SMT-Aided Combinatorial Materials Discovery*

Stefano Ermon[1], Ronan Le Bras[1], Carla P. Gomes[1],
Bart Selman[1], and R. Bruce van Dover[2]

[1] Dept. of Computer Science,
[2] Dept. of Materials Science and Engr.,
Cornell University, Ithaca, NY 14853

Abstract. In combinatorial materials discovery, one searches for new materials with desirable properties by obtaining measurements on hundreds of samples in a single high-throughput batch experiment. As manual data analysis is becoming more and more impractical, there is a growing need to develop new techniques to automatically analyze and interpret such data. We describe a novel approach to the phase map identification problem where we integrate domain-specific scientific background knowledge about the physical and chemical properties of the materials into an SMT reasoning framework. We evaluate the performance of our method on realistic synthetic measurements, and we show that it provides accurate and physically meaningful interpretations of the data, even in the presence of artificially added noise.

Keywords: SMT, Combinatorial Materials Discovery, Automated Reasoning

1 Introduction

In recent years, we have witnessed an unprecedented growth in data generation rates in many fields of science [10]. For instance, in combinatorial materials discovery, one searches for materials with new desirable properties by obtaining measurements on hundreds of samples in a single batch experiment [7, 14]. These are referred to as 'high-throughput' experiments, and are common to many other fields such as molecular biology or astronomy, where there is a need to optimize the data throughput of high-cost equipment [2]. As manual data analysis is becoming more and more impractical, there is a growing need to develop new techniques to automatically analyze and interpret such vast amount of data for important trends and results. Modern statistical machine learning and data-mining approaches have been quite effective in extracting relevant information from the ever increasing streams of raw digital data. However, in scientific data analysis, there is a large amount of rather complex domain-specific background knowledge that needs to be taken into account, such as the physical and chemical properties of the materials in the combinatorial materials discovery domain.

* This work was supported by NSF Grant 0832782.

In this paper, we describe a novel approach to the phase map identification problem, a key step towards understanding the properties of new materials created and examined using the combinatorial materials discovery method. The process of identifying a phase map has been traditionally carried out manually by domain-experts, but a completely automatic solution for the phase map identification problem would open the way for even more automation in the combinatorial approach pipeline. Further, a scalable and reliable automatic data interpretation procedure would allow us to analyze larger datasets that go beyond the capabilities of human experts.

In our approach, we integrate domain-specific scientific background knowledge about the physical and chemical properties of the materials into an SMT reasoning framework based on linear arithmetic. The problem has a hybrid nature, with continuous measurement data, discrete decision variables and combinatorial constraints at the same time. We show that using our novel encoding, state-of-the-art SMT solvers can automatically analyze large synthetic datasets, and generate interpretations that are physically meaningful and very accurate, even in the presence of artificially added noise. Moreover, our approach scales to realistic-sized problem instances, outperforming a previous approach based on Constraint Programming and a set-variables encoding [11]. Further, we show that SMT solving outperforms both Constraint Programming and Mixed Integer Programming translations of our SMT formulation. This suggests that the improvements come from the SMT solving procedure rather than from the new arithmetic-based encoding, opening a novel application area for SMT solving technology beyond the traditional verification domains [4, 5].

We see this work as a first step towards using automated reasoning technology to aid the scientific discovery process. While several aspects of our method are specific to the materials discovery application, the approach we take to scientific data analysis is general. Given the flexibility and reasoning power of modern day SMT solvers, we expect to see more applications of this technology to other fields of science.

2 Combinatorial Materials Discovery

The combinatorial method is a general experimentation setting where many simultaneous experiments are performed and analyzed in batch at each step. This experimental methodology is intended to speed up the scientific discovery process, and is becoming common in a number of areas, including catalyst discovery, drug discovery, polymer optimization, and chemical synthesis. For example, new catalysts have been discovered 10 to 30 times faster using the combinatorial approach rather than conventional methodology [7, 14]. This is an important research direction in the field of Computational Sustainability, for instance because new materials with improved catalytic activity can be used for fuel cell applications [8].

In this paper, we consider a combinatorial materials discovery approach called *composition-spread*, that has been recently applied with success to speed up the discovery of new catalysts [15]. In the composition spread approach, three metals (or oxides) are sputtered onto a silicon wafer using guns pointed at three distinct locations, resulting in a so-called *thin film*. Different locations on the silicon wafer correspond to different concentrations of the sputtered materials, depending on their distance from the gunpoints. During experimentation, a number of locations (samples) on the thin film are examined using an x-ray diffraction technique, obtaining a diffraction pattern for each sampled point that gives the intensity of the electromagnetic waves as a function of the scattering angle of radiation. The observed diffraction pattern is closely related to the underlying crystal structure, which provides important insights into chemical and physical properties of the corresponding composite material.

