A Computational Challenge Problem in Materials Discovery: Synthetic Problem Generator and Real-World Datasets

Ronan Le Bras Richard Bernstein
Computer Science Dept.
Cornell University, Ithaca NY

John M. Gregoire Santosh K. Suram
Joint Center for Artificial Photosynthesis
California Inst. of Technology, Pasadena CA

Carla P. Gomes Bart Selman
Computer Science Dept.
Cornell University, Ithaca NY

R. Bruce van Dover
Materials Science and Engineering Dept.
Cornell University, Ithaca NY

Abstract

Newly-discovered materials have been central to recent technological advances. They have contributed significantly to breakthroughs in electronics, renewable energy and green buildings, and overall, have promoted the advancement of global human welfare. Yet, only a fraction of all possible materials have been explored. Accelerating the pace of discovery of materials would foster technological innovations, and would potentially address pressing issues in sustainability, such as energy production or consumption.

The bottleneck of this discovery cycle lies, however, in the analysis of the materials data. As materials scientists have recently devised techniques to efficiently create thousands of materials and experimentalists have developed new methods and tools to characterize these materials, the limiting factor has become the data analysis itself. Hence, the goal of this paper is to stimulate the development of new computational techniques for the analysis of materials data, by bringing together the complimentary expertise of materials scientists and computer scientists.

In collaboration with two major research laboratories in materials science, we provide the first publicly available dataset for the phase map identification problem. In addition, we provide a parameterized synthetic data generator to assess the quality of proposed approaches, as well as tools for data visualization and solution evaluation.

Introduction

The discovery of new advanced materials has made possible recent technological inventions, from silicon circuits and batteries to solar and fuel cells. As underlined by the Materials Genome Initiative (Patel 2011; White 2012), accelerating the discovery and deployment cycle of new advanced materials is essential to improving human welfare and to achieving sustainable, clean energy.

While most inter-metallic compounds or oxides involving up to two elements have now been studied, it only represents a very small fraction of all possible materials that can be obtained by mixing three elements or more, and by varying the synthesis conditions (e.g. temperature and pressure). Out of billions of candidate materials, there are potentially thousands of materials with interesting physical properties, such as conductivity, light absorbency or catalytic properties, that remain to be uncovered.

The field of materials discovery requires a deep understanding of the underlying crystallographic process that governs a material formation. Experts in materials science have developed elaborate deposition processes to efficiently create so-called composition libraries in a high-throughput regime (Takeuchi, Dover, and Koinuma 2002; Gregoire et al. 2007, e.g.). For example, the High-Throughput Experimentation research program at the Joint Center for Artificial Photosynthesis (JCAP) is capable, at full capacity, of outputting about one million materials a day. Once synthesized, the promising composition libraries need to be characterized using X-ray diffraction and X-ray fluorescence (Chu et al. 2004; Vogt et al. 2004; Gregoire et al. 2009) in order to map the composition and structure of a library. This data collection step requires an X-ray source such as a synchrotron, where the experimentation cost amounts to about $1M for a week of data collection.

Nevertheless, as the data collection reaches a high-throughput regime, the bottleneck of the discovery cycle becomes the data interpretation itself. Indeed, this task remains a laborious manual inspection that relies on materials scientists expertise. The goal of this paper is to encourage computer scientists to propose new computational methods for this data interpretation, by providing synthetic and real data, as well as a problem description and evaluation metrics that transcend materials science research.

This work is in collaboration with two major research centers in materials science. The High-Throughput Experimentation research program at JCAP focuses on automated,

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

high-throughput discovery of materials that can act as light absorbers or catalysts for solar-fuel generation, in collaboration with the Stanford Synchrotron Radiation Lightsource. The Energy Materials Center at Cornell (emc2) aims at improving energy conversion and storage by understanding and exploiting fundamental properties of materials, and conducts experiments at the Cornell High Energy Synchrotron Source (CHESS).

There are several existing sources of tabulated X-ray diffraction data from crystallographic studies of materials. Both the National Institute of Standards and Technology (Bergerhoff and Brown 2002) and the International Centre for Diffraction Data offer libraries (pdf 2004) characterizing hundreds of thousands of individual inorganic crystalline compounds, including X-ray patterns. These libraries can be useful for matching experimental data to previously measured compounds, however they are not suitable for developing methods to analyze composition spreads involving mixtures of compounds and solid solutions. The Materials Project (Jain et al. 2013) and the aflowlib.org repository (Curtarolo et al. 2013) also provide data characterizing inorganic crystalline materials, as well as phase map information, derived using ab-initio and other simulation methods. However, these simulate only low-temperature synthesis, and have limited capabilities to describe solid solutions.

