

Identifying crystal structures beyond known prototypes from x-ray powder diffraction spectra

Abhijith S. Parackal ¹, Rhys E. A. Goodall ², Felix A. Faber ^{2,*} and Rickard Armiento ^{1,†}¹Department of Physics, Chemistry and Biology, [Linköping University](https://www.linkoping.se), SE-581 83 Linköping, Sweden²Department of Physics, [University of Cambridge](https://www.cam.ac.uk), Cambridge CB3 0WA, United Kingdom

(Received 15 November 2023; accepted 17 July 2024; published 2 October 2024)

The large amount of powder diffraction data for which the corresponding crystal structures have not yet been identified suggests the existence of numerous undiscovered, physically relevant crystal structure prototypes. In this paper, we present a scheme to resolve powder diffraction data into crystal structures with precise atomic coordinates by screening the space of all possible atomic arrangements, i.e., structural prototypes, including those not previously observed, using a pre-trained machine learning (ML) model. This involves (i) enumerating all possible symmetry-confined ways in which a given composition can be accommodated in a given space group, (ii) ranking the element-assigned prototype representations using energies predicted using the Wyckoff representation regression ML model [Goodall *et al.*, *Sci. Adv.* **8**, eabn4117 (2022)], (iii) assigning and perturbing atoms along the degree of freedom allowed by the Wyckoff positions to match the experimental diffraction data, and (iv) validating the thermodynamic stability of the material using density-functional theory. An advantage of the presented method is that it does not rely on a database of previously observed prototypes and is, therefore capable of finding crystal structures with entirely new symmetric arrangements of atoms. We demonstrate the workflow on unidentified x-ray diffraction spectra from the ICDD database and identify a number of stable structures, where a majority turns out to be derivable from known prototypes. However, at least two are found not to be part of our prior structural data sets.

DOI: [10.1103/PhysRevMaterials.8.103801](https://doi.org/10.1103/PhysRevMaterials.8.103801)

I. INTRODUCTION

Powder diffraction using x rays [x-ray diffraction (XRD)] is a key characterization technique to resolve crystal structures, which is used in most materials-related fields. Refinement based on diffraction data is generally highly successful in yielding precise and accurate structural information [1,2]. However, there are also many challenges associated with this process [3–5], leading to the existence of numerous synthesized materials with indexed diffraction peaks for which the precise crystal structures remain unknown [6,7]. The discovery of these unknown structures plays a crucial role in addressing materials design challenges [8].

Moreover, the structural information in most theoretical crystal structure databases [9–11] originates from experimental structures [12,13]. Discovering new crystal structures that are not yet included in the existing crystal structure databases is essential for expanding their diversity. Many machine learning (ML) models for materials rely on these theoretical databases for training data [14–17]. With a broader

selection and structurally diversified set of materials to train on, these algorithms can expedite the discovery of materials with desirable properties.

A significant volume of literature addresses the structural characterization of XRD patterns. The first steps usually include separating the contributions from phases present in the sample, locating the Bragg peaks, indexing, and space group classification [18]. Automated tools have long been in common use for indexing [19–24]. Many prior works have explored ML models for these and related tasks. LeBras *et al.* used constraint reasoning and kernel clustering for phase identification [25] and Park *et al.* developed convolutional neural networks to classify the crystal system, extinction group, and space group for a pattern [26]. More recent examples include Bragg peak positioning [27], crystallographic orientation [28], artifact removal [29], novelty detection [30], volumetric segmentation [31], phase identification and background extraction [32–55], and a wide range of structural classifications [56–73]. Hence, there is a rich set of prior works on identifying space group, lattice parameters, and compositions from experimental powder XRD patterns. For further details, see recent reviews covering this topic, e.g., Refs. [74–77].

The focus in the present work is on the next step of structural characterization: taking an indexed and classified pattern and deriving one or more matching crystal structures with specific coordinates for the atomic positions (see discussions in Refs. [8,18,78–82]). The use of ML models to directly solve the inversion problem currently appears to be limited to specific types of systems [83,84]. The more common approach is

*Contact author: ff350@cam.ac.uk†Contact author: rickard.armiento@liu.se

to start from reference structures, selected either for compositional similarity or by matching known diffraction patterns, which are then iteratively refined. The atomic positions within these structures are optimized to align the simulated and experimental XRD patterns, targeting the minimization of the R value [79,85,86], a measure of XRD pattern misalignment which is widely recognized to not produce a unique solution [5]. It is a common practice to rule out some of the matching structures based on thermodynamic instability, which can be estimated, e.g., by calculations using density-functional theory (DFT) [87,88]. Numerous techniques, such as particle swarm optimizers, Monte Carlo methods, and genetic algorithms, address the challenges of simultaneously optimizing the R value and thermodynamic stability [89–95].