A key step towards understanding the chemical and physical properties of the composite materials on a *thin film* is to obtain a so-called *phase map*, that is used to identify regions of the silicon wafer that share the same underlying crystal structure (see Figure 2 for an example). Intuitively, the idea is that the different diffraction patterns observed across the *thin film* can all be explained as combinations of a small number (typically, less than 6) of diffraction patterns called *basis patterns* or *phases.* Finding the phase map corresponds to identifying these *basis patterns* and their location on the silicon wafer. This is a challenging task because we only observe combinations of the *basis patterns*, and the measurements are affected by noise. Furthermore, due to a fairly complicated physical process dealing with the expansion of crystals on the lattice, *basis patterns* can appear scaled (contracted to a smaller or larger frequency range), and they must satisfy a number of physical constraints (for instance, basis patterns must appear in contiguous locations on the *thin film* and there is a maximum number of *basis patterns* that can appear in each sample diffraction pattern).

2.1 Phase map identification

Formally, we are given P diffraction patterns $\mathbf{D}_0, \cdots, \mathbf{D}_{P-1}$, one for each of the P points sampled on the *thin film*, where each vector $\mathbf{D}_i = (d_{0,i}, \cdots, d_{B-1,i}) \in (\mathbb{R}_{\geq 0})^B$ represents the intensity of the electromagnetic waves for a fixed set of B scattering angles of radiation. The sample points are embedded into a graph $\mathcal{G}$, such that there is a vertex for every point and edges connect points that are close on the *thin film* (for instance, based on Delaunay triangulation). Given a norm $||\cdot||$ (for instance, an L_∞ norm), we want to find K basis patterns $\mathbf{B}_0, \cdots, \mathbf{B}_{K-1}$ where $\mathbf{B}_i \in (\mathbb{R}_{\geq 0})^B$, coefficients $a_{i,j} \in \mathbb{R}$ and scaling factors $s_{i,j} \in \mathbb{R}$ for $i = 0, \cdots, P-1$, $j = 0, \cdots, K-1$ that minimize

$$\sum_{i=0}^{P-1} ||\mathbf{D}_i - \sum a_{i,j} S\left(\mathbf{B}_j, s_{i,j}\right)|| \tag{1}$$

where $S(\cdot)$ is an operator modeling the scaling phenomena (see below), and the coefficients $a_{i,j}$ must satisfy

$$\begin{aligned} a_{i,j} \geq 0 \quad & i = 0, \cdots, P-1, j = 0, \cdots, K-1 \\ |\{j | a_{i,j} > 0\}| \leq M \quad & i = 0, \cdots, P-1 \end{aligned}$$

that is, they are non-negative and no more than M basis patterns can be used to explain a point i. Furthermore, the subgraph induced by $\{i|a_{i,j} > 0\}$ must be connected for $j = 0, \cdots, K-1$ (so that the basis patterns appear in contiguous locations on the *thin film*). The scaling operator $S(\cdot)$ models the potential expansion of the crystals on the lattice. Specifically, a peak appearing at scattering angle a in the k-th basis pattern might appear respectively at scattering angles $s_{p,k} \cdot a$ and $s_{p',k} \cdot a$ at points p, p' of the silicon wafer because of the scaling effect. For each basis pattern k, the corresponding scaling coefficients $s_{i,k}$ must be continuous and monotonic as a function of the corresponding location i on the *thin film*. Further, the presence of 3 or more basis patterns in the same point prevents any significant expansion of the crystals, and therefore scalings do not occur.

Notice that this formulation is closely related to a principal component analysis (PCA) of the data, but includes additional constraints needed to ensure that the solution is physically meaningful, such as the non-negativity of eigenvectors, connectivity, and phase usage limitations.

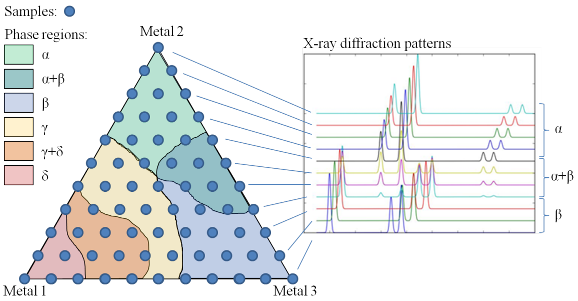

Fig. 1: Left: Pictorial depiction of the problem, showing a set of sampled points on a *thin film*. Each sample corresponds to a different composition, and has an associated measured x-ray diffraction pattern. Colors correspond to different combinations of the basis patterns $\alpha, \beta, \gamma, \delta$. On the right: Scaling (shifting) of the diffraction patterns as one moves from one point to a neighboring one.

