Crystallography companion agent for high-throughput materials discovery

Phillip M. Maffettone,[1, 2, *] Lars Banko,[3] Peng Cui,[2] Yury Lysogorskiy,[4]
Marc A. Little,[2] Daniel Olds,[1] Alfred Ludwig,[3] and Andrew I. Cooper[2, †]

[1] *National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, New York 11973, USA*
[2] *Department of Chemistry and Materials Innovation Factory,*
University of Liverpool, Crown Street, Liverpool L69 7ZD, U.K.
[3] *Institute for Materials, Faculty of Mechanical Engineering,*
Ruhr University Bochum, 44801 Bochum, Germany
[4] *Interdisciplinary Centre for Advanced Materials Simulation (ICAMS), Ruhr University, 44801 Bochum, Germany*

(Dated: March 17, 2021)

The discovery of new structural and functional materials is driven by phase identification, often using X-ray diffraction (XRD). Automation has accelerated the rate of XRD measurements, greatly outpacing XRD analysis techniques that remain manual, time-consuming, error-prone, and impossible to scale. With the advent of autonomous robotic scientists or self-driving labs, contemporary techniques prohibit the integration of XRD. Here, we describe a computer program for the autonomous characterization of XRD data, driven by artificial intelligence (AI), for the discovery of new materials. Starting from structural databases, we train an ensemble model using a physically accurate synthetic dataset, which output probabilistic classifications—rather than absolutes—to overcome the overconfidence in traditional neural networks. This AI agent behaves as a companion to the researcher, improving accuracy and offering significant time savings. It was demonstrated on a diverse set of organic and inorganic materials characterization challenges. This innovation is directly applicable to inverse design approaches, robotic discovery systems, and can be immediately considered for other forms of characterization such as spectroscopy and the pair distribution function.

[*] pmaffetto@bnl.gov
[†] aicooper@liverpool.ac.uk

I. INTRODUCTION

Phase identification using X-ray diffraction (XRD) is a linchpin in the discovery of materials for diverse applications including batteries, catalysis, and pharmaceuticals. Automation has accelerated the rate of XRD measurements, greatly outpacing XRD analysis techniques that remain for the most part manual, time consuming, error prone, and impossible to scale. This prevents the integration of this essential technique with autonomous robotic searches or self-driving labs[1–3]. Artificial intelligence (AI) can assist in the classification of XRD patterns[4–13], but widespread adoption is challenging due to limited reproducibility beyond specific materials systems[5,9,13,14]. Here we report an AI approach for the autonomous phase identification of diffraction patterns that is accurate across both organic and inorganic materials systems. We created a crystallography companion agent (XCA)—an algorithm-powered tool to collaborate with the researcher—that achieves expert accuracy in real-time with the measurements, using both experimental and predicted crystal structures as inputs. XCA overcomes the overconfidence of traditional neural networks through a probabilistic strategy that can incorporate multimodal analysis. This is accomplished without manual human-labelled data and is robust against many sources of complexity in diffraction. It is also extendable to other forms of characterization that can be accurately simulated, such as spectroscopy. This development complements recent advancements in automation[1,2,15–21] and autonomous experimentation[1–3], thus enabling a key step in the accelerated materials analysis.

Even with the help of dedicated software, the analysis of XRD patterns to determine unknown phases is challenging, error prone, and time consuming. Multiple sources of aberration affect experimental XRD patterns, affecting peak shapes, positions, and intensities (Fig. 1a, b), leading to degenerate patterns. This is compounded by the problem of homometrics[22] (Fig. 1a), where multiple unique structures can equivalently explain an XRD pattern. Thus, a given crystal phase can correspond to many unique XRD patterns, and an XRD pattern can correspond to multiple unique phases. Some modern XRD instruments can measure hundreds, if not thousands, of patterns per hour, while the analysis of a single novel pattern can take hours or even days, and this introduces a significant bottleneck to discovery. Furthermore, both inorganic[23] and organic[24] crystal structure prediction (CSP) are increasingly useful tools for the prediction of stable crystalline materials with useful functional properties, but here the XRD data problem is further compounded by the need to match large numbers (1000's) of predicted patterns with experimental data that may not match the theoretical predictions precisely, either because of limitations in the predictions, the measurements, or often both.

Emerging big-data applications in materials science show promise as tools for XRD analysis. Pattern matching is commonly used to compare XRD patterns to reference structures[25–28], either by hand or with peak matching algorithms, sometimes incorporating additional constraints; for example, the Gibbs phase rule[6,29–32]. These methods do not account for common effects such as preferred orientation, peak shifting, or phase mixtures. Unsupervised methods attempt to statistically segregate experimental patterns for further analysis[4–8]. These methods are useful when there are no data on expected phases and they can be combined with traditional forms of structure solution. However, unsupervised methods are highly susceptible to experimental variation, which can lead to an overestimation of the number of distinct phases[5]. More recently, semi-supervised deep learning has been shown effective for inorganic XRD[9,11,14], convergent beam electron diffraction12 and electron backscatter diffraction[33]. Many of these supervised methods are reliant on large proprietary datasets[10,14,33], suffer from combinatorial explosion[10], and remain overconfident in their predictions, offering no measure of uncertainty[34]. These models make use of physical knowledge and are trained successfully on partially[9,35] or completely simulated datasets[10]. To date, these machine learning methods have only been demonstrated as accurate in specific domains where there are available test cases; that is, these approaches are only predictive for certain classes of materials, frequently inorganic oxides[5,9,10,14].

Our objective was to build a computer program to assist in phase identifications from large experimental XRD datasets of diverse materials systems. Such a tool requires rapid predictions, a high degree of automation, and accuracy on par with an expert crystallographer. If a realistic dataset can be synthesized, as is the case for diffraction, then supervised learning can be implemented without manually labelling data. A synthetic dataset needs to capture the underlying physics of the measurement and the diversity of patterns caused by experimental non-idealities (Fig. 1b). This approach is analogous to attempting to train a classification model to recognize photographs based on hand-drawn sketches; it is not feasible for sketches because they are insufficiently realistic and cannot be produced *en masse*, but such an approach is possible in the physical sciences (Fig. 1c). These datasets should be embedded in a model that is accurate, but not overconfident, and capable of integrating other prior information, such as composition. The entire protocol should also be easy to use by researchers who are not specialists in machine learning.

To address this challenge, we developed an autonomous crystallography companion agent (XCA) that learns from fully synthetic data and can predict phases from XRD patterns in real-time. The parameters that govern the dataset construction exploit the same prior knowledge of a researcher that is required for experiment preparation; for example, sample composition and diffraction instrument parameters. Under this paradigm, the scientist remains sovereign over the research while the companion agent autonomously prepares analyses under the researcher's direction.