The iterative optimization of the positions of N atoms in three-dimensional (3D) space to match a powder XRD pattern has a computational complexity of $O(N^3)$. A first step to overcome this combinatorial wall is to restrict the degrees of freedom by leveraging the identified symmetry. The first-principles-assisted structure solution (FPASS) technique [7,96] further leverages symmetry information by biasing the assignment to Wyckoff positions [97] based on occupancies mined from crystal structure databases. This approach allows a high-throughput workflow to analyze XRD data with minimal manual input. Griesemer *et al.* [98] developed a method to use reference structures from the prototypes of known structures within crystal structure databases such as OQMD [10]. A similar approach was used in Ref. [99]. However, approaches that rely, even just in part, on previously observed structures will have difficulties in identifying structures that belong to completely new prototypes that do not yet exist in any databases [100].

In this work, we present an algorithm to invert XRD patterns into crystal structures with specified atomic coordinates without using data from previously identified structures and prototypes. Our approach uses an efficient systematic enumeration of possible reference structures via a coarse-grained representation of crystal symmetry using Wyckoff positions. The enumeration of Wyckoff position assignments is restricted based on the composition, space group, and formula units per unit cell, which are available from indexing and classification. The possible assignments are exhaustively explored up to a space group dependent limit of unit cell complexity. The candidate reference structures are then prioritized by a predicted formation energy using the *Wyckoff representation regression* (Wren) ML model that operates on the coarse-grained representation [16]. While there are prior works that have proposed systematic enumeration of structures, e.g., over integer decision variables of coordinate values [101], our approach can demonstrably reach structures up to relevant complexity since the enumeration over Wyckoff positions aligns well both with the information available from structural analysis in crystallography and the description of crystals in the Wren ML model.

We have implemented the above scheme in a software package HTTK-SYMGGEN, which utilizes GPU-accelerated optimizations to refine a cost function related to the R value and interatomic distances on an ensemble of possible candidate matches for a given XRD pattern. We demonstrate the application of our implementation for inverting XRD to crystal

structures on synthetic and experimental patterns, including the identification of two crystal structures on prototypes that are not present in materials databases.

The rest of the paper is organized as follows. In Sec. II, Background, we define the concepts of Wyckoff prototypes and protostructures, which we consistently use for our coarse-grained description of crystal structures. We also explain how simulated and experimental XRD are matched using the R factor and give an overview of the Wren ML model on which this work is based. Section III, Workflow, describes the steps to invert an XRD pattern. Section IV, Implementation, gives the details on our implementation of the workflow in the HTTK-SYMGGEN software. Section V discusses limitations in the applicability of our algorithm and the provided implementation. Sections VI–VIII present tests on synthetic and experimental XRD. Sections IX and X present our discussion and conclusions.

II. BACKGROUND

A. Wyckoff prototypes and protostructures

Crystal structures are commonly characterized by the symmetry properties of the lattice and the positions of the atoms. This leads to the division of all possible crystal structures into 230 space groups. Each space group has a set of Wyckoff sites, labeled using letters in the Latin alphabet (a, b, c, etc.), distinguished based on how the symmetry operators act on them. The coordinates of a Wyckoff position can be fixed, or there can be one or more freely variable coordinates, allowing a degree of freedom along a line, plane, or in 3D space. We can obtain the coordinates of the atoms by specifying the Wyckoff position and numerical value for each degree of freedom for the given Wyckoff position. The *Volume of International Tables for Crystallography* lists the Wyckoff positions for each space group [102].

The concept of “a crystal structure prototype” does not have a single, strict definition in literature; it can refer to any representation with a one-to-many relationship to a range of crystal structures with precisely specified atomic coordinates. In this paper, we use the term *Wyckoff prototype* to refer to the specific kind of prototype that groups crystal structures by space group and a list of occupied Wyckoff labels, i.e., where distinct atoms of unspecified species reside. Furthermore, we will adopt the name *Wyckoff protostructure* for a Wyckoff prototype with specific elements assigned to all of the occupied Wyckoff sites, but where the precise atomic coordinates allowed according to the degrees of freedom of the occupied Wyckoff sites are not specified.