The paper is structured as follows. The next section presents the phase map identification problem, while Section 3 describes the data format of the datasets. In Section 4, we present the data visualization user interface as well as the solution evaluation tool. Sections 5 and 6 present the synthetic and real world datasets, respectively, while conclusions and comments are given in the last section.

Problem Definition

The goal in the phase map identification problem is to produce a model for the crystal structures that form under equilibrium conditions, as a function of material composition. The model should correspond to certain constraints that describe the physical processes. The experimental generation of the data under consideration can be described as the deposition of material library and subsequent X-ray diffraction measurements.

To generate a given data set, materials scientists create an array of thin-film depositions of two or more chemical elements on one substrate (or multiple substrates), consisting of different mixtures within the composition space of interest. The resulting material library consists of hundreds to thousands of unique composition samples, and typically each sample crystallizes into a collection of millions of small crystallites. Despite the vast number of crystallites present in the library, the number of distinct crystal structures is relatively small and approximately equal to the number of elements. For the present purposes, we consider each sample in the material library as a unique composition of matter deposited at a discrete location on a planar substrate.

The diffraction of an X-ray beam by the thin-film is then measured for each sample. The diffracted X-ray intensity is recorded as a function of scattering direction (angle). Diffraction intensity peaks occur at directions (angles) determined primarily by regular spacing of electrons within individual crystallites. Therefore the angular dependence of the diffracted intensity (diffraction pattern) contains information about the crystal structures stable for a given composition.

Data preprocessing additionally consists of:

1. Removal of background signal originating from the X-ray detector, as well as scattering from the substrate, air and apparatus
2. Integration of 2-dimensional detector signal by diffraction direction
3. Further filtering and diffraction peak detection

The phase map identification problem can then be described as:

Given A set of X-ray diffraction signals (in the form of a diffraction pattern, and/or detected peak parameters) representing different material compositions; and K, the expected number of material phases present.

Find A model for K phases (basis functions that change gradually with composition, in terms of structure and intensity), and their parameters at each sampled composition.

Subject to Physical constraints from the known properties of crystals, such as:

1. Gibbs Phase Rule, which says that (assuming constant temperature and pressure), that $k_s \leq M$, where $k_s \leq K$ is the number of phases present in sample s, and M is the number of different material elements in the system
2. The compositions at which a phase is observed should be a connected region and its parameters should vary smoothly as a function of composition

This definition is intended as a guideline, and could be reasonably modified, for example to treat K as an estimated parameter, or to incorporate other knowledge from the materials science or solid state physics literature. It is important, however, that a solution model or its predictions have a reasonable physical interpretation (potentially after post-processing or reconstruction), and are consistent with the underlying physics.

An output model should be evaluated using prediction accuracy on the diffraction curve or peak parameters of held-out samples. For synthetically generated data, model parameters can be compared directly with those that follow from the generator parameters.

Challenges

Many challenges arise as data is collected and as it is processed. The first challenge to overcome relates to the noise in the data. The uncertainty in the diffracted intensities comes from multiple factors, such as measurement errors, background noise, and detector ghosting. There is also uncertainty in the composition values, arising from the analysis of the X-ray fluorescence data.

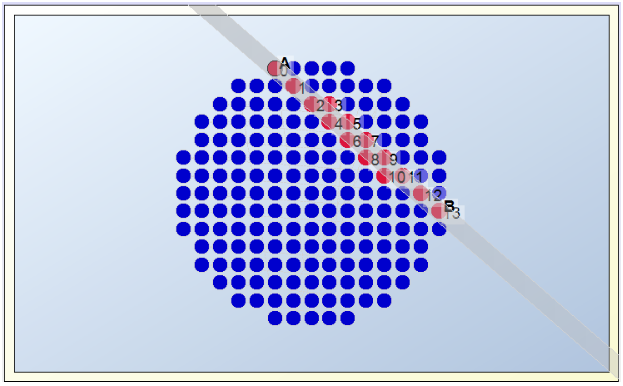

Figure 1: Map of selectable sample points in deposition-space coordinates for the instance *inst5*.

Another challenge relates to the model assumptions. For example, while thermodynamic equilibrium is usually assumed to solve this problem, it is possible that the equilibrium conditions have not been reached. Other issues might arise from imperfections in the thin-film, variable thickness of the deposited film, or preferred orientation adopted by the crystallites in the thin-film. Finally, for approaches that are based on the discretization of the x-ray patterns into peaks, the peak detection algorithm itself is not a trivial task and might lead to additional uncertainties.