3 Prior Work

There have been several attempts to automate the *phase map* identification process. Most of the solutions in the literature are based on unsupervised machine learning techniques, such as clustering and non-negative matrix factorization [13, 12]. While these approaches are quite effective at extracting information from large amounts of noisy data, their major limitation is that it is hard to enforce the physical constraints of the problem at the same time. As a result, the interpretations obtained with these techniques are often not physically meaningful, for instance because regions corresponding to some basis patterns are not connected [11].

To address these limitations, in [11] they used a Constraint Programming approach to enforce the constraints on the phase maps, defining a new problem called *Pattern Decomposition with Scaling*. They propose an encoding based on set variables, but the main limitation of their work is that current state-of-the-art CP solvers cannot scale to realistic size instances (e.g., with at least 40 sample points). To overcome this limitation, the authors used a heuristic preprocessing step based on clustering to fix the value of certain variables before attempting to solve the problem. While the solutions they found are empirically shown to be accurate, their strategy cannot provide any guarantee because it only explores part of the search space.

Our approach is similar to the one proposed in [11], but in this work we introduce a novel SMT encoding based on arithmetic to formulate the phase map identification problem. The SMT formalism nicely captures the hybrid nature of the problem, which involves discrete decision variables and continuous measurement data at the same time. Furthermore, we show that the ability to reason at the level of arithmetic operations of SMT solvers allows our approach to scale to instances of realistic size without need for Machine Learning-based heuristics.

4 SMT-Aided Phase Map Identification

In our first attempt to model the phase map identification problem, we constructed an SMT-based model where we described the entire spectrum of all the unknown basis patterns $\mathbf{B}_0, \cdots, \mathbf{B}_{K-1}$. However, this approach requires too many variables to obtain a sufficiently fine-grained description of the diffraction patterns, and ultimately leads to instances that cannot be solved using current state-of-the art solvers. We therefore use the same approach presented in [11], and we preprocess the diffraction patterns $\mathbf{D}_0, \cdots, \mathbf{D}_{P-1}$ using a peak detection algorithm, extracting the locations of the *peaks* $\mathcal{Q}(p)$ in the x-ray diffraction pattern of each point p (see Figure 1). This is justified by the nature of the diffraction patterns, as constructive interference of the scattered x-rays occurs at specific angles (thus creating peaks of intensities) that characterize the underlying crystal. Furthermore, matching the locations of the peaks is what human experts do when they try to manually solve these problems.

Given the sets of observed peaks $\{\mathcal{Q}(p)\}_{p=0}^{P-1}$ extracted from the measured diffraction patterns $\mathbf{D}_0, \cdots, \mathbf{D}_{P-1}$, our goal is to find a set of peaks $\{\mathcal{E}_k\}_{k=0}^{K-1}$ for the K basis patterns that can explain the observed sets of peaks $\{\mathcal{Q}(p)\}_{p=0}^{P-1}$. The new variables $\{\mathcal{E}_k\}_{k=0}^{K-1}$ therefore replace the original variables $\mathbf{B}_0, \cdots, \mathbf{B}_{K-1}$ in the problem described earlier in Section 2. For each peak $c \in \mathcal{Q}(p)$ we want to have at least one peak $e \in \mathcal{E}_k$ that can explain it, i.e.

$$\forall c \in \mathcal{Q}(p) \exists e \in \mathcal{E}_k \ s.t. \ (a_{p,k} > 0 \wedge \ |c - s_{p,k} \cdot e| \leq \epsilon)$$

where ϵ is a parameter that depends on how accurate the peak-detection algorithm is. Notice that we match the location of the peak, which can be measured accurately, but not its intensity, which can be very noisy. At the same time, we want to limit the number of missing peaks, i.e. peaks that should appear because they belong to some basis pattern but have not actually been measured. Therefore, instead of optimizing the objective in equation (1), we consider an approximation given by

$$\sum_{p=0}^{P-1} \sum_{k=0}^{K-1} \mathbb{1}_{[a_{p,k}>0]} \sum_{e \in \mathcal{E}_k} \mathbb{1}_{[\forall c \in \mathcal{Q}(p), |c - s_{p,k} \cdot e| > \epsilon]}$$

that gives the total number of missing peaks. All the other constraints of the problem previously introduced are not affected and still need to be satisfied. Note that we can avoid the use of expensive non-linear arithmetic by using a logarithmic scale for the x-ray data, so that multiplicative scalings become linear operations. We refer to these effects (corresponding to the scalings in the original problem formulation) as *shifts*. For each point, we therefore define a set $\mathcal{A}(p) = \{\log q, q \in \mathcal{Q}(p)\}$ of peak positions in log-scale and similarly we represent the positions of the peaks of the basis patterns using the same logarithmic scale.

