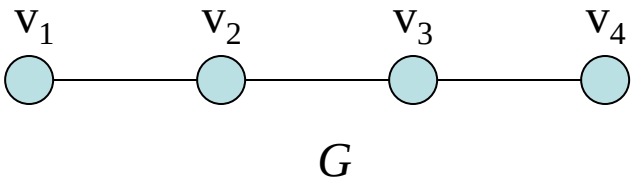
- **At most M (=3) non-zero coefficients** per point

$$|\{j \mid a_{ij} > 0\}| \leq M$$

- Basis patterns appear in **contiguous** locations on silicon wafer

  Build a graph G of the points on the silicon wafer

  The subgraph induced by $|\{i \mid a_{ij} > 0\}|$ is connected



Problem Definition

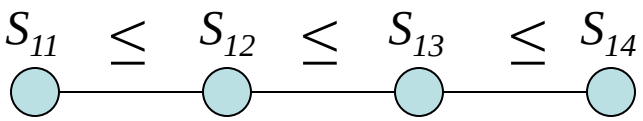
- Basis vector can be **shifted**

Shift operator

Shift coefficients

$$\min \| \boldsymbol{D_i} - (a_{i1} S(\boldsymbol{B_1}, s_{i1}) + \ldots + a_{iK} S(\boldsymbol{B_K}, s_{iK})) \|$$

- Shifts coefficients are **bounded**, **continuous** and **monotonic**



$$|S_{12} - S_{11}| \leq c$$

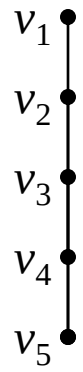
Outline

- ✓ Motivation
- ✓ Physical Characteristics
- ✓ Problem Definition
- • Satisfiability Modulo Theory Approach
- • UDiscoverIt
- • Empirical Validation
- • Conclusions and Future Work

Satisfiability Modulo Theories Approach

- Initial graph G and number K of basis patterns

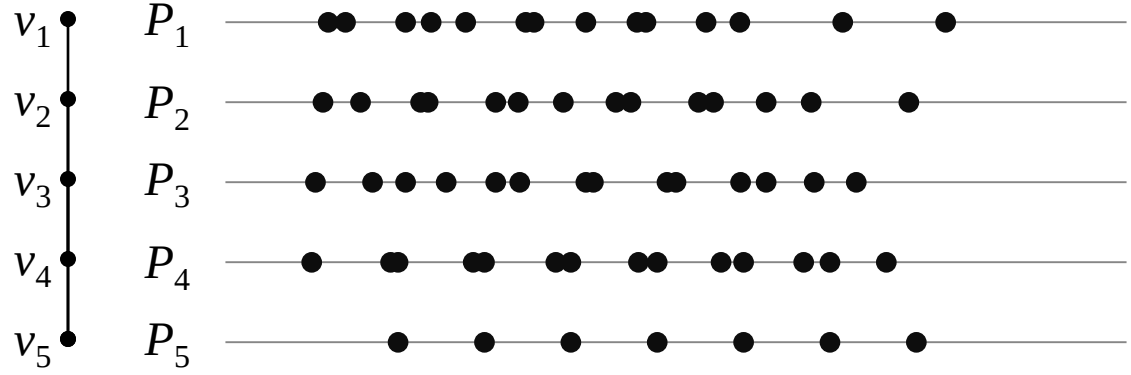
K=2 basis patterns



Satisfiability Modulo Theories Approach

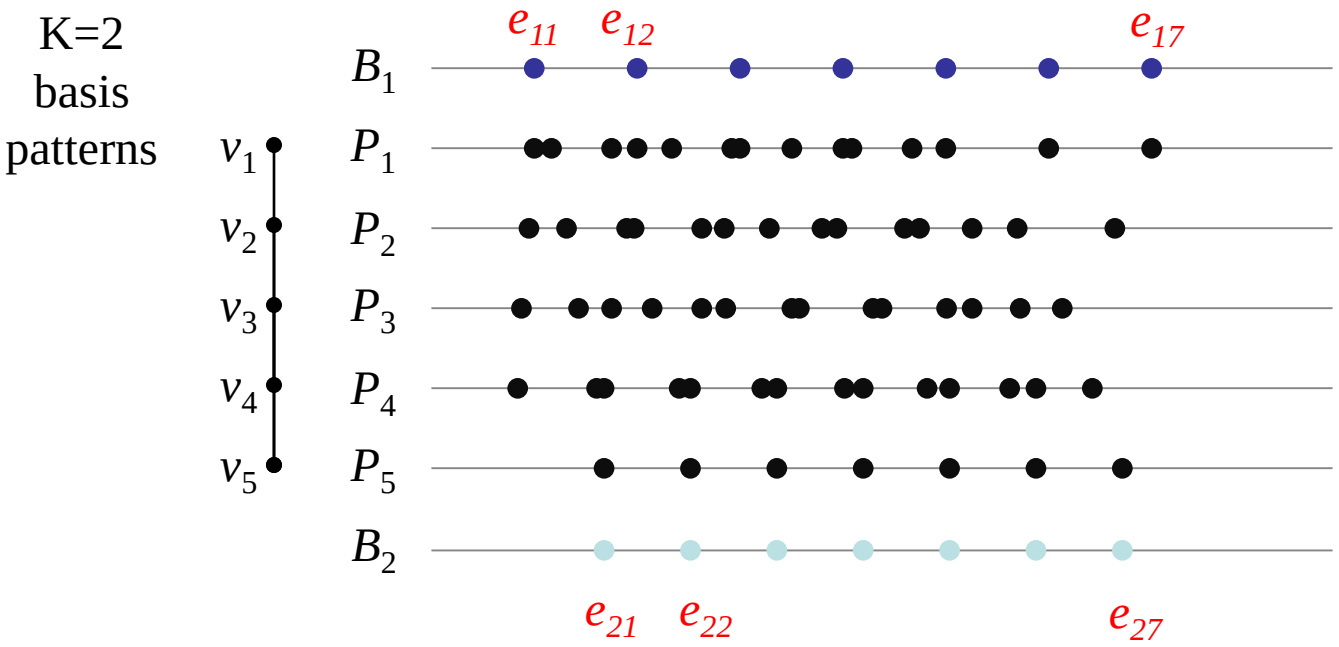
- Initial graph G and number K of basis patterns
- *Peak detection* to extract a set of peaks P_i for each diffraction pattern $\boldsymbol{D_i}$

