Phase-Mapper: Accelerating Materials Discovery with AI

Junwen Bai, Yexiang Xue, Johan Bjorck, Ronan Le Bras, Brendan Rappazzo, Richard Bernstein, Santosh K. Suram, R. Bruce van Dover, John M. Gregoire, Carla P. Gomes

■ *From the stone age to the bronze, iron, and modern silicon ages, the discovery and characterization of new materials has always been instrumental to humanity's development and progress. With the current pressing need to address sustainability challenges and find alternatives to fossil fuels, we look for solutions in the development of new materials that will allow for renewable energy. To discover materials with the required properties, materials scientists can perform high-throughput materials discovery, which includes rapid synthesis and characterization via X-ray diffraction (XRD) of thousands of materials. A central problem in materials discovery, the phase map identification problem, involves the determination of the crystal structure of materials from materials composition and structural characterization data. This analysis is traditionally performed mainly by hand, which can take days for a single material system. In this work we present Phase-Mapper, a solution platform that tightly integrates XRD experimentation, AI problem solving, and human intelligence for interpreting XRD patterns and inferring the crystal structures of the underlying materials. Phase-Mapper is compatible with any spectral demixing algorithm, including our novel solver, AgileFD, which is based on convolutive nonnegative matrix factorization (NMF). AgileFD allows materials scientists to rapidly interpret XRD patterns, and incorporates constraints to capture prior knowledge about the physics of the materials as well as human feedback. With our system, materials scientists have been able to interpret previously unsolvable systems of XRD data at the Department of Energy's Joint Center for Artificial Photosynthesis, including the Nb-Mn-V oxide system, which led to the discovery of new solar light absorbers and is provided as an illustrative example of AI-enabled high-throughput materials discovery.*

The marvels of modern technology can largely be attributed to the discovery and characterization of new materials. The discovery of semiconductors laid the foundation for modern electronics, while the formulation of new molecules allows us to treat diseases previously thought incurable. Looking into the future, some of the largest problems facing humanity now are likely to be solved by the discovery of new materials. In this article, we explore the techniques materials scientists are using and show how our novel artificial intelligence system, Phase-Mapper, allows materials scientists to quickly solve material systems to infer their underlying crystal structures and has led to the discovery of new solar light absorbers.

 ISSN 0738-4602

A powerful strategy employed by materials scientist is high-throughput materials discovery (Green, Takeuchi, and Hattrick-Simpers 2013), the idea being to rapidly synthesize thousands of different materials and quickly screen them for desirable properties. These materials are developed by depositing different elements on a wafer in varying amounts, which is analogous to atomic spray painting, where elements are mixed with different proportions, rendering simple mixtures as well as enabling the emergence of new materials, just as the mixture of primary colors results in both obvious mixtures and secondary colors.

In this article, we address the phase-mapping problem, a central problem in high-throughput materials discovery, which has critically lacked an efficient solution method. At a fundamental level, the phase-mapping problem entails demixing data measurements in terms of a few simple components or crystal structures, each describing a single material, subject to intricate constraints on the solutions induced by the physics of the underlying materials. A material's phase describes a range of elemental composition and other conditions over which its properties and structure, the arrangement of the constituent atoms, change little. X-ray diffraction (XRD) is a ubiquitous technique to characterize crystal phases, as it produces a signal containing a series of peaks that serve as a fingerprint of the underlying atomic arrangement or crystal structure. Using traditional methods, materials scientists can obtain and interpret 1 to 10 XRD measurements per day, and with the recent development of automated, synchrotron-based XRD experiments, the measurement throughput has been accelerated to 10^3 to 10^5 measurements per day (Gregoire et al. 2009; 2014). The creation of a phase-mapping algorithm that generates phase diagrams from these data remains an unsolved problem in materials science despite a series of advancements over the past decade (Hattrick-Simpers, Gregoire, and Kusne 2016). The problem is challenging, given that often the X-ray diffraction patterns correspond to a mixture of crystal structures, some of them not necessarily sampled individually, requiring an algorithm that can demix patterns while simultaneously identifying the basis patterns. The most pertinent need is to generate a physically meaningful phase diagram (one that generates materials science knowledge) for the materials in a given library, or a collection of co-deposited materials on a substrate, which relies on the spectral demixing of the 10^2–10^3 XRD patterns into a small set of basis patterns (typically less than 10).

The traditional analysis workflow relies on iterative manual analysis and heuristics, resulting in the analysis of only a few systems a year. This quickly becomes a bottleneck, as manual analysis cannot keep up with the rate at which data are generated. Automatic analysis becomes imperative in order to analyze the vast amount of data that are generated in high-throughput experiments.

The need for automatic and scalable tools provides unique opportunities to apply cutting-edge techniques in computer science and AI to accelerate the materials discovery process. We developed Phase-Mapper, a comprehensive platform that tightly integrates XRD experimentation, AI problem solving, and human intelligence (figure 1), to address this computational challenge. In this platform, within minutes, an AI solver provides for the phase-mapping problem physically meaningful results, which are examined and potentially further refined by materials scientists interactively and in real time. In addition, the results of Phase-Mapper can be used to further inform future experimental designs. The demixing algorithm is a cornerstone of the Phase-Mapper platform. We have developed a novel solver called *AgileFD*, which is based on convolutive nonnegative matrix factorization (cNMF). Nonnegative matrix factorization (NMF) is commonly used in applications such as computer vision and topic modeling (Lee and Seung 2001), and cNMF extends this method to convolutive mixtures used in blind source separation of audio signals and speech recognition (Smaragdis 2004; Mørup and Schmidt 2006). AgileFD features a computationally efficient, gradient-based search method using lightweight iterative updates of candidate solutions. In addition, AgileFD integrates functionalities beyond cNMF. The extensions of AgileFD, described here, include incorporation of constraints to encode both human input, which capitalizes on a researcher's knowledge of a particular data set, and prior knowledge of the problem related to the underlying physics of phase diagrams. This, as demonstrated below, can be critical in obtaining physically meaningful solutions. In developing the Phase-Mapper platform, careful attention has been given to delivering a rich suite of capabilities while maintaining solver convergence times within minutes, which enables researchers to interact with the solver to refine the solution.

We evaluate Phase-Mapper and several solvers that were proposed in recent years. In general, we observe that the solutions found by AgileFD and its variants better match the ground truth. A vanilla NMF approach performs poorly, as it fails to capture physical constraints. Conversely, constraint programming–based approaches are able to enforce some of the physical constraints but scale poorly. We show empirically that AgileFD with its extensions is able to find solutions that are close to the physical reality.