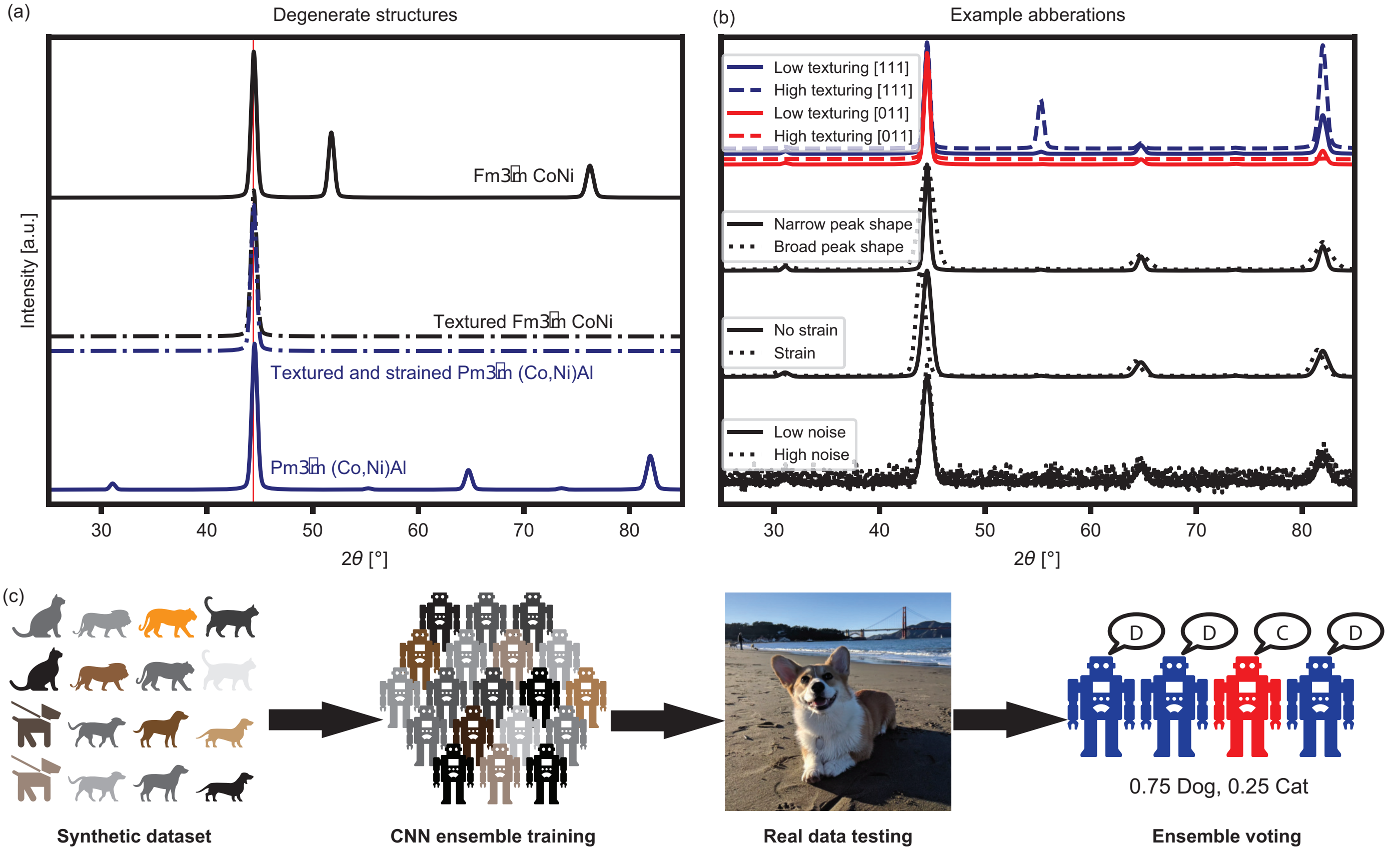

Figure 1. **Experimental XRD complexity and training an ensemble using synthetic data.** **a,** Different crystal phases can create identical XRD patterns under conditions that are common in thin-film experiments. **b,** There are many causes of aberration in an XRD pattern, such as intensity changes from preferred orientation, peak shifting from lattice strain or solid solutions, peak broadening, and background, as illustrated here for the Pm$\bar{3}$m phase of (Ni,Co)Al. **c,** The crystallography companion agent (XCA) statistically solves the problem of simultaneous experimental complexity and data scarcity by automatically building a synthetic dataset and training an ensemble of learners from this data. The dataset covers the scope of variation in XRD patterns and the ensemble model outputs an existence probability of each phase when tested against real data. This protocol is analogous to training a cat-vs-dog classifier on artistic sketches of the animals and testing on photographs. Unlike the sketch analogy, this training approach is possible for XRD because of the speed and accuracy of the simulations.

RESULTS

XCA generates a synthetic dataset from phases within databases that encompasses the range of experimental variation for a given materials system and experimental set-up (Fig. 1b). These data are then used to train an ensemble of fifty convolutional neural networks (CNNs) that output a probability distribution over the input phases, $P(\phi|\text{XRD})$ (Fig. 2). This can be conflated with independent distributions from calculated phase stability, thermodynamic constraints (such as Gibbs phase rule or phase diagram connectivity constraints), or multimodal analysis, exemplified here using energy dispersive X-ray spectroscopy (EDX). Building the dataset and model takes a few hours on a dedicated desktop. Analysis can then be conducted in realtime for each sample, or across each phase for a full experimental dataset.

Analysis proceeds on a phase-by-phase basis by mapping the likelihood of a phase across compositional space, or a sample-by-sample basis by exploring the probability of all phases in a given sample. This process is fully automated for finding pure phases of interest. The output offers a qualitative measure of phase mixing, and it is therefore suited for complete combinatorial phase mapping of multinary materials libraries. Starting from structural databases, XCA learns diffraction asynchronously with high-throughput experiments, and thus provides real-time, probabilistic analysis for the researcher. This creation of an AI tool that translates XRD measurements to probabilistic phase mapsis a necessary innovation to enable autonomous materials experimentation.

We successfully applied XCA to solve three separate materials challenges: detecting subtle symmetry transitions in

