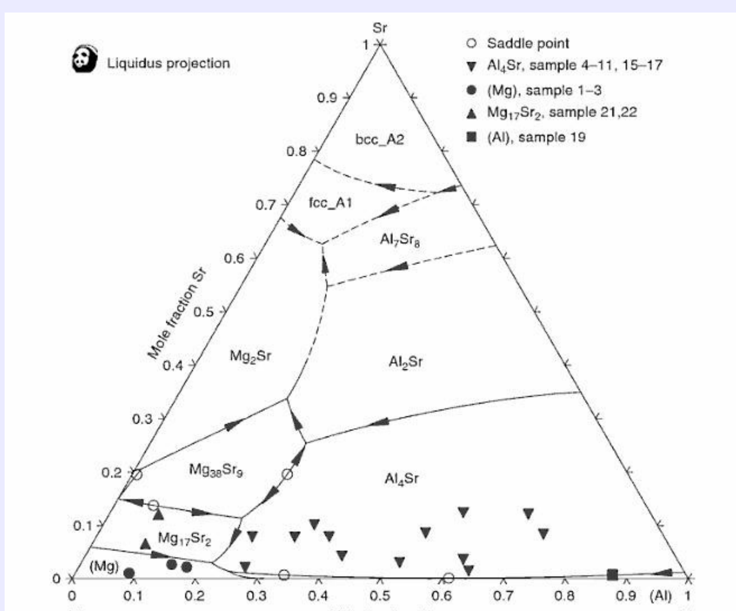
Computational Thinking for Material Discovery: Bridging Constraint Reasoning and Learning

Ronan LeBras, John M. Gregoire, Ashish Sabharwal, Carla P. Gomes, R. Bruce van Dover

Institute for Computational Sustainability

Cornell University Center for a Sustainable Future

Motivation



[Source: Methods for phase diagram determination, Ji-Cheng Zhao, 07]

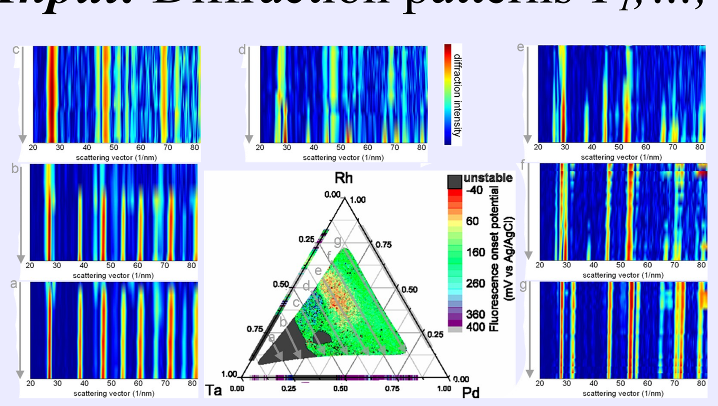
- Finding new products
- Finding product substitutes
- Understanding material properties

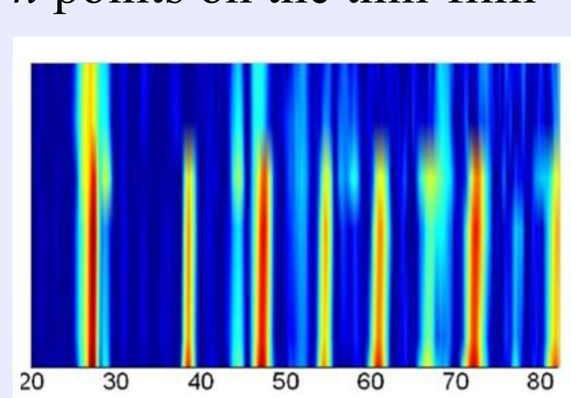
- Automating a laborious task
- Exploiting large amount of newly-available data

Problem Definition

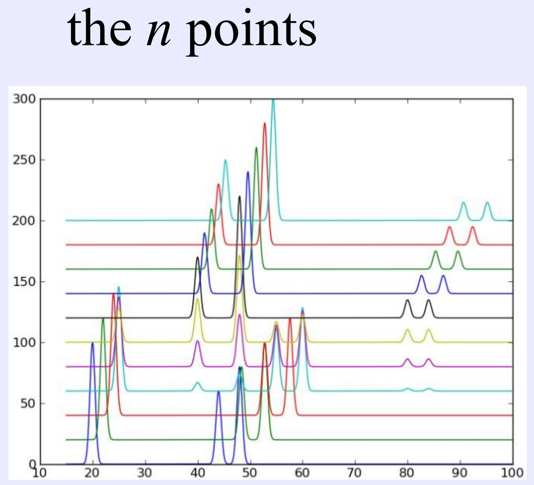
Combinatorial Method: sputtering 3 metals (or oxides) onto a silicon wafer (which produces a *thin-film*) and using x-ray diffraction to obtain structural information about crystal lattice.

