Computational Thinking for Material Discovery: Bridging Constraint Reasoning and Learning*

Ronan LeBras[1] Theodoros Damoulas[1] John M. Gregoire[2]
Ashish Sabharwal[1] Carla Gomes[1] R. Bruce van Dover[2]

[1] Department of Computer Science, Cornell University, Ithaca NY 14853, USA

[2] Department of Materials Science and Engr., Cornell University, Ithaca NY 14853, USA

In material sciences, a combinatorial method for discovering new materials consists in sputtering three metals (or oxides) onto a silicon wafer, resulting in what we call a thin film. The goal is to identify structural regions in thin films, which might lead to material discoveries and to a better understanding of material properties. This is an important direction in the emerging field of *computational sustainability* [4], and aims to achieve the best possible use of our available material resources.

A key incentive directly comes from the industry, as the discovery of a new material might lead to a cheaper substitute to a widely-used (or over-used) material. A cheaper substitute might translate to footprint reduction if its extraction is easier than the original product, especially if the latter gets scarce. Moreover, knowing the crystallographic structure is central to understanding the underlying mechanism of interesting material properties, such as catalyst activity for fuel cells. For example, a recent study of a platinum-tantalum library revealed an important correlation between crystallographic phase and improved catalytic activity for fuel cell applications [6]. Finally, new electromagnetic radiation tools have led to an extensive physical characterization of thin films, and consequently to an immense library of data. However, the analysis of this data remains a laborious manual task.

Any location on a thin film corresponds to a crystal with a particular composition of the three sputtered metals (or oxides); see left pane of Figure 1. The structural information of this crystal lattice is usually characterized by its x-ray diffraction pattern—a continuous waveform obtained by electromagnetic radiation. The diffraction pattern represents the intensity of the electromagnetic waves as a function of the incidence angle of radiation. Typically, a diffraction pattern of a thin film is in fact a combination of a total of about half a dozen of basis patterns (or *phases*). In other words, a thin film involves a small number of basis crystallographic phases, and every crystal corresponds either to one of these structures (i.e., a *pure* phase) or to a combination of them (i.e., a *mixture* of phases).

Given the diffraction patterns for a sample of a few dozen to a few hundred locations, the problem is to compute the most likely phase map, i.e., the set of basis patterns that are involved at any sampled location of the thin film and in which proportion. A subproblem of this is to only cluster the sample locations such that points in a cluster can be explained using the same set of basis patterns.

Previous Work and Challenges

In 2007, Long et al. [9] suggested a *hierarchical agglomerative clustering* (HAC) approach which aims to solve the aforementioned clustering subproblem. In a follow-up paper, Long et al. [8] applied *non-negative matrix factorization*, which approximates (through gradient descent) the observed diffraction patterns with a linear combination of positive basis patterns. The main limitation of both approaches lies in the assumption that peaks of a phase will always appear at the same position in any spectrum. Physical crystallographic property, however, strongly corroborates the presence of *shifts* of peaks within a phase as we move from one sample point to another. These two approaches also assume that the diffraction intensities vary uniformly over the patterns, which often does not hold.

Our goal is to take the actual physics behind the crystallographic process (e.g., the nature of shifts in the patterns) into account in order to design robust algorithms for solving this problem in the presence of

*Supported by NSF Expeditions in Computing award on Computational Sustainability (Grant 0832782).

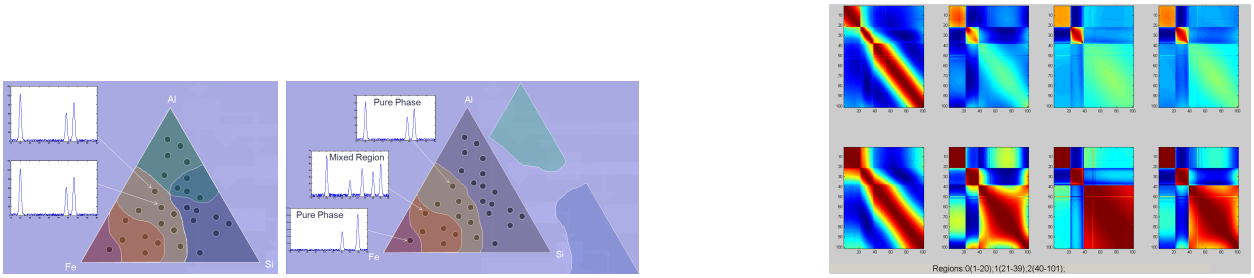

Figure 1: Left: Pictorial depiction of the problem. Right: Kernel-based similarity detection.

experimental noise.

Different Approaches

This section presents the approaches that we successively considered to solve this problem. The first two approaches strongly rely on the following observation: every peak in intensity indicates a preferable orientation of the crystal structure, and therefore *peak locations* capture the underlying parameters of the crystals. As a result, it is more the actual location of a peak that characterizes a crystal than the absolute intensity of the peak. Thereafter, we rely on well-studied peak detection algorithms [3, 10] to first determine the peak locations.

Greedy Approach. Instead of considering the entire diffraction pattern, we discretize any spectrum into a list of detected peaks. The problem is no longer to match the spectra but instead to match the (relatively few) peaks—while taking shifts explicitly into account. We first propose a greedy approach that proceeds as follows. Starting with the pattern with the fewest peaks, it extends this pattern to neighboring points that share the same possibly-shifted pattern. Thus, it creates a pure-phase connected region. It then repeats this step with the remaining pattern with the fewest peaks. Every time a pure-phase region is identified, the algorithm tries to discover any mixture region that might lie between two pure-phase regions.