K=2 basis patterns



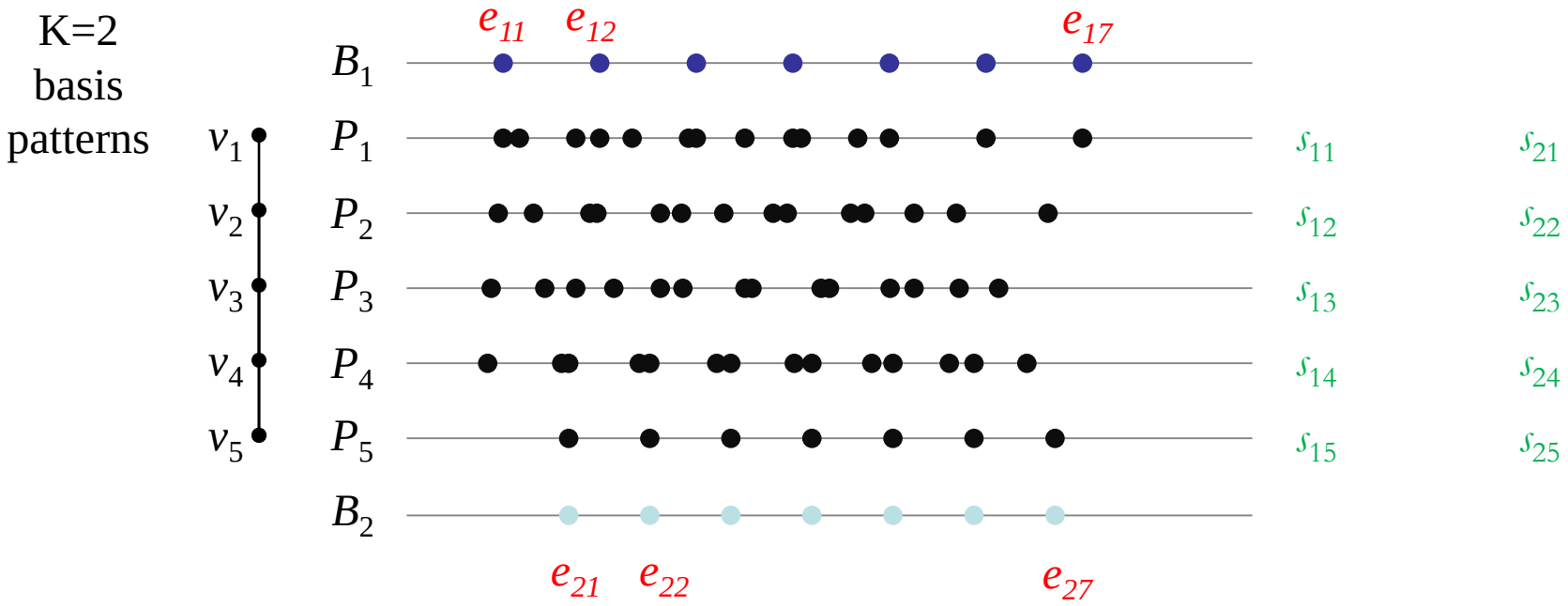
Satisfiability Modulo Theories Approach

- Initial graph G and number K of basis patterns
- *Peak detection* to extract a set of peaks P_i for each diffraction pattern $\boldsymbol{D_i}$
- Real variable e_{jk} for the **location of peak** k in basis B_j



Satisfiability Modulo Theories Approach

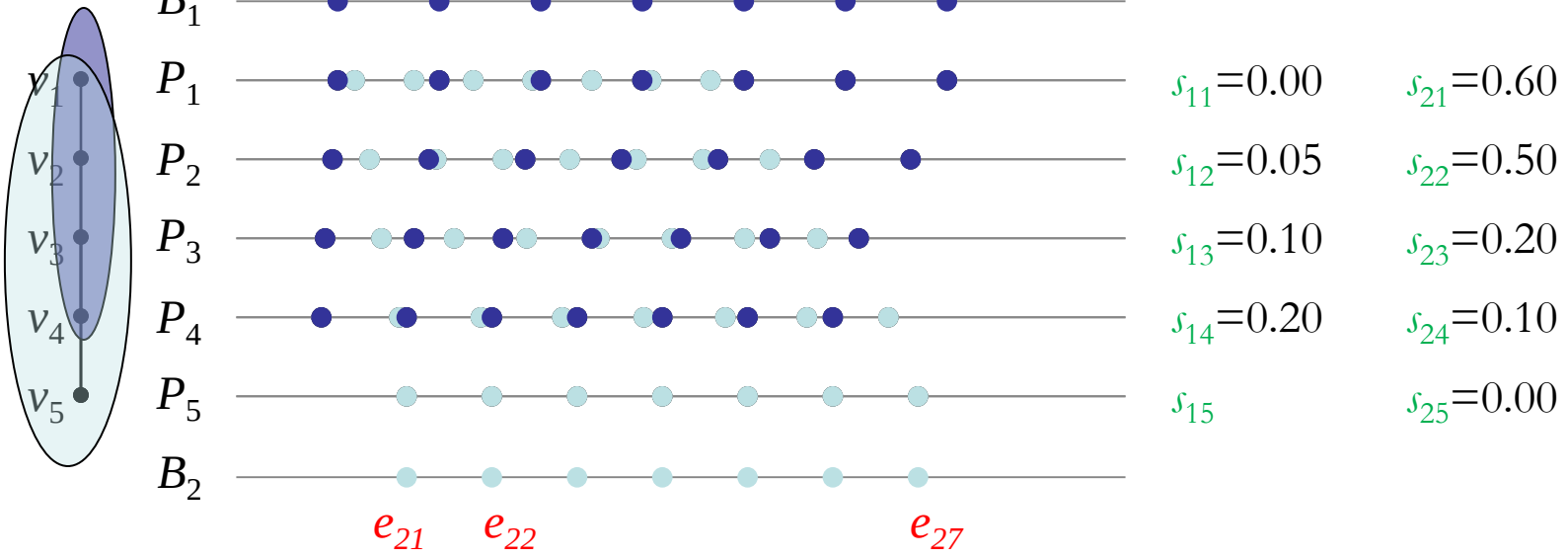
- Initial graph G and number K of basis patterns
- *Peak detection* to extract a set of peaks P_i for each diffraction pattern $\boldsymbol{D_i}$
- Real variable e_{jk} for the **location of peak** k in basis B_j
- Real variable s_{ij} for the **shift** coefficient of basis B_j in point P_i



Satisfiability Modulo Theories Approach

- An observed peak p is "*explained*" if there exists s_{ij} ,e_{jk} s.t. $|p-(s_{ij} + e_{jk})| \leq \varepsilon$
- Every observed peak must be *explained*