We first encountered the phase-mapping problem seven years ago as part of our computational sustainability (Gomes 2009) effort to address pressing problems in renewable energy. Phase-Mapper is the culmination of our work since then, in close collaboration with experts in materials science. Over the course of this collaboration, we have made important contributions to the formal characterization of this problem, developed several synthetic instance generators, and developed several algorithms with theoretical and practical guarantees. We have also continuously developed tools to share experimental instance data, results, and solution visualizations with our collaborators throughout (Le Bras et al. 2011; Ermon et al. 2012; Le Bras et al. 2014; Ermon et al. 2015; Xue et al. 2015). Phase-Mapper is our most successful tool to date in this area: it removes many of the practical barriers to the use of previous methods, including better scalability, runtimes suitable for interactive use, and ease of access.

Phase-Mapper has been used at the Department of Energy's Joint Center for Artificial Photosynthesis (JCAP) to run hundreds of phase-mapping solutions in the JCAP materials discovery pipeline. Prior to Phase-Mapper, the difficulty of interpreting X-ray diffraction data limited JCAP scientists' ability to take full advantage of resources to conduct high-throughput experiments. Since the deployment of Phase-Mapper, thousands of X-ray diffraction patterns have been processed and the results are yielding discovery of new materials for energy applications. These are exemplified by the discovery of a new family of metal oxide light absorbers in the previously unsolved Nb-Mn-V oxide system, which is provided here as a case study and is an illustrative example of the importance of encoding physical constraints to obtain physically meaningful phase diagram solutions. We believe Phase-Mapper will lead to further developments in high-throughput materials discovery by providing rapid and critical insights into the phase behavior of new materials.

Phase-Mapper: AI for Materials Discovery

An experimentation pipeline for rapidly synthesizing, characterizing, and identifying new materials is referred to as *high-throughput materials discovery* or *combinatorial materials discovery*. In this pipeline, a handful of elements are deposited together on a two-dimensional substrate, so that different locations on the substrate receive varying proportions of the elements. This smooth variation in elemental composition across the substrate gives rise to the forming of a discrete set of materials, each of which is present in particular regions of the substrate.

The deposition process is analogous to atomic spray paint, as mentioned earlier. Imagine red, green, and blue spray paint being simultaneously sprayed onto a surface (or wafer) with each color source placed at the vertex of an equilateral triangle. Near these vertices, the deposited color appears simply red, green, or blue, and throughout the area of the triangle a continuum of the possible colors are obtained, where each color on the spectrum exists at a unique point on the wafer. In the same manner, the deposited materials "library" contains a broad spectrum of compositions (given the starting elements), and the atoms in different composition regions may arrange in a unique way to form a unique "phase" whose properties differ from other materials, even other compositions and phases formed from the same elements. It is the hope that one of these new materials will have a composition and phase that exhibit the desired properties, and to fully understand the composition-phase-property relationships, the full phase map must be solved.

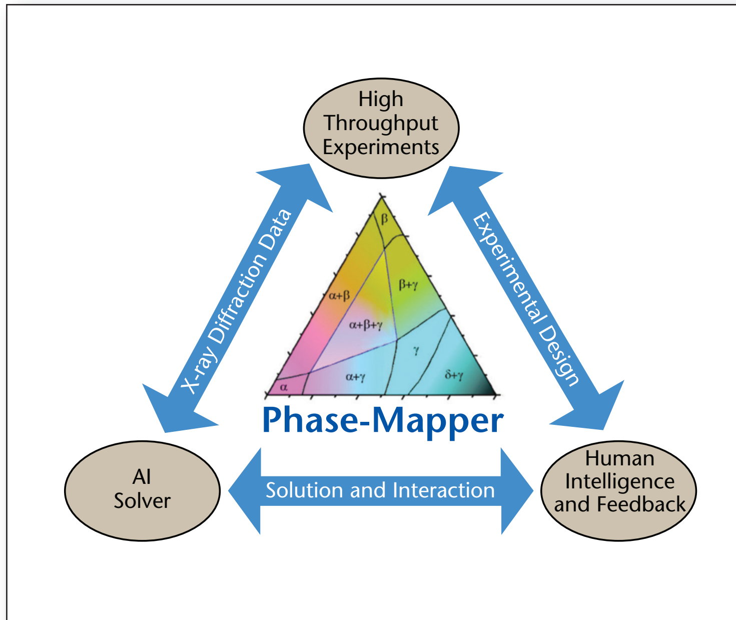

Figure 1. The Phase-Mapper Platform.

The Phase-Mapper platform integrates experimentation, AI problem solving, and human feedback into a platform for high-throughput materials discovery.

In the libraries being studied, the new materials are typically crystalline, meaning that at the atomic scale atoms are arranged in particular lattice structures, and the phase noted earlier is described by the symmetry and composition of the lattice structure. On a larger length scale, typically 5 to 500 nm, the lattice structure may alternate between two or three different structures, constituting a mixed-phase material. Each phase and phase mixture can exhibit unique properties, creating the need for materials scientists to understand, for each material library, how to categorize each material in the library (on the wafer) in term of its phase mixture

What data should materials scientists look at to determine the crystal structures? An indirect way of probing the microscopic structure is through X-ray diffraction. When X-rays are directed against a crystal, atomic layers will reflect the light; and for specific angles determined by the spacing between atomic layers, this reflected light will interfere constructively, giving rise to a strong signal. Thus, by scanning through all angles and measuring the reflected light, materials scientists are able to infer the structure of