After a preliminary investigation where we evaluated the performance of real-valued arithmetic, we decided to discretize the problem and use Integer variables to represent peak locations (with a user-defined discretization step). Since the diffraction data is measured using digital sensors, there is no actual loss of information if we use a small enough discretization step, and it significantly improves the efficiency of the solvers. In the resulting SMT model we therefore use a *quantifier-free linear integer arithmetic theory.*

4.1 Model Parameters

Let P be the number of sampled points on the *thin film*. We define L as the maximum number of peaks per point, i.e. $L = \max_p |\mathcal{A}_p|$. Based on the observed patterns, we precompute an upper and lower bound e_{max} and e_{min} for the positions of the peaks: $e_{max} = \max_p \max_{a \in \mathcal{A}(p)} a$, $e_{min} = \min_p \min_{a \in \mathcal{A}(p)} a$. There are also a number of user-defined parameters. K is the total maximum number of basis patterns used to explain the observed diffraction patterns, while M is the maximum number of basis patterns that can appear in any point p. ϵ is a tolerance level such that two peaks within an interval of size 2ϵ are considered to be overlapping. ϵ_S is a bound on the maximum allowed difference in the shifts of neighboring locations on the thin film, while S_{max} is a bound on the maximum possible shift. Furthermore, the user specifies a parameter T which gives a bound on the total number of peaks that should appear because they belong to some basis pattern but have not actually been measured (we will refer to them as *missing peaks*).

4.2 Variables

We use a set of Boolean variables

$$r_{p,k}, \quad p = 0, \cdots, P-1, k = 0, \cdots, K-1$$

where $r_{p,k} = \mathit{TRUE}$ means that phase (basis pattern) k appears in point p (i.e., $a_{p,k} > 0$). We also have the following *Integer* variables:

$$\begin{aligned} e_{k,\ell}, &\quad k = 0, \cdots, K-1, \ell = 0, \cdots, L-1 \\ S_{p,k}, &\quad p = 0, \cdots, P-1, k = 0, \cdots, K-1 \\ I_{p,k}, &\quad p = 0, \cdots, P-1, k = 0, \cdots, K-1 \\ t_p, &\quad p = 0, \cdots, P-1 \end{aligned}$$

where $e_{k,\ell}$ represents the position of the ℓ-th peak of the k-th basis pattern. $S_{p,k}$ represents the shift of the k-th basis pattern at point p. The variables $I_{p,k}$ are redundant and used to count the number of phases used at point p. The variables t_p represent the number of unexplained peaks at point p, i.e. the number of missing peaks at point p. These are peaks that should appear according to the values of $\{r_{p,k}\}_{k=0}^{K-1}$, $\{e_{k,\ell}\}_{\ell=0}^{L-1}$, and $\{S_{p,k}\}_{k=0}^{K-1}$, but are not present, i.e. they do not belong to $\mathcal{Q}(p)$.

4.3 Constraints

The variables $I_{p,k}$ are Integer indicators for the Boolean variables $r_{p,k}$ that must satisfy

$$\begin{aligned} 0 \leq I_{p,k} \leq 1 &\; k = 0, \cdots, K-1, p = 0, \cdots, P-1 \\ r_{p,k} \Leftrightarrow (I_{p,k} = 1) &\; k = 0, \cdots, K-1, p = 0, \cdots, P-1 \end{aligned}$$

Peak locations $e_{k,\ell}$ in the basis patterns are bounded by what we observe in the x-ray diffraction pattern:

$$e_{min} \leq e_{k,\ell} \leq e_{max}, \quad k = 0, \cdots, K-1, \ell = 0, \cdots, L-1$$

Shifts are bounded by the maximum allowed shift, and can be assumed to be non-negative without loss of generality:

$$0 \leq S_{p,k} \leq S_{max}, \;\; k = 0, \cdots, K-1, p = 0, \cdots, P-1$$

Every peak $a \in \mathcal{A}(p)$ appearing at point p must be explained by at least one peak belonging to one phase k, which can appear shifted by $S_{p,k}$:

$$\bigvee_{k=0}^{K-1} \bigvee_{\ell=0}^{L-1} \left(r_{p,k} \wedge (|e_{k,\ell} + S_{p,k} - a| \leq \epsilon)\right) \forall p, \forall a \in \mathcal{A}(p)$$

Inequalities involving the absolute value of an expression of the form $|e| < c$ where c is a positive constant are encoded as $(e < c) \wedge (e > -c)$.