The Wyckoff prototypes used in this work are based on the formalism established by the AFLOW prototype labels and thus use nearly the same notation [103]. An AFLOW prototype label is a textual representation of a prototype defined as the following series of fields separated by an underscore character (`_`): the anonymous chemical formula, the Pearson symbol, the space group number, and the sequence of occupied Wyckoff letters. For example, Calcite structure (CaCO_3) has the AFLOW prototype label `ABC3_hr10_167_a_b_c`. In this work, we specify *Wyckoff prototype IDs* using AFLOW labels, which can be extended into *protostructure IDs* by

appending a segment consisting of a colon (:) and a list of elements separated by a dash (-) to assign specific elements to each of the Wyckoff sites. For example, the protostructure ID ABC3_hr10_167_a_b_e:Ca-C-O signifies that calcium atoms occupy the orbits of Wyckoff position a, carbon atoms the b position, and oxygen the c position.

We furthermore define that these IDs should be normalized over all possible AFLOW labels that represent the same prototype via transformations of the positions under the coset representatives of the affine normalizer of the space group [104] to use the transformed set whose ID would appear first when sorted on the sum of the alphabetical indices of the Wyckoff letters and, second, lexicographically on the order they appear in the ID. The normalization of our IDs turns them into an origin-independent representation of the respective prototype and protostructure.

B. Matching simulated and experimental XRD

Structure matching with experimental XRD data is done by identifying a suitable crystal structure and computing a simulated XRD of this reference structure. Then, the value of an R factor is optimized by perturbing the atoms in the reference structure [85,86,96]. The R factor is a unitless measure of a relative mismatch between the peaks of the simulated and experimental patterns, but in the context of refining structures, multiple definitions are in use. To allow direct comparisons with Ref. [98] we use their definition:

$$R = \frac{\sum_{\text{peaks}} (I_{\text{expt.}} - I_{\text{sim}})^2}{\sum_{\text{peaks}} I_{\text{sim}}^2}, \quad (1)$$

where I is the intensity of the peaks. Each peak in the simulated pattern is matched to the closest peaks in the experimental pattern as long as they are within 0.15° in 2θ . If multiple peaks are within this range, their intensity is combined before comparison.

The algorithm for generating simulated XRD patterns from a crystal structure is obtained from De Graef and McHenry [105]. To summarize, all points in the limiting sphere given by $2/\lambda$ are calculated from the reciprocal lattice of the crystal structure, where λ is the wavelength of the x-ray beam. Then, for every reciprocal point G_{hkl} in this limiting sphere, we calculate θ using the Bragg's condition [106]. The atomic scattering factor for a given $s = \sin(\theta)/\lambda$ and an element with atomic number Z is given by

$$f(s) = Z - cs^2 \sum_{i=1}^N a_i e^{-b_i s^2}, \quad (2)$$

where a and b are the fitting parameters for each species and c is 41.782 14. The structure factor for given hkl indices is given by

$$F_{hkl} = \sum_{i=1}^N f_i \left(\frac{\sin \theta_{hkl}}{\lambda} \right) e^{2\pi i(hx_i + ky_i + lz_i)}, \quad (3)$$

where N is the number of atoms in the unit cell and (x, y, z) denotes the atoms' relative (fractional) coordinates. The intensity is obtained by multiplying $|F_{hkl}|^2$ with the Lorentz

polarization factor

$$I_{hkl} = |F_{hkl}|^2 \frac{1 + \cos^2(2\theta_{hkl})}{\sin^2(\theta_{hkl}) \cos(\theta_{hkl})}. \quad (4)$$

We use an implementation of the algorithm based on the one in PYMATGEN [107] that uses atomic scattering parameters from a table in De Graef and McHenry [105].

The context largely determines what can be considered a good or bad R value. Comparing R values across different studies is challenging due to differences in the definitions. Previous works that discuss this topic more broadly suggest that a low R value does not necessarily guarantee a correct structure, nor does a high R value disqualify a candidate [98,108,109]. Therefore, it is recommended to visually inspect how well the patterns match [110,111]. In this work, we primarily use the R value as a guide for quick relative comparisons. Our optimization workflow employs a continuous cost function that is inspired by the R factor (see Sec. III B).

