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

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

(Dated: August 4, 2020)

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 unprecedented time savings, and is demonstrated on a diverse set of organic and inorganic materials 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.

Phase identification using X-ray diffraction (XRD) is a linchpin in the discovery of new materials for diverse applications including batteries, catalysis, and pharmaceuticals. Automation has accelerated the rate of XRD measurements, greatly outpacing XRD analysis techniques that remain manual, time consuming, error prone, and impossible to scale, thus prohibiting the integration of this essential technique with autonomous robotic scientists 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]. 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 measurements. XCA overcomes the overconfidence of traditional neural networks through a probabilistic strategy that can incorporate multimodal analysis. This is accomplished without pre-labelled data, is robust against many sources of complexity in diffraction, and is extendable to other forms of characterization that can be accurately simulated (e.g., spectroscopy). This complements recent advancements in automation [11, 12, 15–21] and autonomous experimentation [11–13], and marks a crucial step in the acceleration of materials discovery.

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, altering peak shapes, positions, and intensities [Fig. 1a, b], leading to degenerate patterns. This is compounded by the problem of homometrics [22] [Fig. 1a]: 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. Compounding this challenge is the fact that while modern XRD instruments can measure hundreds if not thousands of patterns per hour, yet the analysis of a single novel pattern can take hours or even days, introducing significant bottleneck to discovery.

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 [22, 24], either by hand or with peak matching algorithms, sometimes incorporating additional constraints, e.g., the Gibbs phase rule, [4] [27, 30]. These methods, however, 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. These are useful when there are no data on expected phases and can be combined with traditional forms of structure solution; however, they are highly susceptible to experimental variation, leading to an overestimation of the number of phases [5]. More recently, semi-supervised deep learning has been shown effective for inorganic XRD [9, 11, 14], convergent beam electron diffraction [12] and electron backscatter diffraction [31]. Many of these supervised methods are reliant on large proprietary datasets [10, 32] suffer from combinatorial explosion [10], and remain over-confident in their predictions, offering no measure of uncertainty [33]. These models make use of physical knowledge and are trained successfully
Figure 1. **Experimental XRD complexity and training an ensemble from synthetic data.** (a) Different crystal phases can create identical XRD patterns under conditions common in thin-film experiments. (b) As shown for the Pm$\bar{3}$m phase of (Ni,Co)Al, there are many causes of aberration to an XRD pattern, such as intensity changes from preferred orientation, peak shifting from lattice strain or solid solutions, peak broadening, and background. (c) 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 analogy, it is possible for XRD due to the speed and accuracy of simulations.

on partially [9, 34] or completely simulated datasets [10]. To date, these machine learning methods have only been demonstrated as accurate in the domain of available test cases; that is, the above 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, automation, and accuracy on par with an expert crystallographer. A theme emerges from recent big-data advancements: if a realistic dataset can be synthesized — as is the case for diffraction — supervised learning can be implemented without manually labelling data. A synthetic dataset needs to capture the underlying physics of the measurement and 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 this 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 (e.g., composition). Moreover, the entire protocol must 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 (e.g., sample composition and diffraction instrument parameters). In this paradigm, the scientist remains sovereign over the research while the companion agent autonomously prepares analyses under the researchers direction.
High-throughput searches for new organic materials applied in pharmaceuticals, electronics, or molecular separations require rapid, but high-quality, analysis of vast amounts of XRD data produced by high-throughput experiments.

From phases within databases, XCA generates a synthetic dataset that encompasses the range of experimental variation for a given materials system and experimental set-up [Fig. 1]. 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 or multimodal analysis, exemplified here using energy dispersive X-Ray spectroscopy (EDX).

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 therefore is 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 maps is a key innovation that accelerates materials discovery and will enable autonomous experimentation.