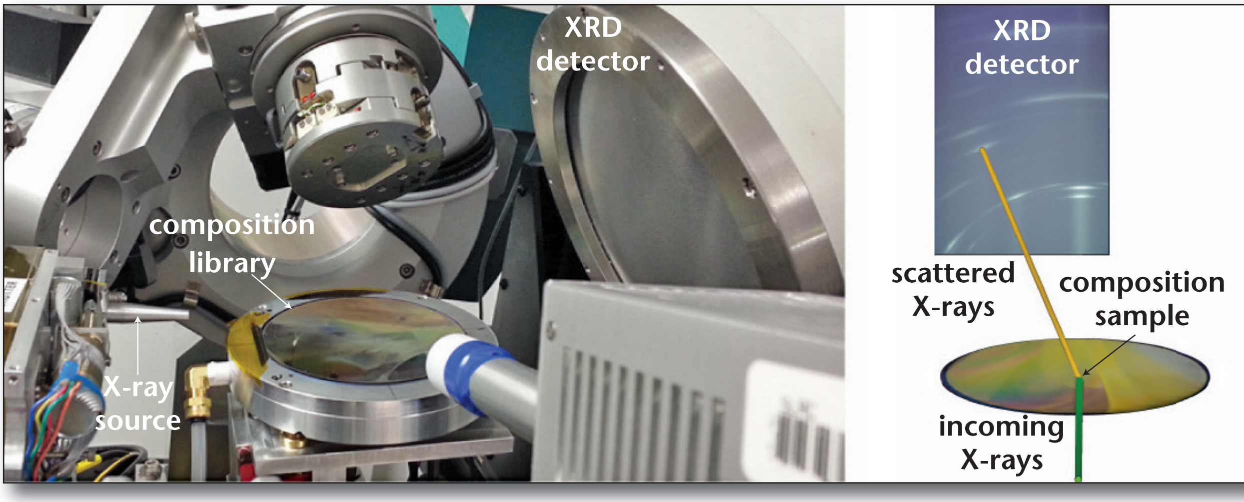

Figure 2. Apparatus for Taking XRD Measurements.

To the left is an image of the X-ray diffraction apparatus being used to characterize a composition library in the high-throughput experiment. On the right is an illustration of X-ray diffraction. Incoming X-rays hit the composition sample. They then diffract and are detected by the XRD detector.

the atomic layers without any microscopic measurements. This process is done for different points on a single wafer and because this process is fast when using synchrotron sources that provide a large flux of X-rays, materials scientists can rapidly characterize hundreds or thousands of materials a day using X-ray diffraction. An example of the apparatus for taking these XRD measurements can be seen in figure 2.

Phase Mapping

A key challenge in high-throughput materials discovery is to solve the phase-mapping problem, which identifies the characteristic XRD patterns of the materials (or basis patterns or crystal structures of the materials) that demix the XRD signals from the high-throughput experiments, some signals of which may be pure in that they represent one phase, while others are a linear combination of the pure phases. A visual description of the phase-mapping problem can be seen in figure 3.

Mathematically, the measured XRD pattern in the j-th sample point can be characterized by a one dimensional signal $A_j(q)$. The scattering vector magnitude (q) is a monotonic transformation of the diffraction angle, and is directly related to the spacing of atoms in a crystal. The phase-mapping problem is to find a small number of phases $W_1(q),\ldots, W_K(q)$, and their corresponding activation coefficients h_{ij}, such that the XRD patterns at each sample point can be explained by a linear combination of phases:

$$A_j(q) \approx \sum_{i=1}^{K} h_{ij} W_i(\lambda_{ij}, q) \tag{1}$$

The physical process of alloying complicates the linear combination by introducing additional scaling factors λ_{ij}. Alloying typically can be approximated by a multiplicative scaling of the XRD pattern of a specific phase in the q domain; we also refer to this process as *peak shifting*. We use the term $W_i(\lambda_{ij},q)$ to allow for the phases to scale slightly according to parameter λ_{ij} at each sample point. In addition to peak shifting, there are a number of other constraints on the solution of the phase-mapping problem, arising from the fact that the solution must describe a system constrained by the laws of physics. One important precept is the Gibbs phase rule, which limits the number of phases present to at most k phases per sample point, in a system involving k elements. Therefore, in a k-element system, no more than k coefficients among h_{ij} for fixed j may be nonzero. Additionally, feasible spatial variation of h_{ij} by composition, as well as the shapes that each W_i may take, are constrained by the relevant physics.

Fundamentally novel techniques are required to solve the phase-mapping problem quickly and accurately. Historically, the phase-mapping problem has been solved by hand, which can take days or months for a single system, and has become the bottleneck of the entire materials discovery workflow. A number of automatic techniques have been developed in recent years, which can be broadly grouped into clustering, constraint reasoning, and factor decomposition approaches. Proposed clustering methods such as hierarchical clustering (Long et al. 2007), dynamic time warping kernel clustering (Le Bras et al. 2011), and mean shift theory (Kusne et al. 2014) produce maps of phase regions, but fail to resolve mixtures or identify basis patterns, and do not necessarily produce results consistent with physics. Constraint reasoning approaches, including satisfiability modulo theory (SMT) methods (Ermon et al. 2012), can provide physically meaningful results, but depend heavily on effective preprocessing, such as peak identification, and are computationally intensive. Approaches based on nonnegative matrix factorization (Long et al. 2009) are computationally efficient, but generally perform poorly when peak-shifting phenomena are present, failing to produce physically meaningful solutions. CombiFD (Ermon et al. 2015) is another factor decomposition approach that uses combinatorial constraints to simultaneously enforce some of the physical rules and accommodate peak shifting, but requires solving a combinatorial problem in each descent step, and is therefore computationally expensive and does not enforce all the physics constraints.

Here, we describe Phase-Mapper, an AI platform for rapidly solving the phase-mapping problem, integrating three key components: (1) cutting-edge AI solvers, (2) human intelligence and feedback, and (3) high-throughput physical experiments. These components form an integrated process (see figure 1).

Phase-Mapper features novel solver AgileFD as a key component of the platform. Motivated by convolutive NMF, AgileFD includes a set of lightweight updating rules, and therefore a very fast gradient descent process. AgileFD is flexible, allowing for the incorporation of additional contraints, as well as human feedback through refinement. AgileFD can also run autonomously, producing physically meaningful solutions.

Phase-Mapper also provides tools for data exploration, visualization, and configuration that allow human experts as well as laypeople to analyze and improve solutions.

Phase-Mapper's solutions, obtained by the interaction between solvers and human users or autonomously, can also shed light on the development of new physical experiments. For example, the results can be incorporated into an active learning system, specifying regions of composition space to sample at higher resolution.