If a phase k is chosen for point p (i.e., $r_{p,k} = \mathit{TRUE}$), then most of the peaks $e_{k,0}, \cdots, e_{k,L-1}$ should belong to $\mathcal{Q}(p)$. We count the number of missing peaks as follows:

$$t_p = \sum_{k=0}^{K-1} \sum_{\ell=0}^{L-1} ITE(r_{p,k} \wedge \neg \left(\bigvee_{a \in \mathcal{A}(p)} (|e_{k,\ell} + S_{p,k} - a| \leq \epsilon) \right), 1, 0), \forall p$$

where ITE is an if-then-else expression. Here we assume that each phase contains at least one peak, but since peaks can be overlapping (e.g., $e_{k,\ell} = e_{k,\ell+1}$) a basis pattern is allowed to contain less than L distinct peaks.

Missing Peaks Bound We limit the number of total missing peaks (across all points p) with the user-defined parameter T

$$\sum_{p=0}^{P-1} t_p \leq T$$

Intuitively, the smaller T is, the better an interpretation of the data.

Phase Usage There is a bound M on the total number of phases that can be used to explain the peaks observed at any location p:

$$\sum_{k=0}^{K-1} I_{p,k} \leq M, p = 0, \cdots, P-1$$

For instance, when three metals or oxides are used to obtain the thin film, we have a *ternary system*, where no more than three phases can appear in each point p, that is $M = 3$.

Shift Continuity Phase shifting is a continuous process over the *thin film*. We therefore have the following constraint:

$$|S_{p,k} - S_{p',k}| < \epsilon_S, \forall p, \forall p' \in \mathcal{N}(p)$$

where $\mathcal{N}(p)$ is the set of neighbors of p according to the connectivity graph $\mathcal{G}$ (i.e., points that lie close to p on the *thin film*).

Shift Monotonicity Let $\mathcal{D} = (d_0, \cdots, d_t)$ where $d_i \in \{0, \cdots, P-1\}$ be a sequence of points that lie in a straight line on the thin film. Shifting is a monotonic process, i.e. it must satisfy the following constraint

$$\left(\bigwedge_{i=0}^{t-1} (S_{d_i,k} \geq S_{d_{i+1},k})\right) \vee \left(\bigwedge_{i=0}^{t-1} (S_{d_i,k} \leq S_{d_{i+1},k})\right), k = 0, \cdots, K-1$$

Since points are usually collected on a grid lattice on the silicon wafer, we enforce shift monotonicity on the lines forming the grid.

Ternary Phases Shift Ternary phases (where 3 basis patterns are used) are not affected by shifting:

$$\left(\left(\sum_{k=0}^{K-1} I_{p,k} = 3\right) \wedge \bigwedge_{k=0}^{K-1} (r_{p,k} \Leftrightarrow r_{p',k})\right) \Rightarrow (S_{p,k} = S_{p',k}), \forall p, \forall p' \in \mathcal{N}(p)$$

where $\mathcal{N}(p)$ is the set of neighbors of p.

Connectivity Constraint Each of the basis patterns must be connected. Formally, for every pair of points p, p' such that $r_{p,k} \wedge r_{p',k}$, there must exist a path $\mathbb{P}$ from p to p' such that $r_{j,k} = \mathit{TRUE}$ for all $j \in \mathbb{P}$. Since it would require too many constraints, we use a lazy approach to enforce connectivity. If we find a solution where a basis pattern k is not connected, i.e. there exists p, p' such that $r_{p,k} \wedge r_{p',k}$ but there is no path $\mathbb{P}$ with p, p' as endpoints such that $r_{j,k} = \mathit{TRUE}$ for all $j \in \mathbb{P}$, then we consider the smallest cut C between p and p' such that $r_{j,k} = \mathit{FALSE}$ for all $j \in C$ and we add a new constraint

$$(r_{p,k} \wedge r_{p',k}) \Rightarrow \bigvee_{c \in C} r_{c,k}$$

Symmetry Breaking Without loss of generality, we can impose an ordering on the peak locations within every phase k:

$$e_{k,\ell} \leq e_{k,\ell+1}, \ell = 0, \cdots, L-2, k = 0, \cdots, K-1$$

Furthermore, notice that the problem is symmetric with respect to permutations of the phase indexes $k = 0, \cdots, K-1$. We therefore enforce an ordering on the way phases are assigned to points

$$\begin{gathered} \textstyle\bigwedge_{k=1}^{K-1} (r_{0,k} \Rightarrow r_{0,k-1}) \\ \cdots \\ \textstyle\bigwedge_{j=1}^{K-1} \left(\left(\bigwedge_{i=0}^{Y} \neg r_{i,j} \right) \Rightarrow \bigwedge_{k=j}^{K-1} (r_{Y+1,k} \Rightarrow r_{Y+1,k-1}) \right) \end{gathered}$$

where we set $Y = 4$.

