Human Computation for Combinatorial Materials Discovery

Ronan Le Bras
Computer Science Dept.
Cornell University
Ithaca, NY 14853
lebras@cs.cornell.edu

Richard Bernstein
Computer Science Dept.
Cornell University
Ithaca, NY 14853
rab38@cornell.edu

Carla P. Gomes
Computer Science Dept.
Cornell University
Ithaca, NY 14853
gomes@cs.cornell.edu

Bart Selman
Computer Science Dept.
Cornell University
Ithaca, NY 14853
selman@cs.cornell.edu

R. Bruce van Dover
Materials Science and Engr. Dept.
Cornell University
Ithaca, NY 14853
vandover@cornell.edu

Abstract

We will show how human computation can dramatically speed up the performance of combinatorial optimization methods. We describe our work in the context of the domain of materials discovery. Our approach leverages the complementary strength of human input, providing global insights into problem structure, and the power of combinatorial solvers to exploit complex local constraints.

1 Introduction

Combinatorial materials discovery involves the rapid, high-throughput synthesis, measurement, and analysis of a large number of different but structurally related materials. In combinatorial materials discovery, materials scientists search for intermetallic compounds with desirable physical properties by obtaining measurements on hundreds of samples from a *thin film* composition spread. This approach has been successfully applied for example to speed up the discovery of new materials with improved catalytic activity for fuel cell applications [1, 2]. Determining the structure of the materials (or *phase map*) formed in a composition spread is key to understanding composition and property relations and can potentially result in a breakthrough discovery. In the set-up we consider in this paper, scientists run several experiments at the Cornell High Energy Synchrotron Source (CHESS) for about one week per year (at an experimentation cost of about $1M) and spend the rest of the year analyzing the data. Our goal is to reduce the processing time of much of the data interpretation task to a timeframe of hours. Such rapid analysis will enable scientists to dynamically optimize their experiments over the days that they have access to the synchrotron, thereby reducing overall experimentation time and significantly accelerating the discovery cycle.

The motivation for considering the materials discovery problem comes from the fact that new materials provide a fundamental basis for solutions to some of the most pressing issues in energy generation, transport, and utilization as well as more general issues in sustainability. In many cases, long-term solutions will depend on breakthrough innovations in materials, such as the development of new materials for more efficient fuel cells, solar cell arrays, or for wind turbines.

Combinatorial materials discovery, in particular the problem of ternary phase-field identification addressed in this paper, provides unique computational and modeling challenges. While statistical methods and machine learning are important components to address this challenge, they fail to

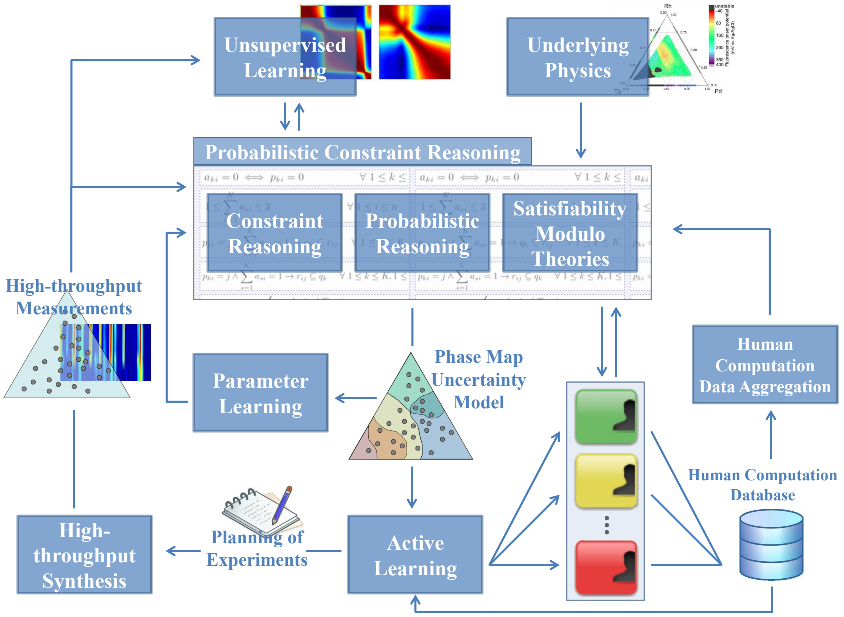

Unsupervised Learning
Underlying Physics
Probabilistic Constraint Reasoning
Constraint Reasoning
Probabilistic Reasoning
Satisfiability Modulo Theories
High-throughput Measurements
Parameter Learning
Phase Map Uncertainty Model
Human Computation Data Aggregation
Human Computation Database
Planning of Experiments
Active Learning
High-throughput Synthesis

