Machine Learning in X-ray Scattering for Materials Discovery and Characterization

Connor Davel[1], Nazanin Bassiri-Gharb[1,2], Juan-Pablo Correa-Baena[1,3],

[1]School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

[2]School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

[3]School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

Corresponding authors: JPCB jpcorrea@gatech.edu, NBG nazanin.bassirigharb@me.gatech.edu

KEYWORDS: Machine Learning, X-ray diffraction, Autonomous Materials Characterization, Combinatorial Characterization

Abstract

X-ray diffraction (XRD) is an immediate and powerful characterization technique that provides detailed information on the lattice structure and long-range order in crystalline materials. In recent decades, the quality and quantity of available crystal structure data has exploded, in large part due to the advent of online crystal structure databases, increased use of *in-situ* and *operando* methodologies, and user-accessible beamlines. The new wealth of data has also spawned an increasing use of machine learning (ML) to either construct high-throughput surrogates of established analysis or extract patterns from large datasets. However, XRD spectroscopy has been for many years solved via Rietveld refinement, while most ML techniques are simply complex statistical evaluation methods that are physics-agnostic. The discrepancy between data analysis and the underlying physics can lead to incorrect conclusions and/or limit the wide-spread adoption of ML techniques. In this review, we bridge the gap between ML and XRD spectroscopy with an introduction designed both for new data scientists and experimentalists interested in problems related to ML-guided spectroscopy analysis. We cover how supervised ML methods are used to predict likely symmetries and phases in pure and mixed samples, including challenges related to experimental artifacts and model interpretation. We also review recent uses of unsupervised methods in the extraction of patterns hidden in high-dimensional data, such as in *in-situ* and microscopic studies. Finally, we discuss the importance of problem formulation, data transferability, and reporting with recent case studies and give various resources throughout to expedite the learning curve for readers new to XRD or ML. We advocate for greater scrutiny of ML methods, how they are presented in the literature, and how to conduct data-driven research responsibly.

1 INTRODUCTION

1.1 X-RAY DIFFRACTION SPECTROSCCOPY

In the century since its birth, X-ray spectroscopy has come to define how material structures are quantified and understood. X-rays can probe the nanostructure of crystalline and amorphous materials with bond-length resolution and with minimum sample damage. These properties make X-ray spectroscopy fundamental to answering questions about bonding and nanoscopic order in hard crystalline, organic, and polymer materials. X-ray spectroscopy enabled landmark studies in materials science from the structure of metals[1] and alloys [2], our understanding of defects, the discovery of ferroelectric domains[3], quasicrystals[4], and the tetrahedral bonding of diamond[5]. This success has resulted in many technological leaps, pushing the boundaries of structural quantification in materials. Synchrotron facilities have steadily advanced the temporal and spatial limits of X-ray spectroscopy with X-ray free electron lasers[6], nano-diffraction[7], coherent XRD imaging[8], and 4th-generation beamlines facilities[9]. At the same time, specialized laboratory diffractometers and beamlines have greatly increased options for high-throughput, *in-situ*, and *operando* experiments[10,11]. However, despite the breakneck pace of recent technological developments, there has been little change to the foundations of X-ray spectroscopy since its invention.

In the late 19th and early 20th centuries, the nature of crystalline materials was still debated. Although the periodic structure of solids had been proposed by crystallographers in the 19th century, atomic scale experimental validation was lacking. In 1912, M. V. Laue, W. Friedrich, and P. Knipping published the first diffraction pattern of a crystalline solid, zinc sulfide[12]. This experiment and the latter interpretations from Bragg, Lau, Sommerfeld, and Ewald, among others, would settle the nature of solids and X-rays: most solids consist of a periodic arrangement of atoms, and X-rays are electromagnetic radiation of sufficiently short wavelength to diffract within such periodic structure. In 1913, W. L. Bragg described the diffraction problem as the reflection of X-rays from planes of atoms[13]. X-rays constructively interfere when the path traversed by adjacent X-rays reflected from the atomic planes is equal to a multiple of their wavelength. However, as Lau and Ewald would contribute, the diffraction phenomenon is not only explained by reflection. Diffraction is also the result of spherically scattered plane waves that constructively interfere at certain angles from incidence[14] (Fig 1a). This diffraction condition is visualized as the "Ewald sphere" constructed in reciprocal space, where points represent the spatial frequency of lattice planes. The diffraction condition is when the path difference of incoming (k_1) and diffracted (k) X-ray wavevectors is equal to an integer number of wavelengths (λ):

$$(k - k_1) \cdot a_i = n\lambda \tag{1}$$

$$proj_{a_i}k - proj_{a_i}k_1 = n\lambda \tag{2}$$

where a_i is any of the 3 lattice vectors. Eqn. 2 is visualized in Fig 1b. If k_1 and k begin and end on the Ewald sphere and reciprocal lattice points, then both wavevectors projected onto any of the 3 principal axes must differ by an integer number of wavelengths expressed in terms of lattice constant (for more details, see chapter 6 of ref[14]). As a result, constructive interference is observed. By precisely measuring the locations of diffracted X-rays with a known source and wavelength, the lattice constant of a crystal could be calculated.

The theory of X-ray diffraction developed greatly in the early and mid-20th century to quantify experimental data. Scherrer and Debye [15] showed that since the crystal lattice and experimental setup are a finite size, diffraction may still occur at points surrounding the Lau-Bragg diffraction condition (Eqn. 1), yielding crystallite size. Similarly, microstrain and the defect structure of the material also disrupt the periodic lattice and broaden peaks[14]. In polycrystalline samples, preferential orientation obscures diffraction in the out-of-plane direction. The Lotgering's factor[16] mathematically describes the change in peak intensity according to this preferential orientation. Importantly, it was found that the intensity of peaks could be predicted from the atomic number and occupancy in the lattice, quantified by the structure factor. The structure factor describes how X-rays scatter according to the scattering power of different atoms and explains how certain peaks are absent due to intermediate diffracting planes of atoms. Each of these historical contributions are summarized in Fig 1c and can be quantitatively extracted from the XRD pattern. This is the basis of the Rietveld refinement method[17]. A crystal structure is guessed, it's theoretical XRD pattern is calculated, and the difference between the theoretical and experimental pattern is iteratively refined until the position of atoms, crystal orientation, strain, and size effects most closely match that of the sample.

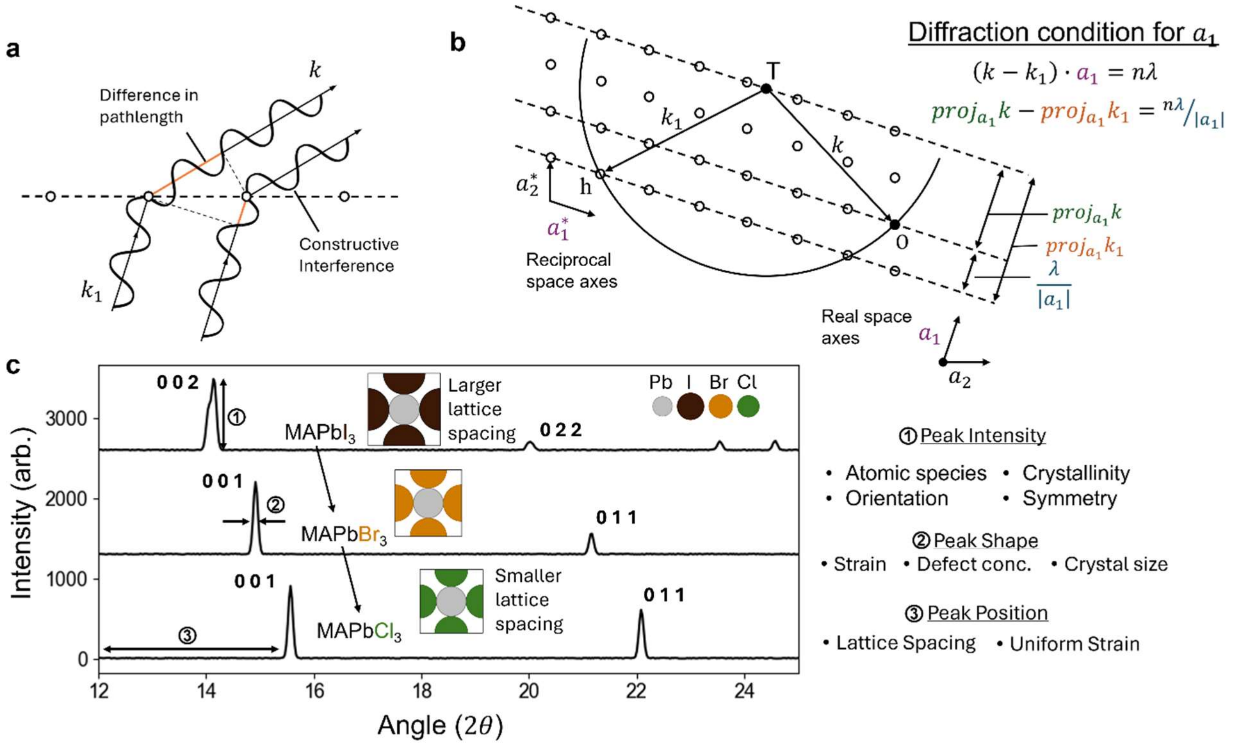

Figure 1. Depiction of (a) Scattering from a single line of atoms, adapted from Fig 6-2 from 50 years of Diffraction (P. P. Ewald) (b) construction of the Ewald sphere and geometric relation to Lau's diffraction condition, and (c) a schematic summary of information obtainable from an XRD pattern as a function of peak intensity, position, and shape using simulated lead halide perovskite XRD patterns. Crystal structure files obtained from Jaffe et al.[18] (orth. $MAPbI_3$, cubic $MAPbBr_3$) and Nandi et al.[19] (cubic $MAPbCl_3$) and calculated at approximately ambient temperatures and pressures. Inset drawings are of the 001 plane of one Pb-X octahedra for all perovskite crystal structures, where atomic radii (1.49Å, 2.06Å, 1.82Å, 1.67Å for Pb^{2+}, I^-, Br^-, and Cl^-) are drawn to scale showing a decrease in lattice spacing.

The choice of XRD method and resulting structural information depends on the existing available data on a material structure. For novel material, single-crystal XRD is used to solve the crystal structure. This involves obtaining all diffracting X-ray geometries to construct the full reciprocal space around 4 axes of rotation (Fig 2a). The smallest unit cell and space group is constructed from the sample stoichiometry and diffracting X-ray angles. For crystals with simple repeating units, the atomic positions can be inferred from the symmetry. However, for more complex materials, atomic positions may vary within the unit cell while still satisfying the symmetry, size, and stoichiometry of the cell. These variable atomic positions are then solved by predicting an XRD pattern that closely matches the measured intensities of individual reflections. When a material is well-known and published on ICSD, COD, or other databases, it is often sufficient to use the simpler Bragg-Brentano geometry which collects intensities about the single 2θ axis of rotation (Fig 2b). This is typically used for powdered samples with inherent spherical symmetry. X-ray scattering is also heavily affected by microstructure of materials, for which area detectors in high-brilliance synchrotron facilities are especially useful. In thin film metals and ceramics, solution and vapor deposition induce preferential orientation in the film grains resulting in preferred scattering angles (Fig 2c,d). Orientation and microstructure are also often studied in polymer materials (Fig 2e), where the orientation of the polymer backbone and ligands with respect to the substrate can heavily influence functional properties. When studying the diffraction of larger structures, such as nanoparticles and porous materials, smaller angles of scattering are studied.

Small angle X-ray scattering yields information of crystal size, shape, and aggregation for nanoparticles in solution as well as pore size and distribution for porous materials depending on the slope, decay, and peaks from 0 to ~0.5Å^{-1} depending on the material.

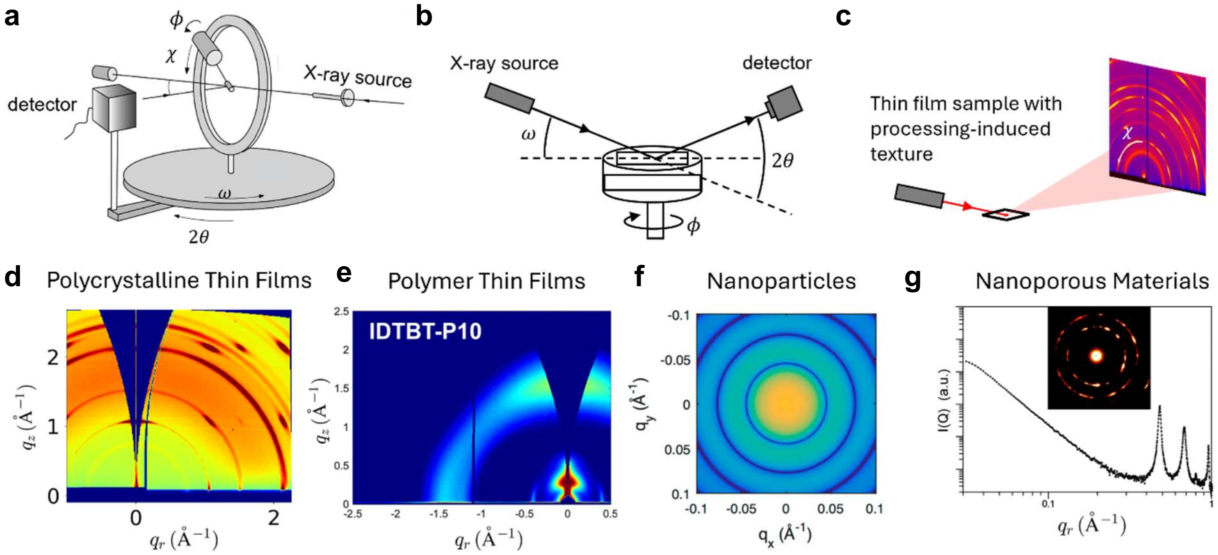

Figure 2: Various source-sample-detector geometries that determine the information obtained from XRD. (a) Single crystal XRD provides the most complete view of the X-ray and sample geometry. Reproduced with permission from Spring Nature[20] (b) $2\theta/\theta$ powder diffractometers obtain primarily 2θ angular information. (c) area detector geometries used in some high-luminance beamline facilities which capture a range of detector-sample-source geometries at a single time. (d-g) Various examples of 2D XRD patterns from materials systems including (left to right) lead halide perovskite thin films with preferential orientation resulting in spots of diffraction, indacenodithiophene-*co*-benzothiadiazole polymer thin films from ref[21], spherical nanoparticles (Adapted with permission from Li et al. [22] Copyright 2016 American Chemical Society), and porous metal organic frameworks (Adapted with permission from Tsao et al.[23] Copyright 2007 American Chemical Society)

1.2 THE EMERGING MATERIALS DATA CHALLENGE

The fundamentals of XRD spectroscopy and Rietveld refinement can uniquely determine the structure of materials from an XRD pattern. However, the quantity of data and high-throughput methodologies in modern science require innovations to interpret trends and extract patterns from large datasets. In the previous century, the number of known structures and complementary XRD patterns could fit in a few bound volumes. As of today, the Crystallography Open Database[24]

(COD) and Inorganic Crystal Structure Database[25] (ICSD) hold hundreds of thousands of crystal structures, escalating the complexity of finding high-quality references. At the same time, beamlines and laboratory-scale instruments increasingly utilize high-throughput measurement equipment and sample holders, collecting sometimes terabytes of data in a single experiment.