5 Experimental Results

We evaluate the performance of our approach on a benchmark set of synthetic instances for which the ground truth is known (namely, what the true basis patterns are and how they are combined to form the observed diffraction patterns). All the systems we consider are ternary, where three metals are combined, so that M is set to 3 in the entire experimental section. For all experiments, two peaks are considered to be overlapping if they are within 1% of each other, and the maximum allowed shift is 15%.

We compare our SMT-based approach with the Constraint Programming based solution presented in [11]. Since their CP-based formulation does not scale to realistic-sized instances, they integrate a Machine Learning based component to simplify the problem that the CP solver needs to solve to improve scalability. Note that by doing this they lose the completeness of the search, because they only explore a subtree (suggested by the ML part) of the original search space. In contrast, our approach scales to instances of realistic size (with over 40 points) without need for the ML component. Note however that if desired, the ML heuristic component could be easily integrated with our method.

Synthetic Data We consider the known Al-Li-Fe system [1] previously used in [11], represented with a ternary diagram in Figure 2. A ternary diagram is a simplex where each point corresponds to a different concentration of the three constituent elements, in this case Al, Li, and Fe. The composition of a point depends on its distance from the corners. For a fixed value of the parameter P, synthetic instances are generated by sampling P points in the ternary diagram, each corresponding to different concentrations of the three constituent elements. For each point, synthetic x-ray diffraction patters are generated starting from known diffraction patterns of the constituent phases (taken from the JCPDS database [1]), that are combined according to the concentrations of the elements in that point. A peak detection algorithm is then used to generate a discrete set of peaks.

We first consider a set of instances without any noise, for which we have the exact location of all the peaks for every sample (the maximum number of peaks per sample is $L = 12$), without any outlier or missing peak. Starting from the diffraction patterns and the corresponding peaks, we generate the corresponding instance using the formulation described in the previous section, encoded in the SMTLibV2 language [3]. In this case, we set $K = 6$, the true number of underlying unknown basis patterns, and we try to recover a solution with $T = 0$ missing peaks. We also consider a set of simplified instances, where we fix some of the six unknown basis patterns to their true values. We solved these instances on a 3 Ghz Intel Core2Duo machine running Windows, using the SMT solvers Z3 [6] and MathSAT5 [9]. However, MathSAT is significantly slower (for instance, it takes over 50 minutes to solve a small instance with $P = 10$ points that Z3 solves in about 15 seconds) and it does not scale to larger problems. We therefore report only times obtained with Z3.

Running time We compare our method with previous CP-based approach presented in [11] on the same set of benchmark instances. The runtime for the CP solver are taken from [11], and were obtained on a comparable 3.8 GHz Intel Xeon machine. In Table 1a we show runtime as a function of the instance size P and the number of basis patterns left unknown K' (e.g., $K' = 3$ when the instance has been simplified by fixing three out of the six unknown basis patterns).

As we can see from the runtimes reported in Table 1a, our approach based on SMT and Z3 is always considerably faster, except for the smallest simplified problems where the difference is in the order of a few seconds. More importantly, our SMT-based approach shows a significantly improved scaling behavior, and can solve problems of realistic size with 6 unknown phases and over 40 points within an hour. In contrast, the previous CP-based approach can only solve simplified problems and cannot solve any problem with 6 unknown basis patterns [11].

Solving Strategy In order to understand whether the improvement comes from the new problem encoding (based on integer arithmetic and not on set variables as the one in [11]) or from the SMT solving strategy, we translated

Dataset	**Z3 (s)**	**ILOG Solver (s)**
P=10 K'=3	8	**0.5**
K'=6	**12**	timeout at 1200
P=15 K'=3	13	**0.5**
K'=6	**20**	timeout at 1200
P=18 K'=3	29	384.8
K'=6	**125**	timeout at 1200
P=29 K'=3	**78**	276
K'=6	**186**	timeout at 1200
P=45 K'=6	518	timeout at 1200

(a) Running time.