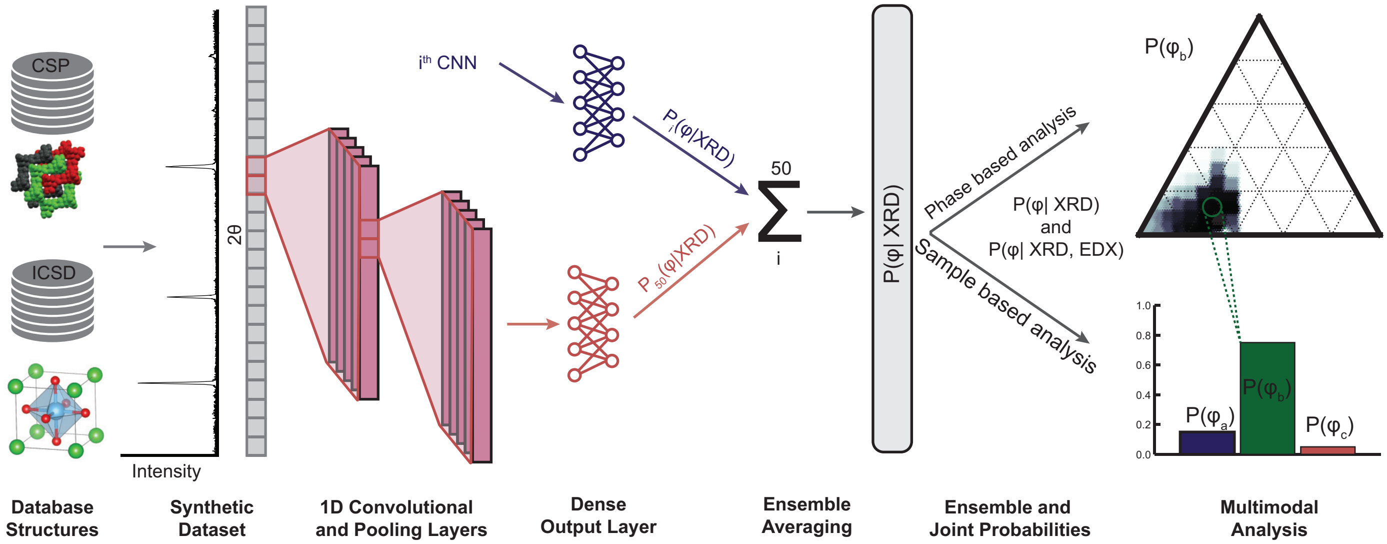

Figure 2. **Schematic of the crystallography companion agent (XCA).** Using only structures from databases and experimental information as inputs, XCA builds a realistic dataset of XRD patterns, and trains an ensemble of convolutional neural networks (CNNs). A single CNN learner is composed of alternating convolutional and pooling layers, followed by a single dense layer. A convolutional layer extracts features from its preceding layer, using filters learned during training, to form feature maps. The pooling down-samples the feature maps to exploit locality. The final feature maps are combined and fed to a dense layer—a simple type of classification model—that computes an output probability that the input diffraction pattern belongs to a given phase. The model averages the output of many learners to solve the problem of overconfidence in an individual CNN. Once trained, XCA will take an XRD pattern and output a probability, P, over all proposed phases in a few milliseconds. This $P(\phi|\mathrm{XRD})$ can then be conflated with other probability distributions; for example, those based on measured composition from EDX or, potentially, relative lattice energies from crystal structure prediction.

an inorganic ferroelectric, discovering organic polymorphs predicted *a priori* by computation, and mapping the phase space of a metal alloy system. These three challenges span a range of technical problems, with varying data quality and resolution. In the first example, classification across the phase transitions of $BaTiO_3$, a canonical ferroelectric (Fig. 3a), is not possible using traditional methods without expert intervention (Fig. 3b)[36], but we achieved this here using the XCA. High-throughput searches for new organic pharmaceuticals and other functional organic materials can be informed by CSP algorithms that predict energetically preferred crystal structures of a candidate molecule (Fig. 3c)[23,24], but XRD patterns for such materials often include a disordered background from amorphous or low-crystallinity impurities (Fig. 3d); also, the CSP-derived patterns may not agree precisely with experiment for the reasons outlined above. In our second example, we applied the XCA to adamantane-1,3,5,7-tetracarboxylic acid (ADTA), which forms a range of hydrogen bonded nets[37,38]. Out third example was phase mapping of a complete ternary alloy system, Ni-Co-Al, which requires high-throughput characterization following combinatorial synthesis (Fig. 3e)[39]. XRD patterns of thin-film samples suffer from significant and varying texturing, as well as peak shifting from the expected positions induced by strain, sample offset, and compositional variation according to Vegard's law (Fig. 3f). The systems are also different in their data quality, as the $BaTiO_3$ data was collected using synchrotron radiation, and the others with distinct in-house powder diffraction configurations. As such, each of these three challenges is technically distinct, but united by a need for rapid but high-quality analysis of large amounts of XRD data produced by high-throughput experiments.

A. Detecting subtle phase transitions in $BaTiO_3$

We first tested XCA with a temperature-dependent XRD experiment across a temperature range of 150 K to 450 K covering four phases of $BaTiO_3$: rhombahedral (R3m), orthorhombic (Amm2), tetrahedral (P4mm), and cubic (Pm$\bar{3}$m). $BaTiO_3$-based materials are a platform for probing the mechanisms of ferroelectric materials because their phase transitions occur at relatively low temperature[36,40]. Cooling from the paraelectric cubic phase induces three phase transitions that involve polarized displacements of the Ti ion along unique axes (Fig. 3a), resulting in nonobvious XRD peak splitting and shifts (Fig. 3b). From four refined initial phases, XCA outputs smoothly varying probabilities across each transition temperature, and successfully identifies the mixed-phase transitions in the dataset

