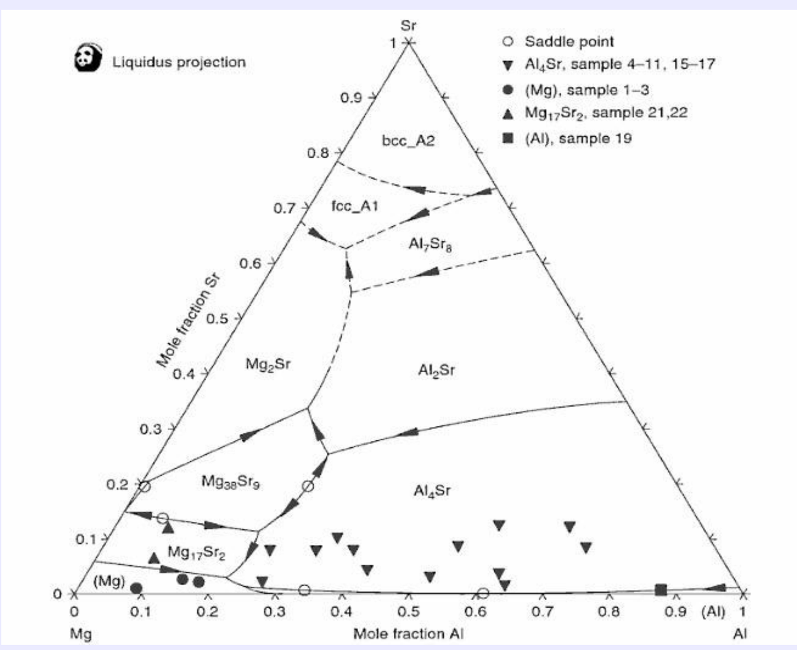
Materials Discovery: New Opportunities at the Intersection of Constraint Reasoning and Learning

Ronan Le Bras, Stefano Ermon, Theodoros Damoulas,
Rich Bernstein, Carla P. Gomes, Bart Selman, R. Bruce van Dover

Motivation



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

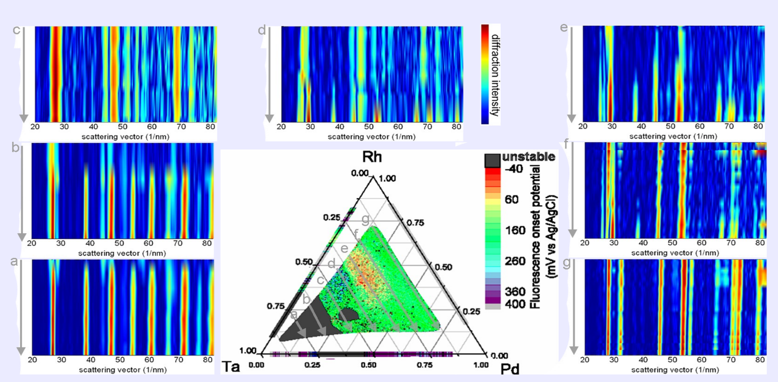
- Finding new products
- Finding product substitutes
- Understanding material properties (such as catalysts for fuel cell technologies)

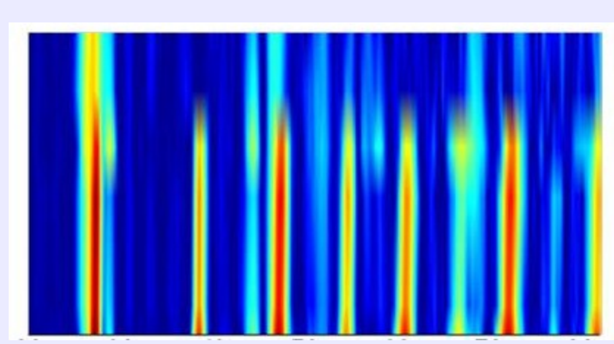
- Automating a laborious task
- Exploiting large amount of newly-available data

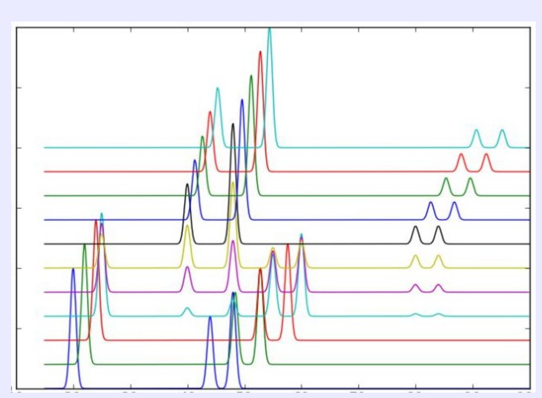
Problem Definition

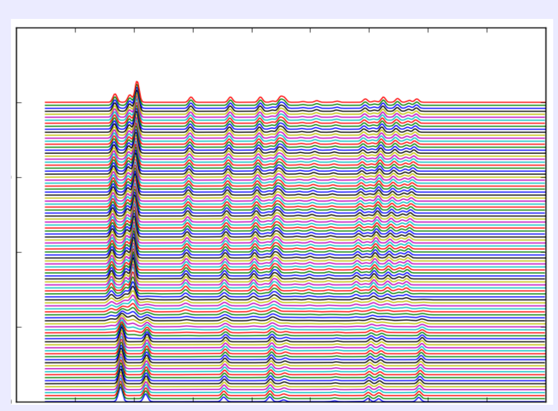
Combinatorial Method: sputtering 3 metals (or oxides) onto a silicon wafer (which produces a *thin-film*) and using x-ray diffraction to obtain structural information about crystal lattice.