Dataset	**Precision** (%)	**Recall** (%)
P=10, e=0	95.8	100
P=15, e=0	96.6	100
P=18, e=0	97.2	96.6
P=28, e=0	96.1	92.8
P=45, e=0	95.8	91.6
P=15, e=1	96.1	99.6
P=15, e=2	96.3	99.3
P=15, e=3	96.7	99.5
P=15, e=4	95.3	98.9
P=15, e=4	94.8	99.7

(b) Accuracy.

Table 1: P is the number of sampled points. K' is the number of basis patterns left unknown. e is the number of peaks removed (simulating measurement errors).

our arithmetic-based encoding as a Constraint Satisfaction Problem and as a Mixed Integer Program. As our SMT model combines logical constraints and linear inequalities exclusively, a Mixed Integer Programming (MIP) approach is particularly appealing. Indeed, one can fairly naturally translate the logical constraints of our model, namely 'Or', 'And', 'Not', 'IfThenElse', into a system of linear inequalities by using additional binary variables, and be left with a MIP formulation. The ability of the MIP to handle continuous variables for both the peak locations and the shifts, as well as to reason in terms of an objective function (e.g., the total number of missing peaks) makes it an attractive option. Nevertheless, the translation of the logical constraints yields a high number of binary variables (e.g., over 23K binary variables for a synthetic instance with $P = 10$), which contrasts with a low total number of continuous variables (about 120 for the same instance) and thus, weakens the potential of the MIP formulation. Empirically, none of the instances could be solved by the MIP formulation within the time limit of one hour. Similarly, we were not able to solve any of the instances (not even when simplified) obtained from translating our SMT formulation (symmetry breaking constraints included) to a CSP using the state-of-the-art IBM ILOG Cplex Solver within one hour. This suggests that the improvement over CP based solutions is not achieved thanks to the different problem encoding, but is due to the SMT solving procedure itself, which is stronger in the reasoning part and can handle well the intricate combinatorial constraints of the problem.

Accuracy We evaluate the accuracy of our method by comparing the solutions we find (i.e., the phase map given by the values of $r_{p,k}$ for $p = 0, \cdots, P-1, k = 0, \cdots, K-1$) with the ground truth in terms of precision/recall scores, reported in Table 1b. Precision is defined as the fraction of the number of points correctly identified as belonging to phase k (true positives), over the total number of points identified as belonging to phase k (true positives + false positives). Recall is defined as the fraction of points correctly identified as belonging to phase k (true positives) over the true number of points belonging to phase k (true positives + false negatives). These values are obtained by comparing with ground truth all $K!$ permutations of the phases we obtain, and taking the one with the smallest number of errors (recall that the problem is symmetric with respect to permutations of the phase indexes k). Further, the values in Table 1b are the precision/recall scores obtained for each single phase k averaged over the $K = 6$ phases. The results show that the phase maps we identify are always very accurate, with precision and recall values always larger than 90%.

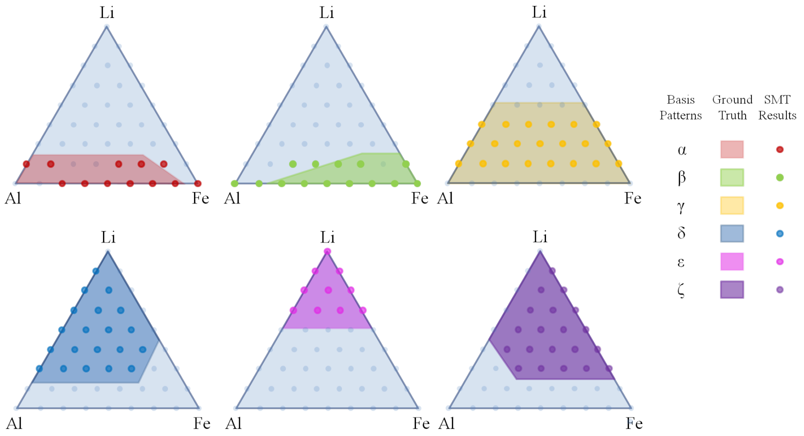

Fig. 2: Phase map for the synthetic Al-Li-Fe system with 45 sampled points, no errors. Each of the six colored areas represents one of the basis patterns $(\alpha, \beta, ..., \zeta)$ of the ground truth, while the colored dots correspond to the solution of our SMT model. The SMT results closely delimit each phase of the ground truth, which is quantitatively validated by the high precision/recall score of our approach.