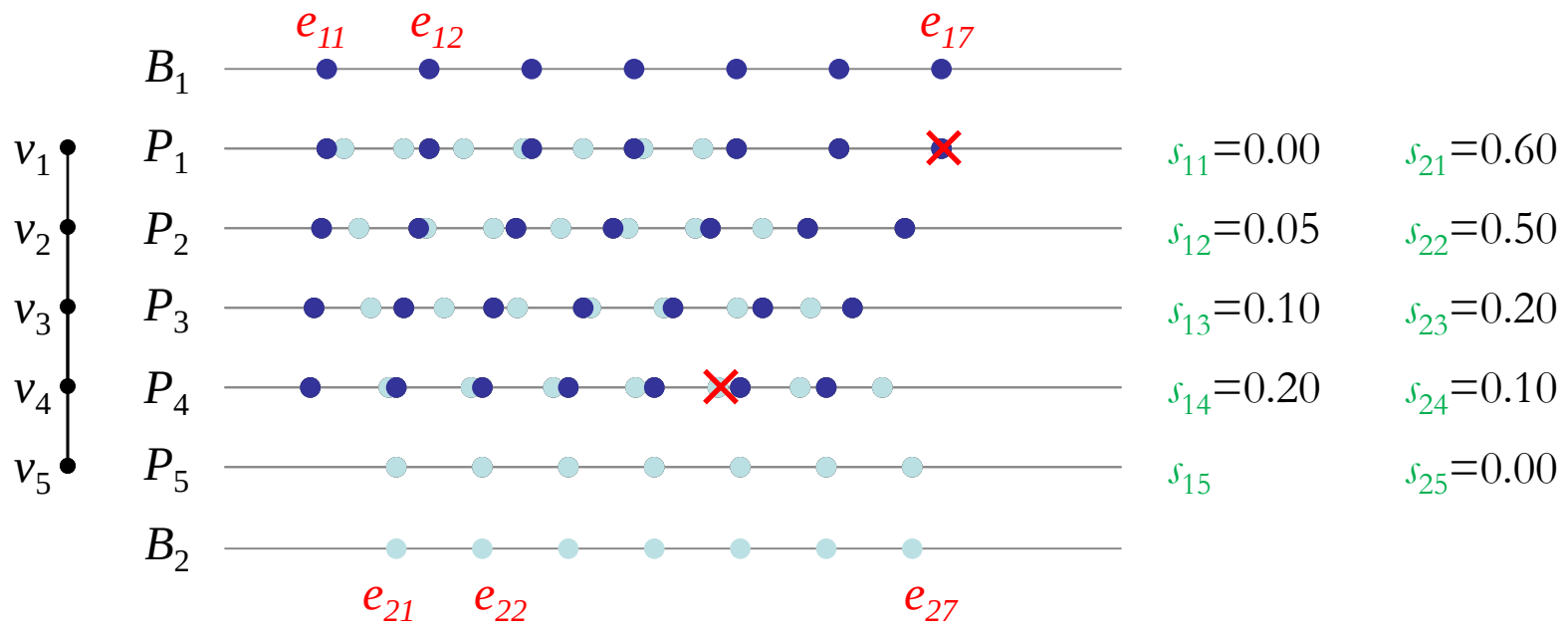
K=2 basis patterns



Satisfiability Modulo Theories Approach

- An observed peak p is "*explained*" if there exists s_{ij} ,e_{jk} s.t. $|p-(s_{ij} + e_{jk})| \leq \varepsilon$
- Every observed peak must be *explained*
- Some peaks might be missing (unobserved)
- Bound the number of missing peaks $\leq T$
- Minimization by (binary) search on T

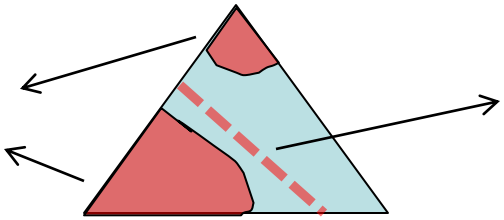
K=2 basis patterns



Satisfiability Modulo Theories Approach

- Linear phase usage constraints (up to M basis patterns per point)
- Linear constraints for shift monotonicity and continuity ($s_{ij} \leq s_{lm}$)
- **Lazy connectivity**: add a cut if current solution is not connected

If disconnected regions explained with phase 1



Then Phase 1 must appear in at least one of these points

- **Symmetry breaking**:
  - Renaming of pure phases
  - Ordering of the peak locations e_{jk} (per basis pattern)

⇨ **Quantifier-free linear arithmetic**

Experimental results for SMT

	Dataset					Time *(s)*
System	P	L^*	K	*#var*	*#cst*	
A/B/C	36	8	4	408	2095	3502
A/B/C	60	8	4	624	3369	17345
Al/Li/Fe	15	6	6	267	1009	79
Al/Li/Fe	28	6	6	436	1864	346
Al/Li/Fe	28	8	6	490	2131	10076
Al/Li/Fe	28	10	6	526	2309	28170
Al/Li/Fe	45	7	6	693	3281	18882
Al/Li/Fe	45	8	6	711	3410	46816

Table: Runtime (seconds) of the SMT solver. P is the number of sample points, L^* is the average number of peaks per phase, K is the number of basis patterns, $\#var$ is the number of variables and $\#cst$ is the number of constraints.

Research Questions

- Can **human input** be used to significantly **boost the performance** of combinatorial **reasoning** and **optimization** methods?.

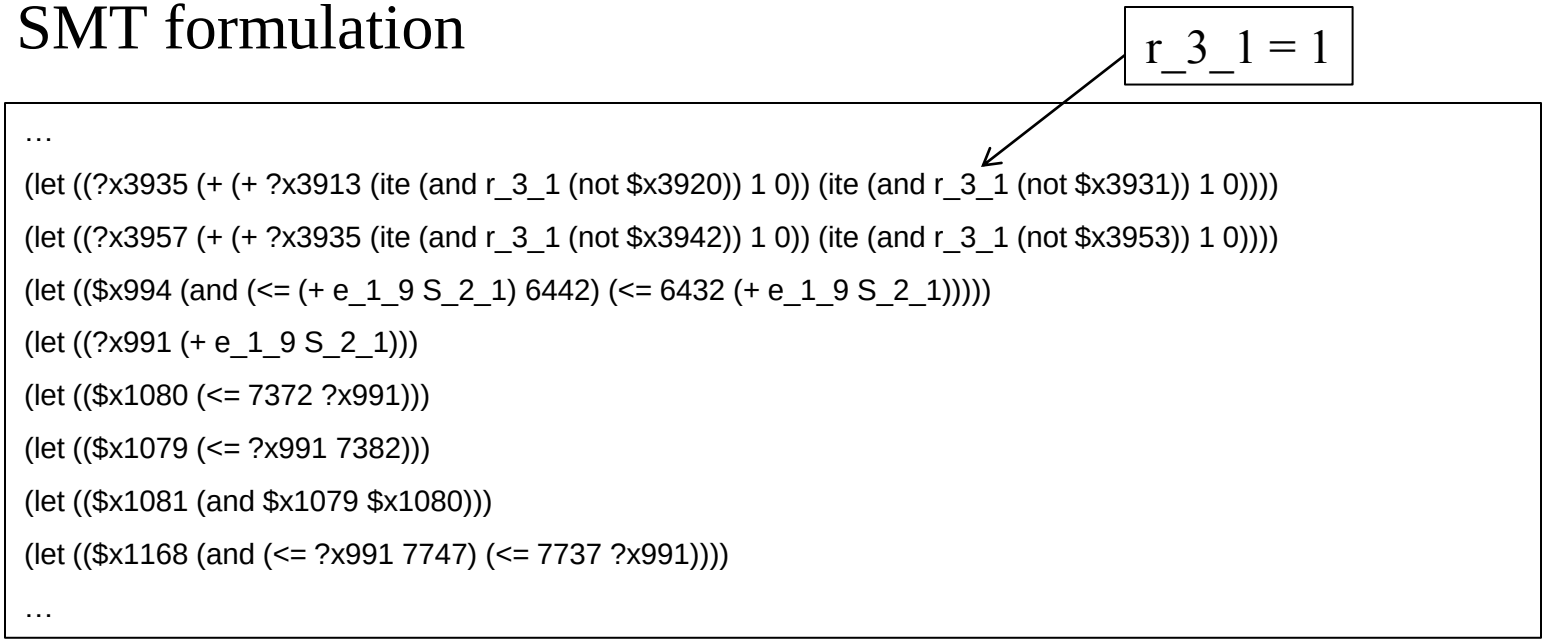
- Can **human input** provide useful global guidance to the solver, by identifying the setting of so-called **backdoor variables** in the SMT model?