Input: Diffraction patterns $D_1, \ldots, D_n$ of n points on the thin-film.









Output: Set of K basis patterns (or *phases*) $B_1, \ldots, B_K$ (along with weights a_{ij} and shifts s_{ij} of basis pattern j in point i).

Physical Constraints

Each diffraction point D_i is explained by the basis patterns:

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

There is experimental noise:

$$\min \| D_i - a_{i1} B_1 + \ldots + a_{iK} B_K \|$$

Non-negative basis patterns and coefficients:

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

At most M non-zero coefficients per point:

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

Basis patterns appear in contiguous locations on the silicon wafer:

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

Basis patterns can be shifted: (Shift operator, Shift coefficients)

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

Shifts coefficients are bounded, continuous and monotonic:

$$s_{11} \leq s_{12} \leq s_{13} \leq s_{14}$$

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

Satisfiability Modulo Theories Approach

Real variables e_{ij} for the **peak locations** in each B_i

Real variables for the shift coefficients s_{ij}

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

Every observed peak must be "*explained*"

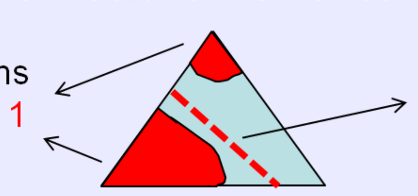
Bound the number of missing peaks $\leq T$

Minimization by (binary) search on T

Linear phase usage constraint (up to M basis patterns per point)

Linear constraint for shift monotonicity and continuity ($s_{ij} \leq s_{lm}$)

Lazy connectivity: add a cut if current solution is not connected



Symmetry breaking: Renaming of pure phases.

Ordering the peaks location e_{ij} (per basis pattern)

Runtime Analysis

# 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!

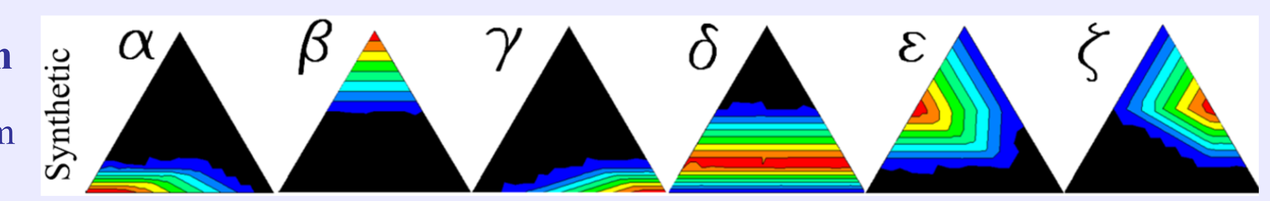
Arithmetic encoding translated to CP and MIP:

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

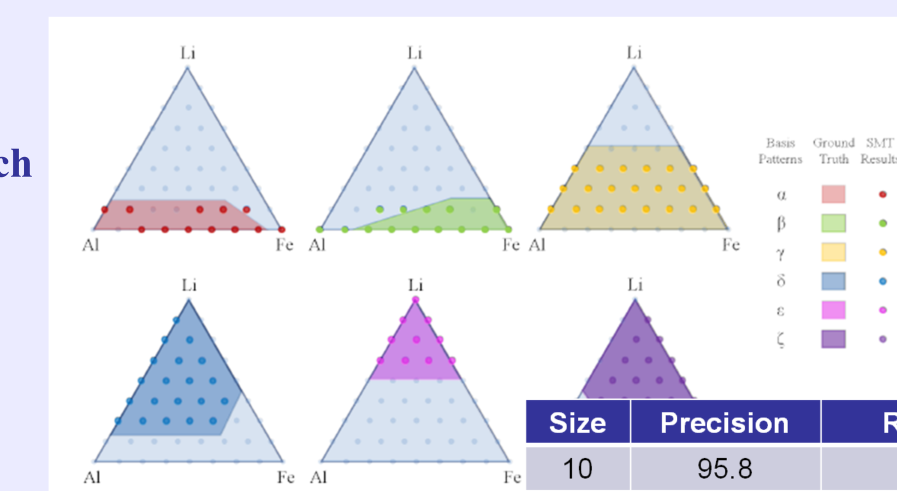
Experimental Results

Ground truth

Al-Li-Fe system



SMT approach



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

New **arithmetic-based** encoding for Materials Discovery

Good performance on synthetic data, exciting results analyzing **real-world data.**

New application domain for the area of Satisfiability Modulo Theories.

Acknowledgments

The authors gratefully acknowledge the support of the **National Science Foundation**, award number 0832782.