The quantity of experimental data is expected to further explode in many emerging highly tunable materials systems. For example, hybrid organic-inorganic semiconductors have highly tunable chemistries and processing parameters that can heavily influence their structure. These highly correlated chemistry-processing-structure-properties relationships have spawned multiple high-throughput studies of the effects of chemistry on structure and resulting electro-optical properties and robotic laboratories that greatly increase productivity and data throughput[26] (Fig 3a). In fact, the throughput processing of inorganic solids[27], nanomaterials[28,29], polymers[30,31], photocatalysts[32], and more[33] has dramatically increased with robotic and microfluidic synthesis approaches that remove humans from part of the analysis workflow. Such laboratories can synthesize up to hundreds of samples per day and explore the complex space of processing conditions to maximize a desired outcome from two or more processing conditions (Fig 3b). Continuous compositional libraries can also be created using co-sputtering and evaporation approaches[34–36] and yield many samples, depending on how the compositional space is probed. (Fig 3c). Although the existing XRD refinement tools are very robust, the analysis of high-throughput data can take between minutes and hours per XRD pattern depending on complexity and experience, even with the help of database searching and automatic peak-search and refinement algorithms. Therefore, a limitation of (semi-)autonomous materials science labs is not the information that can be extracted from individual XRD patterns for each composition but what structural data and trends can be distilled from thousands of patterns in a timely fashion.

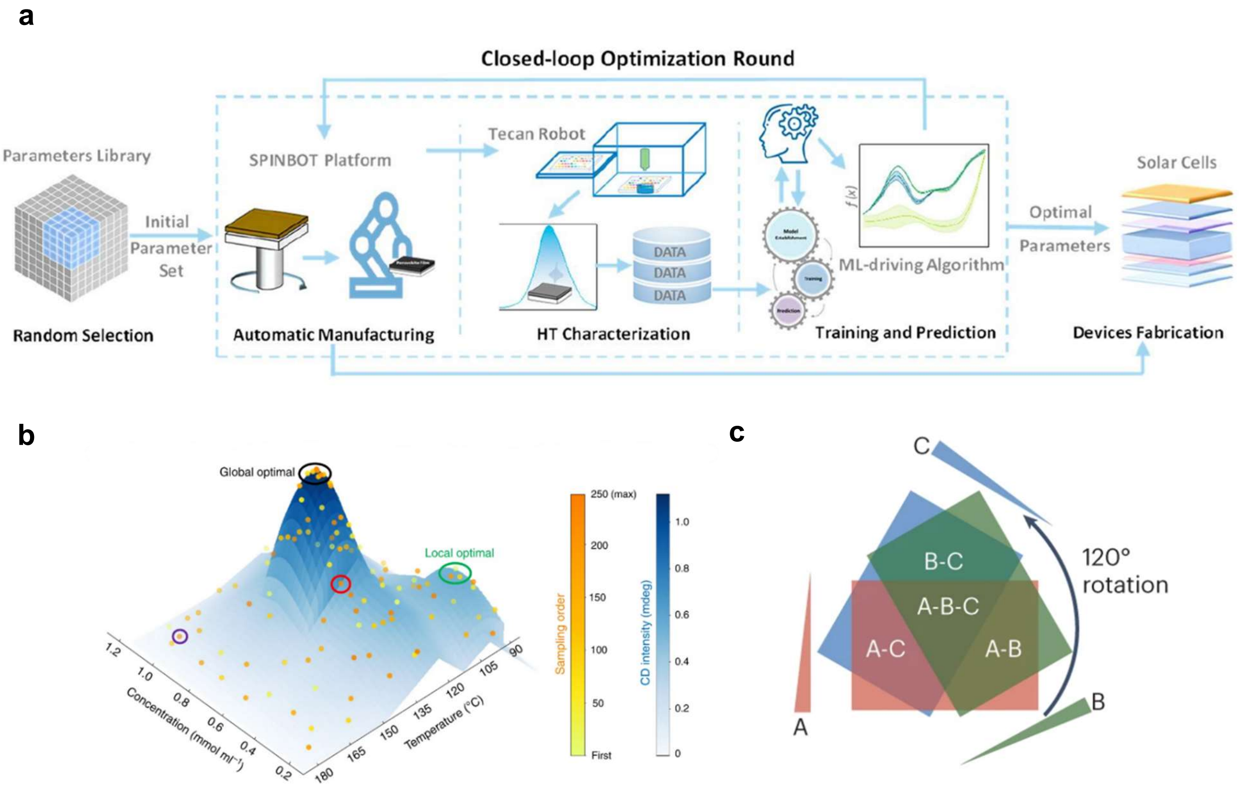

Figure 3. Examples from the field of HT experimentation, including a) the methodology of self-driven laboratories enabling the synthesis and characterization of hundreds of thin films per day[26], b) a successful optimization of photocatalytic nanoparticles in an autonomous microfluidic system[37], and c) composition gradient libraries created by use of masks and shutters in the co-evaporation of metals; commonly used for many evaporation-processable materials[36].

Here, we focus on the applications of ML in solid-state X-ray structural data visualization and structure prediction, providing researchers with a comprehensive overview of the current state-of-the-art and challenges in harnessing ML for structural analysis. Importantly, we hope to bridge the conceptual divide between known physics and purely data-driven ML to promote understanding and sensible implementation of ML methods. We aim to supplement existing reviews on the application of ML in photovoltaic device design[38–42], semiconductor property design[43–45], microstructure design[46–48], materials property design[49–51], device engineering[50,52] , and the broader field of materials science[53–55], focusing primarily on the representation and prediction of structural X-ray data. Finally, we discuss best practices in data curation, preprocessing, training, and reporting to avoid common pitfalls and expedite communication. The summarized approaches should enable more robust analysis in high-throughput studies and increase confidence in adopting ML analysis in closed-loop experimental protocols.

2 BACKGROUND IN MACHINE LEARNING

The mathematical and statistical foundations of Machine Learning (ML) have long roots in physics, chemistry, and biology to discover trends and inform theory. For example, linear regression, principal component analysis, and discriminant analysis were used throughout the 20th century to understand trends in spectroscopy data[56]. The growth of ML, as it is thought of today, would come during the mid-20th century when researchers questioned how machines can be programmed to learn inductively from data (rather than deductively from pre-programmed rules and logic). This led to many early successes such as Marvin Minsky's "SNARC" analog neural network[57], Frank Rosenblatt's perceptron modeled after the neuron[58], Hopfield networks[59] and Boltzmann machines[60], and support vector machines[61]. In the late-20th century, much of the theory of connectionism discussed in this review was formalized, including feed-forward and convolutional neural networks and methods of parameter initialization and training[62]. However, it was not until the 2000s that researchers had the computational power necessary to train larger ML models and apply them at a broader scale. Primed by the influx of data concomitant with the information age and developing internet, ML saw a surge in use cases such as search engines and online behavior analytics. It was around this time that the freshly marketed field of ML saw use in materials science, increasingly cited in materials research publications starting from the early 2010s [63,64]. Open-source ML software libraries like scikit-learn, PyTorch, and TensorFlow were developed to make ML more accessible, meaning ML could be more easily implemented in materials science problems. At the same time, materials data infrastructure like the Materials Project[65] made a wealth of materials data available to the wider research community, resulting in the online "materials intelligence ecosystem"[54] we know today.

In general, ML seeks to establish relationships within a defined domain of data. When these relationships are known in advance, a surrogate model can be trained to map the data with known properties. The model is thus "supervised", using known input-output data to replicate the same patterns later. In our case, the input domain encompasses all or some defined subset of the experimentally obtained or simulated XRD patterns. The output domain may comprise any of the useful parameters extracted from XRD patterns, such as lattice constant, strain, crystal size, symmetry, phase identity, etc. (Fig 1c). The output domain can be segmented into continuous (lattice constant, strain, crystal size, phase fraction) and discrete (symmetry, phase identity) variables. The challenges of analyzing these types of output domains are called regression and classification tasks, respectively (Fig 4a). Deciding the input and output domain is an important step in producing usable ML models since both define the limits for confident model predictions.

A second common situation is when a specific input-output relationship is not known in advance, in which case the model is "unsupervised" and tasked with finding relationships based on how data points are distributed relative to each other. Clustering problems identify data that aggregate together due to an underlying similarity in composition or processing (Fig4a). This is useful when analyzing datasets where points are expected to fall into few discrete categories (such as phases, domains, etc). Dimensionality reduction models seek to identify hidden or "latent" variables in data that describe structures and patterns of similarity. When correctly trained, the latent variables of a dataset can reconstruct the original without loss of information.

A central distinction between supervised and unsupervised learning methods for XRD analysis is that supervised methods require the prior analysis of the structural data of interest, whereas unsupervised models may be applied directly to XRD data. When obtaining crystal structures from databases like ICSD, pattern fitting or single crystal analysis has already been done and is ready for training of supervised learning models. If the goal is to apply ML to analyze experimental data, then a small representative subset of the original data must be selected and analyzed to train an effective surrogate model. Alternatively, a representative dataset must be constructed from simulated data. Supervised models are the first ML algorithms discussed in this Review. However, when structural analysis or pattern fitting is difficult—such is often the case with novel beamline experiments, *in-situ* studies, and high-throughput datasets, or when methods do not exist to simulate a representative dataset—unsupervised methods are chosen to visualize the relationships obscured by the dimensionality of the data. Unsupervised methods are the second class of ML models discussed. Often, supervised and unsupervised methods can be complimentary, especially to reduce the dimensionality of data before surrogate model training or to quantity model certainty.

2.1 Supervised Models

Although neural networks are only one ML model among hundreds commonly utilized in data science, they have come to dominate much of the recent literature on ML-based XRD analysis. As such, they are briefly discussed here using a diffraction pattern as an example. This example is adapted from ref[66] ch. 18. We consider XRD patterns $k = 1 \ldots N$ where each pattern is represented by intensities vector x_k of length M. This dataset is represented by an N x M matrix X where each row is an XRD pattern x_y. Assume each pattern corresponds to a value y_1 or y_2 representing a majority phase identity. If the XRD patterns consist of well-defined phases with non-overlapping peaks, an optimal decision boundary plane exists that divides vectors $x_1 \dots x_N$ into two groups. The equation of this plane has the form $b + Xw = 0$, where w is the M-length vector normal to the dividing plane. The goal of this problem is to solve for w. This is the basis of the single layer perceptron model with linear activation function: if $y(x) = b + wX$ is greater than 0, we classify the pattern as y_1, and otherwise y_2. Consider then a more complex problem of a dataset with many phases $y_1 \dots y_p$. A single decision boundary can only distinguish 2 phases and cannot generalize to multiple phases. However, this problem *can* be solved by the combination of many simple decision boundaries. This is the conceptual foundation of neural networks: that complex decisions can be made from the agglomeration and abstraction of simple computational units.

Feed-forward neural networks (FFNNs) are composed of N stacked layers of M neurons. The width of each layer approximates complex functions through the addition of multiple simple linear combinations of the input (each perceptron giving a single linear combination weighted by the activation function). One might note that such a network can only construct linear relationships. However, nonlinear functions may be approximated by using an appropriate nonlinear activation function $\alpha\big(y(x)\big)$ that acts on the output, $b + wX$, of each neuron. Furthermore, the depth of the neural network allows for abstraction of the input. It has been shown that the early layers in a deep feed-forward neural network are primarily responsible for representing simple data patterns and gradients while the later layers are responsible for building abstraction and complexity. It may seem difficult to rationalize that large ensembles of perceptrons can be manipulated to produce a coherent output. However, the idea of complex behaviors originating from simple first principles calculations is not new. In fact, many of the same tools used to study the macroscopic behavior of elementary operators from theoretical physics have been applied to study DL models[67]. The biggest challenge in pattern recognition when using a FFNN is the number of parameters that can rapidly become intractable, since every data point is input to every node (a total of $M(P+1)$ parameters for the first layer alone!). A major reduction in computational complexity can be achieved by using convolutional layers. A convolution layer consists of a kernel that moves across 2θ, computing the integral of the multiple of the kernel with the pattern. Fig 4b illustrates this process for an experimental $MAPbBr_3$ patten and a manually generated kernel consisting of its first 5 simulated peaks. When the kernel most overlaps with the pattern, the output of the convolution operator is maximized, extracting an important feature from the pattern. Multiple convolutional layers in sequence construct convolutional neural networks (CNNs). However, this is only a toy example of a single convolution operation. It is difficult to visualize convolution layers, since each layer successively builds in complexity and abstraction to construct the final inference. A method to visualize the mechanics and choices of CNN models with class activation mapping is discussed later in Section 3.