One main challenge is to capture the physical rules that govern the underlying crystallographic process (Long et al. 2007; 2009; Le Bras et al. 2011; Ermon et al. 2012) while allowing the solution approach to scale up to real-sized instances (Le Bras et al. 2011; 2013; Finger et al. 2013). Finally, while structural changes in the crystals typically occur in a linear fashion in composition space, there might be non-linear changes in shifts, peak widths or relative intensities (peak heights).

Data format

Each data set includes 3 basic types of information:

1. **Metadata**, such as the chemical **elements** used in the system. This metadata is useful to retrieve the **atomic numbers** of the elements, as well as to correlate the proposed phases of a solution with powder patterns of **known phases** involving these elements.
2. **Composition** data for each experimental sample, in one of two forms:
   (a) **Deposition location**, x and y coordinates which are topologically related to the relative concentrations of the constituent elements
   (b) **Amount of each element** present in the sample (estimated), in scaled absolute units reflecting the thickness of the film as well as relative concentration
3. **Diffraction intensity** for each sample, in one or both of the following forms. In both cases, the intensity values are scaled by the total amount of X-ray exposure.
   (a) **Integrated X-ray counts** at each scattering direction in a given range. We provide the filtered curves, which represent a better overall estimate of the signal. Neverthess, we also provide the unfiltered curves, as it might be more suitable for methods sensitive to artifacts in the shapes of the curves.
   (b) **Diffraction peak** locations, heights and widths extracted using a wavelet-based peak detection algorithm (Gregoire, Dale, and van Dover 2011).

Listing 1 provides a high-level description of the data format. The metadata contains the system dimension M, the elements involved, the number of samples N, as well as experimentation setup and calibration data such as the beam wavelength λ or detector distance d. Next, the composition data provides the location of the samples on the thin-film (here in the format "$[X, Y]$"). For the diffraction intensity for each sample, the vector Q corresponds to the range of $qvalues$ for which the beam intensities are reported (i.e. x-axis of Fig. 4), and each X-ray pattern is given with respect to these values (i.e. y-axis of Fig. 4). Finally, the detected peaks are listed as triplets $(location, height, width)$.

Listing 1: Data format example for the instance *inst5*

```
// Metadata
M=3
Elements=Fe,Bi,V
N=177
Lambda=0.9185
...

// Composition data
X=-42.0,-36.0,-30.0,-24.0,-18.0,...
Y=-12.0,-12.0,-12.0,-12.0,-12.0,...
...

// Integrated counts data
Q=10.0,10.1,10.2,10.3,10.4,...
I1=222.27,167.79,163.02,99.733,...
I2=177.57,161.73,177.19,123.15,...
I3=153.22,189.00,124.71,56.34,...
...

// Diffraction peaks data
P1=[12.6,8079.6,0.12],...
P2=[12.5,3604.9,0.14],...
P3=[12.6,2767.9,0.11],...
...
```

Visualization and Evaluation Tools

We have developed a graphical user-interface application for exploring and visualizing input datasets as well as solutions to the phase map identification problem[1]. For visualizing input data, it provides:

1. A map of selectable sample points arranged by deposition or composition-space coordinates (Fig. 3)
2. An interface for viewing diffraction curves from individual sample points (Fig. 4)

[1] Available at http://www.udiscover.it

3. An interface for viewing diffraction curves from multiple sample points together at once (Fig. 5)

The application can also load solutions provided by a solver.

Listing 2 gives a description of the format of a solution. First, it provides the pattern of each one of the K phases as a vector of intensities. Next, it provides the phase concentrations, namely the fraction of the phases that are involved in each one of the N samples, and how the phase patterns should be shifted in each sample.

Listing 2: Solution format example for the instance *inst5*

```
// Number of phases
K=3

// Phase patterns (basis)
Q=10.0,10.1,10.2,10.3,10.4,...
B1=153.02,164.52,127.68,87.34,...
B2=166.70,178.54,153.51,98.02,...
...

// Phase concentrations at each sample
C1=0.00,0.72,0.85
C2=0.00,0.87,0.79
C3=0.00,0.92,0.71
...

// Phase shifts at each sample
S1=0.00,1.00,1.02
S2=0.00,1.01,1.02
S3=0.00,1.01,1.01
...
```

The solution can then be visualized with:

1. A map showing the estimated concentration (and other parameters) of each phase, arranged by deposition or composition-space coordinates (Fig. 6)
2. An interface showing the actual, estimated, component, and residual diffraction curves for each sample point (Fig. 4)

The application can calculate an evaluation measure based on absolute error, however other error models are possible, and therefore it is preferable for quantitative evaluation to design a custom measure (incorporating model complexity) to accompany each particular solution structure.