AgileFD: A Novel Phase-Mapping Solver

The Phase-Mapper platform features the AgileFD solver for the phase-mapping problem. AgileFD uses iterative updates of candidate solutions that are significantly faster than previously proposed methods. Human experts can interact with the algorithm in real time, and this speed is due to an efficient problem representation. Let the XRD patterns for all samples be represented by a matrix A, where each column corresponds to one sample point and each row corresponds to $A_j(q)$ for a particular value of q. Under the assumptions of no noise and no shifting, meaning that $\lambda_{ij} = 1$ for all i and j, describing A as a linear combination of a few basis patterns $W_i(q)$ is equivalent to factoring A as a product of two matrices W and H:

$$A \approx W \cdot H = R \tag{2}$$

Here, R denotes the approximate reconstruction of A. In this formulation, the columns of W form a set of basis patterns $W_i(q)$, and the columns of H correspond to the values hij in equation 1. We enforce nonnegativity for W and H, which is required for the solutions to be physically meaningful. Previous approaches to solve the phase-mapping problem based on NMF have been unsuccessful in handling peak shifting, where $\lambda_{ij} \neq 1$. The first contribution of AgileFD is to circumvent the shifting problem by a log space resampling. Under the variable transformation q into $\log q$, our signal becomes $W_i(\log q)$. More importantly, the shifted phase $W_i(\log \lambda q)$ becomes $W_i(\log \lambda + \log q)$, which transforms the multiplicative shift in the q domain into a constant additive offset. This allows the problem to be formulated in terms of convolutive nonnegative matrix factorization. After this variable substitution, we discretize the values of allowed λ and interpolate the signals at the corresponding geometric series of q values. The problem can then be written:

$$A \approx \sum_{m=1}^{M} W^{\downarrow m} \cdot H^m = R \tag{3}$$

With the columns of W representing the basis pat-

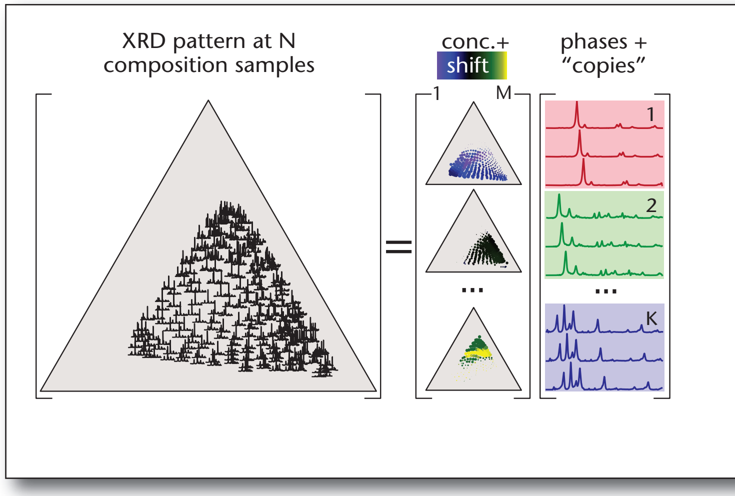

Figure 3. An Illustration of the Phase-Mapping Problem.

Given a material system with XRD data read at discrete points, find a set of basis phases, such that every point's XRD data can be made by a linear combination of the basis phases. Here, the left image is the original data, the right image is the found basis phases, and the middle image represents how much of a particular phase (the "phase concentration") is present at each data point along with the composition-dependent shifting.

terns, $W^{\downarrow m}$ is the result of shifting the rows of the W matrix down m rows, and filling the displaced rows with zeros. This represents the basis patterns with a constant offset in the log q domain, and is equivalent to the original multiplicative shift in the q domain. The columns of H^m act as the activation of basis patterns for the basis patterns shifted down m units. Note that when $M = 1$, this formulation is equivalent to NMF, aside from the log transformation.

We can think of AgileFD as a family of algorithms that can be adapted to use different loss functions, regularization, and certain imposed constraints. Equation 2 is adapted from convolutive NMF, which was first proposed to analyze audio signals (Smaragdis 2004). The phase-mapping problem differs from previous applications of cNMF for blind source separation as the log q domain is substituted for the time domain, and each source (phase) is expected to appear at most once per sample with a relatively small offset. As in cNMF, AgileFD uses a gradient descent approach to fit W and H. When a generalized Kullback-Leibler (KL) divergence is used in the objective function, gradient updates can be written multiplicatively, and are applied iteratively until convergence. See Xue et al. (2017) for further details.

Lightweight Update Rules

AgileFD's linear gradient update rules solutions typically converge within minutes. This is orders of magnitude faster than CombiFD, which uses a similar problem formulation but with combinatorial constraints explicitly enforced globally, using a mixed-integer programming (MIP) representation. This increased efficiency of AgileFD enables high-throughput analysis and also makes it possible for a human to interact with the system in almost real time.

Further Extensions of AgileFD for Materials Discovery

The ultimate aim of the phase-mapping problem is to find a physically meaningful decomposition of the signal. In the next few sections, we provide a number of novel modifications to the basic AgileFD algorithm, in order to impose prior knowledge or additional constraints derived from user interpretation of a proposed solution.

AgileFD with Frozen Values

In the Phase-Mapper platform, the user is provided with the opportunity to freeze individual values in the W and H matrices. For example, a user might specify a known pattern or part of a previous solution as a basis pattern a priori, freezing the corresponding row or part thereof of W. Or the user might specify that a certain set of samples contain only a single phase and set the corresponding H values to zero. The result is interactive, iterative matrix factorization.

Custom Initialization

By initializing basis patterns or coefficients to values close to the expected solution, rather than random values, the user can direct the search to the correct solution space. We allow the user to specify basis patterns that can be taken from previous solutions, from data samples, or be provided manually, to use as an initial value. Similarly, initial values for the activation matrix can be specified.

Sparsity Regularization

Sparse solutions are more consistent with the underlying physics and are also usually more easily interpreted. The Phase-Mapper system provides the option to introduce a soft penalty term for sparsity in H, which can vary by index according to a human expert's preferences.

The Gibbs Phase Rule

In general, correct solutions to the phase-mapping problem should follow the Gibbs phase rule, which specifies that the number of observed phases at a given chemical composition is no more than the number of chemical elements N_{el}:

$$\sum_i I_{i,j} \leq N_{el} \tag{4}$$

Here, $I_{i,j}$ is an indicator of whether phase i is present at sample location j. Materials scientists might also know a priori, or infer from previous proposed solutions, that certain regions contain fewer phases than the usual limit.

Such combinatorial constraints cannot be encoded directly in the update rules of AgileFD, which has been used in previous methods such as CombiFD. However, these encodings result in a slow update process, as we have to solve a MIP problem in each iteration. As a novel routine, we apply the Gibbs phase rule by first solving the relaxed problem, then choose the best values to set to zero in H, and then refine the solution by applying the update rules until convergence. Because the update rules are multiplicative, the zeroed values will remain zero.