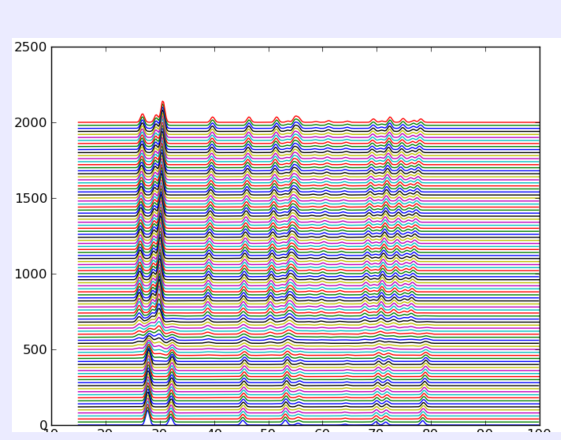
Input: Diffraction patterns $Y_1, \ldots, Y_n$ of n points on the thin-film





Output: Set of k basis patterns (or *phases*) $X_1, \ldots, X_k$.
Weights $A_1, \ldots, A_n$ and shifts $B_1, \ldots, B_n$ of these basis patterns in the n points





Problem Complexity

Theorem: This problem is NP-complete.

Proof: Reduction from the *Normal Set Basis Problem* (which is itself reduced from the *Vertex Cover Problem*)

Limitations of previous work

	[Long et al., '07]	[Long et al., '09]
Requires sampling of pure phases	x	
Detects regions but not phases	x	
Fails to ensure sparseness of A	x	x
Assumes constant peak intensity	x	x
Assumes there is no phase shift	x	x

First Approach - CP

Variables	Description	Type
p_{ki}	Normalizing peak for phase k in pattern c_i	Decision
a_{ki}	Whether phase k is present in pattern c_i	Auxiliary
q_k	Set of normalized peak locations of phase k	Auxiliary

$$a_{ki}=0 \iff p_{ki}=0 \qquad \forall\, 1\le k\le K, 1\le i\le n \tag{1}$$

$$1 \le \sum_{s=1}^{K} a_{si} \le 3 \qquad \forall\, 1\le i\le n \tag{2}$$

$$p_{ki}=j \wedge \sum_{s=1}^{K} a_{si}=1 \rightarrow q_k \subseteq r_{ij} \qquad \forall\, 1\le k\le K, 1\le i\le n, 1\le j\le |c_i| \tag{3}$$

$$p_{ki}=j \wedge \sum_{s=1}^{K} a_{si}=1 \rightarrow r_{ij} \subseteq q_k \qquad \forall\, 1\le k\le K, 1\le i\le n, 1\le j\le |c_i| \tag{4}$$

$$P(k,k',i,j,j') \rightarrow \left\{ \begin{array}{l} member(r_{ij}[j''], q_k) \\ \vee \\ member(r_{ij'}[j''], q_{k'}) \end{array} \right. \qquad \forall\, 1\le k< k'\le K, 1\le i\le n, 1\le j,j',j''\le |c_i| \tag{5}$$

where $P(k,k',i,j,j')$ is the proposition: $p_{ki}=j \wedge p_{k'i}=j' \wedge \sum_{s=1}^{K} a_{si}=2$.

$$p_{ki}=j \rightarrow p_{ki'}\neq j' \qquad \forall\, 1\le k\le K, (i,j,i',j')\in \Phi \tag{6}$$

$$phaseConnectivity(\{a_{ki} | 1\le i\le n\}) \qquad \forall\, 1\le k\le K \tag{7}$$