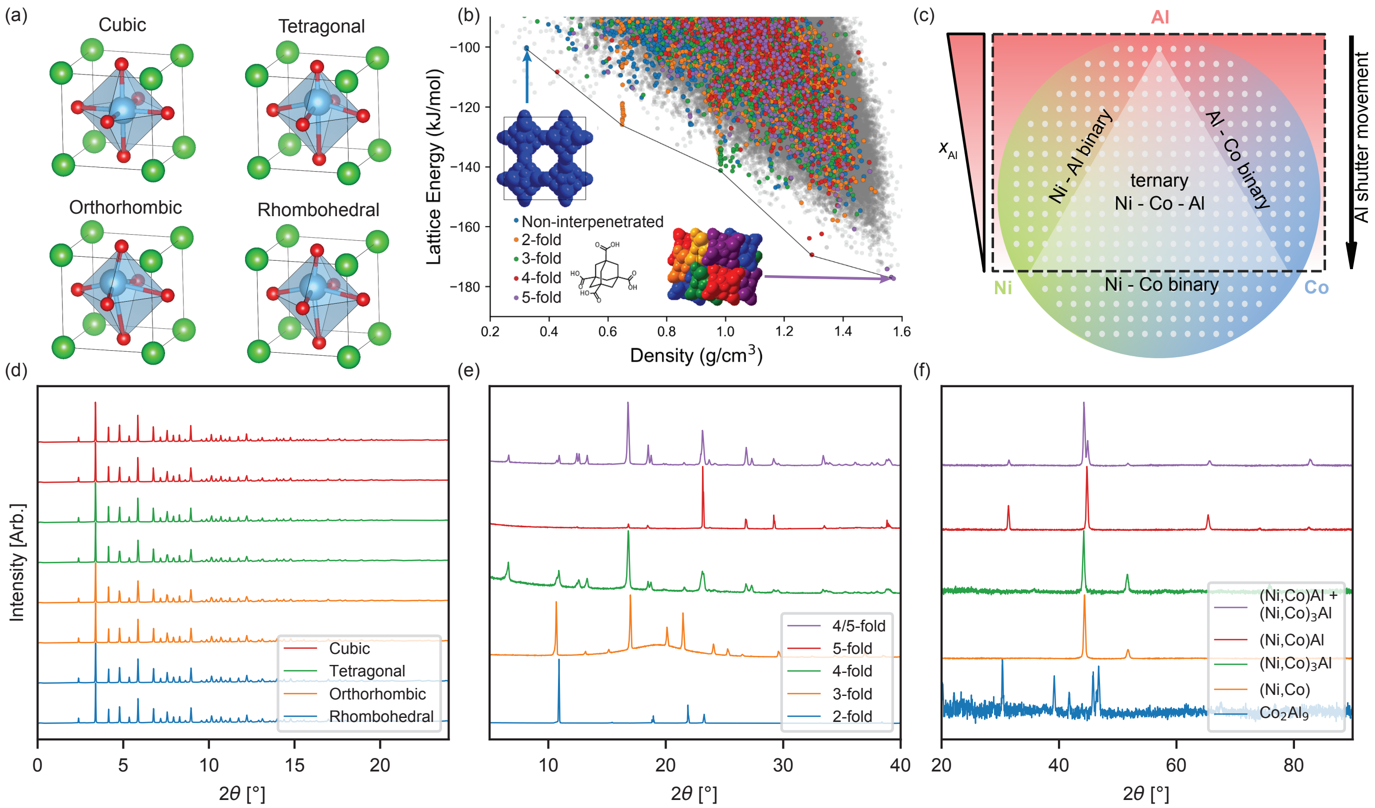

Figure 3. **Testing XCA against three different inorganic and organic materials challenges. a,** Phase transitions in BaTiO$_3$ involve symmetry breaking from a Ti translation. **d,** These phases are hard to distinguish by XRD. **b,** Crystal structure prediction for ADTA[35] identified five low-energy phases (end members shown here) with increasing degrees of interpenetration that were used as input for the XCA, along with XRD data from a high-throughput crystallization screen. **e,** The experimental XRD patterns from this organic polymorph screen are low symmetry in comparison the inorganic phases studied here and they are also often complicated by amorphous or low-crystallinity impurities. **c,** A combinatorial materials library comprising a complete ternary system and binary sub-systems prepared in a single experiment by multilayer wedge-type nanoscale film deposition and annealing. **f,** Thin film XRD patterns from the library suffer from preferred orientation, phase mixing, peak shifts according to Vegard's law, and variable noise from oxide and library edge effects.

(Fig. 4a). Almost all (56 of 60) of the classifications match the expectation, the only differences being accounted for by a temperature-lag between the measured and actual sample temperature during ramping. The smooth variations and automatic classification represent a major improvement on current Rietveld refinement procedures, which cannot automatically identify the phases and often require additional expert scrutiny (Fig. S2).

B. Searching for predicted phases in a CSP database

We next used XCA to search for predicted organic polymorphs in a CSP structural database. Polymorphism in organic crystals is important because different polymorphs of active pharmaceuticals, electronic molecules, and porous molecules can exhibit profoundly different physicochemical or physisorption properties[41]. In our previous study[38], high-throughput crystallization screening and XRD for a tetrahedral molecule, ADTA, generated 228 XRD patterns of varying quality. ADTA typically crystallizes to form a hydrogen-bonded network with a diamondoid topology, but CSP predicted[38] that polymorphism was likely because of the relatively small calculated energy gaps between the 5-, 4-, 3-, and 2-fold interpenetrated structures (Fig. 3c). To find these phases in our earlier study[38], the XRD patterns were searched iteratively by eye using the CSP dataset as a structural guide. This manual analysis of 228 experimental XRD patterns against the five lowest lattice energy CSP-derived patterns for the 5-, 4-, 3-, 2-, and 0-fold interpenetrated structures (labelled in Fig. 3c) required weeks of effort, yielding a labeled test set of 187 labeled patterns, and 41 patterns which could not be labeled. Starting from the same CSP dataset, and using the five labelled low-energy phases as inputs, XCA classified the pure phases and phase mixtures for which an expert classification was available with an experimental test accuracy of 0.952 in just a single day. The cosine similarity, which is a measure of

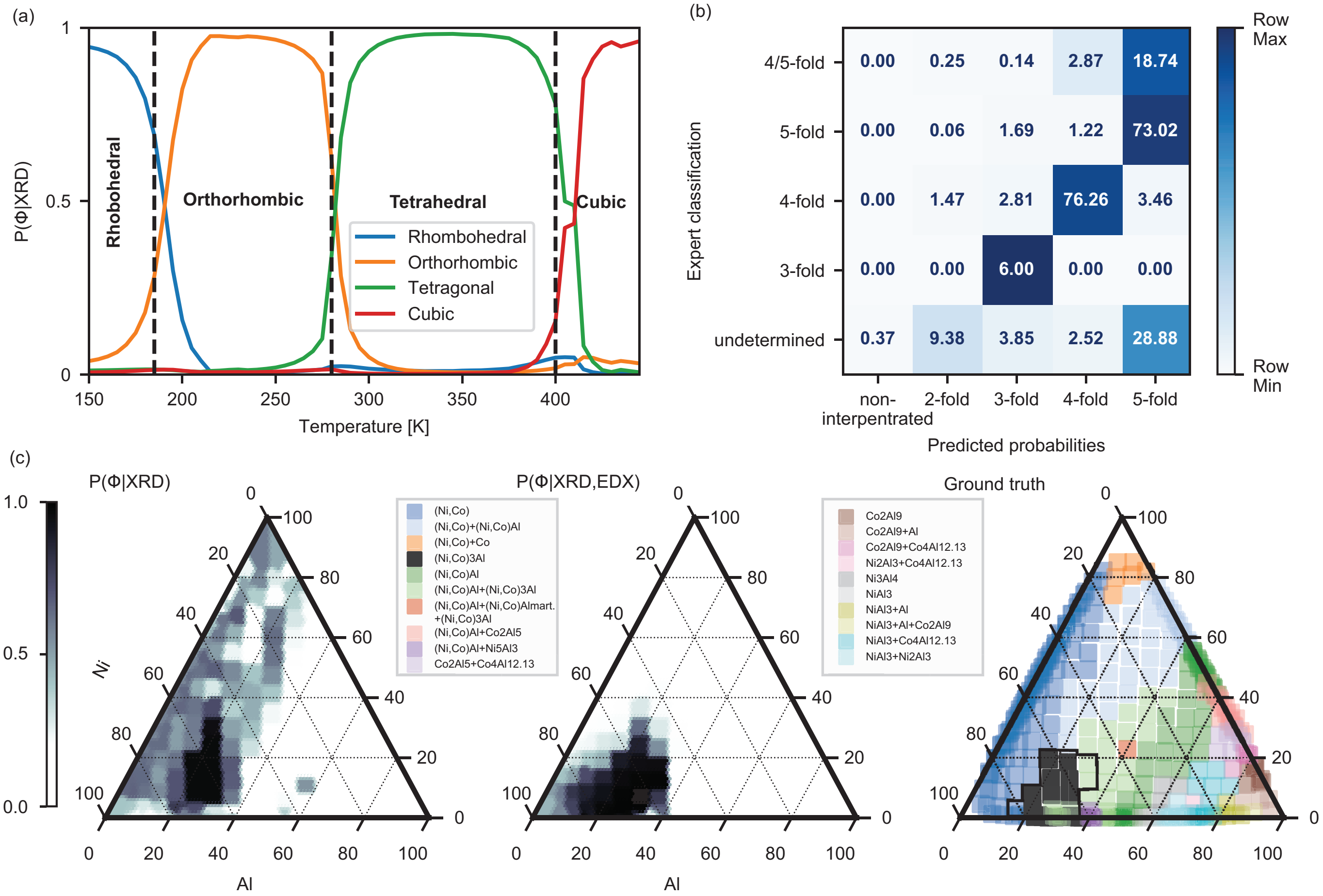

Figure 4. **Autonomous XRD analysis results from XCA. a,** XCA rapidly produces a probabilistic temperature-dependent phase mapping of $BaTiO_3$ that is more accurate than current refinement techniques. Dotted lines show the expected transition temperatures, and each colored line corresponds to the probability of a given phase existing. **b,** Confusion matrix showing the sum of predicted phase probabilities for each expert-classified phase of ADTA. **c,** Phase mapping for cubic Ni_3Al compared with the ground truth phase diagram with a black line outlining the probability region of $P(\phi_i|\mathrm{XRD},\mathrm{EDX}) >= 0.85$ (right). The XRD-based probability (left) captures the uncertainty associated with classifying this phase against another cubic phase with similar peak positions. The joint probability (centre) reduces the uncertainty by conflating prior information from composition.