Outline

- ✓ Motivation
- ✓ Physical Characteristics
- ✓ Problem Definition
- ✓ Satisfiability Modulo Theory Approach
- • UDiscoverIt
- • Empirical Validation
- • Conclusions and Future Work

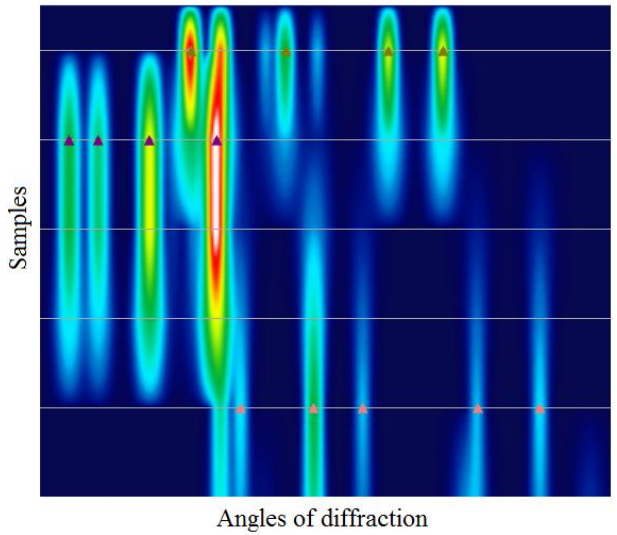
UDiscoverIt

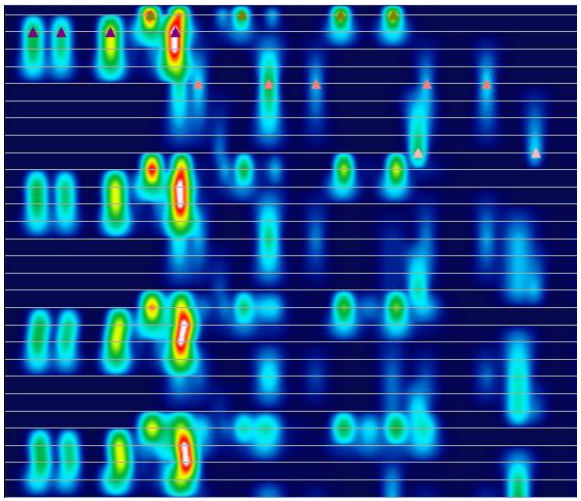
- Graphical User Interface for providing user input to an SMT solver
  - **Visualization** of the X-ray diffraction data
  - User provides **partial assignments** of the variables of the SMT formulation



UDiscoverIt

- Graphical User Interface for providing user input to an SMT solver
  - **Visualization** of the X-ray diffraction data
  - User provides **partial assignments** of the variables of the SMT formulation
  - **Representation** inspired from how **materials scientists analyze** the data and address this problem