Robustness To evaluate the robustness of our method to experimental noise, we also consider another dataset from [11] where peaks are removed from the observed diffraction patterns with probability proportional to the square of the inverse peak height, in order to simulate the fact that low-intensity peaks might not be detected or they can be discarded by the peak detection algorithm. This situation is common for real-world instances, where measurements are affected by noise. We consider instances generated by removing exactly e peaks from the observed diffraction patterns, and we solve them by setting the upper bound T on the number of missing peaks equal to e. In figure 3 we see the median running time as a function of the number of missing peaks T. This is averaged over 10 instances with $P = 15$ points, and 20 runs per instance, with a timeout set at 1 hour. As shown in figure 3, the problem becomes significantly harder as we introduce missing peaks, because the constraint on the total number of missing peaks allowed T becomes less and less effective at pruning the search space as T grows. However, the median running time appears to increase linearly, and we are still able to recover a phase map efficiently even for instances affected by noise.

In table 1b we show precision recall values for these instances affected by noise. We see that the phase maps we identify are still very accurate even in presence of noise, with precision/recall scores over 95%.

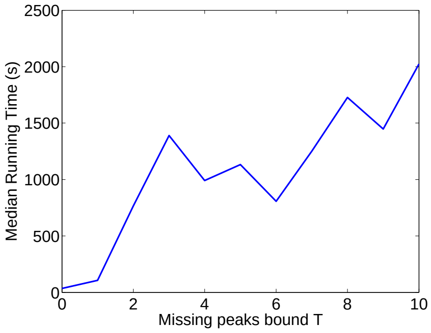

Fig. 3: Median running time as a function of the bound on the total number of missing peaks allowed T.

6 Conclusions

We described a novel approach to the phase map identification problem, a key step towards automatically understanding the properties of new materials created and examined using the *composition spread* method. In our approach, we integrate domain-specific scientific background knowledge about the physical and chemical properties of the materials into an SMT reasoning framework based on linear arithmetic. Using state-of-the-art SMT solvers, we are able to automatically analyze large synthetic datasets, generating interpretations that are physically meaningful and very accurate, even in the presence of artificially added noise. Moreover, we showed that our solution outperforms in terms of scalability both Constraint Programming and Mixed Integer Programming approaches, allowing us to solve instances of realistic size. Our experiments show a novel application area for SMT technology, where we can exploit its reasoning power in a hybrid setting with continuous measurement data and rather intricate combinatorial constraints.

As there is an ever-growing amount of data in many fields of science, the grand challenge for computing and information science is how to provide efficient methods for interpreting such data, a process that generally requires integration with domain-specific scientific background knowledge. As a first step towards this goal, in this work we demonstrated the use of automated reasoning technology to support the scientific data analysis process in materials discovery. While several aspects of our method are specific to the phase map identification problem, the approach we take for the data analysis problem is quite general. Given the flexibility and ever-growing reasoning power of modern day SMT solvers, we expect to see more applications of this technology to other areas of scientific exploration that require sophisticated reasoning to interpret experimental data.

References

1. *Powder Diffract. File, JCPDS Internat. Centre Diffract. Data, PA*, 2004.
2. S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
3. C. Barrett, A. Stump, and C. Tinelli. The SMT-LIB Standard: Version 2.0. In A. Gupta and D. Kroening, editors, *Proceedings of the 8th International Workshop on Satisfiability Modulo Theories (Edinburgh, England)*, 2010.
4. A. Biere. Sat, smt and applications. *Logic Programming and Nonmonotonic Reasoning*, pages 1–1, 2009.
5. R. Brummayer and A. Biere. Boolector: An efficient smt solver for bit-vectors and arrays. *Tools and Algorithms for the Construction and Analysis of Systems*, pages 174–177, 2009.
6. L. De Moura and N. Bjørner. Z3: An efficient smt solver. *Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340, 2008.
7. D. Ginley, C. Teplin, M. Taylor, M. van Hest, and J. Perkins. Combinatorial materials science. In *AccessScience*. McGraw-Hill Companies, 2005.
8. 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.
9. A. Griggio. A Practical Approach to Satisfiability Modulo Linear Integer Arithmetic. *JSAT*, 8:1–27, January 2012.
10. A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.
11. R. Le Bras, T. Damoulas, J. M. Gregoire, A. Sabharwal, C. Gomes, and R. B. van Dover. Constraint reasoning and kernel clustering for pattern decomposition with scaling. In *CP*, 2011.
12. 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.
13. 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. Inst.*, 78, 2007.
14. B. Narasimhan, S. Mallapragada, and M. Porter. *Combinatorial materials science.* John Wiley and Sons, 2007.
15. R. B. Van Dover, L. Schneemeyer, and R. Fleming. Discovery of a useful thin-film dielectric using a composition-spread approach. *Nature*, 392(6672):162–164, 1998.