Figure 1: Proposed Framework for Combinatorial Materials Discovery

incorporate relationships that are inherent due to the basic physics and chemistry of the underlying materials. In fact, a successful approach to materials discovery requires *a tight integration of **statistical machine learning methods**, to deal with noise and uncertainty in the measurement data*, and ***optimization and inference techniques**, to incorporate a rich set of constraints arising from the underlying materials physics and chemistry*. (See Fig. 1.) In [3], we showed that a constraint reasoning and optimization approach, as performed by a state-of-the-art Satisfiable Modulo Theory (SMT) solver, can effectively solve small- to medium-scale synthetic instances. The challenge we consider here is how to signficantly scale up this approach, including the significant measurement noise level present in real-world data. In particular, our ultimate objective is to obtain an analysis turn-around time of under 12 hours. This would enable us have the analysis guide further experiments during the time period scheduled for experimentation.

The broader underlying question that we consider is whether human input can be used to significantly boost the performance of combinatorial reasoning and optimizaton methods. The project is close in spirit to the seminal FoldIt project [4] for protein folding. In FoldIt, human computation is the main driving force, complemented with a limited amount of local computation (e.g., "shaking" of structures). We are proposing a much tighter integration between our computational framework and the human computation [5] component. In our approach, the human input and the SMT solver are highly complementary: the complex local physical constraints require a sophisticated optimization approach, whereas the global human insights are used to guide the solver. In particular, we will show how we can boost the perfomance of the SMT solver by providing additional information from human input. The task at hand, which involves the interpretation of complex high-intensity X-ray diffraction patterns, appears to be well-suited for a human computation approach. As we will see, the human input provides useful *global guidance* to the solver, by identifying the setting of so-called backdoor variables in the SMT model, critical variables that when assigned a value, enable highly efficient constraint reasoning and inference, leading to orders of magnitude speedups for the SMT solver. Overall, the results show that our hybrid human-computer approach presents us with unique opportunities for tackling hard combinatorial optimization challenges.

2 Phase-Map Identification Problem – Description and Formulation

In the so-called composition spread approach, three metals (or oxides) are deposited onto a silicon wafer using sputter guns pointed at three distinct locations, resulting in a *thin film* (Fig. 2). Different sample points on the thin film have different concentrations of the sputtered metals, based on their distance from the gunpoints. X-ray diffraction (XRD) is then used to characterize a number of samples on the thin film. For each sample point, it provides the intensity of the electromagnetic waves as a function of the scattering angle. Constructive interference of the scattered X-rays occurs, by nature, at *specific* angles, thus creating *peaks* of intensity. The observed diffraction pattern is closely related to the underlying crystalline structure, which provides important insights into chemical and physical properties of the corresponding composite material.

The goal of the *phase-map identification problem* is to identify regions of the thin film that share the same underlying crystalline structure. Intuitively, the XRD patterns observed across the *thin film* can be explained as combinations of a small set of basis patterns called *phases*. Finding the phase map corresponds to identifying these *phases* as well as their concentration on the thin film. The main challenge is to model the complex crystallographic process that these phases are subject to (such as the expansion of the lattice, which results in a shifting or scaling of the XRD pattern), while taking into account the imperfection of the silicon wafer as well as experimental noise of the data. It can take several weeks for a human expert to interpret the diffraction patterns from a single thin film experiment.

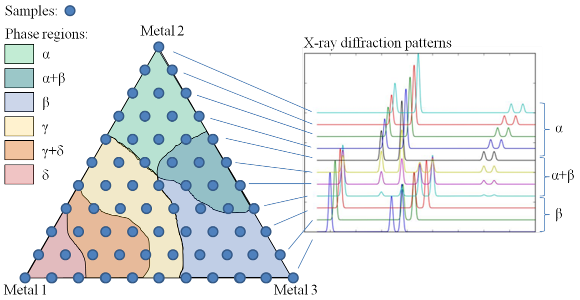


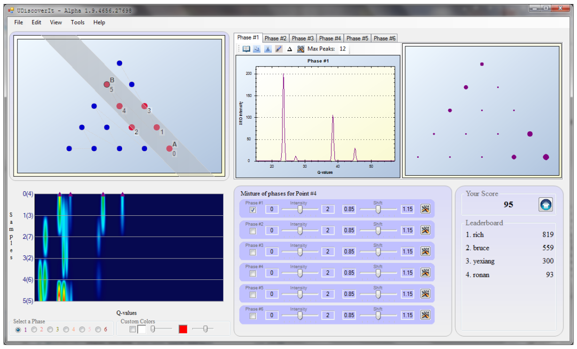
Figure 2: Left: Depiction of the problem, showing a set of samples on a *thin film*. Each sample corresponds to a different composition, and has an associated measured x-ray diffraction pattern. Colors correspond to different combinations of the basis patterns $\alpha, \beta, \gamma, \delta$. Right: Diffraction patterns shifting and combining along the thin film.