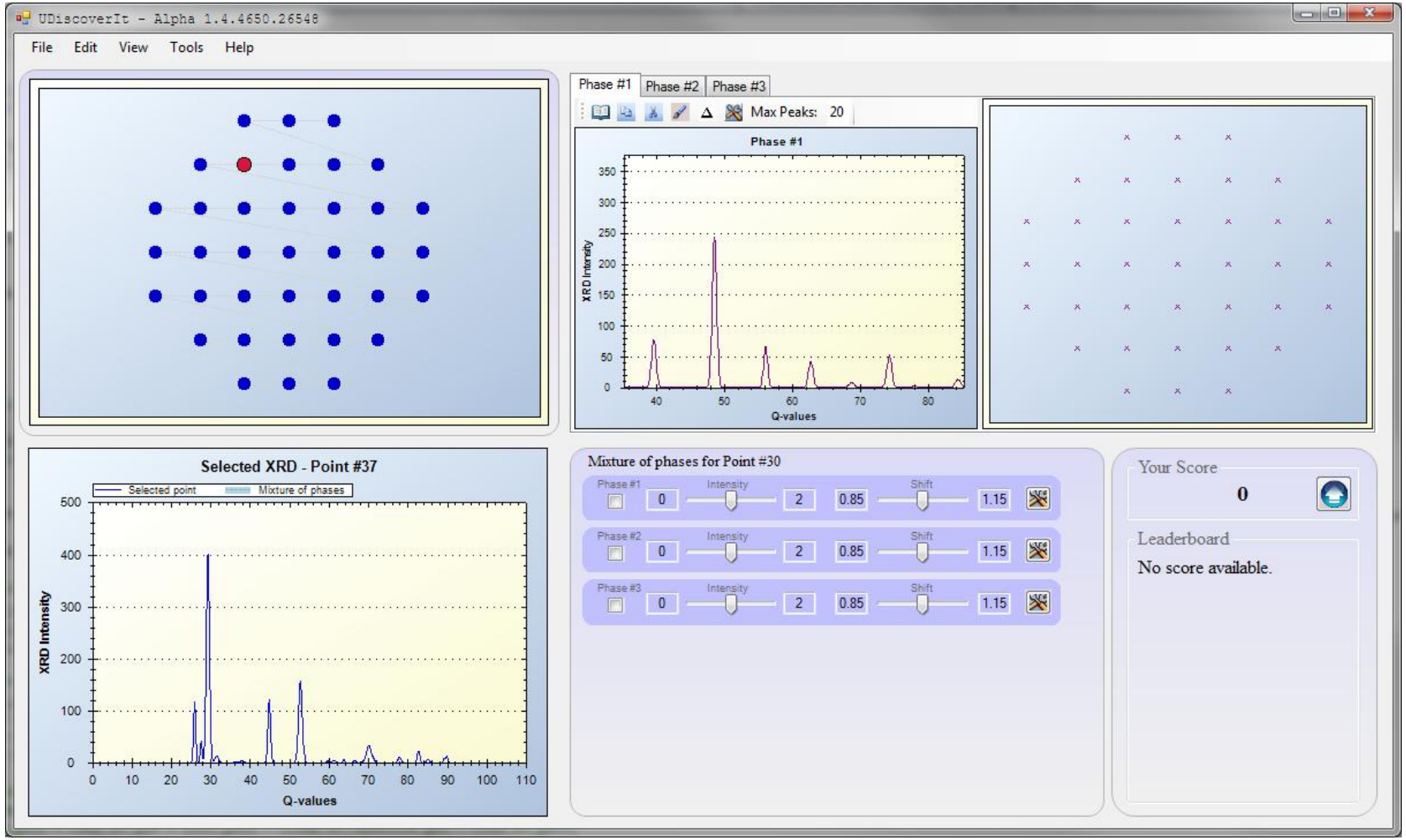




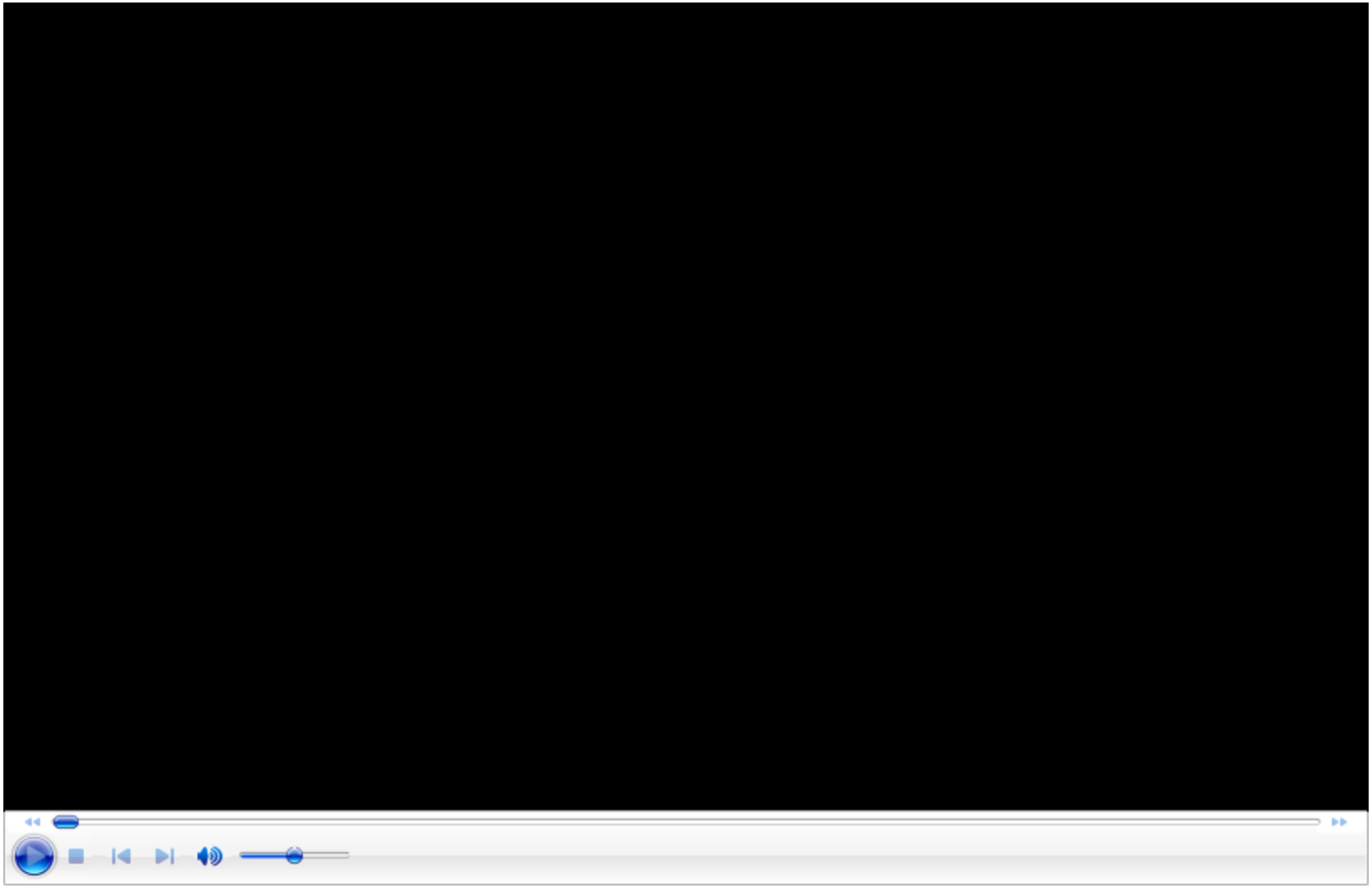
UDiscoverIt

- Graphical User Interface for providing user input to an SMT solver
  - **Visualization** of the X-ray diffraction data
  - User provides **partial assignments** of the variables of the SMT formulation
  - **Representation** inspired from how **materials scientists analyze** the data and address this problem
  - **No knowledge** requirement about the underlying **phase** structure, and only very **limited knowledge** of **diffraction** methods.

UDiscoverIt



UDiscoverIt



Outline

- ✓ Motivation
- ✓ Physical Characteristics
- ✓ Problem Definition
- ✓ Satisfiability Modulo Theory Approach
- ✓ UDiscoverIt
- • Empirical Validation
- • Conclusions and Future Work

Experimental results for SMT with user input

Dataset						Time w/o user input *(s)*	Time w/ user input *(s)*	# assigned var. by user
System	P	L^*	K	*#var*	*#cst*			
A/B/C	36	8	4	408	2095	3502	**150**	19 (4.6%)
A/B/C	60	8	4	624	3369	17345	**261**	18 (2.9%)
Al/Li/Fe	15	6	6	267	1009	79	**27**	6 (2.2%)
Al/Li/Fe	28	6	6	436	1864	346	**83**	12 (2.7%)
Al/Li/Fe	28	8	6	490	2131	10076	**435**	26 (5.3%)
Al/Li/Fe	28	10	6	526	2309	28170	**188**	23 (4.3%)
Al/Li/Fe	45	7	6	693	3281	18882	**105**	28 (4.0%)
Al/Li/Fe	45	8	6	711	3410	46816	**74**	30 (4.2%)

Table: Runtime (seconds) of the SMT solver with and without user input. P is the number of sample points, L^* is the average number of peaks per phase, K is the number of basis patterns, $\#var$ is the number of variables and $\#cst$ is the number of constraints.

Conclusions

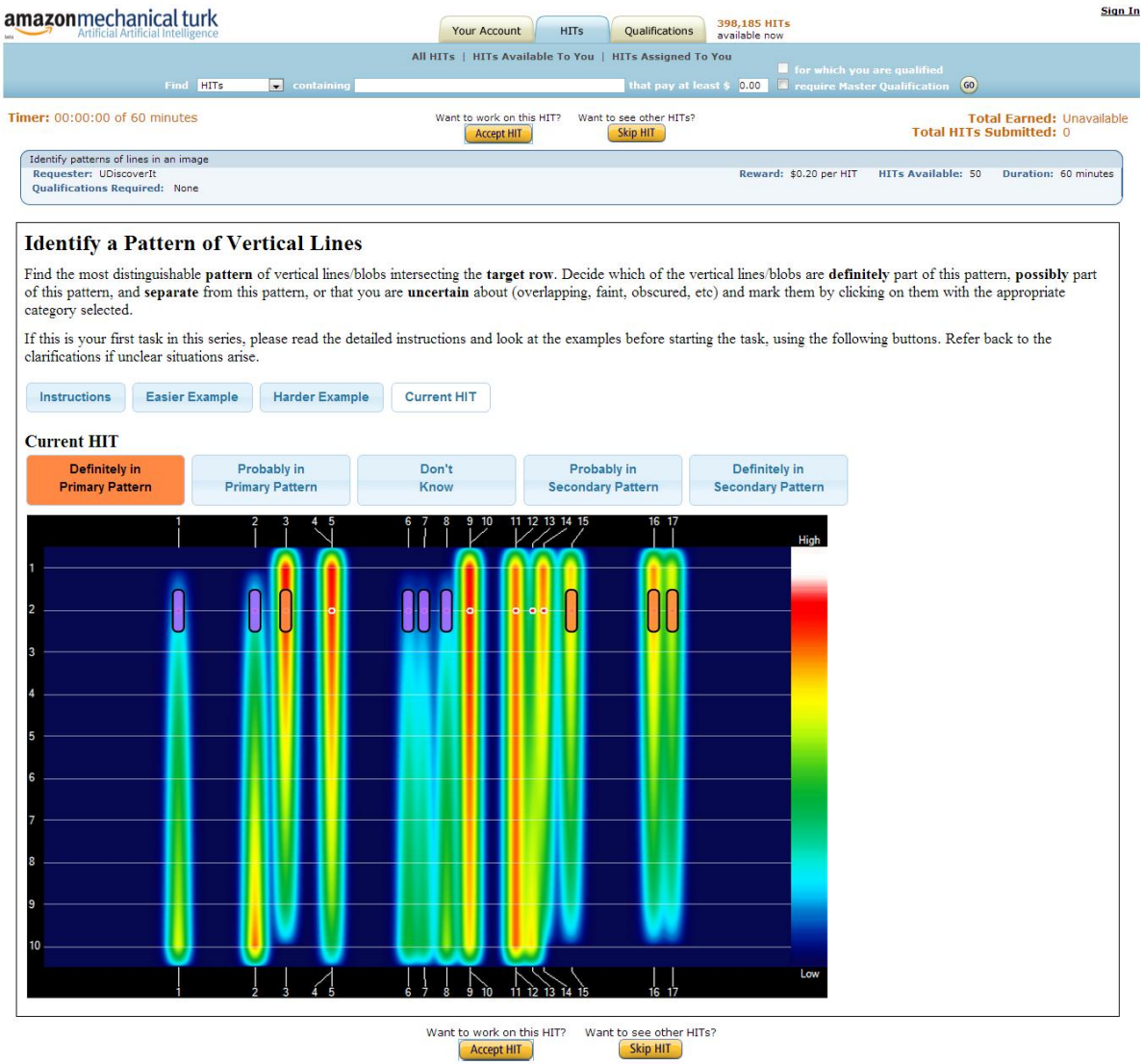
- With **limited effort** and **input**, a user can provide insightful information about the structure of the problem, and dramatically **speed up** the performance of the SMT solver.