The value to set to zero in H is independent for each sample point j. This can be solved greedily if a faster solution is desired, or using a MIP formulation, which results in a small MIP program, or successive rounds of constraints and refinement, if a more precise solution is desired with a not much longer wait time. In addition, in the presence of alloying, the number of possible phases is further reduced by one. This alloying rule can also be captured with a small MIP program. These extensions are particularly useful when the unconstrained algorithm recovers a solution that is nearly correct except for relatively small violations of phase limits. See details in Bai et al. (2017).

Figure 4. Screenshot of the Phase-Mapper Web Application.

Displayed in the top left is the Nb-Mn-VO system in composition space. Each dot corresponds to an XRD measurement. The dots highlighted in red correspond to selected XRD measurements. The top right plot displays the heatmap of the selected data points' XRD patterns, that is, it visualizes the XRDs for all red sample locations in the top left panel. The bottom left plot shows the basis phases found in the loaded solution of a particular data point. The bottom right plot shows the phase fields for the system.

Phase-Mapper: A Human-Machine Integrated Platform

In this section, we present the Phase-Mapper platform workflow, which includes visualizing and analyzing an instance file, setting the solver framework, analyzing the solution, and using that analysis to update the solver framework. The design objectives were simple: create a practical application that seamlessly connects a visualization system with a powerful solver that allows for interactive and large-scale use. The main features of Phase-Mapper are the visualization tools and the solver interface. The interface of the application can be seen in figure 4.

With Phase-Mapper, both the input materials systems and generated solutions can be visualized in the same application. When an instance file of a materials system is uploaded to the system, the visualizer will generate a composition map, which illustrates the varying compositions of elements for all sample points. The user can freely inspect the XRD patterns of each sample point, as well as the heatmap of XRD patterns for a slice of sample points. A slice heatmap example is shown in the top left plot of figure 4, where a selected slice is indicated by the red data points. The heatmap plot on the top right of figure 4 represents the XRD patterns of the sample points in the slice.

When the solution files are loaded into the application, either uploaded by the user or generated by the solver, two new plots are generated: (1) the basis patterns that were found as solutions and (2) a composition map displaying the mixture proportions.

Connection to Solver

The solving feature of Phase-Mapper enables users to interact with the AI solver behind the scenes. The user can specify many solver parameters, such as how much to enforce sparsity, how many phases the solution should have, and how much shift between basis patterns the solver should allow. The user can also specify initial or frozen values to use as basis patterns. Incorporating user inputs helps the solver improve efficiency and accuracy. We provide tools for expert users to start the solver off closer to a solution, or to distort the solution space so the solver finds a more accurate solution.

Algorithm Effectiveness

In this section, we evaluate many solvers that were proposed in recent years against our Phase-Mapper system. We tested NMF (implemented as AgileFD with $M = 1$), AgileFD, AgileFD with sparsity regularization (AgileFD-Sp), AgileFD with sparsity and the Gibbs phase rule enforced (AgileFD-Sp-Gibbs), and CombiFD (Ermon et. al. 2015). We generated synthetic ternary metallic systems using data provided by the Materials Project (Jain et al. 2013), which provides crystal structure information and energy of formation using density functional theory for each phase. We applied a stylized model of solid solubility and used structure interpolation to simulate modified phase diagrams that include the additional degrees of freedom from alloying. We calculated XRD patterns for each modified constituent, including their interpolated structures, using pymatgen (Ong et al. 2013).

The quality of a solution is judged by how well each sample's reconstructed signal matches the corresponding measured signal. We find the permutation of the phases in the solution to best match the ground truth. In general, we observe that the solutions found by AgileFD (including AgileFD-Sp and AgileFD-Sp-Gibbs) better match the ground truth when compared with NMF and CombiFD. NMF underperforms because it cannot model peak shifting. Despite the fact that CombiFD also captures some of the physical constraints, it does not scale well because it formulates the physical constraints using mixed-integer programming. AgileFD with extensions (AgileFD-Sp and AgileFD-Sp-Gibbs) outperform vanilla AgileFD. They are able to find solutions that better match the physical constraints.

Illustrative Example: Discovery of Nb-V-Mn Oxides Light Absorbers for Energy Applications

The integration of the rapid solver with visualization tools enables materials scientists to interact with the data in a variety of ways. The web-accessible visualization tools enable rapid data exploration by materials scientists, which empowers materials scientists to inject their expert knowledge into the solution, for example by specifying the number of phases, the extent of alloying-based peak shifting, or the known existence of a phase in a certain composition region. In this way, Phase-Mapper can run in unsupervised or semisupervised modes per the availability of prior knowledge. To demonstrate the phase-mapping capabilities and the importance of the Gibbs constraint, figure 5 contains solutions for the phase map of 317 XRD patterns in the Nb-V-Mn oxide composition space using $M = 10$ shifted versions, which corresponds to approximately 2 percent alloying-based peak shifting. Although the phase behavior of binary subcompositions (for example, Nb-V oxides) has been previously studied, the ternary compositions are being explored for the first time to discover solar light absorbers for energy applications. Materials researchers were unable to obtain a meaningful phase diagram using manual analysis of this data set, even with advanced visualization tools, primarily because there are a number of phases with somewhat similar basis patterns, and most basis patterns contain dozens of peaks, yielding a collection of XRD patterns that are rich in information, but that exceed human conceptualization.