We demonstrated the utility of XCA through its successful application in solving three unique materials challenges: detecting subtle symmetry transitions in an inorganic ferroelectric, discovering organic polymorphs predicted a priori by calculation, and mapping the phase space of an alloy system. Classification across the phase transitions of BaTiO$_3$, a canonical ferroelectric [Fig. 3a], is not possible using traditional methods without expert intervention [Fig. 3d] [35]. High-throughput searches for new organic materials applied in pharmaceuticals, electronics, or molecular separations — such as adamantane-1,3,5,7-tetracarboxylic acid (ADTA) — are informed by crystal structure prediction (CSP) algorithms that predict energetically preferred crystal structures of a candidate molecule [Fig. 3b] [36]. XRD patterns for these experiments often contain many peaks from relatively low symmetry materials and a disordered background from amorphous or low-crystallinity impurities [Fig. 3c]. Lastly, phase mapping of a complete ternary alloy system, Ni-Co-Al, requires high-throughput characterization following combinatorial synthesis [Fig. 3c] [37]. 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 Vegards law [Fig. 3f]. All of these challenges require rapid, but high-quality, analysis of vast amounts of XRD data produced by high-throughput experiments.

We first tested XCA with a temperature-dependent XRD experiment across a range of 150 K to 450 K covering four phases of BaTiO$_3$: rhombahedral (R3m), orthorhombic (Amm2), tetrahedral (P4mm), and cubic (Pm3m). BaTiO$_3$-...
Figure 3. Testing XCA with organic and inorganic materials challenges. (a) Phase transitions in BaTiO$_3$ involve symmetry breaking from a Ti translation. (d) The XRD patterns across these transitions are prohibitively difficult to distinguish. (b) The CSP of ADTA predicted five low-energy phases with increasing degrees of interpenetration that were searched for using a high-throughput crystallization screen. (e) Experimental XRD patterns from the organic polymorph screen are low symmetry and contain amorphous or low-crystallinity impurities. (c) 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 Vegards law, and variable noise from oxide and library edge effects.

Based materials are a research platform for probing the mechanisms of ferroelectric materials because their phase transitions exist at relatively low temperature [35, 38]. 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 [Fig. 4a]. Almost all (56 of 60) of the classifications match the expectation (the only differences accounted by a temperature-lag between the measured and actual sample temperature during ramping). The smooth variations and automatic classification represent a vast improvement on current refinement procedures, which cannot automatically identify the phases and often require additional expert scrutiny [Fig. S2].

We then 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 [39].

High-throughput crystallization screening and XRD of ADTA generated 228 XRD patterns [36]. ADTA typically crystallizes to form a hydrogen-bonded network with a diamondoid shaped topology, but CSP predicted that polymorphism was likely [Fig. 3b], due to the calculated energy rankings of 5-, 4-, 3-, and 2-fold interpenetrated structures. Initially to find these phases, the XRD patterns were iteratively searched by eye using the CSP dataset as a structural guide, over the course of months. Starting from the same CSP dataset, XCA classified both pure phases and mixtures with an accuracy of 0.952 in a single day. The cosine similarity — a measure of alignment between the output probability the ground truth — and F1-score — a metric accounting for the effects of class imbalance — reflect this accuracy (0.941 and 0.946, respectively). This is the first demonstration of accurate AI on organic crystal XRD [Fig. 4b].
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) The confusion matrix shows the sum of predicted phase probabilities for each expert classified phase of ADTA. Four of the undetermined structures suggested by the XCA as 2-fold were found to be this phase by a single crystal XRD experiment. (c) Phase mapping for cubic Ni$_3$Al compared with the ground truth phase diagram with a black line outlining the probability region of $P(\phi_i|\text{XRD, EDX}) > 0.85$ (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.

In the reported dataset, the 2-fold phase could not be found, yet some patterns could not be characterized by the experts: varied peak position and intensity provided insufficient information for confident classification. As shown in Figure 4b, 11 of these undetermined phases were suggested by XCA as the 2-fold phase. Four of these were reproduced for a single crystal XRD experiment, and their structures were classified as the 2-fold interpenetrated phase (Two other previously unidentified structures were found to be this phase via single crystal XRD, albeit these were missed by XCA [Fig. S4]). Using the metastable structures identified by CSP, XCA discovered an elusive phase in the XRD data that could not be confidently classified by the research team.

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] [40]. Phases in Ni-Co-Al are of interest for different applications such as superalloys and ferromagnetic shape memory applications [40, 41]. 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, $P(\phi_i|\text{EDX})$, to yield a joint probability, $P(\phi_i|\text{XRD, EDX})$.