2.2 UNSUPERVISED MODELS

Unsupervised models are trained and evaluated in a fundamentally different way to supervised models. A successful unsupervised model should represent the available data, either through grouping (e.g., cluster assignment) or discovering hidden representations (e.g., latent variables) with minimal loss of important information. For example, K-means clustering finds data clusters of smallest variances or spreads, and hierarchical clustering successively merges clusters to build a nested representation of the data. Many clustering algorithms rely on the careful choice of similarity metric, which determines how similar two points are in space. Euclidian distance is the most intuitive metric, but data may also be clustered according to their correlation via the Pearson correlation coefficient. For example, the Pearson correlation coefficient would recognize an XRD pattern as identical to the same pattern with all intensities doubled (say, by changes in source luminance) since all peaks are correlated. In contrast to clustering, dimensionality reduction techniques focus on uncovering latent variables—underlying factors or features in data that are often continuous or overlapping—allowing for a compact representation without explicit grouping as in clustering. Principal Component Analysis (PCA) and Non-negative Matrix Factorization (NMF) are two widely used methods, each with distinct approaches. PCA transforms data into a set of orthogonal (uncorrelated) components that capture the most variance, producing a global view of data structure with both positive and negative values, which may be less intuitive for interpretability. NMF decomposes data into additive, non-negative parts, which is particularly useful when interpretability and sparse component representation matter, such as identifying single phases in mixed XRD patterns. Both PCA and NMF are linear methods, making them particularly useful for X-ray diffraction (XRD) analysis, where different phases contribute to linearly increasing intensities in the diffraction pattern. Where nonlinear relationships are expected, autoencoders, a type of neural network, learn compressed latent representations that faithfully reconstruct the input—effectively solving for $f(x) = x$, where f is the learned transformation. PCA, NMF, and autoencoders are all evaluated via the reconstruction error, or how accurately the reduced information approximates the original data. Reconstruction error is closely related to the information content that is lost during training, meaning that reconstruction error can be expected to be high when encountering data outside the training domain.

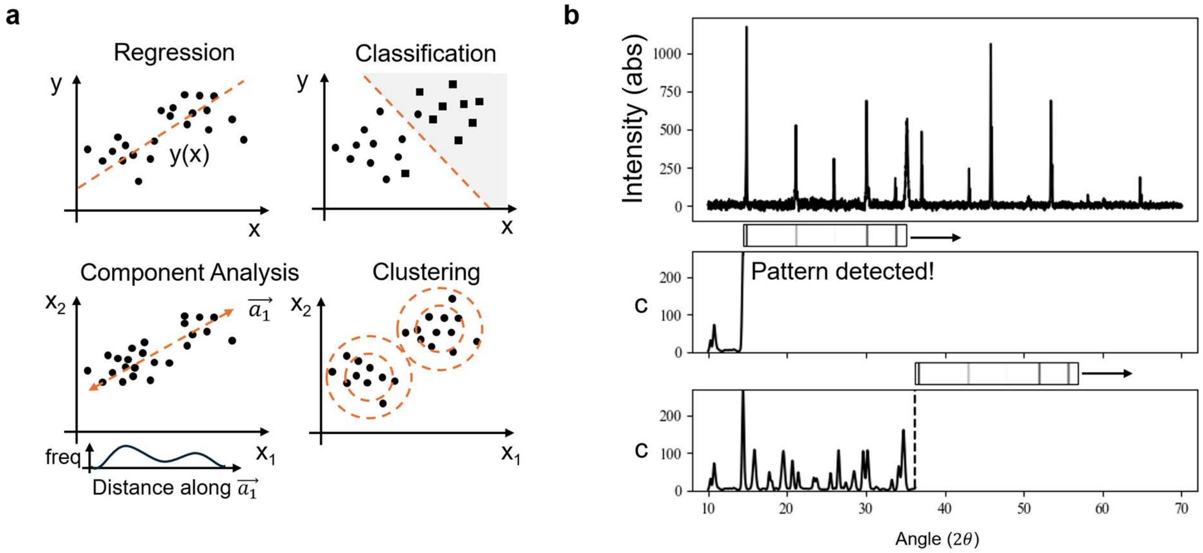

Figure 4. (a) Visualization of supervised (regression and classification) and unsupervised (component analysis and clustering) learning problems. (b) effect of a pre-chosen convolution, the first 5 peaks of MAPbBr3 pattern, on an experimental XRD pattern of the same compound.

3 ML Analysis of X-ray Scattering and Diffraction

3.1 Supervised Methods: Expediting Manual Analysis with Predictive Modeling

With the advent of online databases of solved crystal structures, it is possible to construct large datasets of simulated XRD reference patterns using the theory discussed previously. It is the reverse process, calculating crystallographic information from an XRD pattern, that supervised methods then approximate. One such problem is the global (i.e., models that work for all materials) prediction of the 7 crystal systems or 230 crystallographic space groups from an input XRD pattern. In 2017, Park et al.[68] solved this problem with a CNN. In this work the model architecture begins with a series of convolution layers, which, as previously discussed, distinguish dominant features in the input with increasing abstraction. Pooling or dropout layers are placed in between convolution layers to subsample the most important information and reduce the number of model parameters. Finally, fully connected layers of the NN construct the final functional form of the model from the abstracted features, resulting in an output of 230, 101, and 7-point vectors corresponding to the number of space groups, extinction groups, and crystal systems respectively. This model achieved a classification accuracy of 81%, 84%, and 95% for these parameters, with accuracy increasing with decreasing output vector size. Following this work, multiple studies using CNNs have been performed on global crystal structure and space group identification as summarized in Table 1. A recent notable entry is Corriero et al.[69] who published their model with an online graphical user interface.

Table 1. Summary of some studies predicting crystal system and space group from most or all XRD spectra found in ICSD. Accuracies are given according to the accuracy of predicting the single crystal system or space group on simulated data from ICSD.

Year	Num. Input	Train/ Test (%)	Acc. Crystal System (%)	Acc. Space Group (%)	Ref
2017	150000	80/20	95	81	[68]
2019	128404	90/10	85	76	[70]
2020	169563	80/20	92	80	[71]
2020	192004	80/20	90	80	[72]
2021	12000	-	96*	-	[73]
2022	-	80/20	92	80	[74]
2023	187131	95/5	92	84	[75]
2023	280000	-	70*	-	[69]

*Cross-validation accuracy, which overestimates the accuracy expected on unseen data.

Most work in predicting crystal systems has shown that accuracies above 90% are achievable for simulated data. However, models tend to perform less optimally when classifying space groups given that peak absences do not uniquely determine the space group, and single crystal XRD is needed for a confident classification. Lee et al. [74] note that the currently achievable accuracies of 80% are on-par with manual XRD analysis and that model accuracy is inherently limited when only considering powder XRD spectra.

A rather significant limitation of current ML models for XRD symmetry classification is data availability and quality, a challenge particularly impacting the classification accuracy of uncommon space groups. Classification accuracy is heavily dependent on the availability of training data across space groups: as the number of training data decreases, the classification accuracy decreases as well [70,72] (Fig 5a). Besides the number of training data, the complexity of XRD spectra increases with decreasing space group symmetry. Suzuki et al.[71] demonstrated that a classifier trained on the first N peaks required more peaks in the pattern to classify lower symmetry crystal systems. Cubic crystal systems required two or three peaks to accurately classify their crystal system identity, whereas trigonal and lower symmetry crystal systems required 4 or more peaks for accurate classification (Fig 5b). This follows the physical intuition that lower-symmetry space groups exhibit more peaks and require more independent parameters to describe their crystal structure. It is therefore imperative for symmetry prediction models that the classification performance and model certainty within each space group and crystal system should be reported to users and considered on a case-by-case basis. Unfortunately, the only certain solution to the data scarcity problem is to discover novel compounds in under-represented symmetries, a slow and tedious process. However, Schopmans et al.[76] reported a method to pre-train a neural network on theoretical, yet-demonstrated crystal structures-of which there are infinitely many for each space group-achieving a minimum space group prediction accuracy of 80%. This strategy of pre-training or “transfer-learning” (transferring knowledge gained from one domain to another) is one of the under-explored topics of current literature.

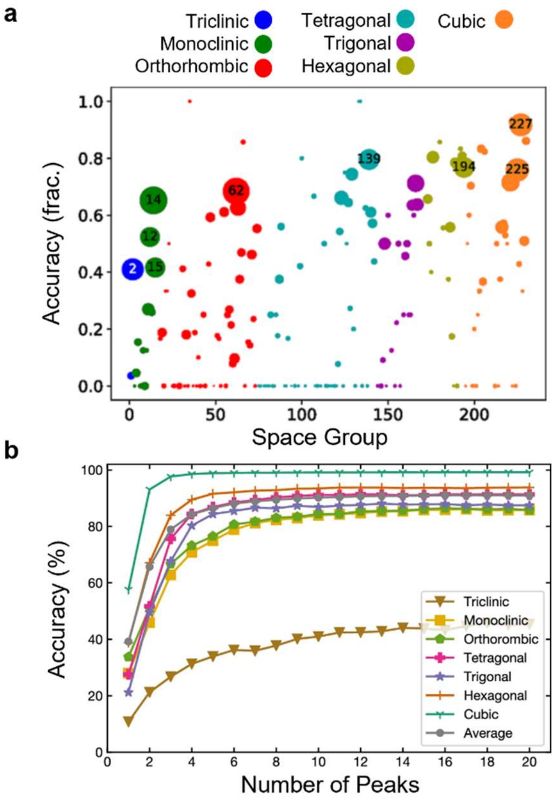

Figure 5. (a) ML model accuracy a function of space group (x-axis), crystal system (color), and number of training samples (circle size). Accuracy increases with more training samples and when classifying higher symmetry crystal systems such as cubic materials. Reprinted with permission from Vecsei et al.[70] Copyright 2019 by the American Physical Society. (b) The training classification accuracy across different crystal systems (Triclinic, Monoclinic, etc) as a function of the number of peaks allowed to represent each pattern during training from Suzuki et al.[71]

Data quality and generalizability also affect model performance on experimental data, which typically includes artifacts and noise. Model performance decreases significantly when classifying experimental data with models exclusively trained on simulated data. Salgado et al.[77] observed a significant decrease in model space group prediction accuracy down to 12% and 22% for CNN and MLP architectures trained with synthetic data. These models were evaluated using data from the RRUFF[78] experimental database. Vecsei et al. [70] similarly reported a decrease in CNN crystal system prediction accuracy from 85% (Table 1) 56% when evaluated with the RRUFF database and similar reductions in accuracy for space group classification.

Two general solutions have been devised to increase model accuracy for experimental data. First, data augmentation strategies to mimic experimental noise, preferential orientation, peak broadening, background, or a combination of artifacts in simulated data (Fig. 6) can increase accuracy in experimental pattern classification[75,77,79–83]. Data augmentation strategies result in simulated training data that more closely aligned with real-world experimental data, which improves model performance. For example, Lee at al. [75] increased the accuracy of their model from 33% to 56% on experimental data when incorporating peak shift, broadening, and background noise into the simulated training dataset and slightly improved accuracy to 59% when introducing texture (varying peak heights). Vecsei et al. increased model accuracy for RRUFF experimental crystal system classification from 22% to 64% for a FFNN model by training with a diversity of simulated peak shapes and synthetic instrumental noise[77]. Second, supplementing simulated training data with experimental data can increase model accuracy by similarly conditioning the training dataset to experimental artifacts. For example, Oviedo et al.[80] incorporated pre-analyzed experimental data during model training to boost space group classification accuracy from 80% to 89% for a subset of space groups. They also incorporated data augmentation strategies such as peak scaling, peak elimination, and pattern shifting, showing that a combination approach of physics-based data augmentation and experimental data supplementation is highly effective at decreasing overfitting. Vecsei et al. used a similar strategy, improving their best model accuracy from 74% to 86% accuracy on RRUFF experimental data with no change in model architecture.

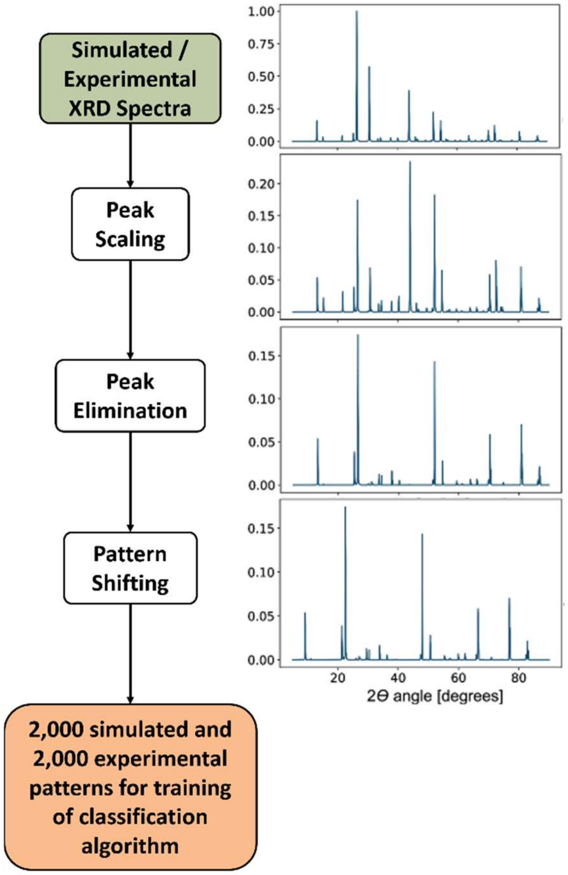

Figure 6. General data augmentation procedure to increase accuracy on experimental data from Oviedo et al.[84] After simulating a series of intensities, peaks are scaled and eliminated to simulate orientation and site occupancy effects and shifted to simulate the effects of strain in synthesized samples.