- The approach leverages the **complementary strength** of **human input**, providing **global insights** about problem structure, and the power of **combinatorial solvers** to exploit **complex local constraints**.

Future Work

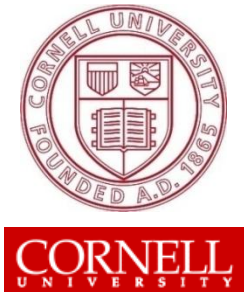
- Aggregating input from multiple users
- Guiding the search as a variable/value ordering heuristics, as opposed to pre-assignments of variables
- Providing the user with explanation and feedback about inconsistencies
- Adapting the GUI to a more ‘FoldIt’ spirit
- Adapting this method to Mechanical Turk

Future Work


amazonmechanical turk
Artificial Artificial Intelligence
Sign In
Your Account
HITs
Qualifications
398,185 HITs available now
All HITs | HITs Available To You | HITs Assigned To You
Find HITs containing that pay at least $ 0.00 for which you are qualified require Master Qualification GO
Timer: 00:00:00 of 60 minutes
Want to work on this HIT? Accept HIT
Want to see other HITs? Skip HIT
Total Earned: Unavailable
Total HITs Submitted: 0
Identify patterns of lines in an image
Requester: UDiscoverIt
Qualifications Required: None
Reward: $0.20 per HIT
HITs Available: 50
Duration: 60 minutes
Identify a Pattern of Vertical Lines
Find the most distinguishable pattern of vertical lines/blobs intersecting the target row. Decide which of the vertical lines/blobs are definitely part of this pattern, possibly part of this pattern, and separate from this pattern, or that you are uncertain about (overlapping, faint, obscured, etc) and mark them by clicking on them with the appropriate category selected.
If this is your first task in this series, please read the detailed instructions and look at the examples before starting the task, using the following buttons. Refer back to the clarifications if unclear situations arise.
Instructions
Easier Example
Harder Example
Current HIT
Current HIT
Definitely in Primary Pattern
Probably in Primary Pattern
Don't Know
Probably in Secondary Pattern
Definitely in Secondary Pattern
High
Low
Want to work on this HIT? Accept HIT
Want to see other HITs? Skip HIT


IISI
Intelligent Information
Systems Institute


Institute for Computational Sustainability
ICS


CORNELL UNIVERSITY
FOUNDED A.D. 1865
CORNELL
UNIVERSITY

THANK YOU!

Extra slides

Outline

- Motivation
- Problem Definition (Part I)
- Prior Work: Non-negative Matrix Factorization
- Problem Definition (Part II)
- Our Work: Satisfiability Modulo Theories Approach
- Conclusion and Future work

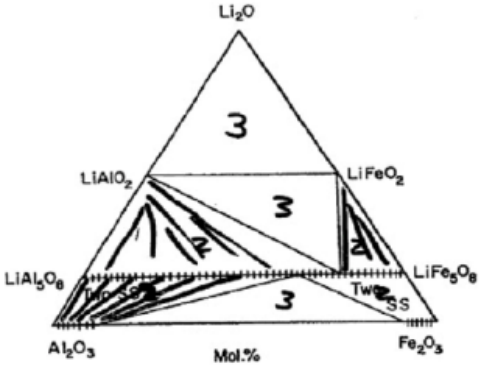
Guidelines

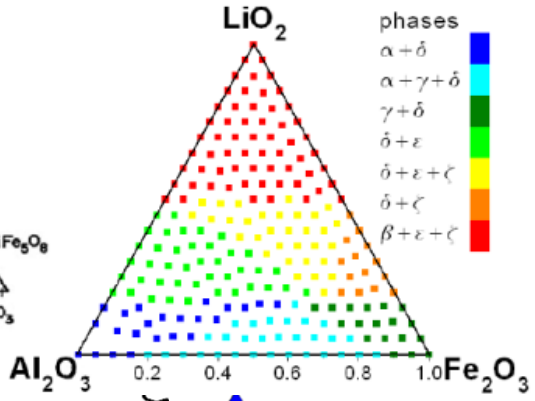
Non-negative Matrix Factorization [Long et al., 2009]

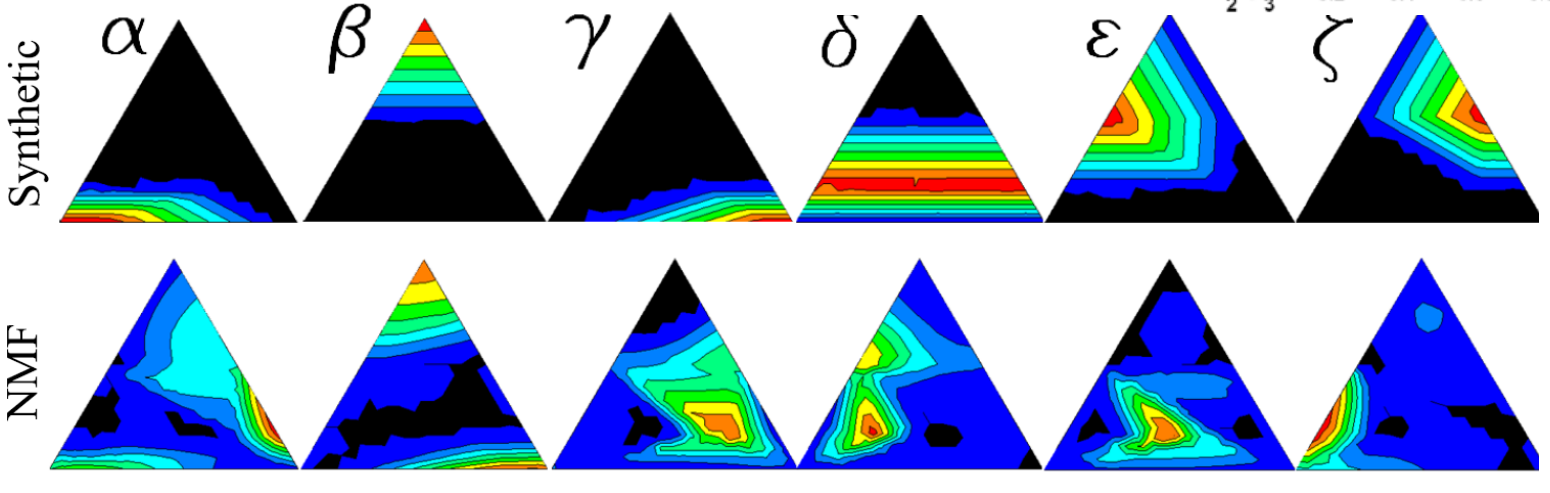
Advantages: scales up very well, accurately solves simple systems

Drawbacks: overlooks critical physical behavior, making the results physically meaningless for more complex svstems.