We show in the paper by Suram et al. (2016) that without accounting for alloying-based peak shifting, solutions are not meaningful in a number of ways, most notably the basis patterns do not correspond to individual phases because the intensity for a phase whose patterns shift across the data set is spread out over multiple basis patterns, creating phase-mixed basis patterns that are as difficult to interpret as the mixed-phase patterns in the raw data. The overlapping features in the basis patterns amplify this problem and result in its persistence even when alloying-based peak shifting is taken into account. When two phases have overlapping features in their basis patterns and the phases coexist in a range of compositions, approximately equal data reconstructions can be obtained using basis patterns that each contain one phase or that each contain a mixture of phases. To empower the algorithm to overcome this degeneracy in phase map solutions, we additionally apply the Gibbs constraint on the number of phases that can coexist in each composition sample. Figure 5 shows the basis patterns without (top) and with (bottom) application of this constraint, with one basis pattern on top highlighted to show that it contains a mixture of phases. So although this constraint is applied on the activations of the basis patterns, it indirectly makes the basis patterns more physically meaningful. The Phase-Mapper solution also exhibits excellent composition space connectivity for each phase concentration map, as expected for equilibrium phase behavior, and it exhibits systematic com-

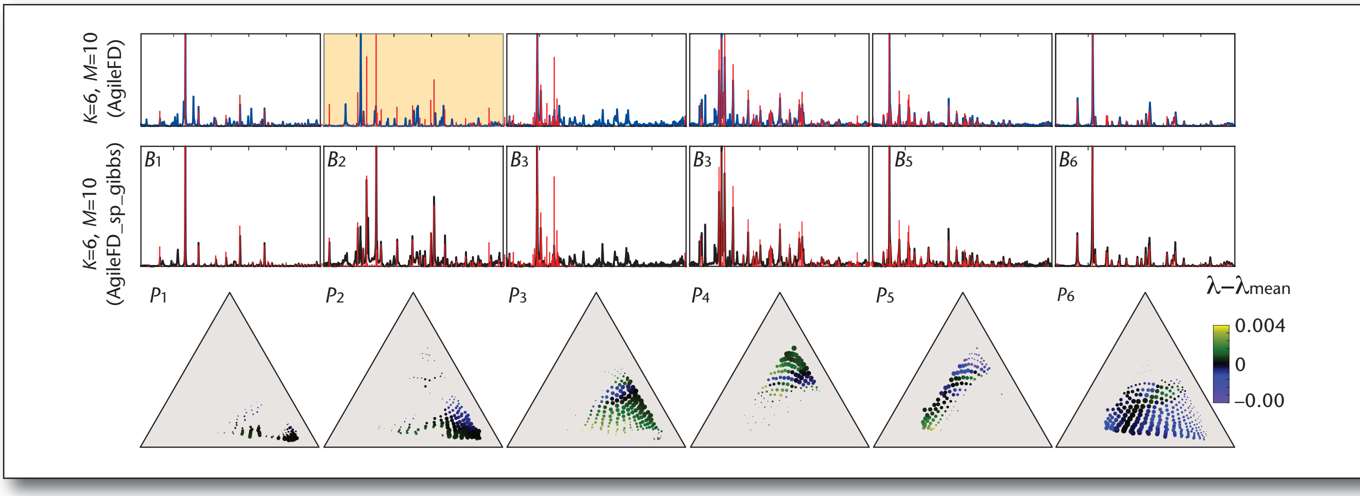

Figure 5. Solutions for the Phase Map of 317 XRD Patterns in the Nb-V-Mn Oxide Composition Space.

The 317 XRD patterns measured in the Nb-V-Mn oxide composition space were analyzed to produce phase-mapping solutions. On top, the six basis patterns obtained using AgileFD (blue) are shown along with the peak pattern (red sticks) for the phases identified by materials scientists. For pattern 2, this phase could be identified only after applying the Gibbs phase constraint, which, even though applied only to the basis pattern activations, results in the procurement of more meaningful (phase-pure) basis patterns. The middle row shows the six basis patterns obtained with AgileFD using Gibbs phase constraints (black), with the bottom row showing the composition map of activations of each basis pattern. There is a point for each of the 317 composition samples, with point size corresponding to the phase concentration and color corresponding to the alloying-based peak shifting. Nonshifted (nonalloyed) samples show as black.

positional variation in the shift parameter λ, demonstrating alloying within phases 3, 5, and 6. Incorporating both an alloying-based peak shifting model and the Gibbs phase constraint resulted in basis patterns and phase concentration maps that are physically meaningful, which is emblematic of a general strategy for AI-enhanced scientific discovery — injecting scientific knowledge where possible has trickle-down effects that result in scientifically meaningful solutions.

In addition to the broader implications of our algorithms for scientific discovery, the solutions in figures 5 and 6 are also emblematic of broader trends in materials science. As new technologies are conceived, new materials are needed, and often these materials are required to simultaneously exhibit a variety of properties and perform a variety of functions. At JCAP, researchers are pursuing the discovery of several materials, including photoanodes, which are materials that must absorb sunlight and harness its energy to oxidize water into oxygen, freeing protons and electrons to be utilized in fuel synthesis. Metal oxides, such as the compositions in the Nb-V-Mn oxide composition library discussed earlier, are excellent photoanode candidates because of their generally good stability under these conditions. Among the challenges in identifying and tailoring metal oxides to be effective photoanodes is the general difficulty in tuning their optical properties. The band gap energy of a given material dictates the range of sunlight that can be utilized by the material, and although materials with a variety of band gaps are available in the photovoltaic and light-emitting diode fields, such band gap tuning is quite rare in metal oxides. Pattern 4 in figure 5 corresponds to the MnV_2O_6 crystal structure, and as indicated by the activation map for this phase, it exists over a range of compositions where the peaks shift due to alloying. By combining this data with band gap measurements, figure 6 shows that within this single phase, the band gap can be tuned from approximately 1.9 to 2.1 eV. At the higher band gap range, the materials do not absorb orange or red light, but lowering the band gap enables these absorptions and thus increases the potential efficiency of the material. This discovery of alloying-based band gap tuning in this crystal structure is one component of the much broader portfolio of materials science needed to design and create photoanode materials. More generally, the discovery of a material for new technology is typically the culmination of a suite of smaller discoveries, and with AI algorithms such as those provided by Phase-Mapper, these discoveries are being accelerated and compiled to create a more comprehensive understanding of the underlying science, thus changing the arc of scientific discovery.