C. Wren

As outlined in the Introduction, our algorithm uses formation energies estimated using an ML model to guide the choice of reference structures to investigate further. Most available energy-predicting models, such as machine learning interatomic potentials, will not work in the present context since they require knowledge of (at least approximate) atomic coordinates, which are not known at the stage when reference structures are selected.

Hence, an essential component of our work is the Wren ML model [16]. This model takes as input precisely what we defined above as a Wyckoff protostructure and uses a message-passing neural network to predict the lowest formation energy representation of that protostructure allowed by the degrees of freedom for the occupied Wyckoff positions. An ensemble of models is trained using different random initializations, which increases accuracy and enables an internal representation of the uncertainty of the model as a combination of aleatoric and epistemic uncertainty.

The present work uses the pretrained model delivered with Ref. [16], which consists of ten ensemble members trained on the union of the datasets of Materials Project (MP) [9] and Wang, Botti, and Marques (WBM) [112]. This combined dataset contains 322 915 materials after cleanup, with 19 312 unique Wyckoff prototypes. The resulting validation error for formation energy prediction was shown to be approximately 30 meV/atom. For more details on the ML model and its performance, see Ref. [16].

III. WORKFLOW

Our workflow to invert XRD into crystal structures with coordinates takes the compositional and symmetry information (retrieved after indexing) as input. Enumerating the allowed occupations of Wyckoff positions for that space group (see the formulation provided in Sec. III A), yields a list of Wyckoff protostructures. These are subsequently ranked and short-listed based on predicted formation energies using the Wren model. Each Wyckoff protostructure can be used to construct a crystal structure by populating the degrees of freedom the

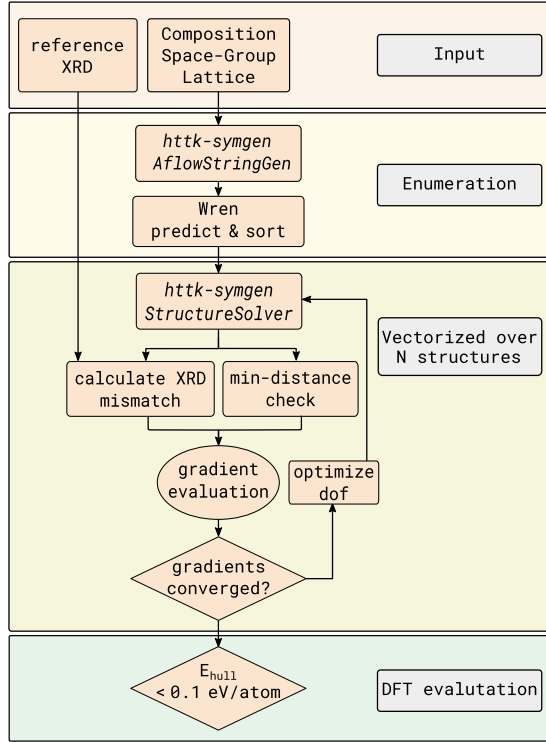


FIG. 1. Overview of the workflow. HTTK-SYMGEM contains three major components: the `AflowStringGenerator` method takes composition and space group information and enumerates valid Wyckoff protostructures. This is then ranked and sorted using the Wren model. After sorting, the `StructureSolver` method takes a Wyckoff protostructure and lattice parameters and generates crystal structures by assigning values to the available degrees of freedom. Synthetic XRD is generated using this structure and compared with the experimental XRD. A structure that matches the experimental XRD and passes the minimal distance checker will be further evaluated using DFT.

Wyckoff site allows (detailed in Sec. IV B). This representation can be passed to an optimization workflow, where the values are filled to minimize a cost function (discussed in Sec. III B) so that the resultant structure produces a simulated XRD that agrees with the experimental XRD. An overview of the implementation is provided in Fig. 1.

A. Enumeration

To enumerate all possible Wyckoff positions that a given composition χ_η can accommodate, we query the multiplicities of each Wyckoff position in a given space group. Let it be $M = \{m_1, m_2, \dots, m_n\}$ where m is the multiplicity for a given Wyckoff site and n is the number of Wyckoff positions in that space group. For a given number of atoms η , we reduce the space of M to M' by only taking the sites that match the condition $m_j \leq \eta$ for every $j \in \{1, \dots, n\}$.

This problem can be formulated as a generalized case of the subset sum problem [113], where we are to find every subset $S' \subseteq M'$ such that

$$\sum_{i=1}^{n'} km_i = \eta, \quad (5)$$

where n' is the size of M' , and k is zero or some positive number.