Synthetic Data Generator

In order to assess the quality of proposed approaches and validate them, we have created a synthetic data generator. This generator can create basic artificial binary systems ($M = 2$), as well as complex artificial ternary systems ($M = 3$), and a theoretical ternary Aluminum-Lithium-Iron (Al-Li-Fe) system, where the component phases (the patterns and their parameters) have been theoretically calculated.

The generator has a set of user-specified parameters that allow controlling the complexity of the generated instances, as follows:

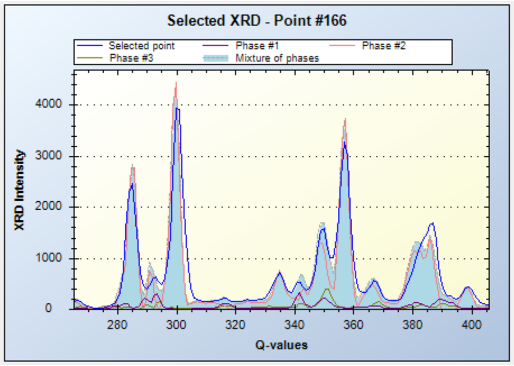

Figure 2: Diffraction curve (in blue) of a selected sample point (sample point #166) for the instance *inst5*. The other curves (purple, red, and green) correspond to the component phases involved in that point for a given solution, while the shaded-blue area represents the mixture of these phases, and approximates the diffraction curve.

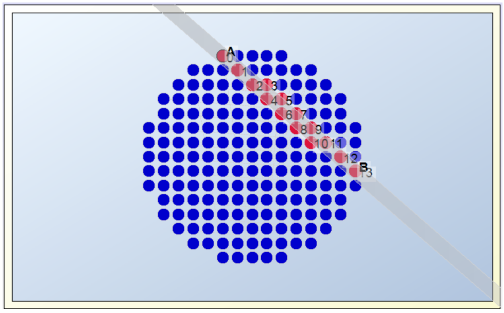

Figure 3: Map of selectable sample points in deposition-space coordinates for the instance *inst5*.

1. **Underlying system** that governs the number of phases K and their concentration on the film.
2. **Spacing** (in atomic percent) of the data points, which determines the total number N of points.
3. **Total number of peaks** of the component phases, up to the theoretically defined number of peaks.
4. **Number of diffraction angles** of the X-ray patterns, which corresponds to the x-axis precision of the patterns.
5. **Noise level** as a total number (or total amount) of removed peaks from the original constructed patterns.

Using this generator, we provide a benchmark of 100 instances, with varying complexity[2].

Real World Data

While the previous datasets are made of artificial and theoretical diffraction patterns, the datasets we present in this section has been collected empirically. Nonetheless, for standardization purposes, it follows the same format as described in the previous section. The instances of this dataset

[2] Available at http://www.udiscover.it

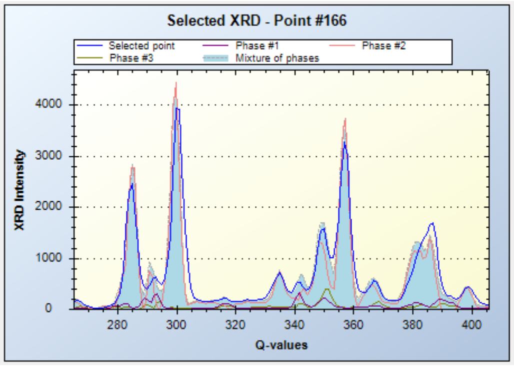

Figure 4: Diffraction curve (in blue) of a selected sample point (sample point #166) for the instance *inst5*. The other curves (purple, red, and green) correspond to the component phases involved in that point for a given solution, while the shaded-blue area represents the mixture of these phases, and approximates the diffraction curve.