We formulate the phase-map identification problem as a Satisfiability Modulo Theories (SMT) encoding, using a *quantifier-free linear integer arithmetic theory*, as proposed in [6]. In this model, each XRD pattern p is discretized into a set of peaks $\mathcal{A}(p)$ using a peak detection algorithm. For a given number K, the goal is to find a set of peaks $\{\mathcal{E}_k\}_{k=0}^{K-1}$ for the K basis patterns that provide the best possible interpretation of the observed sets of peaks $\{\mathcal{A}(p)\}_{p=0}^{P-1}$. We refer the reader to [6] for the full SMT encoding.

3 Experimental Results

In our experimental setting, we provide a group of users with an interface for the visualization of the X-Ray diffraction (XRD) patterns. The user provides input about partial diffraction patterns that identify peaks that are likely to belong to the same materials phase. Our user group consisted of students who had no knowledge of the underlying phase structure present in the samples, and had only very limited knowledge of diffraction methods. In fact, the required user input can be phrased fully in terms of a search for patterns that follow certain geometric properties. For example, in Fig. 4, the user grouped (correctly) four vertical, colored, "skinny" ovals on the left side as diffraction peaks belonging to the same crystalline phase on the sample. In this manner, the user provides global insights into a partial interpretation of the overall XRD pattern. These inputs are saved to a compute cloud that subsequently generates an SMT encoding with the partial pattern instantiated, and uses the SMT solver to try to find the best globally consistent interpretation.

Table 3 shows the runtime of the SMT solver (Z3, version 4.1) on 8 instances of various sizes, depending on whether user input was provided. It shows that user input allows a significant improvement in runtime on each instance, with at least one order of magnitude improvement on 6 instances, and two orders of magnitude improvement on 3 instances. Interestingly, the level of user input needed to reach such performance is quite minimal with respect to the instance size. The input corresponds to the assignments of about 20 variables, which represents barely 5% of all the variables of the SMT encoding.



Mixture of phases for Point #4

	Intensity			Shift	
Phase #1 ☑	0	2	0.85		1.15
Phase #2 □	0	2	0.85		1.15
Phase #3 □	0	2	0.85		1.15
Phase #4 □	0	2	0.85		1.15
Phase #5 □	0	2	0.85		1.15
Phase #6 □	0	2	0.85		1.15

Your Score
95

Leaderboard

1. rich	819
2. bruce	559
3. yexiang	300
4. ronan	93

Figure 3: Snapshot of UDiscoverIt, a graphical user interface for providing human input to an SMT solver for the phase-map identification problem.

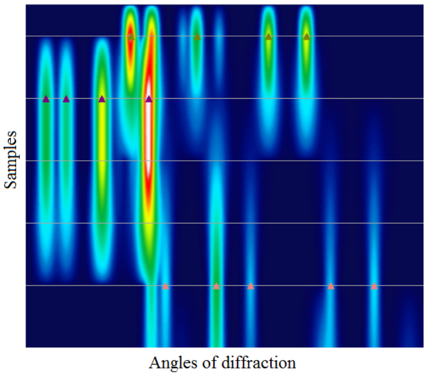


Figure 4: Example of user input on the heat map of XRD patterns.

	Dataset					Time w/o	Time w/	# assigned
System	P	L^*	K	*#var*	*#cst*	**user input** *(s)*	**user input** *(s)*	**var. by user**
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 1: 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.

4 Conclusions

Our experiments show how human computation can dramatically speed up the performance of combinatorial optimization methods. We described our work in the context of the domain of materials discovery. Our approach leverages the complementary strength of human input, providing global insights into problem structure, and the power of combinatorial solvers to exploit complex local constraints.

Acknowledgments

We thank Stefano Ermon for his insights into the SMT formulation, and Yexiang Xue and Bruce van Dover's students for providing user input. This work was supported by an NSF Expeditions in Computing Award on Computational Sustainability.

References

[1] R. B. Van Dover, LF Schneemeyer, and RM Fleming. Discovery of a useful thin-film dielectric using a composition-spread approach. *Nature*, 392(6672):162–164, 1998.

[2] 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.

[3] R. Le Bras, T. Damoulas, J. M. Gregoire, A. Sabharwal, C. P. Gomes, and R. B. van Dover. 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, pages 508–522, 2011.

[4] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and Foldit Players. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, August 2010.

[5] E. Law and L. von Ahn. *Human Computation*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

[6] S. Ermon, R. Le Bras, C. P. Gomes, B. Selman, and R. B. van Dover. Smt-aided combinatorial materials discovery. In *Proc. of the Conference on Theory and Applications of Satisfiability Testing*, SAT'12, 2012.