Another particularly challenging problem in XRD analysis is multiphase identification. Work in the 20th century focused on the extension of Klug and Alexander's[85] equation for quantitative X-ray diffraction analysis using pure and unpure standards[86] (which should also be considered in addition to ML strategies). Starting from the 1980s, linear regression found success in phase fraction prediction which attempt to find components to represent pure phases from impure samples without standards[87]. In recent years, CNNs have been used for multiphase fraction regression or pattern identification using simulated XRD patterns from online databases as standards. Wang et al.[82] demonstrated a CNN model to classify metal organic frameworks (MOFs) from a set of simulated and augmented XRD spectra to achieve 57% accuracy. This was increased to 97% when predicting the top-5 most likely phases present to accommodate phases that may exhibit very similar spectra. Lee et al. reported on CNNs for predicting inorganic crystal phases in the Sr/Li/Al[79] and later Li/La/Zr[88] oxide chemical space, trained on large combinatorial libraries spanning the entire composition range at regular intervals. A more scalable approach to multiphase prediction was developed by Szymanski et al.[81] for the Li/Mn/Ti oxyfluoride chemical space to identify if a *single* phase was present in an XRD pattern. Once identified, the phase was subtracted from the pattern, and the same classification-subtraction analysis was performed until no phases were left. This work was later implemented in an autonomous phase identification diffractometer experiment as shown in Fig. 7[89].

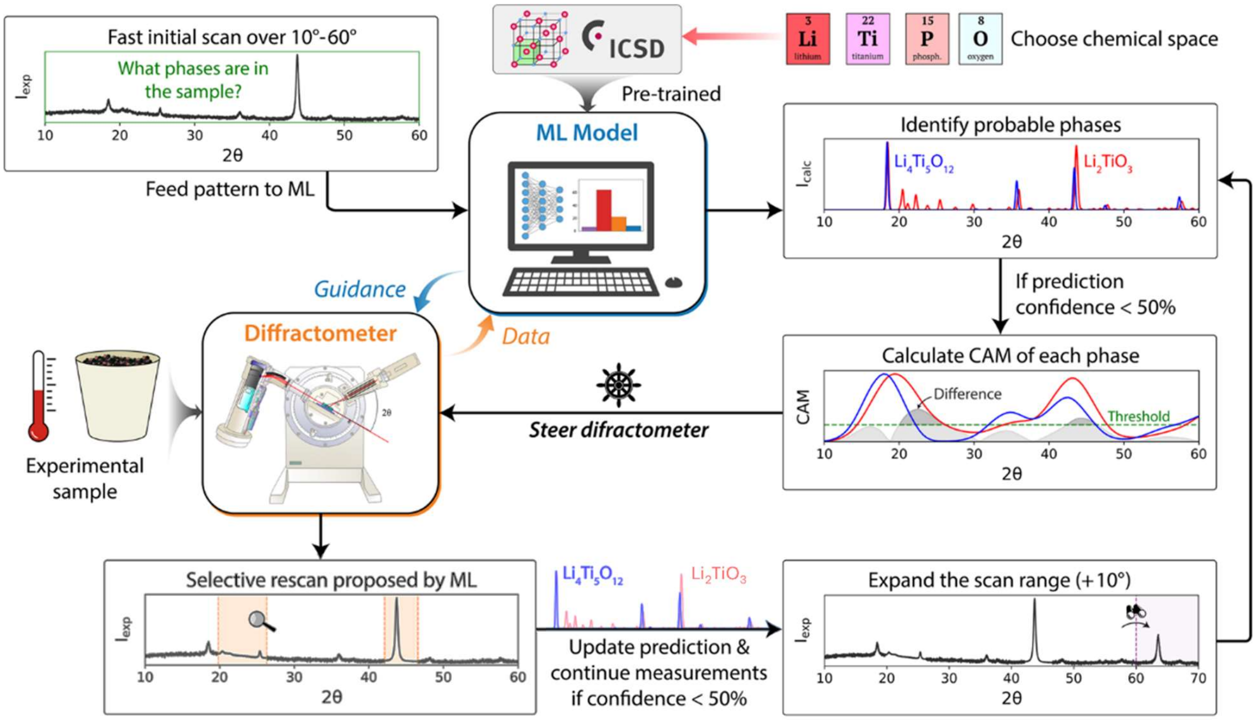

Figure 7. Example of a full experimental workflow incorporating real-time ML analysis from Szymanski et al. [89]. First, models are pretrained from a subset of patterns in the ICSD database. Probable phases are identified, and class activation maps (CAMs) are used to identify regions of uncertainty in the XRD pattern. The diffractometer is automatically programmed to increase the resolution and range of the scan until the final phases are predicted with confidence.

An important aspect of DL models, especially critical to the symmetry and phase classification problems discussed above, is model interpretability and certainty. Interpretable models are vital to experimentalist since confidently false and unjustified predictions can cost days of work. Class activation maps (CAMs) are one method to justify model outcomes. CAM methods generally compute the spatial location of features most important in effecting the final classification[73,90,91]. For example, CAMs were used by Szymanski et al[89] (Fig 6) in their automatic phase classification experiment to identify angle ranges most important to distinguishing two similar phases. When the model identified two plausible existing phases, each with regions of interest in the XRD pattern represented by the CAM, the XRD diffractometer steered towards regions where the CAMs differed to resolve distinguishing features and peaks. Generally, CAMs allow users (and autonomous laboratories) to fact-check the CNN model. If the CAM does not indicate that certain key peaks are being used in the phase identification, then it can be inferred that the CNN is not learning the correct pattern-symmetry or pattern-phase relationships. Another method to visualize regions with important features is feature sub-selection, which involves reducing the available data to all but the most important peaks and intensities. Massuyeau et al.[92] used this strategy to indicate which angles of the XRD pattern are most important to distinguishing a perovskite phase. For example, regions 2 and 3 in Fig 8a, centered around the well-known 15° perovskite peak were confirmed to be critical regions for CNNs to identify perovskite-like structures.

Model certainty can be quantified through proper choice of activation function. Sun et al.[83] classified perovskite systems according to dimensionality with a training set of 2000 experimental and 2000 simulated patterns (Fig 8b). The output layer included a softmax activation function normalizing the output over a probability distribution of assignment confidence. Ensemble learning is another probabilistic strategy that aggregates the predictions of many models to quantify certainty. Maffettone et al.[93] used this strategy in a multi-phase prediction model with 50 shallow CNNs, where the combined ensemble "votes" to elect a most probable phase by some margin of error. The previously discussed work by Szymanski et al[81,89] used another ensemble strategy where 1000 CNNs were generated by randomly removing 60% of the fully connected layers of the CNN regardless of weight. This created a large ensemble of poor model estimators whose percentage vote approximated model certainty.

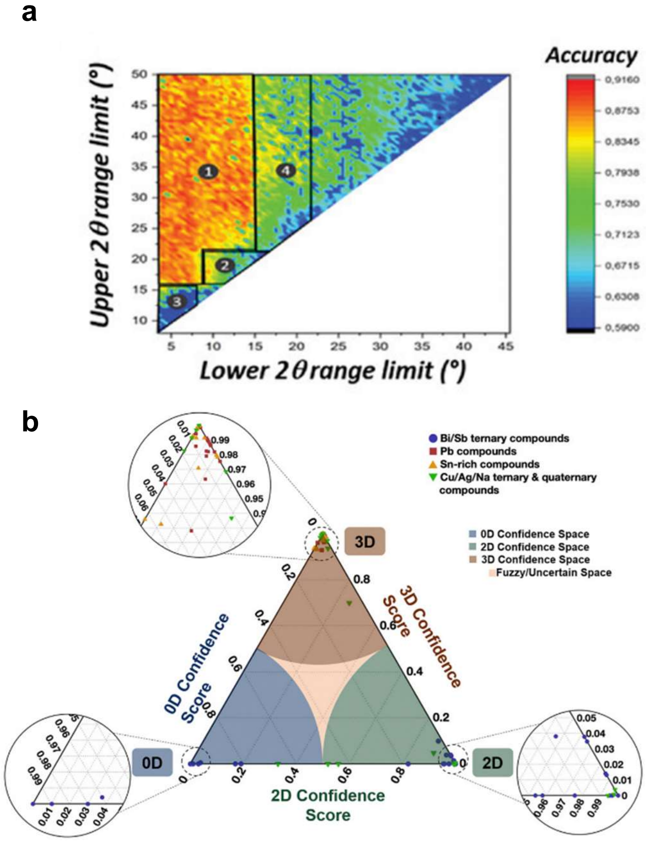

Figure 8. (a) Extraction of most impactful regions in increasing CNN accuracy in classifying whether a XRD pattern corresponds to a perovskite or non-perovskite phase. Reproduced with permission from John Wiley and Sons from Massuyeau et al.[92]. The sharp decrease in accuracy around regions 2 and 3 helps distinguish important low-order peak regions (ie. the highest magnitude 15 degree perovskite-phase peak) (b) Results of FFNN model classification of perovskite dimensionality as 0D, 2D, or 3D. Reproduced with permission from Elsevier from Sun et al.[83] (Note: 1D structures not identified in this composition range).

Other applications of ML have been for attempts to shortcut later steps in structural analysis and extract information beyond symmetry and phase identity. For example, Chitturi et al.[94] trained a CNN to predict lattice parameters for any XRD pattern, completing the full XRD analysis workflow from peak indexing to Rietveld refinement in a single step. However, differences between real and predicted lattice parameters exceeded 2 angstroms for some systems. Dong et al.[95] also developed a CNN to estimate lattice parameters (along with the scale factor and crystallite size) and achieved greater accuracies by considering the Ni-Pd/CeO_2-ZrO_2/Al_2O_3 materials space rather than all materials. Lee et al.[96] demonstrated a multi-step optimization process that differs significantly from the CNN-based methods described in earlier section. They developed an "Evolv&Morph" algorithm which uses a genetic algorithm to produce a crystal structure that closely replicate the target XRD before Rietveld refinement. ML may also find applications in linking chemistry to structure, such as the work by Dong et al.[97] which predicted the likely XRD spectra from the elemental composition of a material in a single step for specific and well-defined chemical domains.

All studies discussed till this point have used the raw XRD intensities as input, albeit with occasional data augmentation. However, data pre-processing strategies have also been explored. Artursson et al[98]. demonstrated that select wavelet transforms could improve model (partial least squares) prediction accuracy on a binary phase fraction regression problem while also significantly reducing the number of data points and experimental noise. They also noted that, among others, the Fourier transform and Savitzky-Golay filter improved prediction accuracy. For small datasets (i.e., not applicable to large datasets such as ICSD), Karstang and Eastgate[99] have reported a method to reduce an XRD pattern to only the angles that distinguish samples. Such an approach greatly reduces the input data size and maximizes information density. Preprocessing methods can be particularly impactful for CNN performance. Szymanski et al.[100] recently reported that training with the Fourier transform of an XRD pattern can increase a CNN's ability to differentiate between subtle variations in low-intensity peaks, as would be expected with varying site occupancy and order. This creates a "virtual pair distribution function (PDF)" of the XRD data, which highlights the local arrangement of atoms. The utility of using CNNs to predict crystallographic information from PDF data has been explored in the literature[101,102], validating the approach of preprocessing XRD data via the Fourier transform. The choice of input representation used (i.e., the original pattern or a transformation of the pattern) impacts which features are represented most strongly and is an important application-specific decision.

3.2 Unsupervised Methods: Pattern Analysis and Data Visualization

Unsupervised ML methods have been a mainstay analysis tool for data visualization in complex datasets. The uniting assumption behind all unsupervised learning problems is that data within an experiment or domain will follow common patterns that are independent of noise and specific outliers. This implies that data can be disentangled from those outliers and noise. Applications in XRD analysis enabled by this approach include improving signal-to-noise ratio [64,65], identifying buried peaks[103], identifying outlier patterns[104], and identifying outlier detector pixels[105]. A second implication of this assumption is that some features of the data are redundant: indeed, an N-point spectra is condensed to only a few parameters (lattice constants, orientation, microstrain, etc.) after Rietveld refinement or similar analysis. Hence, unsupervised methods are also often used as feature-extractors, which visualize trends in data that are obscured by their high dimensionality. For example, Matos et al. [106] used principal component analysis (PCA) to visualize the independent effect of crystallite size and orientation on an XRD pattern. PCA identifies linear transformations of the data with high variance, such as select peaks decreasing in intensity with increasing orientation. The effect of this transformation is to reduce the dataset to only the most important dimensions (e.g., the size and degree of orientation). CNN autoencoders have been similarly used as feature extractors. Banko et al.[107] used autoencoders to understand the distribution of data among different space groups, which form distinct clusters in the information dense latent space. Similarly, Utimula et al. [108] used an autoencoder architecture to extract features to aide in phase fraction prediction. A useful feature of unsupervised learning models is that they quantify error according to how much of the original information in the dataset is retained after encoding. For example, Banko et al.[107] showed that the autoencoder reconstruction error (how closely f(x) matches x after encoding) can be used as a proxy for model certainty which decreases the chance of "confidently false" predictions. Reconstruction error can be computed for PCA or any other dimensionality reduction technique, making these powerful complementary techniques to the previously discussed supervised methods.