Illustration on synthetic instances from the *Al-Li-Fe* ternary system





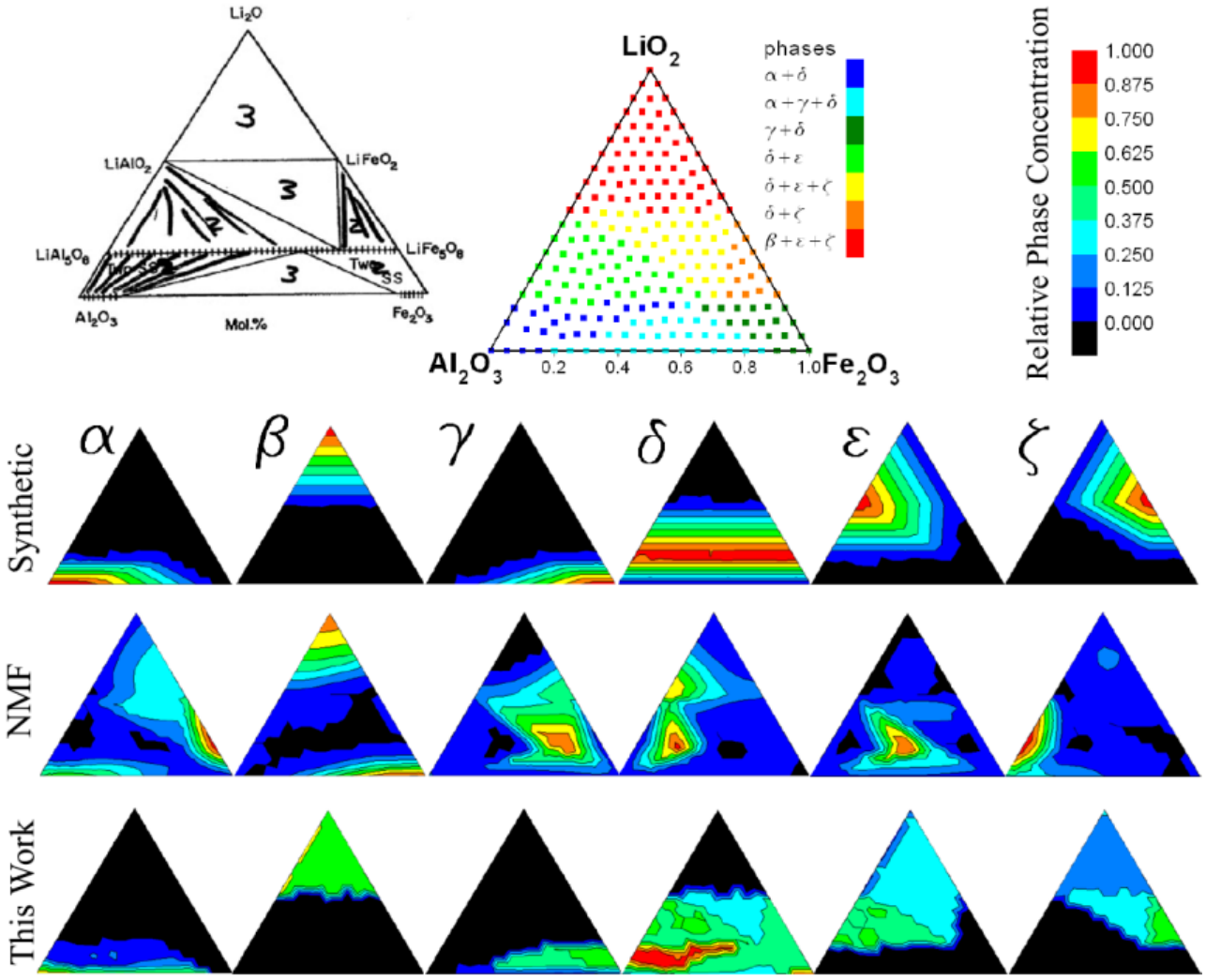


Outline

- Motivation
- Problem Definition (Part I)
- Prior Work: Non-negative Matrix Factorization
- Problem Definition (Part II)
- Our Work: Satisfiability Modulo Theories Approach
- Conclusion and Future work

Experimental Results

- Illustration on Al-Li-Fe system



Runtime

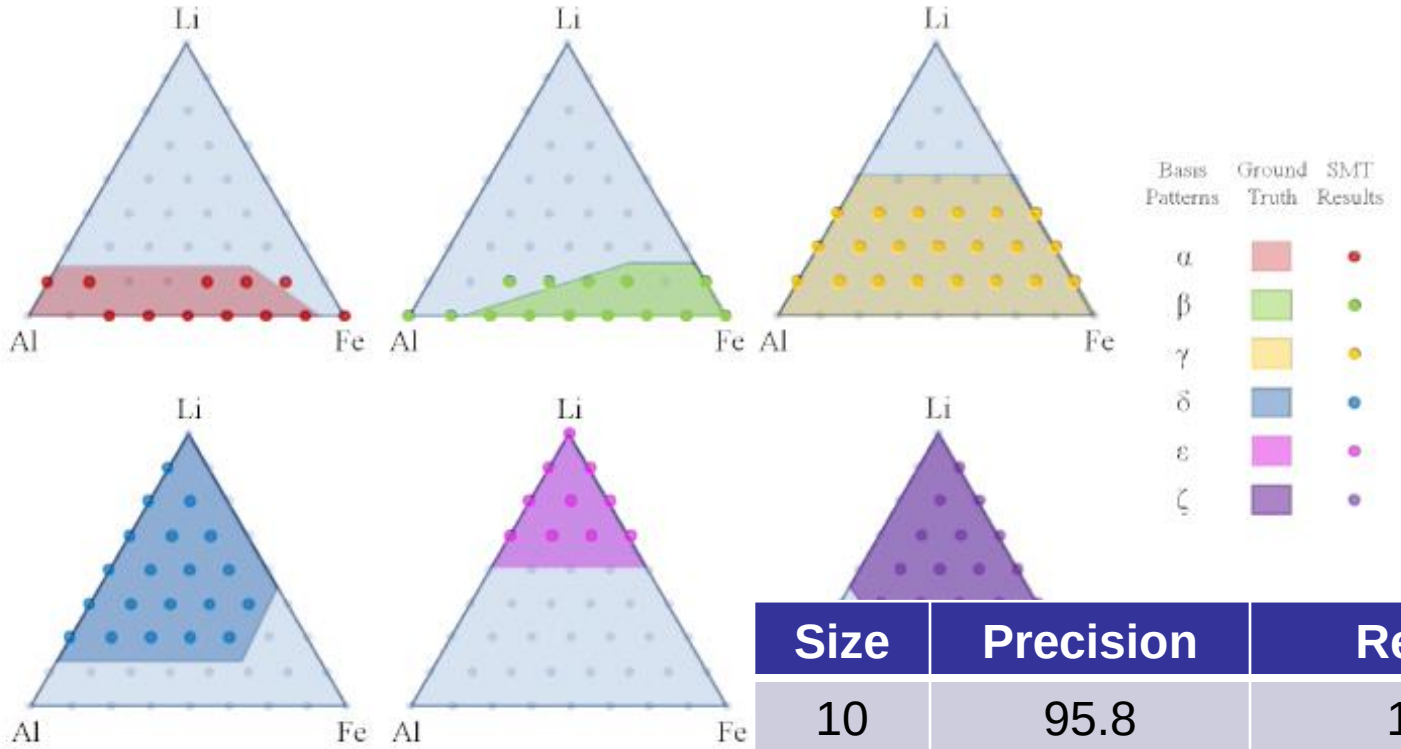
# Points	Unknown Phases	Arithmetic + Z3 (s)	Set-based + CPLEX (s)
10	3	8	**0.5**
	6	**12**	Timeout
15	3	13	**0.5**
	6	**20**	Timeout
18	3	**29**	384.8
	6	**125**	Timeout
29	3	**78**	276
	6	**186**	Timeout
45	6	**518**	Timeout