When testing using the 12 phases found in the experiment, $P(\phi_i|\text{XRD, EDX})$ approached the ground truth, with most misclassifications still assigning high but not highest probability to the correct phase (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 ICSD41 (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); however, > 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. Previously taking weeks to months of manual effort, this characterization task is now accelerated in hours of computer time.

Where XCA and the expert disagree, additional information helps to make the correct classification. An example of this 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. S26]: values of $P(\phi_i|\text{XRD})$ and $P(\phi_i|\text{XRD,EDX})$ align with expert opinion, but are not informed by the literature or predicted phase stability. Since $P(\phi_i|\text{EDX})$ only captures the average sample composition, both $P(\phi_i|\text{XRD,EDX})$ and $P(\phi_i|\text{XRD})$ should be considered in tandem for a full phase map. As an example, Figure 3 compares the outputs for a representative phase (Ni$_3$Al) in the ternary composition space: while $P(\phi_i|\text{XRD})$ extends to encompass degeneracy [Fig. 1b] and mixed phase regions, the $P(\phi_i|\text{XRD,EDX})$ is confined by the expected composition of pure Ni$_3$Al. Similar to BaTiO$_3$, this produces a probabilistic solution to instantiate a precise refinement, emphasizing XCA-researcher collaboration.

Overall, 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 the cutting edge, XCA is significantly more accurate across materials systems [Fig. S31]. For alloys, 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 performance of XCA on the synthetic data is independent of learner architecture, as observed elsewhere [9,10]. Testing performance does not depend strongly on ensemble size, and improves with increasing dataset size, but with diminishing returns around $10^5$ patterns per phase [Fig. S32]. The onset of this limit depends on the diversity of experimental patterns.

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 (XPS, XANES, PDF, PL, NMR, mass spectroscopy, 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. It can further be extended to pair with unsupervised clustering methods to enable materials discovery in the absence of predicted phases. The collaboration between XCA and other AI agents will allow for advanced materials characterization to be incorporated into adaptive learning approaches for autonomous, data-guided experimentation.

In conclusion, we presented 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 to a challenge 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, this approach 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 [12] and complex solid solution electrocatalysts [13]. The innovation is directly applicable to inverse design approaches [14], new robotic discovery systems [3], and can be immediately considered for other forms of characterization such as spectroscopy and the pair distribution function.

**METHODS**

**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) [15]. 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) [16]. 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 [16]. 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.

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 [47] for 10 epochs. A discrete Bayesian optimization scheme was applied to optimize the hyperparameters and learner architecture [3]. As 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, $\phi$, given the EDX data,

$$ P(\phi|\text{EDX}) = \prod_\alpha \exp \left( \frac{(x_\alpha - x^\alpha_\phi)^2}{\sigma^2} \right), $$

where $x_\alpha$ is the measured mass fraction of component $\alpha$, $x^\alpha_\phi$ is the mass fraction of component $\alpha$ in phase $\phi$, and $\sigma$ is set such that the full-width-at-tenth-max is 0.5. Treating the output from the ensemble neural net as a probability, $P(\phi|\text{XRD})$, and given that these distributions are independent, a joint probability, $P(\phi|\text{XRD, EDX})$ 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(a)]. This emulates the thought process of a metallurgist in an explicitly probabilistic way, analyzing 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).

Acknowledgements. We acknowledge financial support from the Engineering and Physical Sciences Research Council (EPSRC) (grant number EP/N004884/1), BNL Laboratory Directed Research and Development (LDRD) projects 20-032 Accelerating materials discovery with total scattering via machine learning, the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design, 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. 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.

Data availability. The exact datasets are available from the authors on reasonable request, with the code used for construction available on the Github below.

Code availability. To facilitate the impact of this tool, the approach is kept entirely open-source and being embedded into data acquisition frameworks at central facilities blueskyproject.io. A release at the time of publication can be found at github.com/maffettone/xca.

[1] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, and L. Cronin, Nature 559, 377 (2018).
[2] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, and C. P. Berlinguette, Sci. Adv. 6 (2020).