A common use-case of unsupervised methods is the analysis of structural data with varying chemistry to construct experimental phase diagrams (Fig 9a). High-throughput studies have previously relied on effort-intensive manual analysis[109], data visualization[110], or reduction of the XRD patterns to single key features such as peak amplitude[111]; however, clustering and dimensionality reduction offer alternative approaches that enable reduced bias and simplified data interpretation. From the late 2000s, metric multidimensional data scaling[112] and later non-negative matrix factorization (NMF)[113] have been used to reduce a dataset of hundreds of patterns to 9 or fewer representative patterns used for database searching. Physical constraints can be imparted during phase analysis by augmenting with prior data from ICSD, which incorporates domain knowledge into model training [114]. As previously discussed, experimental artifacts such as peak shifting, broadening, and grain texture often motivate model choices. Hierarchical clustering has been demonstrated to identify constituent phases in a Ni/Ti/Co thin film library[115], which is a clustering technique that can distinguish large changes in data (different phases) from small variation (peak shifts). Additionally, the use of specific similarity metrics can be chosen to accommodate experimental artifacts. For example, clustering according to the Pearson correlation coefficient is robust for identifying separate phases even with orientation artifacts which change relative peak intensities[116]. Comparatively, clustering according to constrained dynamic time warping (DTW) has been shown to identify phases even when the majority of variance in the data is explained by peak shifts[116,117]. DTW finds the optimal shift and warp-corrected correlation between data points, making it efficient at recognizing shifted XRD patterns but unable to assess peak broadening. Other solutions to peak shifting include the use of a logarithm kernel to the angle/q axis to transform increasing peak shifts at higher angles into a constant shift factor across the transformed pattern [118] and removal of peak-shifted patterns of the same phase through a cross-correlation analysis of shifted copied of each constituent phase after training [119].

The fact that unsupervised learning requires little or no data processing has made it popular among beamline *in-situ* and microscopy studies. Autoencoders and gaussian mixture models have been used successfully to identify phase transformations during *in-situ* temperature dependent XRD experiments [120,121] (Fig 9b). In synchrotron X-ray nano-diffraction microscopy, autoencoders and clustering may extract relevant features from the detector to project onto the real-space. Song et al. [122] used a combination of DL autoencoder feature extraction and clustering to automatically visualize the domain variants from X-ray microdiffraction scans in a ferroelectric material (Fig 9c). Salameh et al.[123] alternatively used clustering to identify peaks most correlated to existing phases. The local maxima of intensity of these peaks was then an indication of separate phases and lattice distortions. Borobtsov et al.[124] used a similar approach to quantify the local concentration of phases and phase transitions across the surface of a Ca_2RuO_4 thin film (also see their earlier work in ref [125]).

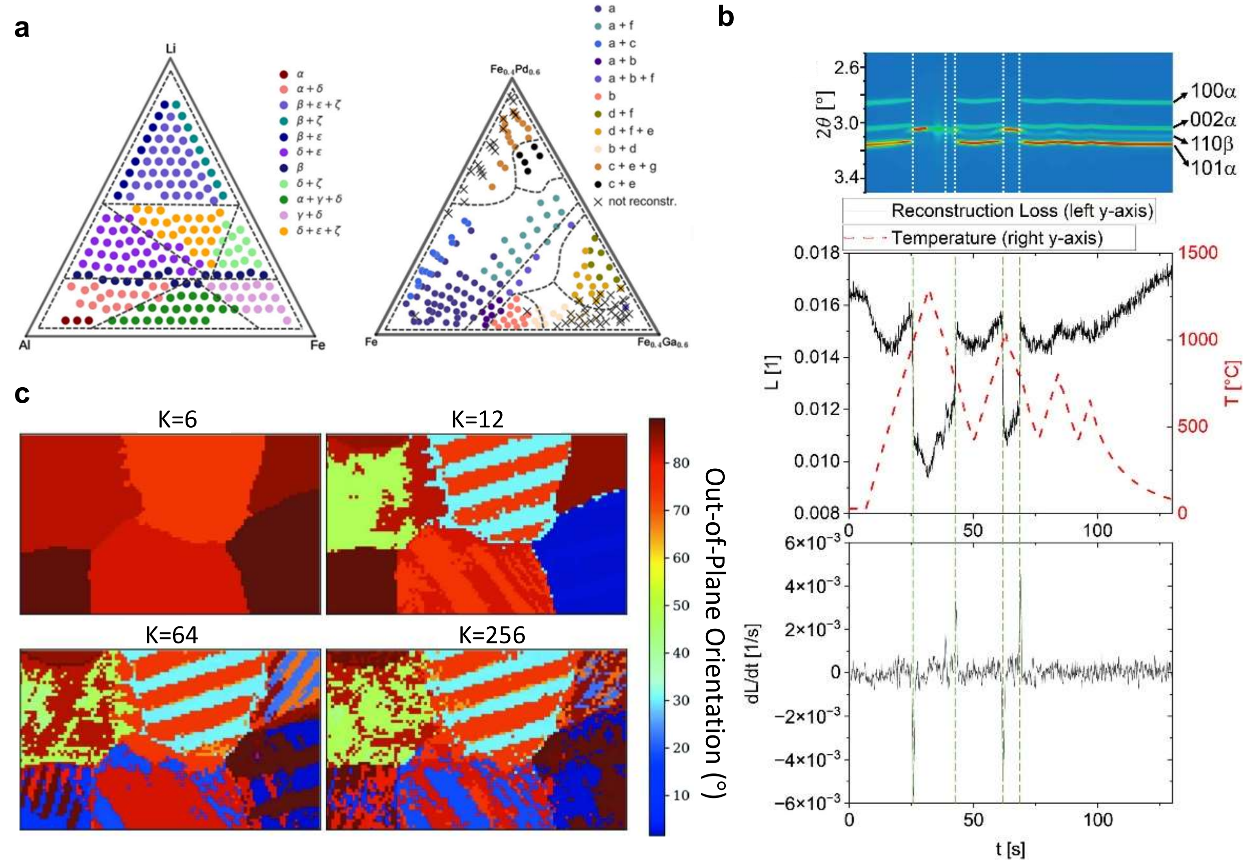

Figure 9. (a) constructed ternary phase diagrams from identified end-member phases from a simulated (left) and experimental (right) materials library.[119] (b) [Top] XRD intensity as a function of 2θ and time in an *in-situ* study. [Middle] autoencoder reconstruction loss as a function of time (black) and experimental temperature (red). [Bottom] Change in the autoencoder reconstruction loss as a function of time. Large peaks correspond to times in which the autoencoder received new XRD spectra that could not be encoded with high accuracy. Reprinted with permission from Strohmann et al. [120] Copyright 2023 by Elsevier, (c) K-means clustering analysis of spatial distribution of domains in a $BaTiO_3$ sample, where K denotes the number of clusters and the color bar is the calculated out-of-plane crystallographic orientation within a cluster. Reprinted with permission from Song et al.[122] by the International Union of Crystallography.

4 Challenges in AI Analysis of X-ray Scattering and Diffraction: Perspectives and Best Practices From the Fields of AI and Materials Science

The existing literature on the ML analysis of X-ray scattering and diffraction offers a strong basis for future expansions and applications in autonomous and high-throughput experimentation. However, as ML transitions from a novel approach to a common practice for data analysis in the sciences, the accurate curation, training, deployment, and final reporting of data analysis outcomes should be emphasized. To aid this effort, we have compiled perspectives and lessons learned from the fields of data science and materials science for emerging ML and high-throughput studies.

Data Quality and Problem Formulation: A commonly overlooked aspect of data science from the existing scientific literature is data curation. The amount, quality, and diversity of data, as well as proposed source of training data and target domain are parameters that may bottleneck the performance and generalizability of the final model. To date, ICSD, the most common database used in literature, contains over 300,000 crystal structures. While data may be supplemented with other databases such as Crystallography Open Database or Powder Diffraction File[126], serious consideration must be given to data quality, the presence of experimental artifacts, noise, secondary phases, and accidental classifications[69]. Most experimentalists are familiar with the frustration of sifting through multiple patterns and source publications to find reliable references. Similarly, only verified data (via preprocessing or manual inspection) should be used to train supervised models. The ICSD "quality" parameter quantifies data robustness for this purpose: its use could reduce the number of available training data by more than half depending on the material. Additionally, structure databases contain data duplicates—of either identical compounds or very similar—often also true of experimental data. Such source domains are termed "stratified" and typically require data handling to prevent data leakage[127], or the accidental duplication of data in the training and testing datasets resulting in overoptimistic models[128]. For example, Vecsei et al. [70] and Park et al.[68] deleted duplicate entries in the source ICSD domain, and Suzuki et al.[71] proposed a method to separate testing and training datasets chronologically to reduce the chance of leakage into the testing dataset.

The analysis of a specific target domain, or the nature of the data to be analyzed by the model, is critical after the evaluation of available training data. The nature and diversity of data in the source domain should ideally match that of the target domain for models to learn the anticipated data distribution. The effects of data diversity may become especially apparent when using different equipment, environments, and processing conditions which result in deviations of the collected pattern from simulated data. Data pre-processing and augmentation steps, such as those already covered to introduce experimental artifacts, may be used to supplement training data in such cases. A recent example of an inaccuracy in adapting the source to the target domain of experimental XRD spectra involved an autonomous materials discovery laboratory[27]. The lab synthesized 36 compositionally ordered compounds using data from Google DeepMind's GNoME database which were thought to be novel[129]. However, an analysis of the synthesized ordered compounds found that more than two thirds were identical to existing disordered analogues of the same compounds not represented in the source domain[130]. The construction of the training dataset and problem formulation is arguably the most technical and important step in ML analysis and should rely heavily on existing literature and subject matter experts in XRD.

Model Selection: Despite the wealth of literature showcasing large-parameter CNN or autoencoder architectures for classification and regression tasks, the choice of model should be dictated by the specific goals of each experiment and the anticipated domain. Particularly when data is limited, such as is common in experimental work, the best models are generally regarded as those that achieve high accuracy and precision with the fewest number of trainable parameters. General heuristics recommend the use of 10 times the quantity of training data than model parameters and more for deep learning architectures. Another consideration is model interpretability. Since the analysis of DL models is notoriously difficult, conventional models (such as support vector machines and naïve bayes models [131]) are often more interpretable and should be considered depending on the application. When deep learning is preferred, it is still possible to interpret training results using class activation mappings and feature sub-sampling. Additionally, the choice of activation function and probabilistic models discussed can avoid wasting experimental time on "confidently false" predictions. Finally, unsupervised models are complementary to supervised methods. As previously discussed, supervised methods are sensitive to the quality, diversity, outliers, and stratification of the training dataset, all parameters that can be understood first through clustering and dimensionality reduction. Measures of reconstruction error in PCA, NMF, and autoencoder models can be used to determine if a sample is new to the dataset and if supervised learning is expected to give an accurate result.

Model Training and Accuracy Scoring: The choice of training protocol and accuracy scoring can impact the model's ability to generalize to unseen data. Hyperparameter tuning is an often neglected yet critical step of model development. Model parameters that are not typically modified during training, such as the number and depth of neural network layers, are called hyperparameters. The choice of hyperparameters significantly impacts the predictive accuracy and precision of ML models and is often required documentation in data science. Standard protocol when evaluating performance during hyper-parameter tuning is to create a separate "validation" dataset separate from the testing datasets to avoid artificially inflating the model accuracy on the test dataset. The choice of testing protocol is also very important. The common protocols of train-test split (where a separate fraction of the dataset is reserved for testing) and k-fold cross validation (training using multiple subsets of the same data) can significantly impact the perceived accuracy of a predictive model: accuracies obtained from the former generalize better to unseen data in similar target domains, while those from the latter over-estimate model performance[132]. A general rule is to use the train-test split when the model training is not limited by the size of the training dataset, and to use k-fold cross validation when data is scarce or is expected to be unbalanced. Finally, after choosing a model and training protocol, the effect of training data quantity should always be reported using learning curves. Learning curves plot the model accuracy as a function of training time or training data samples and are used to evaluate model fitting and ability to generalize to new data.

Model and Data Reporting Best Practices: The reporting for ML protocols and data has received attention in many scientific fields including chemistry[133], physics[134], and materials science[127,135]. It is generally advised that all pieces of the data analysis pipeline are thoroughly documented with analysis methods spawning from initial hypotheses. These include dataset selection and availability, proven data quality metrics, the target domain and problem formulation, data preprocessing and feature representation, model selection, training protocols (if supervised), model results, and final deployment to unseen data. The careful selection of data and models, including their parameters and architecture, to achieve a hypothesized outcome should not be confused with a blind "dredging" of models and processing routes to achieve the highest accuracy. Hypotheses for the methodology and performance of each ML model should be well documented, and post hoc analysis of model performances should be avoided. For readers beginning their journey in data science curious about how best to curate and document an ML problem, we recommend the book by Aurélien Géron[136] for step-by-step tutorials as well as select papers[127,137,138]. Finally, when organizing bodies of literature or experimental training data, the FAIR[139] (Findable, Accessible, Interoperable, Reusable) data principles are now a requirement for most government funded research and have been increasingly used[140–144]. Recent materials property databases implementing FAIR data include the Perovskite Database Project[145], NOMAD[146], Materials Cloud[147], TFADB[148], and AFlow[149], which serve as inspiration for the continued development of materials structure databases.

CONCLUSIONS

The rise of ML and high-throughput characterization techniques have led to highly comprehensive and thorough analyses of solid-state structure with less bias. It is now possible to quickly analyze thousands of patterns to develop broad chemistry-structure-property relationships, an exciting prospect when coupled with the development of autonomous laboratories. However, the curation, training, deployment, and final reporting of data analysis outcomes of each new study should be carefully evaluated against known best practices to avoid pitfalls and wasted experimental resources. The biggest challenge currently facing the broader implementation of ML in solid-state structural analysis is the availability of high-quality, diverse data representative of experimental patterns, requiring innovations in the analysis, preprocessing, or standardized storage of XRD data. Following the FAIR data principles and model reporting guidelines previously discussed, researchers can ensure the transparent dissemination of their data and models in a way that maximizes their impact on the wider scientific community.

5 Acknowledgements

We acknowledge Diana K. LaFollette[1], Carlo Andrea Riccardo Perini[1], and Ruipeng Li[2] for collecting experimental data used in Fig 2c,d at 11-BM (Complex Materials Scattering) beamline at the Brookhaven national laboratory.