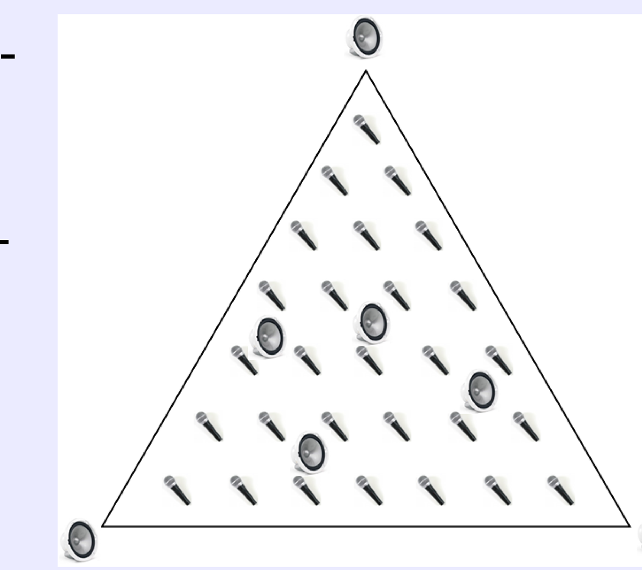
Advantage: Captures physical properties and relies on peak location rather than height.
Drawback: Does not scale to realistic instances; poor propagation if experimental noise.

Second Approach - Learning

[Analogy to the *Cocktail Party Problem*]

Every point [microphone] receives signals from various phases [speakers].

[ICA] We solve $Y=ABX$ using stochastic gradient ascent, and assuming signals to be independent.

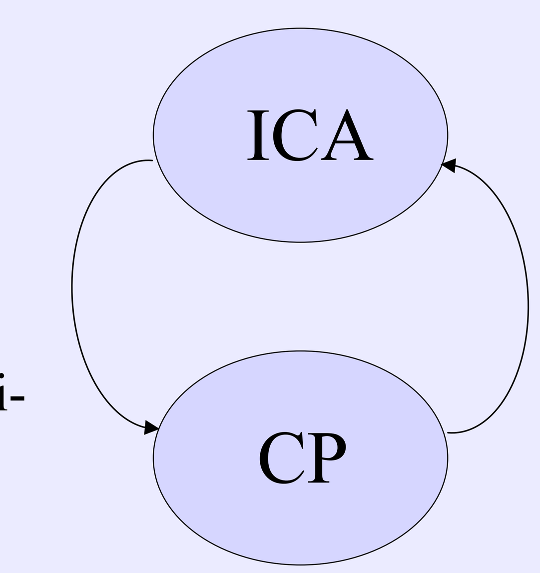


Advantage: Efficient.
Drawback: Assumes constant peak intensity.

Third Approach - Bridging I & II

We merge ICA and CP, alternating between the two approaches. One provides input for the other.

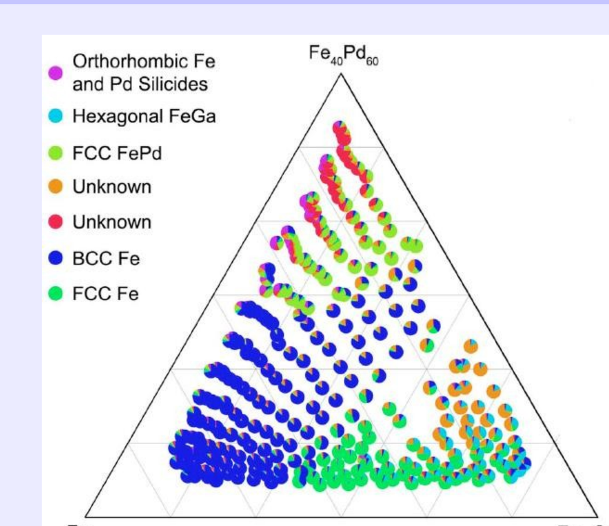
Generates basis patterns & minimizes experimental noise s.t. physical properties



Feedback loop sets ICA parameters w.r.t. the most violated constraint

Advantage: Basis patterns enhance propagation in CP model. Feedback loop tends to overcome limitations of previous work.
Drawback: No longer an exact method.

Experiments & Evaluation



[Source: Rapid identification of structural phases in combinatorial thin-film lib. using x-ray diff. and non-neg. matrix factorization, Long C.J. et al., 09]

- 2 sets of instances: real and synthetic data.
- 2 evaluations: numerical score (based on KL-divergence) and blind tests where physicists assess the quality of the results.