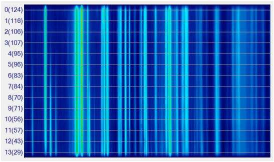

Figure 5: Heat map of the diffraction curves of the 14 selected points of Fig. 3, where bright colors mean high beam intensity. This visualization helps identify the component phases and their boundaries. The x-axis corresponds to the diffraction angles, while the y-axis refers to the samples in the selected slice.

have been acquired under various condition. Table 1 shows the dataset of real instances and their parameters. Each instance is characterized by its system dimension M, which represents the number of different elements, its design stoichiometry at the center of the film (i.e. the relative proportion of the M elements), the type of substrate it was obtained on, the total number N of sample points, and the gas used in the vacuum chamber during the deposition, if any.

While this dataset is meant to be a stable benchmark and a first building block towards a unified and standardized discovery cycle methodology, it is also meant to be expanded as more data gets collected.

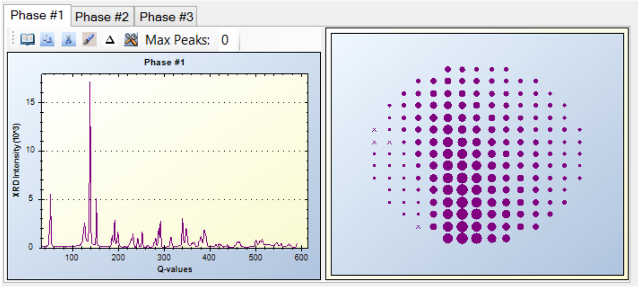

Figure 6: Diffraction pattern of phase 1 in the loaded solution (left panel) and its concentration on the film (right panel) for the instance *inst5*.

Conclusions

We provide the first publicly available dataset for the phase map identification problem. This problem is central to the discovery cycle of new materials, as it aims to provide structure and composition maps that can be correlated with interesting physical properties within an inorganic library. In addition, we provide a synthetic generator in order to evaluate the quality of proposed approaches. Finally, we propose a graphical user interface for the visualization of the data and of its solutions. We hope this paper will motivate computer scientists to propose new computational methods for the phase map identification problem.

Acknowledgments

The authors would like to thank the reviewers for their positive feedback and constructive comments. This work was supported by the National Science Foundation (NSF Expeditions in Computing award for Computational Sustainability, grant 0832782, NSF Inspire grant 1344201 and NSF Eager grant 1258330). The experiments were run on an infrastructure supported by the NSF Computing research infrastructure for Computational Sustainability grant (grant 1059284). This material is based upon work performed by the Joint Center for Artificial Photosynthesis, a DOE Energy Innovation Hub, supported through the Office of Science of the U.S. Department of Energy (Award No. DE-SC0004993). Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. Research conducted at the Cornell High Energy Synchrotron Source (CHESS) is supported by the National Science Foundation and the National Institutes of Health/National Institute of General Medical Sciences under NSF award DMR-0936384.

References

Bergerhoff, G., and Brown, I. 2002. Crystallographic databases, international union of crystallography, chester, 1987 search pubmed; a. belsky, m. hellenbrandt, vl karen and p. luksch. *Acta Crystallogr., Sect. B: Struct. Sci* 58:364.

Chu, Y. S.; Tkachuk, A.; Vogt, S.; Ilinski, P.; Walko, D. A.; Mancini, D. C.; Dufresne, E. M.; He, L.; and Tsui, F. 2004. Structural investigation of comnge combinatorial epitaxial thin films using microfocused synchrotron x-ray. *Applied surface science* 223(1):175–182.

Curtarolo, S.; Hart, G. L.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; and Levy, O. 2013. The high-throughput highway to computational materials design. *Nature materials* 12(3):191–201.

Ermon, S.; Le Bras, R.; Gomes, C. P.; Selman, B.; and van Dover, R. B. 2012. Smt-aided combinatorial materials discovery. In *Proceedings of the 15th International Conference*