Z3 scales to realistic sized problems!

Arithmetic encoding translated to CP and MIP:

- MIP is appealing because it can optimize the objective
- They don't scale → SMT solving strategy

Precision/Recall



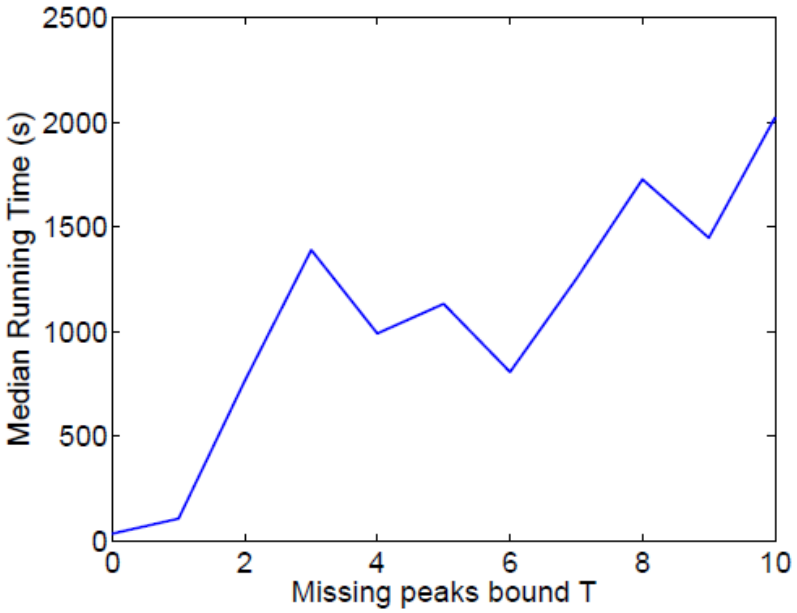
Recovers ground truth

Size	Precision	Recall
10	95.8	100
15	96.6	100
18	97.2	96.6
29	96.1	92.8
45	95.8	91.6

Robustness

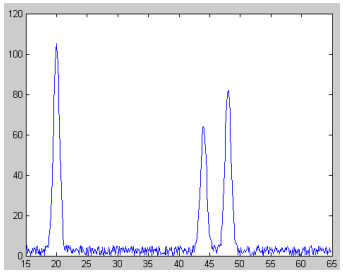
- Remove some peaks to simulate experimental noise
- Size = 15 points

Missing Peaks	Precision	Recall
1	96.1	99.6
2	96.3	99.3
3	96.7	99.5
4	95.3	98.9
5	94.8	99.7



Solutions are still **accurate**. Runtime increases approx **linearly**.

Previous Work 1: Cluster Analysis [Long et al., 2007]

$x_i =$ 

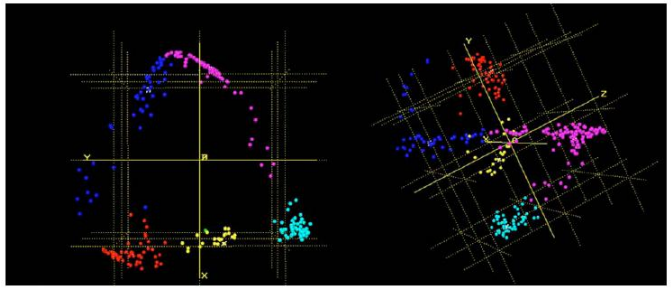
$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{1/2}}$$

$$D=(1-C)/2$$

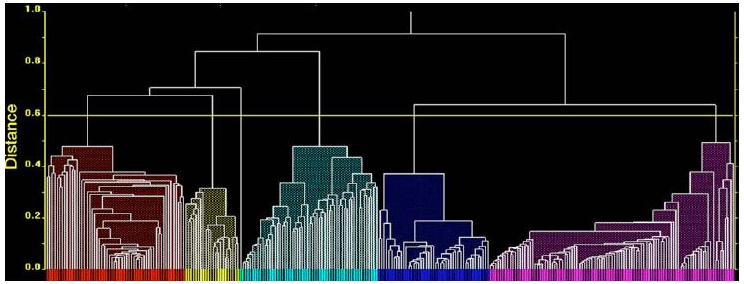
Feature vector

Pearson correlation coefficients

Distance matrix



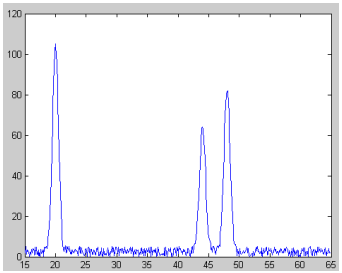
PCA – 3 dimensional approx



Hierarchical Agglomerative Clustering

Drawback: Requires sampling of pure phases, detects phase regions (not phases), overlooks peak shifts, may violate physical constraints (phase continuity, etc.).

Previous Work 2: NMF [Long et al., 2009]

$x_i =$ 

$\longrightarrow$ $X = A.S + E$ $\longrightarrow$ $Min \left\| E \right\|$

Feature vector

Linear positive combination (A) of basis patterns (S)

Minimizing squared Frobenius norm

Drawback: Overlooks peak shifts (linear combination only), may violate physical constraints (phase continuity, etc.).

SMT formulation

- **Parameters**
  - Number of pure phases K, tolerance ε
- Key components
  - Variables peak positions per base
  - Shifts per point
  - Point p is explained by base k

SMT formulation

- New arithmetic-based encoding:
  - Real variables e_{ij} for the peak locations in each B_i
  - Real variables for the shift coefficients s_{ij}
    (per base, per point)
  - An observed peak p is explained if $|p\text{-}s_{ij} - e_{ij}| \leq \varepsilon$
    (Match the height of the peaks)
  - Bound the number of missing peaks $\leq T$