[1]School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA
[2]National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, NY 11973-5000, USA

6 References

1. Hull, A.W. (1921). X-Ray Crystal Analysis of Thirteen Common Metals. Physical Review *17*, 571. https://doi.org/10.1103/PhysRev.17.571.

2. Westgren, A.F., and Phragmén, G. (1929). X-ray studies on alloys. Transactions of the Faraday Society *25*, 379–385. https://doi.org/10.1039/TF9292500379.

3. Ubbelohde, A.R., and Woodward, I. (1945). Sub-crystalline Changes of Structure Accompanying Thermal Transitions in Rochelle Salt, and in Potassium Dihydrogen Orthophosphate. Nature 1945 155:3928 *155*, 170–171. https://doi.org/10.1038/155170a0.

4. Shechtman, D., Blech, I., Gratias, D., and Cahn, J.W. (1984). Metallic phase with long-range orientational order and no translational symmetry. Phys Rev Lett *53*, 1951–1953. https://doi.org/10.1103/PhysRevLett.53.1951.

5. Bragg, W.H., and Bragg, W.L. (1913). The Structure of the Diamond. Nature 1913 91:2283 *91*, 557–557. https://doi.org/10.1038/091557a0.

6. Chapman, H.N., Fromme, P., Barty, A., White, T.A., Kirian, R.A., Aquila, A., Hunter, M.S., Schulz, J., Deponte, D.P., Weierstall, U., et al. (2011). Femtosecond X-ray protein nanocrystallography. Nature 2011 470:7332 *470*, 73–77. https://doi.org/10.1038/nature09750.

7. Hartman, Ya., Snigireva, I., Snigirev, A., Engstrom, P., Riekel, C., Hartman, Ya., Snigireva, I., Snigirev, A., Engstrom, P., and Riekel, C. (1993). First testing and applications of Bragg-Fresnel crystal optics at the ESRF microfocus beamline. AcCrA *49*, c375–c376. https://doi.org/10.1107/S0108767378089461.

8. Isaacs, E.D. (2006). X-ray nanovision. Nature 2006 442:7098 *442*, 35–35. https://doi.org/10.1038/442035a.

9. Li, P., Allain, M., Grünewald, T.A., Rommel, M., Campos, A., Carbone, D., and Chamard, V. (2022). 4th generation synchrotron source boosts crystalline imaging at the nanoscale. Light: Science & Applications 2022 11:1 *11*, 1–12. https://doi.org/10.1038/s41377-022-00758-z.

10. Hidalgo, J., Kaiser, W., An, Y., Li, R., Oh, Z., Castro-Méndez, A.F., LaFollette, D.K., Kim, S., Lai, B., Breternitz, J., et al. (2023). Synergistic Role of Water and Oxygen Leads to Degradation in Formamidinium-Based Halide Perovskites. J Am Chem Soc *145*, 24549–24557. https://doi.org/10.1021/jacs.3c05657.

11. An, Y., Perini, C.A.R., Hidalgo, J., Castro-Méndez, A.F., Vagott, J.N., Li, R., Saidi, W.A., Wang, S., Li, X., and Correa-Baena, J.P. (2021). Identifying high-performance and durable methylammonium-free lead halide perovskites via high-throughput synthesis and characterization. Energy Environ Sci *14*, 6638–6654. https://doi.org/10.1039/D1EE02691G.

12. Friedrich, W., Knipping, P., and Laue, M. (1913). Interferenzerscheinungen bei Röntgenstrahlen. Ann Phys *346*, 971–988. https://doi.org/10.1002/ANDP.19133461004.

13. Bragg, W.H., Bragg Apr, W.L., H Bragg, B.W., and Professor of Physics, C. (1913). The reflection of X-rays by crystals. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character *88*, 428–438. https://doi.org/10.1098/RSPA.1913.0040.

14. Ewald, P., Pauling, L., Robertson, J., Hume-Rothery, W., Wyckoff, R., Lonsdale, K., and Siegbahn, M. (1962). 50 Years of X-ray Diffraction 1st ed. P. Ewald, ed. (International Union of Crystallography).

15. Patterson, A.L. (1939). The Scherrer Formula for X-Ray Particle Size Determination. Physical Review *56*, 978. https://doi.org/10.1103/PhysRev.56.978.

16. Lotgering, F.K. (1959). Topotactical reactions with ferrimagnetic oxides having hexagonal crystal structures—I. Journal of Inorganic and Nuclear Chemistry *9*, 113–123. https://doi.org/10.1016/0022-1902(59)80070-1.

17. Rietveld, H.M. (1969). A profile refinement method for nuclear and magnetic structures. J Appl Crystallogr *2*, 65–71. https://doi.org/10.1107/S0021889869006558.

18. Jaffe, A., Lin, Y., Beavers, C.M., Voss, J., Mao, W.L., and Karunadasa, H.I. (2016). High-pressure single-crystal structures of 3D lead-halide hybrid perovskites and pressure effects on their electronic and optical properties. ACS Cent Sci *2*, 201–209. https://doi.org/10.1021/acscentsci.6b00055.

19. Nandi, P., Giri, C., Swain, D., Manju, U., and Topwal, D. (2019). Room temperature growth of CH3NH3PbCl3 single crystals by solvent evaporation method. CrystEngComm *21*, 656–661. https://doi.org/10.1039/C8CE01939H.

20. Marcos, C. (2022). Methods and Applications of X-ray Diffraction in Crystallography and Mineralogy. 383–436. https://doi.org/10.1007/978-3-030-96783-3_17.

21. Holzer, I., Lemaur, V., Wang, M., Wu, H.Y., Zhang, L., Marcial-Hernandez, R., Gilhooly-Finn, P., Cavassin, P., Hoyas, S., Meli, D., et al. (2024). Side chain engineering in indacenodithiophene-co-benzothiadiazole and its impact on mixed ionic–electronic transport properties. J Mater Chem C Mater *12*, 3686–3697. https://doi.org/10.1039/D3TC04738E.

22. Li, T., Senesi, A.J., and Lee, B. (2016). Small Angle X-ray Scattering for Nanoparticle Research. Chem Rev *116*, 11128–11180. https://doi.org/10.1021/acs.chemrev.5b00690.

23. Tsao, C.S., Yu, M.S., Chung, T.Y., Wu, H.C., Wang, C.Y., Chang, K. Sen, and Chen, H.L. (2007). Characterization of pore structure in metal-organic framework by small-angle X-ray scattering. J Am Chem Soc *129*, 15997–16004. https://doi.org/10.1021/ja0752336.

24. Gražulis, S., Chateigner, D., Downs, R.T., Yokochi, A.F.T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., and Le Bail, A. (2009). Crystallography Open Database – an open-access collection of crystal structures. J Appl Crystallogr *42*, 726–729. https://doi.org/10.1107/S0021889809016690.

25. Zagorac, D., Müller, H., Ruehl, S., Zagorac, J., and Rehme, S. (2019). Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. J Appl Crystallogr *52*, 918–925. https://doi.org/10.1107/S160057671900997X.

26. Zhang, J., Hauch, J.A., and Brabec, C.J. (2024). Toward Self-Driven Autonomous Material and Device Acceleration Platforms (AMADAP) for Emerging Photovoltaics Technologies. Acc Chem Res *57*, 1434–1445. https://doi.org/10.1021/acs.accounts.4c00095.

27. Szymanski, N.J., Rendy, B., Fei, Y., Kumar, R.E., He, T., Milsted, D., McDermott, M.J., Gallant, M., Cubuk, E.D., Merchant, A., et al. (2023). An autonomous laboratory for the accelerated synthesis of novel materials. Nature *624*, 86–91. https://doi.org/10.1038/S41586-023-06734-W.

28. Salley, D., Keenan, G., Grizou, J., Sharma, A., Martín, S., and Cronin, L. (2020). A nanomaterials discovery robot for the Darwinian evolution of shape programmable gold nanoparticles. Nature Communications 2020 11:1 *11*, 1–7. https://doi.org/10.1038/s41467-020-16501-4.

29. Epps, R.W., Bowen, M.S., Volk, A.A., Abdel-Latif, K., Han, S., Reyes, K.G., Amassian, A., Abolhasani, M., Epps, R.W., Bowen, M.S., et al. (2020). Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot. Advanced Materials *32*, 2001626. https://doi.org/10.1002/ADMA.202001626.

30. Reis, M., Gusev, F., Taylor, N.G., Chung, S.H., Verber, M.D., Lee, Y.Z., Isayev, O., and Leibfarth, F.A. (2021). Machine-Learning-Guided Discovery of 19F MRI Agents Enabled by Automated Copolymer Synthesis. J Am Chem Soc *143*, 17677–17689. https://doi.org/10.1021/jacs.1c08181.

31. Salley, D.S., Keenan, G.A., Long, D.L., Bell, N.L., and Cronin, L. (2020). A Modular Programmable Inorganic Cluster Discovery Robot for the Discovery and Synthesis of Polyoxometalates. ACS Cent Sci *6*, 1587–1593. https://doi.org/10.1021/acscentsci.0c00415.

32. Burger, B., Maffettone, P.M., Gusev, V. V., Aitchison, C.M., Bai, Y., Wang, X., Li, X., Alston, B.M., Li, B., Clowes, R., et al. (2020). A mobile robotic chemist. Nature 2020 583:7815 *583*, 237–241. https://doi.org/10.1038/s41586-020-2442-2.

33. Abolhasani, M., and Kumacheva, E. (2023). The rise of self-driving labs in chemical and materials sciences. Nature Synthesis 2023 2:6 *2*, 483–492. https://doi.org/10.1038/s44160-022-00231-0.

34. Gregoire, J.M., Zhou, L., and Haber, J.A. (2023). Combinatorial synthesis for AI-driven materials discovery. Nature Synthesis 2023 2:6 *2*, 493–504. https://doi.org/10.1038/s44160-023-00251-4.

35. Zarnetta, R., Buenconsejo, P.J.S., Savan, A., Thienhaus, S., and Ludwig, A. (2012). High-throughput study of martensitic transformations in the complete Ti–Ni–Cu system. Intermetallics (Barking) *26*, 98–109. https://doi.org/10.1016/J.INTERMET.2012.03.044.

36. Ludwig, A. (2019). Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. npj Computational Materials 2019 5:1 *5*, 1–7. https://doi.org/10.1038/s41524-019-0205-0.

37. Li, J., Li, J., Liu, R., Tu, Y., Li, Y., Cheng, J., He, T., and Zhu, X. (2020). Autonomous discovery of optically active chiral inorganic perovskite nanocrystals through an intelligent cloud lab. Nature Communications 2020 11:1 *11*, 1–10. https://doi.org/10.1038/s41467-020-15728-5.

38. Liu, Y., Tan, X., Liang, J., Han, H., Xiang, P., Yan, W., Liu, Y., Tan, X., Liang, J., Xiang, P., et al. (2023). Machine Learning for Perovskite Solar Cells and Component Materials: Key Technologies and Prospects. Adv Funct Mater *33*, 2214271. https://doi.org/10.1002/ADFM.202214271.

39. Wang, Z., Yang, M., Xie, X., Yu, C., Jiang, Q., Huang, M., Algadi, H., Guo, Z., and Zhang, H. (2022). Applications of machine learning in perovskite materials. Adv Compos Hybrid Mater *5*, 2700–2720. https://doi.org/10.1007/S42114-022-00560-W/METRICS.

40. Hui, Z., Wang, M., Yin, X., Wang, Y., and Yue, Y. (2023). Machine learning for perovskite solar cell design. Comput Mater Sci *226*, 112215. https://doi.org/10.1016/J.COMMATSCI.2023.112215.

41. Zhang, L., He, M., and Shao, S. (2020). Machine learning for halide perovskite materials. Nano Energy *78*, 105380. https://doi.org/10.1016/J.NANOEN.2020.105380.

42. Tao, Q., Xu, P., Li, M., and Lu, W. (2021). Machine learning for perovskite materials design and discovery. npj Computational Materials 2021 7:1 *7*, 1–18. https://doi.org/10.1038/s41524-021-00495-8.

43. Chen, A., Zhang, X., and Zhou, Z. (2020). Machine learning: Accelerating materials development for energy storage and conversion. InfoMat *2*, 553–576. https://doi.org/10.1002/INF2.12094.

44. Chen, C., Zuo, Y., Ye, W., Li, X., Deng, Z., Ping Ong, S., Chen, C., Zuo, Y., Ye, W., Li, X., et al. (2020). A Critical Review of Machine Learning of Energy Materials. Adv Energy Mater *10*, 1903242. https://doi.org/10.1002/AENM.201903242.

45. Gu, G.H., Noh, J., Kim, I., and Jung, Y. (2019). Machine learning for renewable energy materials. J Mater Chem A Mater *7*, 17096–17117. https://doi.org/10.1039/C9TA02356A.

46. Chowdhury, A., Kautz, E., Yener, B., and Lewis, D. (2016). Image driven machine learning methods for microstructure recognition. Comput Mater Sci *123*, 176–187. https://doi.org/10.1016/J.COMMATSCI.2016.05.034.

47. Holm, E.A., Cohn, R., Gao, N., Kitahara, A.R., Matson, T.P., Lei, B., and Yarasi, S.R. (2020). Overview: Computer vision and machine learning for microstructural characterization and analysis. Metall Mater Trans A Phys Metall Mater Sci *51*, 5985–5999. https://doi.org/10.1007/s11661-020-06008-4.

48. Baskaran, A., Kautz, E.J., Chowdhary, A., Ma, W., Yener, B., and Lewis, D.J. (2021). The Adoption of Image-Driven Machine Learning for Microstructure Characterization and Materials Design: A Perspective. JOM *73*, 3639–3657. https://doi.org/10.1007/s11837-021-04805-9.

49. Zhang, L., He, M., and Shao, S. (2020). Machine learning for halide perovskite materials. Nano Energy *78*, 105380. https://doi.org/10.1016/J.NANOEN.2020.105380.