alignment between the output probability the ground truth and the F1-score—a metric accounting for the effects of class imbalance—both reflect this accuracy (0.941 and 0.946, respectively). This demonstration of using accurate AI for organic crystal XRD classification is unique because of these samples' lower symmetry and amorphous backgrounds (Fig. 4b).

In the previously reported XRD dataset,[38] a precise match with the low-energy 2-fold phase from the CSP could not be found, but there were 41 experimental patterns that could not be characterized by the experts in reference to the CSP data. This was because variations in peak position and intensity provided insufficient information for confident classification. As shown in Figure 4b, 11 of these 41 patterns were classified by XCA as the same phase. Of these 11 samples, 4 were able to be produced as single crystals for more extensive XRD experiments. These structures were confirmed to be the 2-fold interpenetrated phase[38]. XCA was therefore able to correctly suggest the predicted, elusive 2-fold phase in the experimental XRD dataset, which could not be confidently classified by the research team. This result suggests wider opportunities for the discovery of low-energy polymorphs that are predicted by CSP but where the experimental data do not agree exactly with the computational predictions; for example, because of solvent inclusion, in the case of porous materials[42], or where the CSP does not fully capture molecular flexibility[43].

C. Using XCA to assist in phase mapping

Lastly, we used XCA for the phase mapping of a complete ternary inorganic system, Ni-Co-Al, where composition-structure-property relationships were previously identified across 21 phase regions (Fig.4c), see Figs. S5 to S24 for representative XRD patterns)[44]. Phases in Ni-Co-Al are of interest for different applications such as superalloys and ferromagnetic shape memory applications[44,45]. Identification of the compositional existence ranges of the phases and phase mixtures requires extensive analytical effort but is critical for these materials. Here, the XCA output probability was conflated with an independent probability, based on chemical composition from EDX,$P(\phi_i|\mathrm{EDX})$, to yield a joint probability, $P(\phi_i|\mathrm{XRD}, \mathrm{EDX})$.

When testing using the 12 phases found in the experiment, this $P(\phi_i|\mathrm{XRD}, \mathrm{EDX})$ approached the ground truth, with most misclassifications still assigning high—but not the highest—probability to the existence of phases in a sample for both pure regions and mixtures (*SI*). To demonstrate robustness of XCA when the existing phases are unknown, we tested XCA on all unique and experimentally accessible structures of all single element, binary and ternary combinations of Ni-Co-Al in the ICSD46 (31 phases, Table S4). Since nearly two thirds of these phases do not exist in the experiment, this approach under-performs (cosine similarity = 0.735, accuracy = 0.763, F1-score = 0.788); nonetheless, $> 90\%$ of the classifications contain the correct phase in the top three probabilities. This behaviour is similar to the pattern matching approaches that propose plausible phases, but here it is effective with non-ideal thin-films and phase mixtures. A task that previously took weeks to months of manual effort is now accelerated to take place within hours of computer time.

Where XCA and the expert disagree, additional information helps to make the correct classification. An example is shown by a low symmetry phase ($NiAl_3$) being fully textured along a specific axis, such that its pattern is commensurate with NiAl (Fig. S25): values of $P(\phi_i|\mathrm{XRD})$ and $P(\phi_i|\mathrm{XRD}, \mathrm{EDX})$ align with expert opinion, but are not informed by the literature or predicted phase stability. Since $P(\phi_i|\mathrm{EDX})$ only captures the average sample composition, both $P(\phi_i|\mathrm{XRD}, \mathrm{EDX})$ and $P(\phi_i|\mathrm{XRD})$ should be considered in tandem for a full phase map. As an example, Figure 4c compares the outputs for a representative phase (Ni_3Al) in the ternary composition space: while $P(\phi_i|\mathrm{XRD})$ extends to encompass degeneracy (Fig. 1a) and mixed phase regions, the $P(\phi_i|\mathrm{XRD}, \mathrm{EDX})$ is confined by the expected composition of pure Ni_3Al. As for $BaTiO_3$, this produces a probabilistic solution to instantiate a precise refinement, emphasizing XCA-researcher collaboration.

II. DISCUSSION

The strength of XCA stems from its combination of a probabilistic model for addressing uncertainty and use of physically relevant synthetic datasets, thus allowing for applications across physics, chemistry, and biology. Compared to cutting edge approaches, XCA is more accurate across materials systems We compared XCA's performance directly against the AutoXRD approach developed by Oviedo et al[9] considering all combinations of dataset synthesis and modelling (Fig. 5). Details of this experiment can be found in the supplementary information. For inorganic systems, this performance stems from the physically accurate training data, learner ensembling, and ability to incorporate additional probability. In the case of organic polymorphs, the data production pipeline is most important because there is less XRD degeneracy between phases.

The model comparison highlights the utility of different features of the XCA. When considering just the pure phases in the alloy dataset, the importance of ensembling and a physically accurate dataset is clear. The information scarcity is also on display: the lower variance offered by a smaller model for each agent in the ensemble prevents overfitting and allows for marginal gains in accuracy. The ensembling outperforms other models, especially when combined with an independent probability. The XCA approach is more accurate in the limiting case of textured phases producing degenerate patterns that we encounter here. In the case of the organic ADTA polymorphs—where the problem of degenerate solutions is less prevalent and there was no secondary probability distribution—there is less sensitivity to the architecture used. Here, it is clear that the data production pipeline is most impactful. This is unsurprising, as the AutoXRD approach was designed against high symmetry perovskite structures that produce fewer peaks in the powder patterns, and organic systems often tend to crystallize in lower symmetry. The relatively lower values for F1-score in most ADTA tests is a result of a class imbalance and the macro-averaging approach that is appropriate for pure phases. Overall, the strength of the XCA stems from its combination of a probabilistic model in the case of an uncertain problem, and using fully physically relevant synthetic datasets, thus allowing for applications across domains in the physical sciences.