Instance	M	Elements	Design Stoich.	Substrate	N	Insitu Gas
inst1	2	Bi, V	1 : 1.4	Si	51	
inst2	3	Fe, Bi, V	0.12 : 1 : 1.4	Si/SiO$_2$	185	
inst3	2	Bi, V	1 : 1.4	SiO$_2$	51	
inst4	3	Fe, Bi, V	0.1 : 1 : 1.4	Si/SiO$_2$	177	O$_2$
inst5	3	Fe, Bi, V	0.1 : 1 : 1.4	Si/SiO$_2$/Ti	177	O$_2$
inst6	2	Bi, V	1 : 1.4	Si/SiO$_2$/Ti	51	O$_2$
inst7	3	Sn, Si, Zn	1 : 1 : 2	Si/SiO$_2$/Ti	249	N$_2$
inst8	3	W, Bi, V	0.25 : 1 : 1	Si/SiO$_2$	177	
inst9	3	Cu, Bi, V	2 : 1 : 1	Si/SiO$_2$	254	
inst10	3	Ag, Bi, V	1 : 1 : 1	Si/SiO$_2$	177	
inst11	4	Ag, Cu, Bi, V	1 : 1 : 1 : 1	Si/SiO$_2$	317	
inst12	3	Ag, Bi, V	2 : 1 : 1	Si/SiO$_2$	261	
inst13	3	Mo, Bi, V	0.25 : 1 : 1	Si/SiO$_2$	177	
inst14	3	Pt, Ti, Ni	1 : 1 : 1	Si/SiO$_2$	132	
inst15	3	V, Ti, Cr	1 : 1 : 1	Si/SiO$_2$	88	
inst16	3	Ni, Ti, Al	1 : 1 : 1	Si/SiO$_2$	132	
inst17	3	Pt, Ti, Ni	1 : 1 : 1	Si/SiO$_2$	132	
inst18	3	V, Ti, Cr	1 : 1 : 1	Si/SiO$_2$	132	
inst19	3	Ni, Ti, Al	1 : 1 : 1	Si/SiO$_2$	132	
inst20	3	Si, Bi, Ti	0.25 : 1 : 1	Si/SiO$_2$	15	O$_2$
inst21	2	Pt, Ti	0.25 : 1 : 1	Si/SiO$_2$	15	O$_2$
inst22	2	Bi, V	0.25 : 1 : 1	Si/SiO$_2$	15	O$_2$

Table 1: Real instances and their parameters, where M is the number of different material elements in the system (the system dimension), and N is the number of sample points. The design stoichiometry corresponds to the relative proportion of the M elements at the center of the film. The table also provides the type of substrate the data was obtained on, as well as the gas used in the vacuum chamber during the deposition, if any.

on Theory and Applications of Satisfiability Testing (SAT), SAT'12, 172–185.

Finger, M.; Le Bras, R.; Gomes, C. P.; and Selman, B. 2013. Solutions for hard and soft constraints using optimized probabilistic satisfiability. In *Proceedings of the 16th International Conference on Theory and Applications of Satisfiability Testing (SAT)*, SAT'13, 233–249.

Gregoire, J. M.; van Dover, R. B.; Jin, J.; DiSalvo, F. J.; and Abruña, H. D. 2007. Getter sputtering system for high-throughput fabrication of composition spreads. *Review of Scientific Instruments* 78(7):072212–072212.

Gregoire, J. M.; Dale, D.; Kazimirov, A.; DiSalvo, F. J.; and van Dover, R. B. 2009. High energy x-ray diffraction/x-ray fluorescence spectroscopy for high-throughput analysis of composition spread thin films. *Review of Scientific Instruments* 80(12):123905–123905.

Gregoire, J. M.; Dale, D.; and van Dover, R. B. 2011. A wavelet transform algorithm for peak detection and application to powder x-ray diffraction data. *Review of Scientific Instruments* 82(1):015105–015105.

Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials* 1(1):011002.

Le Bras, 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. In *Proceedings of the 17th international conference on Principles and practice of constraint programming*, CP'11, 508–522. Berlin, Heidelberg: Springer-Verlag.

Le Bras, R.; Bernstein, R.; Gomes, C. P.; and Selman, B. 2013. Crowdsourcing backdoor identification for combinatorial optimization. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, IJCAI'13.

Long, C.; Hattrick-Simpers, J.; Murakami, M.; Srivastava, R.; Takeuchi, I.; Karen, V.; and Li, X. 2007. Rapid structural mapping of ternary metallic alloy systems using the combinat. approach and cluster analysis. *Rev. Sci. Inst.* 78.

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

Patel, P. 2011. Materials genome initiative and energy. *MRS bulletin* 36(12):964–966.

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

Takeuchi, I.; Dover, R. B. v.; and Koinuma, H. 2002. Combinatorial synthesis and evaluation of functional inorganic materials using thin-film techniques. *MRS bulletin* 27(04):301–308.

Vogt, S.; Chu, Y. S.; Tkachuk, A.; Ilinski, P.; Walko, D. A.; and Tsui, F. 2004. Composition characterization of combinatorial materials by scanning x-ray fluorescence microscopy using microfocused synchrotron x-ray beam. *Applied surface science* 223(1):214–219.

White, A. 2012. The materials genome initiative: One year on. *MRS Bulletin* 37(08):715–716.