50. Liu, T., Wang, S., Shi, Y., Wu, L., Zhu, R., Wang, Y., Zhou, J., and Choy, W.C.H. (2023). Machine-Learning Accelerating the Development of Perovskite Photovoltaics. Solar RRL *7*, 2300650. https://doi.org/10.1002/SOLR.202300650.

51. Yılmaz, B., and Yıldırım, R. (2021). Critical review of machine learning applications in perovskite solar research. Nano Energy *80*, 105546. https://doi.org/10.1016/J.NANOEN.2020.105546.

52. Datta, S., Baul, A., Sarker, G.C., Sadhu, P.K., and Hodges, D.R. (2023). A Comprehensive Review of the Application of Machine Learning in Fabrication and Implementation of Photovoltaic Systems. IEEE Access *11*, 77750–77778. https://doi.org/10.1109/ACCESS.2023.3298542.

53. Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., Park, C.W., Choudhary, A., Agrawal, A., Billinge, S.J.L., et al. (2022). Recent advances and applications of deep learning methods in materials science. npj Computational Materials 2022 8:1 *8*, 1–26. https://doi.org/10.1038/s41524-022-00734-6.

54. Batra, R., Song, L., and Ramprasad, R. (2020). Emerging materials intelligence ecosystems propelled by machine learning. Nature Reviews Materials 2020 6:8 *6*, 655–678. https://doi.org/10.1038/s41578-020-00255-y.

55. Pilania, G. (2021). Machine learning in materials science: From explainable predictions to autonomous design. Comput Mater Sci *193*, 110360. https://doi.org/10.1016/J.COMMATSCI.2021.110360.

56. Kowalski, B.R. ed. (1984). Chemometrics. https://doi.org/10.1007/978-94-017-1026-8.

57. Minsky, M. (1952). A neural-analogue calculator based upon a probability model of reinforcement. Harvard University Psychological Laboratories, Cambridge, Massachusetts.

58. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev *65*, 386–408. https://doi.org/10.1037/H0042519.

59. Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences *79*, 2554–2558. https://doi.org/10.1073/PNAS.79.8.2554.

60. Hinton, G.E., and Sejnowski, T.J. Learning and Relearning in Boltzmann Machines. https://doi.org/10.7551/mitpress/3349.001.0001.

61. Cortes, C. (1995). Support-Vector Networks. *20*, 273–297.

62. Lecun, Y., Kavukcuoglu, K., and Farabet, C. Convolutional Networks and Applications in Vision.

63. Hey, T., Butler, K., Jackson, S., and Thiyagalingam, J. (2020). Machine learning and big scientific data. Philosophical Transactions of the Royal Society A *378*. https://doi.org/10.1098/RSTA.2019.0054.

64. Rupp, M., Von Lilienfeld, O.A., and Burke, K. (2018). Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. Journal of Chemical Physics *148*. https://doi.org/10.1063/1.5043213.

65. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. (2013). Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Mater *1*, 11002. https://doi.org/10.1063/1.4812323/119685.

66. Stuart, R., and Peter, N. (2016). Artificial Intelligence - A Modern Approach 3rd Ed.

67. Roberts, D.A., Yaida, S., and Hanin, B. (2021). The Principles of Deep Learning Theory. The Principles of Deep Learning Theory. https://doi.org/10.1017/9781009023405.

68. Park, W.B., Chung, J., Jung, J., Sohn, K., Singh, S.P., Pyo, M., Shin, N., and Sohn, K.S. (2017). Classification of crystal structure using a convolutional neural network. IUCrJ *4*, 486–494. https://doi.org/10.1107/S205225251700714X.

69. Corriero, N., Rizzi, R., Settembre, G., Del Buono, N., and Diacono, D. (2023). CrystalMELA: a new crystallographic machine learning platform for crystal system determination. urn:issn:1600-5767 *56*, 409–419. https://doi.org/10.1107/S1600576723000596.

70. Vecsei, P.M., Choo, K., Chang, J., and Neupert, T. (2019). Neural network based classification of crystal symmetries from x-ray diffraction patterns. Phys Rev B *99*, 245120. https://doi.org/10.1103/PhysRevB.99.245120.

71. Suzuki, Y., Hino, H., Hawai, T., Saito, K., Kotsugi, M., and Ono, K. (2020). Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach. Scientific Reports 2020 10:1 *10*, 1–11. https://doi.org/10.1038/s41598-020-77474-4.

72. Zaloga, A.N., Stanovov, V. V., Bezrukova, O.E., Dubinin, P.S., and Yakimov, I.S. (2020). Crystal symmetry classification from powder X-ray diffraction patterns using a convolutional neural network. Mater Today Commun *25*, 101662. https://doi.org/10.1016/J.MTCOMM.2020.101662.

73. Chakraborty, A., and Sharma, R. (2022). A deep crystal structure identification system for X-ray diffraction patterns. Visual Computer *38*, 1275–1282. https://doi.org/10.1007/s00371-021-02165-8.

74. Lee, B. Do, Lee, J.-W., Park, W.B., Park, J., Cho, M.-Y., Singh, S.P., Pyo, M., and Sohn, K.-S. (2022). Powder X-Ray Diffraction Pattern Is All You Need for Machine-Learning-Based Symmetry Identification and Property Prediction. Advanced Intelligent Systems *4*, 2200042. https://doi.org/10.1002/AISY.202200042.

75. Lee, B. Do, Lee, J.-W., Ahn, J., Kim, S., Park, W.B., and Sohn, K.-S. (2023). A Deep Learning Approach to Powder X-Ray Diffraction Pattern Analysis: Addressing Generalizability and Perturbation Issues Simultaneously. Advanced Intelligent Systems *5*, 2300140. https://doi.org/10.1002/AISY.202300140.

76. Schopmans, H., Reiser, P., and Friederich, P. (2023). Neural networks trained on synthetically generated crystals can extract structural information from ICSD powder X-ray diffractograms. Digital Discovery *2*, 1414–1424. https://doi.org/10.1039/D3DD00071K.

77. Salgado, J.E., Lerman, S., Du, Z., Xu, C., and Abdolrahim, N. (2023). Automated classification of big X-ray diffraction data using deep learning models. npj Computational Materials 2023 9:1 *9*, 1–12. https://doi.org/10.1038/s41524-023-01164-8.

78. Lafuente, B., Downs, R.T., and Stone, N. (2015). The power of databases: the RRUFF project. In: Highlights in Mineralogical Crystallography. W. De Gruyter, 1–30. https://doi.org/10.1515/9783110417104-003.

79. Lee, J.W., Park, W.B., Lee, J.H., Singh, S.P., and Sohn, K.S. (2020). A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. Nature Communications 2020 11:1 *11*, 1–11. https://doi.org/10.1038/s41467-019-13749-3.

80. Oviedo, F., Ren, Z., Sun, S., Settens, C., Liu, Z., Hartono, N.T.P., Ramasamy, S., DeCost, B.L., Tian, S.I.P., Romano, G., et al. (2019). Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. npj Computational Materials 2019 5:1 *5*, 1–9. https://doi.org/10.1038/s41524-019-0196-x.

81. Szymanski, N.J., Bartel, C.J., Zeng, Y., Tu, Q., and Ceder, G. (2021). Probabilistic Deep Learning Approach to Automate the Interpretation of Multi-phase Diffraction Spectra. Chemistry of Materials *33*, 4204–4215. https://doi.org/10.1021/acs.chemmater.1c01071.

82. Wang, H., Xie, Y., Li, D., Deng, H., Zhao, Y., Xin, M., and Lin, J. (2020). Rapid Identification of X-ray Diffraction Patterns Based on Very Limited Data by Interpretable Convolutional Neural Networks. J Chem Inf Model *60*, 2004–2011. https://doi.org/10.1021/acs.jcim.0c00020.

83. Sun, S., Hartono, N.T.P., Ren, Z.D., Oviedo, F., Buscemi, A.M., Layurova, M., Chen, D.X., Ogunfunmi, T., Thapa, J., Ramasamy, S., et al. (2019). Accelerated Development of Perovskite-Inspired Materials via High-Throughput Synthesis and Machine-Learning Diagnosis. Joule *3*, 1437–1451. https://doi.org/10.1016/J.JOULE.2019.05.014.

84. Oviedo, F., Ren, Z., Sun, S., Settens, C., Liu, Z., Hartono, N.T.P., Ramasamy, S., DeCost, B.L., Tian, S.I.P., Romano, G., et al. (2019). Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. npj Computational Materials 2019 5:1 *5*, 1–9. https://doi.org/10.1038/s41524-019-0196-x.

85. Klug, H., and Alexander, L. (1974). X-ray Diffraction Procedures: For Polycrystalline and Amorphous Materials, 2nd Edition. Willey, New York, EUA, 992.

86. Zevin, L.S. (1977). A method of quantitative phase analysis without standards. urn:issn:0021-8898 *10*, 147–150. https://doi.org/10.1107/S0021889877013144.

87. Fiala, J. (1980). Powder Diffraction Analysis of a Three-Component Sample. Anal Chem *52*, 1272–1275. https://doi.org/10.1021/ac50058a034.

88. Lee, J.W., Park, W.B., Kim, M., Pal Singh, S., Pyo, M., and Sohn, K.S. (2021). A data-driven XRD analysis protocol for phase identification and phase-fraction prediction of multiphase inorganic compounds. Inorg Chem Front *8*, 2492–2504. https://doi.org/10.1039/D0QI01513J.

89. Szymanski, N.J., Bartel, C.J., Zeng, Y., Diallo, M., Kim, H., and Ceder, G. (2023). Adaptively driven X-ray diffraction guided by machine learning for autonomous phase identification. npj Computational Materials 2023 9:1 *9*, 1–8. https://doi.org/10.1038/s41524-023-00984-y.

90. Amamoto, Y., Kikutake, H., Kojio, K., Takahara, A., and Terayama, K. (2021). Visualization of judgment regions in convolutional neural networks for X-ray diffraction and scattering images of aliphatic polyesters. Polymer Journal 2021 53:11 *53*, 1269–1279. https://doi.org/10.1038/s41428-021-00531-w.

91. Nawaz, S., Rahmani, V., Pennicard, D., Setty, S.P.R., Klaudel, B., and Graafsma, H. (2023). Explainable machine learning for diffraction patterns. urn:issn:1600-5767 *56*, 1494–1504. https://doi.org/10.1107/S1600576723007446.

92. Massuyeau, F., Broux, T., Coulet, F., Demessence, A., Mesbah, A., and Gautier, R. (2022). Perovskite or Not Perovskite? A Deep-Learning Approach to Automatically Identify New Hybrid Perovskites from X-ray Diffraction Patterns. Advanced Materials *34*, 2203879. https://doi.org/10.1002/ADMA.202203879.

93. Maffettone, P.M., Banko, L., Cui, P., Lysogorskiy, Y., Little, M.A., Olds, D., Ludwig, A., and Cooper, A.I. (2021). Crystallography companion agent for high-throughput materials discovery. Nature Computational Science 2021 1:4 *1*, 290–297. https://doi.org/10.1038/s43588-021-00059-2.

94. Chitturi, S.R., Ratner, D., Walroth, R.C., Thampy, V., Reed, E.J., Dunne, M., Tassone, C.J., and Stone, K.H. (2021). Automated prediction of lattice parameters from X-ray powder diffraction patterns. J Appl Crystallogr *54*, 1799–1810. https://doi.org/10.1107/S1600576721010840.

95. Dong, H., Butler, K.T., Matras, D., Price, S.W.T., Odarchenko, Y., Khatry, R., Thompson, A., Middelkoop, V., Jacques, S.D.M., Beale, A.M., et al. (2021). A deep convolutional neural network for real-time full profile analysis of big powder diffraction data. npj Computational Materials 2021 7:1 *7*, 1–9. https://doi.org/10.1038/s41524-021-00542-4.

96. Lee, J., Oba, J., Ohba, N., and Kajita, S. (2023). Creation of crystal structure reproducing X-ray diffraction pattern without using database. npj Computational Materials 2023 9:1 *9*, 1–9. https://doi.org/10.1038/s41524-023-01096-3.

97. Dong, R., Zhao, Y., Song, Y., Fu, N., Omee, S.S., Dey, S., Li, Q., Wei, L., and Hu, J. (2022). DeepXRD, a Deep Learning Model for Predicting XRD spectrum from Material Composition. ACS Appl Mater Interfaces *14*, 40102–40115. https://doi.org/10.1021/acsami.2c05812.

98. Artursson, T., Hagman, A., Björk, S., Trygg, J., Wold, S., and Jacobsson, S.P. (2000). Study of preprocessing methods for the determination of crystalline phases in binary mixtures of drug substances by X-ray powder diffraction and multivariate calibration. Appl Spectrosc *54*, 1222–1230. https://doi.org/10.1366/0003702001950805.

99. Karstang, T. V., and Eastgate, R.J. (1987). Multivariate calibration of an x-ray diffractometer by partial least squares regression. Chemometrics and Intelligent Laboratory Systems *2*, 209–219. https://doi.org/10.1016/0169-7439(87)80098-9.

100. Szymanski, N.J., Fu, S., Persson, E., and Ceder, G. (2024). Integrated analysis of X-ray diffraction patterns and pair distribution functions for machine-learned phase identification. npj Computational Materials 2024 10:1 *10*, 1–9. https://doi.org/10.1038/s41524-024-01230-9.

101. Liu, C.H., Tao, Y., Hsu, D., Du, Q., and Billinge, S.J.L. (2019). Using a machine learning approach to determine the space group of a structure from the atomic pair distribution function. urn:issn:2053-2733 *75*, 633–643. https://doi.org/10.1107/S2053273319005606.

102. Lan, L., Liu, C.H., Du, Q., and Billinge, S.J.L. (2022). Robustness test of the spacegroupMining model for determining space groups from atomic pair distribution function data. J Appl Crystallogr *55*, 626–630. https://doi.org/10.1107/S1600576722002990.