Other approaches have attempted to use synthetic XRD data to train models[10,13,14]. Since these methods are proprietary, they cannot be compared directly; however, the study by Lee et al[10] purposefully avoided textured data, which is imperative to thin film diffraction. Recent work[12] incorporating constraints into variational autoencoders has been used to solve a phase diagram problem working from a pattern synthesis similar to that used by Oviedo et

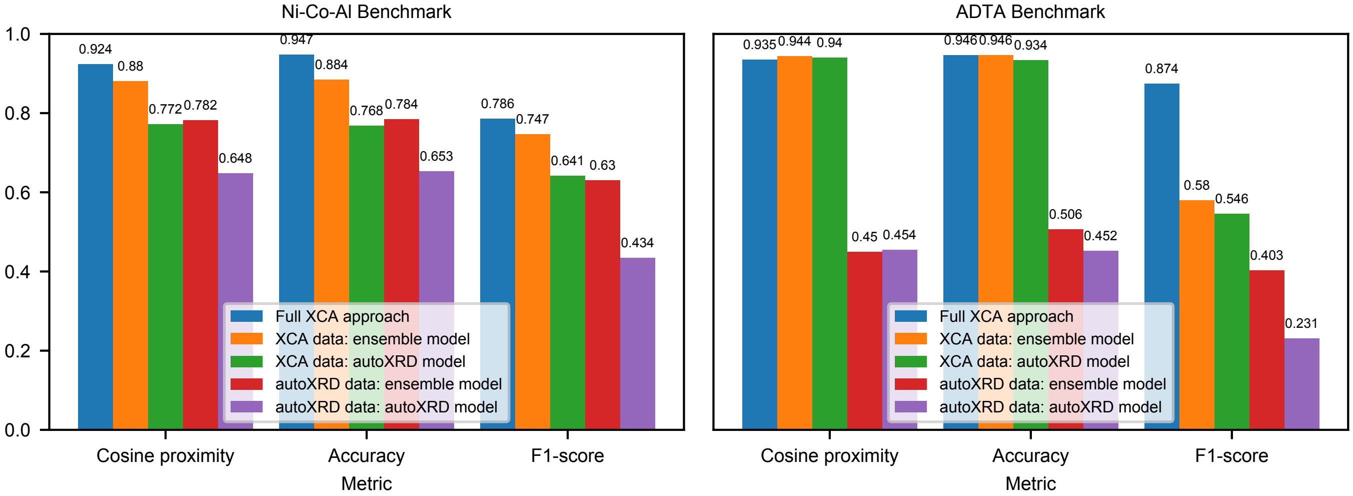

Figure 5. **Comparing different approaches to building a synthetic dataset and classifier.** The cosine proximity, accuracy, and F1-score (macro) are shown for (left) Ni-Co-Al and (right) ADTA pure phases using the full XCA approach, the XCA dataset with an ensemble of AutoXRD classifiers, the XCA dataset with a single AutoXRD classifier, the AutoXRD dataset with an ensemble of AutoXRD classifiers, and the AutoXRD dataset with a single AutoXRD classifer. The ensemble classification includes a joint probability distribution for the alloy system.

al.[9] This approach to demixing is a promising unsupervised approach for high symmetry materials, when there is no prior knowledge of phases and texturing is not a dominant challenge. The development of XCA builds on insights from AI, materials science, and crystallography, and is shown effective across organic and inorganic materials systems facing a wide array of experimental complexity.

A crucial component of XCA is the ensemble of learners, and their respective uncertainty. This uncertainty is necessary in cases where a single pattern has multiple plausible structural solutions. In the Figures S32-S36 and Figures S38-S40, we explore how the Shannon entropy (a measure of uncertainty) of the posterior distribution from a single learner and ensemble of learners is effected when the synthetic dataset doesn't capture the diversity of patterns produced by the experiment. Critically, uncertainty is also a requirement when the training data is perfect. As is the case with $BaTiO_3$, all four phases can be fit to every pattern with a Rietveld refinement (Fig. S2), so even an expert should express some scrutiny in how this decision is made. The ensembles in XCA accomplish this while still producing the correct classifications (Fig. S37). The problem is more complicated for the alloy system, since real uncertainty can arise from homometrics and inadequacies in the training data; however, the XCA methodology produces a rich phase map of uncertainty to accompany the existence phase map (Fig. S41).

Our methodology could be extended to any 1D response function in high-throughput materials research that requires classification and can be simulated at low cost (pair distribution function, X-ray photoelectron spectra, X-ray absorption near edge spectra, photoluminescence spectra, nuclear magnetic resonance spectra, mass spectra, etc.). Moreover, when XCA encounters unseen phases, it will tend to broaden its output probability distribution and maximize information entropy. To enhance this feature, future developments should diversify the architecture of individual learners and their data exposure. To enable materials discovery in the absence of predicted phases, XCA should be extended to pair with unsupervised methods that are not necessarily conditioned on a prior. The collaboration between a federation of agents, including XCA and other agents fit to task, will allow for advanced materials characterization to be incorporated into adaptive learning approaches for autonomous, data-guided experimentation.

In conclusion, we present an autonomous companion agent for the rapid, accurate classification of XRD datasets that is effective across materials domains, requires no labelling of experimental data, and is robust despite varying degrees of texture, peak shifting, peak broadening, phase mixing, and amorphous disorder. The agent was designed as a probabilistic approach for challenges with substantial uncertainty. It outputs phase maps over compositional space and discrete probability distributions per sample. It avoids the combinatorial explosion over mixtures by probabilistically learning about pure phase existence. The success of this approach is underpinned by ensembling ML models and the direct use of expert insight in the dataset development. As such, it can be extended to any analysis method where a rapid, accurate simulation is available. The XCA takes less than a day to train and enables real-time analysis during XRD measurements. It scales effectively for more data intensive challenges involving larger multidimensional search spaces, such as developing high entropy alloys[46] and complex solid solution electrocatalysts[47].

The innovation is directly applicable to inverse design approaches[48], new robotic discovery systems[1,2,23,49], and can be immediately considered for other forms of characterization such as spectroscopy and the pair distribution function.

III. ACKNOWLEDGEMENTS

We acknowledge financial support from the Engineering and Physical Sciences Research Council (EPSRC) (grant number EP/N004884/1) (P.M.M., M.A.L., A.I.C.), BNL Laboratory Directed Research and Development (LDRD) projects 20-032 "Accelerating materials discovery with total scattering via machine learning" (P.M.M., D.O.), the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design (P.C., A.I.C.), and German Research Foundation (DFG) as part of the Collaborative Research Centre TRR87/3 "Pulsed high power plasmas for the synthesis of nanostructured functional layers" (SFB-TR 87), project C2 (L.B., Y.L., A.L.). This research utilized the PDF (28-ID-1) Beamline and resources of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. We thank ZGH (Zentrum für Grenzflächendominierte Höchstleistungswerkstoffe, Ruhr-Universität Bochum) and Diamond Light Source for access to beamlines I19 (MT15777) and I11 (EE17193) for XRD measurements.

IV. AUTHOR CONTRIBUTIONS

P.M.M., L.B. and Y.L. conceived the project. P.M.M. led the development of XCA and coordinated the research teams. L.B. contributed to development, prepared the alloy dataset, and guided the inorganic dataset synthesis. P.C. and M.L. crystallized ADTA, and measured XRD data. Y.L. advised the machine learning. D.O. measured the $BaTiO_3$ and advised the relevant studies. A.L. supervised the development and the alloy studies. A.I.C. supervised the development and organic materials studies. Data was interpreted by all authors and the manuscript was prepared by all authors.