Conclusion

In this article, we show that the combination of high-throughput experimentation, AI problem solving, and human intelligence can yield rich scientific discoveries, with an application in materials science. A major, critically missing component of the high-

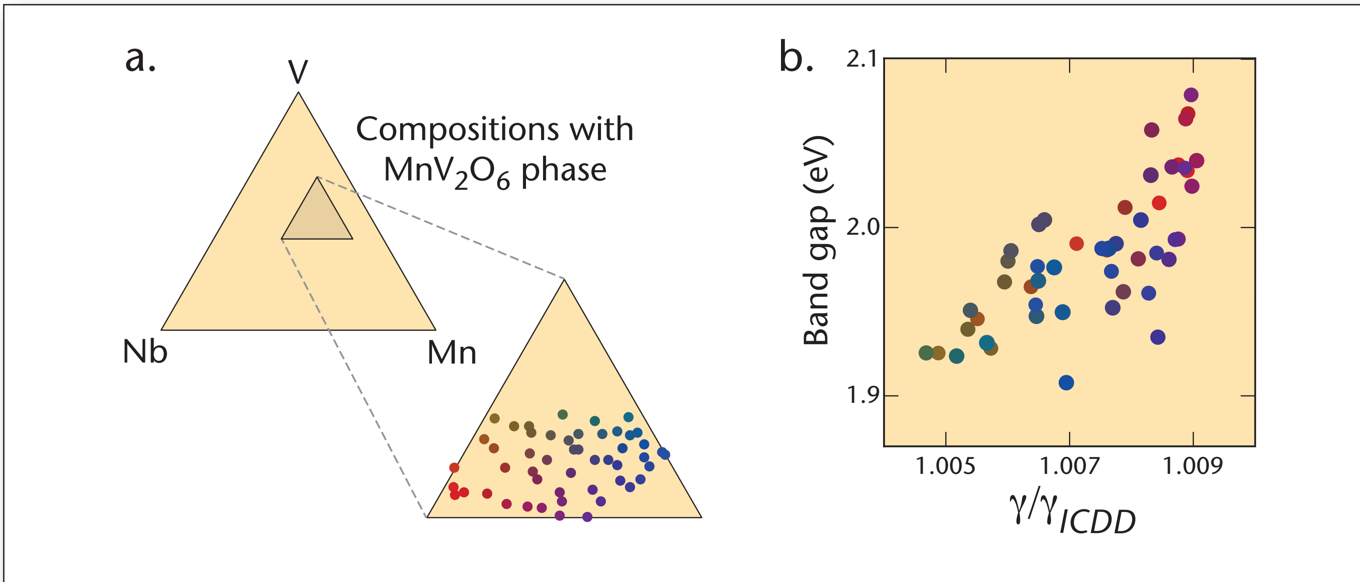

Figure 6. Band-Gap Tuning.

The noted composition region (a) of the Nb-V-Mn oxide composition library contains high-phase concentration of phase 4 from figure 5. This basis pattern was matched to the MnV_2O_6 structure and in (b) the composition points are plotted using the alloying-based shift parameter from the AgileFD with Gibbs constraint solution. This plot also notes the separately measured band gap energy of each sample, which determines the amount of solar light that can be absorbed. As the material composition changes, the composition in this phase changes and causes the volume of the phase and the band gap to change. This type of "band gap tuning" is quite rare in metal oxides, and this discovery was enabled by the detailed phase map provided by Phase-Mapper.

throughput materials discovery pipeline is the ability to rapidly solve the phase map identification problem, which involves the determination of the underlying phase diagram of a family of materials from their composition and structural characterization data. To address this challenge, we developed Phase-Mapper, a comprehensive platform that tightly integrates XRD experimentation, AI problem solving, and human intelligence. The AI solvers in Phase-Mapper provide high-quality solutions to the phase-mapping problem within minutes. These solutions can then be examined and further refined by materials scientists interactively and in real time. We have developed a novel solver, AgileFD, that features lightweight iterative updates of candidate solutions and a suite of adaptations to the multiplicative update rules. In particular, we have developed the ability to incorporate constraints that capture the physics of materials as well as human feedback, enabling functionalities well beyond traditional demixing techniques and producing physically meaningful solutions. Phase-Mapper has been deployed at the Department of Energy's Joint Center for Artificial Photosynthesis for materials scientists to solve a wide variety of real-world phase diagrams. Since the deployment of Phase-Mapper, thousands of X-ray diffraction patterns have been processed and the results are yielding the discovery of new materials for energy applications, as exemplified by the discovery of a new family of metal oxide solar light absorbers, among the previously unsolved Nb-Mn-V oxide system, which is provided here as an illustrative example. We believe Phase-Mapper will lead to further developments in high-throughput materials discovery by providing rapid and critical insights into the phase behavior of new materials.

Acknowledgements

This material is supported by NSF awards CCF-1522054 and CNS-0832782 (Expeditions), CNS-1059284 (Infrastructure), and IIS-1344201 (INSPIRE); and ARO award W911-NF-14-1-0498. Materials experiments are supported through the Office of Science of the U.S. Department of Energy under 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.

References

Bai, J.; Bjorck, J.; Xue, Y.; Suram, S.; Gregoire, J.; and Gomes, C. 2017. Relaxation Methods for Constrained Matrix Factorization Problems: Solving the Phase Mapping Problem in Materials Discovery. In *Proceedings of the 14th International Conference on Integration of Artificial Intelligence and Operations Research Techniques in Constraint Programming,* 104–

112. Cham, Germany: Springer. doi.org/10.1007/978-3-319-59776-8_9

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 on Theory and Applications of Satisfiability Testing,* 172–185. Cham, Germany: Springer. doi.org/10.1007/978-3-642-31612-8_14

Ermon, S.; Le Bras, R.; Santosh, S.; Gregoire, J. M.; Suram, S. K.; Gomes, C. P.; Selman, B.; and van Dover, R. B. 2015. Pattern Decomposition with Complex Combinatorial Constraints: Application to Materials Discovery. In *Proceedings of the 29th International Conference on Artificial Intelligence,* 636–643. Palo Alto, CA: AAAI Press.

Gomes, C. P. 2009. Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society. *The Bridge* 39(4): 5–13.