For a set of elements in M' , numerous solutions can satisfy Eq. (5). Each atom type in the composition allows independent solutions. When chained together, ensuring no two atoms fill a Wyckoff site with zero degrees of freedom (detailed in Sec. IV A), these solutions lead to a series of Wyckoff protostructures. The enumeration process can be computationally demanding, especially for compositions with numerous atoms, and its complexity majorly depends on the selected space group. This problem can be framed as the non-negative integer solutions of the linear diophantine equation [114].

B. Cost function

We use the R value to report how well our simulated patterns match the experimental ones. However, the optimization part of the workflow requires a more continuous and well-behaved cost function to be stable. Therefore we use an alternative cost function that measures the overlap between two XRD profiles by mapping a pseudo-Voigt profile over the experimental and simulated peaks (denoted as \mathbf{f} and \mathbf{g} , respectively) so that peaks are represented as a continuous vector and calculating

$$C_{\text{XRD}} = 1 - \frac{\mathbf{f} \cdot \mathbf{g}}{\|\mathbf{f}\|_2 \|\mathbf{g}\|_2}. \quad (6)$$

One of the advantages of this cost function is that it is bounded between (0, 1), allowing us to use a fixed threshold to assess the similarity.

We also define another objective function, which ensures that the distance between atoms is not closer than a defined threshold. We use the average of the sum of Wigner-Seitz radii [115] of two atoms as this threshold, denoted as D_{\min} . The minimum distance cost function can then be defined as

$$C_{\text{distance}} = \sum_{i=1}^n \sum_{j=i+1}^n \max(0, D_{(i,j)}^{\min} - D_{(i,j)})/2, \quad (7)$$

where $D_{i,j}$ is the distance between two atoms i and j in the crystal structure.

In our implementation, a gradient descent optimization is used to identify the minima of the combined XRD and distance cost functions. The minima are reached when the gradients with respect to the degrees of freedom converge, mirroring the convergence to a local structural minimum. This search is executed multiple times with varied initializations, and the outcomes are contrasted in order to identify the global minima. More details are provided in Sec. IV C.

IV. IMPLEMENTATION

We have implemented the workflow described in Sec. III in a subpackage HTTK-SYMGEM of the high-throughput toolkit (`httk`) software package [116]. This software package has two significant components. `AflowStringGenerator` handles the enumeration and serialization of Wyckoff positions and `StructureSolver` handles operations on a given Wyckoff protostructure.

A. Generating candidate protostructures

AflowStringGenerator takes the chemical composition and space group and returns a list of Wyckoff protostructures. This decoration is constructed by formulating the combinatorial search as a modified subset sum problem. A recursive search function is implemented, which iterates over the multiplicities of Wyckoff sites for a given space group and checks whether any combination of summed multiplicity values would be equal to the desired composition (number of atoms). More details on the algorithm for the enumeration are given in the Supplemental Material, Ref. [117].

As an example, consider the case where we have to find the number of ways to accommodate eight atoms in space group 61. Space group 61 can accommodate three Wyckoff positions, denoted by the Wyckoff letters a , b , and c . Sites a and b have four and c has eight multiplicities, respectively. There are two different ways to put eight atoms from this space group, “ ab ” [4 + 4] and “ c ” [8]. One has to be careful of the added constraint that if a Wyckoff site has no degrees of freedom, it cannot be reused again in the solution. For example, we cannot take “ aa ” because the Wyckoff site a does not have any degrees of freedom. Therefore, trying to put eight atoms in two a sites would mean that more than one atom would sit on exactly the same coordinates.

The HTTK-SYMGEM implementation includes lookup information of symmetry operations, Wyckoff letters and multiplicities, and affine transformations for the origin choices for the standard settings as listed in Volume A of the *International Tables for Crystallography* [102]. Our lookup tables are based on the data available via PYTHON library PyXtal [118], which we have modified slightly with the help of the *Bilbao Crystallographic Server* [104] to match the information in the International Tables.

The generated assigned Wyckoff prototype strings are passed through the pretrained Wren model. As explained in Sec. II C, the pretrained Wren ensemble comprises ten models trained to converge from different random initializations. A standard deviation is calculated from the ensemble predictions, which is then subtracted from the mean predicted formation energy to yield an uncertainty-adjusted value. This adjusted value sets a baseline for the formation energy of the protostructures and is used to shortlist the enumerated protostructures.