Constraint Programming Formulation. The greedy algorithm, unfortunately, breaks as soon as a pure-phase is not sampled at all. In fact, we prove that this discretized version is actually $\mathcal{NP}$-complete, using a reduction from the *Normal Set Basis Problem* [7] (which is itself reduced from the *Vertex Cover Problem* [2]). We then propose a constraint satisfaction formulation, with the desired number of phases as a parameter. The key of this representation is that it relies on a single normalizing peak for every phase, and computes possible shifts relative to this peak. We show that this information is enough to maintain connectivity within a region, and to bound the number of phases involved in any point as well as the shifts. The other peaks are only used to match the set of peaks of any pattern. The main advantage of this approach is its ability to capture the underlying physical properties that characterize the behavior of the crystallographic phases.

Kernel Methods from Machine Learning. The pure CP approach however does not perform very well when there is noise in the data, for example in the form of missing peaks or measurement errors in the location of peaks. In order to tackle this issue, we turn to methods from machine learning, specifically for designing similarity matrices using kernels. The right pane of Figure 1 shows a sample "heat map" of similarities for a problem instance with two pure phases and one mixed phases in-between. Red denotes high similarity (usually along the diagonal) while blue shows dis-similarity. Moving rightwards, we obtain better identification of the 3 regions by taking shifts and normalization into account. The bottom row shows the "kernelized" version of the top row, depicting further improvement in the identification of the three regions. A traditional clustering such as *k-means* then exploits this similarity matrix and attempts to cluster points that belong to the same phase region. The main advantage of this approach is that it provides a data-driven rough global picture of the problem and tends to incorporate complex dependencies of the data. However, this approach might also miss critical details (e.g. connectivity or the fact that small peaks can be in fact discriminating).

Integrating CP and ML: a New Methodology

The motivation that guides us into a fusion of CP and ML stems from the aforementioned distinct strengths of these separate streams of research. A very specialized and detailed view of the problem (CP) can be

guided by a data-driven statistical view (ML). This is especially attractive and suitable for the material discovery scenario that effectively has a two-fold goal: *Identify* generic phase regions and *reconstruct* the component phases. The intuition behind our methodology is depicted in Figure 2. The clustering approach generates clusters of points, where each cluster defines a sub-problem that the CP model attempts to solve. Each of these sub-problems is significantly smaller than the original problem both in terms of the number of points and the number of phases that compose these points. After solving all sub-problems, the resulting phases are merged and any replicate one is removed, leading us to the global solution.

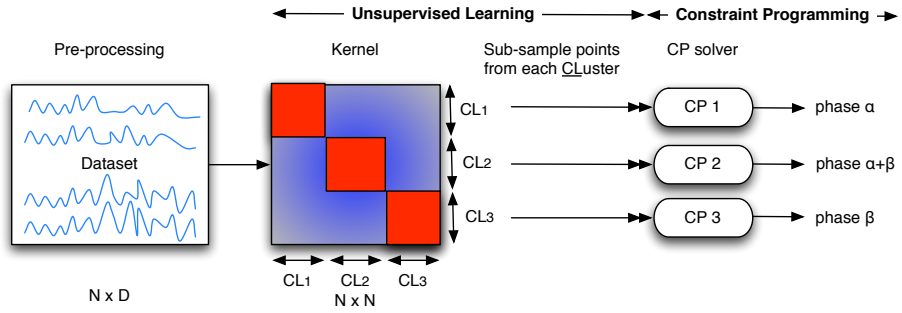

Figure 2: Schematic representation of the integration of CP and ML.

Empirical Validation

For the investigation of the algorithm, synthetic xray diffraction data was generated for the Al_2O_3-Li_2O-Fe_2O_3 phase diagram using diffraction patterns from the JCPDS database [1] with parameter reflecting those of a recently developed combinatorial crystallography technique [5]. This sample experiment indicates that the NMF algorithm fails to capture the underlying phases and tends to combine similar but distinct (and disconnected) phases. On the other hand, our algorithm produces phase concentration maps that are connected in composition space and match closely with those of the synthetic phase maps.

References

[1] *Powder Diffract. File, JCPDS Internat. Centre Diffract. Data, PA*, 2004.

[2] H. Björklund and W. Martens. The tractability frontier for NFA minimization. In *ICALP '08*, pp. 27–38, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-70582-6.

[3] P. Du, W. A. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, 2006. ISSN 1367-4803.

[4] C. P. Gomes. Computational Sustainability: Computational methods for a sustainable environment, economy, and society. *The Bridge, National Academy of Engineering*, 39(4), Winter 2009.

[5] J. M. Gregoire, D. Dale, A. Kazimirov, F. J. DiSalvo, and R. B. van Dover. High energy x-ray diffraction/x-ray fluorescence spectroscopy for high-throughput analysis of composition spread thin films. *Rev. Sci. Instrum.*, 80 (12):123905, 2009.

[6] J. M. Gregoire, M. E. Tague, S. Cahen, S. Khan, H. D. Abruna, F. J. DiSalvo, and R. B. van Dover. Improved fuel cell oxidation catalysis in pt1-xtax. *Chem. Mater.*, 22(3):1080, 2010.

[7] L. T. Kou and C. K. Wong. A note on the set basis problem related to the compaction of character sets. *Commun. ACM*, 18(11):656–657, 1975. ISSN 0001-0782.

[8] C. J. Long, D. Bunker, V. L. Karen, X. Li, and I. Takeuchi. Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. *Rev. Sci. Instruments*, 80(103902), 2009.

[9] C. J. Long, J. Hattrick-Simpers, M. Murakami, R. C. Srivastava, I. Takeuchi, V. L. Karen, and X. Li. Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. *Rev. Sci. Instruments*, 78(072217), 2007.

[10] L. Smrč Ok, M. Durík, and V. Jorík. Wavelet denoising of powder diffraction patterns. *Powder Diffraction*, 14: 300–304, Dec. 1999.