103. Niitsu, N., Mitani, M., Ishii, H., Kobayashi, N., Hirose, K., Watanabe, S., Okamoto, T., and Takeya, J. (2024). Powder x-ray diffraction analysis with machine learning for organic-semiconductor crystal-structure determination. Appl Phys Lett *125*. https://doi.org/10.1063/5.0208919.

104. Ke, T.W., Brewster, A.S., Yu, S.X., Ushizima, D., Yang, C., and Sauter, N.K. (2018). A convolutional neural network-based screening tool for X-ray serial crystallography. urn:issn:1600-5775 *25*, 655–670. https://doi.org/10.1107/S1600577518004873.

105. Sadri, A., Hadian-Jazi, M., Yefanov, O., Galchenkova, M., Kirkwood, H., Mills, G., Sikorski, M., Letrun, R., De Wijn, R., Vakili, M., et al. (2022). Automatic bad-pixel mask maker for X-ray pixel detectors with application to serial crystallography. urn:issn:1600-5767 *55*, 1549–1561. https://doi.org/10.1107/S1600576722009815.

106. Matos, C.R.S., Xavier, M.J., Barreto, L.S., Costa, N.B., and Gimenez, I.F. (2007). Principal component analysis of X-ray diffraction patterns to yield morphological classification of brucite particles. Anal Chem *79*, 2091–2095. https://doi.org/10.1021/ac061991n.

107. Banko, L., Maffettone, P.M., Naujoks, D., Olds, D., and Ludwig, A. (2021). Deep learning for visualization and novelty detection in large X-ray diffraction datasets. npj Computational Materials 2021 7:1 *7*, 1–6. https://doi.org/10.1038/s41524-021-00575-9.

108. Utimula, K., Yano, M., Kimoto, H., Hongo, K., Nakano, K., and Maezono, R. (2023). Feature Space of XRD Patterns Constructed by an Autoencoder. Adv Theory Simul *6*, 2200613. https://doi.org/10.1002/ADTS.202200613.

109. An, Y., Perini, C.A.R., Hidalgo, J., Castro-Méndez, A.F., Vagott, J.N., Li, R., Saidi, W.A., Wang, S., Li, X., and Correa-Baena, J.P. (2021). Identifying high-performance and durable methylammonium-free lead halide perovskites via high-throughput synthesis and characterization. Energy Environ Sci *14*, 6638–6654. https://doi.org/10.1039/D1EE02691G.

110. Takeuchi, I., Long, C.J., Famodu, O.O., Murakami, M., Hattrick-Simpers, J., Rubloff, G.W., Stukowski, M., and Rajan, K. (2005). Data management and visualization of x-ray diffraction

spectra from thin film ternary composition spreads. Review of Scientific Instruments *76*, 62223. https://doi.org/10.1063/1.1927079/352847.

111. Hasegawa, K., Ahmet, P., Okazaki, N., Hasegawa, T., Fujimoto, K., Watanabe, M., Chikyow, T., and Koinuma, H. (2004). Amorphous stability of HfO2 based ternary and binary composition spread oxide films as alternative gate dielectrics. Appl Surf Sci *223*, 229–232. https://doi.org/10.1016/S0169-4332(03)00903-6.

112. 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*, 72217. https://doi.org/10.1063/1.2755487/380219.

113. Long, C.J., Bunker, D., Li, X., Karen, V.L., and Takeuchi, I. (2009). Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. Review of Scientific Instruments *80*, 103902. https://doi.org/10.1063/1.3216809/354187.

114. Kusne, A.G., Keller, D., Anderson, A., Zaban, A., and Takeuchi, I. (2015). High-throughput determination of structural phase diagram and constituent phases using GRENDEL. Nanotechnology *26*, 444002. https://doi.org/10.1088/0957-4484/26/44/444002.

115. Al Hasan, N.M., Hou, H., Gao, T., Counsell, J., Sarker, S., Thienhaus, S., Walton, E., Decker, P., Mehta, A., Ludwig, A., et al. (2020). Combinatorial Exploration and Mapping of Phase Transformation in a Ni-Ti-Co Thin Film Library. ACS Comb Sci *22*, 641–648. https://doi.org/10.1021/acscombsci.0c00097.

116. Iwasaki, Y., Kusne, A.G., and Takeuchi, I. (2017). Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. npj Computational Materials 2017 3:1 *3*, 1–9. https://doi.org/10.1038/s41524-017-0006-2.

117. Utimula, K., Hunkao, R., Yano, M., Kimoto, H., Hongo, K., Kawaguchi, S., Suwanna, S., and Maezono, R. (2020). Machine-Learning Clustering Technique Applied to Powder X-Ray Diffraction Patterns to Distinguish Compositions of ThMn12-Type Alloys. Adv Theory Simul *3*, 2000039. https://doi.org/10.1002/ADTS.202000039.

118. Suram, S.K., Xue, Y., Bai, J., Le Bras, R., Rappazzo, B., Bernstein, R., Bjorck, J., Zhou, L., Van Dover, R.B., Gomes, C.P., et al. (2017). 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. https://doi.org/10.1021/acscombsci.6b00153.

119. Stanev, V., Vesselinov, V. V., Kusne, A.G., Antoszewski, G., Takeuchi, I., and Alexandrov, B.S. (2018). Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. npj Computational Materials 2018 4:1 *4*, 1–10. https://doi.org/10.1038/s41524-018-0099-2.

120. Strohmann, T., Barriobero-Vila, P., Gussone, J., Melching, D., Stark, A., Schell, N., and Requena, G. (2023). Can unsupervised machine learning boost the on-site analysis of in situ synchrotron diffraction data? Scr Mater *226*, 115238. https://doi.org/10.1016/J.SCRIPTAMAT.2022.115238.

121. Venderley, J., Mallayya, K., Matty, M., Krogstad, M., Ruff, J., Pleiss, G., Kishore, V., Mandrus, D., Phelan, D., Poudel, L., et al. (2022). Harnessing interpretable and unsupervised machine learning to address big data from modern X-ray diffraction. Proc Natl Acad Sci U S A *119*, e2109665119. https://doi.org/10.1073/pnas.2109665119.

122. Song, Y., Tamura, N., Zhang, C., Karami, M., and Chen, X. (2019). Data-driven approach for synchrotron X-ray Laue microdiffraction scan analysis. urn:issn:2053-2733 *75*, 876–888. https://doi.org/10.1107/S2053273319012804.

123. Christiansen-Salameh, J., Yang, M., Rippy, G., Li, J., Cai, Z., Holt, M., Agnus, G., Maroutian, T., Lecoeur, P., Matzen, S., et al. (2021). Understanding nanoscale structural distortions in Pb(Zr0.2Ti0.8)O3by utilizing X-ray nanodiffraction and clustering algorithm analysis. J Synchrotron Radiat *28*, 207–213. https://doi.org/10.1107/S1600577520013661.

124. Gorobtsov, O.Y., Miao, L., Shao, Z., Tan, Y., Schnitzer, N., Goodge, B.H., Ruf, J., Weinstock, D., Cherukara, M., Holt, M.V., et al. (2024). Spontaneous Supercrystal Formation During a Strain-Engineered Metal–Insulator Transition. Advanced Materials *36*, 2403873. https://doi.org/10.1002/ADMA.202403873.

125. Luo, A., Gorobtsov, O.Y., Nelson, J.N., Kuo, D.Y., Zhou, T., Shao, Z., Bouck, R., Cherukara, M.J., Holt, M. V., Shen, K.M., et al. (2022). X-ray nano-imaging of defects in thin film catalysts via cluster analysis. Appl Phys Lett *121*, 153904. https://doi.org/10.1063/5.0125268.

126. Surdu, V.A., and Győrgy, R. (2023). X-ray Diffraction Data Analysis by Machine Learning Methods—A Review. Applied Sciences 2023, Vol. 13, Page 9992 *13*, 9992. https://doi.org/10.3390/APP13179992.

127. Karande, P., Gallagher, B., and Han, T.Y.J. (2022). A Strategic Approach to Machine Learning for Material Science: How to Tackle Real-World Challenges and Avoid Pitfalls. Chemistry of Materials *34*, 7650–7665. https://doi.org/10.1021/acs.chemmater.2c01333.

128. Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S.B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J., Love, B.C., et al. (2023). On Leakage in Machine Learning Pipelines. ArXiv. https://doi.org/10.48550/arXiv.2311.04179.

129. Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M., Cheon, G., and Cubuk, E.D. (2023). Scaling deep learning for materials discovery. Nature 2023 624:7990 *624*, 80–85. https://doi.org/10.1038/s41586-023-06735-9.

130. Leeman, J., Liu, Y., Stiles, J., Lee, S.B., Bhatt, P., Schoop, L.M., and Palgrave, R.G. (2024). Challenges in High-Throughput Inorganic Materials Prediction and Autonomous Synthesis. PRX Energy *3*, 011002. https://doi.org/10.1103/PRXEnergy.3.011002.

131. Greasley, J., and Hosein, P. (2023). Exploring supervised machine learning for multi-phase identification and quantification from powder X-ray diffraction spectra. J Mater Sci *58*, 5334–5348. https://doi.org/10.1007/s10853-023-08343-4.

132. Morgan, D., and Jacobs, R. (2020). Opportunities and Challenges for Machine Learning in Materials Science. Annu Rev Mater Res *50*, 71–103. https://doi.org/10.1146/ANNUREV-MATSCI-070218-010015/1.

133. Artrith, N., Butler, K.T., Coudert, F.X., Han, S., Isayev, O., Jain, A., and Walsh, A. (2021). Best practices in machine learning for chemistry. Nature Chemistry 2021 13:6 *13*, 505–508. https://doi.org/10.1038/s41557-021-00716-z.

134. Bedolla, E., Padierna, L.C., and Castañeda-Priego, R. (2020). Machine learning for condensed matter physics. Journal of Physics: Condensed Matter *33*, 053001. https://doi.org/10.1088/1361-648X/ABB895.

135. Wang, A.Y.T., Murdock, R.J., Kauwe, S.K., Oliynyk, A.O., Gurlo, A., Brgoch, J., Persson, K.A., and Sparks, T.D. (2020). Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. Chemistry of Materials *32*, 4954–4965. https://doi.org/10.1021/acs.chemmater.0c01907.

136. Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Unsupervised learning techniques. O'Reilly Media, 510.

137. Paleyes, A., Urma, R.G., and Lawrence, N.D. (2022). Challenges in Deploying Machine Learning: A Survey of Case Studies. ACM Comput Surv *55*. https://doi.org/10.1145/3533378.

138. Zhang, Y., and Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. NPJ Comput Mater *4*. https://doi.org/10.1038/S41524-018-0081-Z.

139. Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data *3*. https://doi.org/10.1038/SDATA.2016.18.

140. Wunder, N., Guba, N., Sivaraman, C., Van Allsburg, K.M., Dinh, H., and Pailing, C. (2021). Energy Material Network Data Hubs: Software Platforms for Advancing Collaborative Energy Materials Research. IJACSA) International Journal of Advanced Computer Science and Applications *12*, 2021. https://doi.org/10.14569/IJACSA.2021.0120677.

141. Shanahan, H., Hoebelheinrich, N., and Whyte, A. (2021). Progress toward a comprehensive teaching approach to the FAIR data principles. https://doi.org/10.1016/j.patter.2021.100324.

142. Aggour, K.S., Kumar, V.S., Gupta, V.K., Gabaldon, A., Cuddihy, P., and Mulwad, V. (2024). Semantics-Enabled Data Federation: Bringing Materials Scientists Closer to FAIR Data. Integr Mater Manuf Innov, 1–15. https://doi.org/10.1007/s40192-024-00348-4.

143. Draxl, C., and Scheffler, M. (2020). Big-Data-Driven Materials Science and its FAIR Data Infrastructure https://doi.org/10.1007/978-3-319-44677-6_104.

144. Scheffler, M., Aeschlimann, M., Albrecht, M., Bereau, T., Bungartz, H.J., Felser, C., Greiner, M., Groß, A., Koch, C.T., Kremer, K., et al. (2022). FAIR data enabling new horizons for materials research. Nature *604*, 635–642. https://doi.org/10.1038/S41586-022-04501-X.

145. Jacobsson, T.J., Hultqvist, A., García-Fernández, A., Anand, A., Al-Ashouri, A., Hagfeldt, A., Crovetto, A., Abate, A., Ricciardulli, A.G., Vijayan, A., et al. (2021). An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nature Energy 2021 7:1 *7*, 107–115. https://doi.org/10.1038/s41560-021-00941-3.

146. Scheidgen, M., Himanen, L., Ladines, A.N., Sikter, D., Nakhaee, M., Fekete, Á., Chang, T., Golparvar, A., Márquez, J.A., Brockhauser, S., et al. (2023). NOMAD: A distributed web-based platform for managing materials science research data. J Open Source Softw *8*, 5388. https://doi.org/10.21105/JOSS.05388.

147. Talirz, L., Kumbhar, S., Passaro, E., Yakutovich, A. V., Granata, V., Gargiulo, F., Borelli, M., Uhrin, M., Huber, S.P., Zoupanos, S., et al. (2020). Materials Cloud, a platform for open computational science. Sci Data *7*. https://doi.org/10.1038/S41597-020-00637-5.

148. Song, J., Jo, H., Kim, T., and Lee, D. (2023). Experimental data management platform for data-driven investigation of combinatorial alloy thin films. APL Mater *11*, 91117. https://doi.org/10.1063/5.0162158/2912941.

149. Esters, M., Oses, C., Divilov, S., Eckert, H., Friedrich, R., Hicks, D., Mehl, M.J., Rose, F., Smolyanyuk, A., Calzolari, A., et al. (2023). aflow.org: A web ecosystem of databases, software and tools. Comput Mater Sci *216*, 111808. https://doi.org/10.1016/J.COMMATSCI.2022.111808.