Green, M. L.; Takeuchi, I.; and Hattrick-Simpers, J. R. 2013. Applications of High Throughput (Combinatorial) Methodologies to Electronic, Magnetic, Optical, and Energy-Related Materials. *Journal of Applied Physics* 113: 231101. doi.org/10.1063/1.4803530

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. doi.org/10.1063/1.3274179

Gregoire, J. M.; Van Campen, D. G.; Miller, C. E.; Jones, R. J. R.; Suram, S. K.; and Mehta, A. 2014. High-Throughput Synchrotron X-Ray Diffraction for Combinatorial Phase Mapping. *Journal of Synchrotron Radiation* 21(6): 1262–1268. doi.org/10.1107/S1600577514016488

Hattrick-Simpers, J. R.; Gregoire, J. M.; and Kusne, A. G. 2016. Perspective: Composition-Structure-Property Mapping in High-Throughput Experiments: Turning Data into Knowledge. *APL Materials* 4(5): 053211. doi.org/10.1063/1.4950995

Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; and Persson, K. A. 2013. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Materials* 1(1): 011002. doi.org/10.1063/1.4812323

Kusne, A.; Gao, T.; Mehta, A.; Ke, L.; Cuong Nguyen, M.; Ho, K.-M.; Antropov, V.; Wang, C.-Z.; Kramer, M. J.; Long, C.; and Takeuchi, I. 2014. On-the-Fly Machine-Learning for High-Throughput Experiments: Search for Rare-Earth-Free Permanent Magnets. *Scientific Reports* 4: 6367. doi.org/10.1038/srep06367 doi.org/10.1038/srep06367

Le Bras, R.; Bernstein, R.; Gregoire, J. M.; Suram, S. K.; Gomes, C. P.; Selman, B.; and van Dover, R. B. 2014. A Computational Challenge Problem in Materials Discovery: Synthetic Problem Generator and Real-World Datasets. In *Proceedings of the 28th International Conference on Artificial Intelligence, [page numbers].* Palo Alto, CA: AAAI Press.

Le Bras, R.; Damoulas, T.; Gregoire, J.; Sabharwal, A.; Gomes, C.; and van Dover, R. 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,* 508–522. Cham, Germany: Springer.

Lee, D. D., and Seung, H. S. 2001. Algorithms for Nonnegative Matrix Factorization. In *Proceedings of the 15th Conference of Neural Information Processing Systems,* 556–562. Cambridge, MA: The MIT Press.Long, C. J.; Bunker, D.; Karen, V. L.; Li, X.; and Takeuchi, I. 2009. Rapid Identification of Structural Phases in Combinatorial Thin-Film Libraries Using X-Ray Diffraction and Nonnegative Matrix Factorization. *Review of Scientific Instruments* 80(10): 103902. doi.org/10.1063/1.3216809

Long, C. J.; Hattrick-Simpers, J.; Murakami, M.; Srivastava, R. C.; Takeuchi, I.; Karen, V. L.; and Li, X. 2007. Rapid Structural Mapping of Ternary Metallic Alloy Systems Using the Combinatorial Approach and Cluster Analysis. *Review of Scientific Instruments* 78(7): 072217. doi.org/10.1063/1.2755487

Mørup, M., and Schmidt, M. N. 2006. Sparse Nonnegative Matrix Factor 2-D Deconvolution. Technical Report, Department of Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark.

Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; and Ceder, G. 2013. Python Materials Genomics (Pymatgen): A Robust, Open Source Python Library for Materials Analysis. *Computational Materials Science* (68): 314–319. doi.org/10.1016/j.commatsci.2012.10.028

Smaragdis, P. 2004. Nonnegative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs. In *Proceedings of the 5th IEEE International Symposium on Independent Component Analysis and Blind Signal Separation*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. doi.org/10.1007/978-3-540-30110-3_63

Suram, S. K.; Xue, Y.; Bai, J.; Le Bras, R.; Rappazzo, B. H.; Bernstein, R.; Bjorck, J.; Zhou, L.; van Dover, R. B.; Gomes, C. P.; and Gregoire, J. M. 2016. Automated Phase Mapping with AgileFD and its Application to Light Absorber Discovery in the V-Mn-Nb Oxide System. *ACS Combinatorial Science* 19(1): 37–46. doi.org/10.1021/acscombsci.6b00153

Xue, Y.; Bai, J.; Le Bras, R.; Rappazzo, B.; Bernstein, R.; Bjorck, J.; Suram, S. K.; van Dover, R. B.; Gregoire, J.; and Gomes, C. P. 2017. Phase-Mapper: An AI Platform to Accelerate High Throughput Materials Discovery. In *Proceedings of the 31st International Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Xue, Y.; Ermon, S.; Gomes, C. P.; and Selman, B. 2015. Uncovering Hidden Structure through Parallel Problem Decomposition for the Set Basis Problem: Application to Materials Discovery. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence,* 146–154. Palo Alto, CA: AAAI Press.

Junwen Bai is a PhD student in the Department of Computer Science at Cornell University. He received his bachelor's degree in computer science from Zhiyuan College, Shanghai Jiao Tong University.

Yexiang Xue is a PhD student in the Department of Computer Science at Cornell University. His research focuses on developing intelligent systems that tightly integrate decision making with machine learning and probabilistic reasoning under uncertainty.

Johan Bjorck is a PhD student in the Department of Computer Science at Cornell University. He received his bachelor's degree in engineering physics from Chalmers University of Technology.

Ronan Le Bras is a research scientist at the Allen Institute for Artificial Intelligence (AI2). His research interests include computational methods for large-scale combinatorial optimization, reasoning, machine learning, and human computation.

Brendan Rappazzo is a research assistant at the Institute for Computational Sustainability, in the Department of Computer Science at Cornell University. He received his BS in bioengineering and biomedical engineering from University of Maryland, College Park.

Richard Bernstein is a research programmer/analyst at the Institute for Computational Sustainability, in the Department of Computer Science at Cornell University, and the project coordinator and IT manager for the Computational Sustainability Network (CompSustNet).

Santosh K. Suram is a research scientist IV at Toyota Research Institute and a member of the Computational Sustainability Network. Suram's expertise is in combining materials informatics and high-throughput experimentation for accelerated materials discovery, specifically for energy materials. Suram strongly believes that AI will play a pivotal role in accelerating materials discovery.

R. Bruce van Dover is the Walter S. Carpenter, Jr., professor of engineering and director of the Department of Materials Science and Engineering at Cornell University, a senior member of the IEEE, and a fellow of the American Physical Society. His research currently focuses on the synthesis and characterization of thin films, and on developing and using high-throughput techniques for the discovery of ionic conductors, fuel cell catalysts, and other energy-related materials.

John Gregoire leads the High Throughput Experimentation group at the California Institute of Technology. He is also the thrust coordinator for photoelectrocatalysis in the Joint Center for Artificial Photosynthesis and an associate director in the Computational Sustainability Network. Gregoire is an expert in the discovery of energy-related materials and a strong advocate for further adoption of AI in materials research.

Carla P. Gomes is a professor of computer science at Cornell University and the director of the Cornell Institute for Computational Sustainability. Gomes is the lead PI of an NSF Expeditions in Computing, establishing and nurturing the new field of computational sustainability. She is a fellow of AAAI, AAAS, and ACM.