V. METHODS

A. Synthetic dataset preparation

Dataset preparation proceeded by collecting a set of proposed phases—from the accessible composition space in the ICSD or the local energetic minima of a CSP landscape—as crystallographic information files, and developing a large experimentally relevant set of diffraction patterns that correspond to each pure phase for the given experiment. From the structural information (symmetry, lattice parameters, atomic positions, occupancies, and thermal displacement parameters), the multiplicities and structure factor can be calculated using the open-source computational crystallography toolbox (CCTBX)[50]. From the experimental geometry and set-up, Lorentz polarization and an optional extinction correction can be applied. We next applied preferred orientation randomly to each pattern. This was done by choosing a reflection plane contained in the experimental 2θ domain, and randomly varying the degree of texturing (described by the March parameter)[51]. This drastically increases the size of the dataset and allows for a full scope of texturing to be applied to a given phase. The peak shape is varied for each pattern using a pseudo-Voigt profile function with a random choice of mixing parameters and Caglioti parameters[51]. A background function is randomly varied using a Laurent series with degree 6 and order -2. The dataset generation is thus directly relevant to the experiment, encompassing the same parameters that would need to be refined during a Rietveld refinement, and depends only on a few user-defined bounds for random sampling that are inferred from experimental system: those for the background function, peak shape, and noise. For all three materials systems, 100,000 XRD patterns were simulated for each phase. Example dataset synthesis parameters are provided in the available code. For the $BaTiO_3$ experiment, the experimental data was high quality with good powder averaging, so the synthetic dataset included low noise, narrow peak width, moderate peak shift (even with high quality data, moderate peak shift enables robustness if the input phase has slightly incorrect dimensions), a linear background, and limited texturing. For the ADTA experiment, the primary concerns were preferred orientation, peak shift from solvation expansion, noise from background and diffuse scatter, divergent behavior at low 2θ (accommodated by the negative Laurent series terms), and broad peak shape from diffuse scatter. The alloy dataset synthesis was focused on texturing and noise, with a relatively constant background, and other terms equivalent to the ADTA synthesis.

B. Probabilistic model architecture

In order to limit the overconfidence of the model, we used an ensemble of 50 shallow CNN learners. Each learner contained 3 convolutional layers (8, 8, and 4 filters, respectively) with a fixed kernel size of 5 and stride of 2, followed by a dense layer the size of the number of phases. Dropout at a rate of 40% was applied to the penultimate layer during training. The final dense layers are averaged, yielding a discrete probability distribution. The networks are trained using the Adam optimizer[52] for 10 epochs. A Bayesian optimization scheme was applied to optimize the hyperparameters and learner architecture. Inputs for the optimization were the number of convolutional layers, the number of initial filters, the initial stride, the initial kernel size, the number of dense nodes, and dropout rate, as well as the rate of change for each parameter of the convolutional layer between layers. Since the validation dataset used to produce metrics for Bayesian optimization is a randomly sampled subset of the synthetic dataset, there was limited variation between training and validation. We abstained from using the test data in the Bayesian optimization scheme to avoid any 'data leakage'. As such XCA results are not significantly impacted by learner architecture.

In the case of Ni-Co-Al, EDX measurements were available for each of the 342 samples in the experimental library. These measurements are used to construct a probability, P, of phases, ϕ, given the EDX data,

$$P(\phi_i|\text{EDX}) = \prod_{\alpha} \exp\frac{(x_\alpha - x_\alpha^i)^2}{\sigma^2}, \tag{1}$$

where x_α is the measured mass fraction of component α, x_α^i is the mass fraction of component α in phase i, and σ is set such that the full-width-at-tenth-max is 0.5. Treating the output from the ensemble neural net as a probability, $P(\phi_i|\text{XRD})$, and given that these distributions are independent, a joint probability, $P(\phi_i|\text{XRD}, \text{EDX})$ is formed by a normalized product. This allows the model to probabilistically differentiate between two phases appearing similar in the XRD. For example, two cubic structures of different composition that may have significant preferred orientation and strain would have similar marginal probability given the diffraction pattern, yet inclusion of the auxiliary measurement dramatically reduces the likelihood of a nonexistent phase (Fig. 2). This emulates the thought process of a metallurgist in an explicitly probabilistic way, analysing a sample by considering all of the experimental information available. This can be extended to materials systems using spectroscopic measurements where key regions of interest can be mapped to the likelihood of an intermolecular configuration.

The models yield a discrete probability vector, which can than be compared against the expert classification. The test sets were not refined as quantitative mixtures, so pure classifications are converted to 1-hot vectors and multiclass classifications are converted to multi-hot vectors. We use three metrics to measure the utility of the approach. Since we are more interested in probabilities than absolute predictions (*i.e.* argmax) and we need to understand the handling of mixtures, we first use a cosine proximity between the prediction and the ground truth as a measurement of accuracy. In a pure system with a fully confident prediction this converges to the traditional accuracy metric, which is the fraction of maximum predicted probabilities which match the ground truth. In the case of mixtures, this would be the fraction of the of maximum predicted probabilities that at least appear in the mixture. Lastly, the F1-score is calculated from the global true positives, false negatives, and false positives for test sets limited to pure phases (macro-average), and aggregated from the contributions of all classes for test sets including multi-class mixtures (micro-average).

C. Data availability

The experimental datasets and code used for constructing the synthetic datasets are available as examples with the source code. Source Data for Figures 1, 3, and 4 is available with this manuscript.

D. Code availability

To facilitate the impact of this tool, the approach is kept entirely open-source under the BSD 3-clause license, and being embedded into data acquisition frameworks at central facilities (blueskyproject.io). Ongoing development of this tool is located at github.com/maffettone/xca. A release at the time of publication and example code for the results contained here can be found at github.com/bnl/pub-Maffettone_2020_08[53]. The Bayesian optimization code can be found at github.com/maffettone/bayes_opt.

E. Competing interests

The authors declare no competing interests.

[1] Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).

[2] MacLeod, B. P. *et al.* Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, 20 (2020).

[3] Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237–241 (2020).

[4] Iwasaki, Y., Kusne, A. G. & Takeuchi, I. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *npj Comput. Mater.* **3**, 1–9 (2017).

[5] Stanev, V. *et al.* Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. *npj Comput. Mater.* **4**, 1–10 (2018).

[6] Xiong, Z., He, Y., Hattrick-Simpers, J. R. & Hu, J. Automated phase segmentation for large-scale X-ray diffraction data using a graph-based phase segmentation (gphase) algorithm. *ACS Comb. Sci.* **19**, 137–144 (2017).

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

[8] Takeuchi, I. *et al.* Data management and visualization of X-ray diffraction spectra from thin film ternary composition spreads. *Rev. Sci. Instrum.* **76**, 062223 (2005).

[9] Oviedo, F. *et al.* Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **5**, 1–9 (2019).

[10] Lee, J.-W., Park, W. B., Lee, J. H., Singh, S. P. & Sohn, K.-S. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic xrd powder patterns. *Nat. Commun.* **11**, 86 (2020).

[11] Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **9**, 2775 (2018).

[12] Aguiar, J. A., Gong, M. L., Unocic, R. R., Tasdizen, T. & Miller, B. D. Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning. *Sci. Adv.* **5**, 10 (2019).

[13] Di Chen *et al.* Deep reasoning networks for unsupervised pattern de-mixing with constraint reasoning. In Bach, F. & Blei, D. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research* (PMLR, 2020).

[14] Park, W. B. *et al.* Classification of crystal structure using a convolutional neural network. *IUCrJ* **4**, 486–494 (2017).

[15] King, R. D. Rise of the robo scientists. *Scientific American* **304**, 72–77 (2011).

[16] Li, J. *et al.* Synthesis of many different types of organic small molecules using one automated process. *Science* **347**, 1221–1226 (2015).

[17] Dragone, V., Sans, V., Henson, A. B., Granda, J. M. & Cronin, L. An autonomous organic reaction search engine for chemical reactivity. *Nat. Commun.* **8**, 1–8 (2017).

[18] Buenconsejo, P. J. S. & Ludwig, A. Composition–structure–function diagrams of Ti–Ni–Au thin film shape memory alloys. *ACS Combi. Sci.* **16**, 678–685 (2014).

[19] Langner, S. *et al.* Beyond ternary opv: High-throughput experimentation and self-driving laboratories optimize multicomponent systems. *Adv. Mater.* **32**, 1907801 (2020).

[20] Steiner, S. *et al.* Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363** (2019).

[21] Bédard, A.-C. *et al.* Reconfigurable system for automated optimization of diverse chemical reactions. *Science* **361**, 1220–1225 (2018).

[22] Patterson, A. L. Homometric structures. *Nature* **143**, 939 (1939).

[23] Collins, C. *et al.* Accelerated discovery of two crystal structure types in a complex inorganic phase field. *Nature* **546**, 280–284 (2017).

[24] Pulido, A. *et al.* Functional materials discovery using energy–structure–function maps. *Nature* **543**, 657 (2017).

[25] Ivanisevic, I., Bugay, D. E. & Bates, S. On pattern matching of X-ray powder diffraction data. *J. Phys. Chem. B* **109**, 7781–7787 (2005).

[26] Huang, T. C. & Parrish, W. A new computer algorithm for qualitative X-ray powder diffraction analysis. *Advances in X-ray Analysis* **25**, 213–219 (1981).

[27] Gregoire, J. M., Dale, D. & van Dover, R. B. A wavelet transform algorithm for peak detection and application to powder X-ray diffraction data. *Rev. Sci. Instrum.* **82**, 015105 (2011).

[28] Stein, H. S., Jiao, S. & Ludwig, A. Expediting combinatorial data set analysis by combining human and algorithmic analysis. *ACS Comb. Sci.* **19**, 1–8 (2017).

[29] Ermon, S. *et al.* Pattern decomposition with complex combinatorial constraints: Application to materials discovery. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 636–643 (AAAI Press, 2015).

[30] Xue, Y. *et al.* Phase-mapper: An ai platform to accelerate high throughput materials discovery. In *29th Conference on Innovative Applications of Artificial Intelligence* (2017). URL `https://aaai.org/ocs/index.php/IAAI/IAAI17/paper/view/14799`.

[31] Kusne, A. G., Keller, D., Anderson, A., Zaban, A. & Takeuchi, I. High-throughput determination of structural phase diagram and constituent phases using grendel. *Nanotechnology* **26**, 444002 (2015).

[32] Suram, S. K. *et al.* Automated phase mapping with agilefd and its application to light absorber discovery in the V–Mn–Nb oxide system. *ACS Comb. Sci.* **19**, 37–46 (2017).

[33] Kaufmann, K., Zhu, C., Rosengarten, A. S. & Vecchio, K. S. Deep neural network enabled space group identification in EBSD. *Microscopy and Microanalysis* 1–11 (2020).

[34] Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural network. In Bach, F. & Blei, D. (eds.) *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37 of *Proceedings of Machine Learning Research*, 1613–1622 (PMLR, Lille, France, 2015).

[35] Wang, H. *et al.* Rapid identification of X-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *J. Chem. Inf. Model.* **60**, 2004–2011 (2020).

[36] Page, K., Proffen, T., Niederberger, M. & Seshadri, R. Probing local dipoles and ligand structure in $BaTiO_3$ nanoparticles. *Chem. Matter.* **22**, 4386–4391 (2010).

[37] Ermer, O. Five-fold diamond structure of adamantane-1,3,5,7-tetracarboxylic acid. *J. Am. Chem. Soc.* **110**, 3747–3754 (1988).

[38] Cui, P. *et al.* Mining predicted crystal structure landscapes with high throughput crystallisation: old molecules, new insights. *Chem. Sci.* **10**, 9988–9997 (2019).

[39] Ludwig, A. Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. *npj Comput. Mater.* **5**, 70 (2019).

[40] Wegner, M., Gu, H., James, R. D. & Quandt, E. Correlation between phase compatibility and efficient energy conversion in Zr-doped barium titanate. *Scientific Reports* **10**, 3496 (2020).

[41] Bernstein, J. *Polymorphism in Molecular Crystals* (Oxford University Press, 2010).

[42] Slater, A. G. *et al.* Computationally-guided synthetic control over pore size in isostructural porous organic cages. *ACS Cent. Sci.* **3**, 734–742 (2017).

[43] Cui, P. *et al.* An expandable hydrogen-bonded organic framework characterized by three-dimensional electron diffraction. *J. Am. Chem. Soc.* **142**, 12743–12750 (2020).

[44] Decker, P., Naujoks, D., Langenkämper, D., Somsen, C. & Ludwig, A. High-throughput structural and functional characterization of the thin film materials system Ni-Co-Al. *ACS Comb. Sci.* **19**, 618–624 (2017).

[45] Naujoks, D. *et al.* Phase formation and oxidation behavior at 500 °C in a Ni-Co-Al thin-film materials library. *ACS Comb. Sci.* **18**, 575–582 (2016).

[46] Miracle, D. B. & Senkov, O. N. A critical review of high entropy alloys and related concepts. *Acta Materialia* **122**, 448–511 (2017).

[47] Löffler, T. *et al.* Toward a paradigm shift in electrocatalysis using complex solid solution nanoparticles. *ACS Energy Letters* **4**, 1206–1214 (2019).

[48] Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).

[49] Li, Z. *et al.* Robot-accelerated perovskite investigation and discovery. *Chem. Matter.* **32**, 5650–5663 (2020).

[50] Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. The *Computational Crystallography Toolbox*: crystallographic algorithms in a reusable software framework. *J. Appl. Cryst.* **35**, 126–136 (2002).

[51] Giacovazzo, C. (ed.) *Fundamentals of Crystallography* (Oxford University Press, 2011), 3rd edn.

[52] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. & LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015). URL http://arxiv.org/abs/1412.6980.

[53] Maffettone, P. M. *et al.* bnl/pub-maffettone_2020_08. https://doi.org/10.11578/dc.20210316.6 (2021). URL https://doi.org/10.11578/dc.20210316.6.