For efficient evaluation, protostructures with uncertainty-adjusted predicted formation energy exceeding 40 meV/atom above the lowest prediction are excluded. This threshold, while user adjustable, is chosen based on the mean absolute error of Wren. Wren demonstrates better accuracy the closer the prediction is to the convex hull of thermodynamical stability in a sufficiently sampled composition [16]. This is a particularly useful property in the context of XRD inversion, where candidates are predominantly located in proximity to the convex hull.

B. Parsing assigned Wyckoff protostructures

A core feature of HTTK-SYMGEM is the ability to realize a Wyckoff protostructure and create an internal representation that returns a filled cell from a list of values that fills the degrees of freedom allowed by each Wyckoff position, provided

by the method StructureSolver. As an example, consider the protostructure ID ABC3_oP40_61_c_ab_3c:As-I-0.

Referring to the previous example, space group 61 accommodates three Wyckoff positions: $4a$, $4b$, and $8c$, with the number indicating multiplicity. $4a$ and $4b$ do not have any degrees of freedom, while $8c$ is a general position with three degrees of freedom. Referring to the protostructure ID, the inference to be made here is that 8 arsenic atoms occupy Wyckoff site c , 8 iodine atoms split between sites a and b (4 atoms in site a + 4 atoms in site b), and 24 oxygen atoms are distributed across three c sites (8 + 8 + 8), summing up to 40 atoms in the unit cell. Therefore, one would have to specify (3) + (0 + 0) + (3 + 3 + 3) or 12 numbers between [0, 1] to define the positions of every atom in this structure uniquely.

Every atomic arrangement permitted by this prototype can be explored by adjusting these 12 values. The desired structure is determined by using this list as input for an optimizer and minimizing specific objective functions, which in this case are the XRD mismatch and distance cost function.

C. Structure refinement using XRD data

Our starting point for XRD inversion is an experimental XRD pattern for which the unit cell parameters, the reduced composition, scaled intensity, and 2θ values of the diffraction peaks are known. This is the information commonly available in databases of otherwise unidentified patterns.

After obtaining these unit cell parameters and symmetry information, we pass this to AflowStringGenerator. This yields a list of Wyckoff protostructure IDs, which are passed through Wren for energy prediction. The sorted list of Wyckoff protostructure labels is then truncated using the cutoff criteria, and the shortlisted candidates are parsed using the StructureSolver method, which identifies the degrees of freedom for each Wyckoff site in the candidate prototype. For each candidate prototype, n different initial structures are generated, using *Latin-hypercube* [119] spacing implemented in Scipy [120] for optimal sampling of the search space.

The cost functions described in Eqs. (6) and (7), as well as the code for generating simulated XRD, Eq. (4), is written in the numerical analysis library, JAX [121,122]. Since the problem is framed and written in an ML library that targets parallel computation, we can evaluate these n random structures in parallel on a GPU.

This capability facilitates the simultaneous identification of the global minima for cost functions C_{distance} and C_{XRD} simultaneously over a grid (n , dof), where (dof) denotes the degrees of freedom associated with the Wyckoff prototype. We employed a gradient descent methodology [123] for the optimization of atomic positions in each structure, leveraging the Adam optimizer [124] available in OPTAX [122] with a set learning rate of 0.001.

The gradient convergence, as detailed in Sec. III B, Workflow, serves as our termination criterion. Post convergence, we filter for unique atomic configurations and rank them by their R values before proceeding with DFT calculations for formation energy.

We use the electronic structure code VASP (version 5.4) [125,126] for DFT calculations, using the INCAR settings and pseudopotentials chosen to comply with Materials Project

[9] settings. Projector augmented wave pseudopotentials [127,128] were used with the Perdew-Burke-Ernzerhof generalized gradient approximation (PBE/GGA) functional [129]. The MaterialsProject2020Compatibility [130] correction scheme implemented in PYMATGEN [107] was applied to the final DFT energy. We used the High-Throughput Toolkit (*httk*) [116] to manage the calculations.

V. LIMITATIONS AND CHALLENGES

In this section, we clarify the role of this work as a component of frameworks for large-scale automated structural characterization using XRD by discussing some of the limitations and challenges of applying our work in that context. This kind of use has recently seen increased interest as part of the transition into highly automatized synthesis and characterization (see, e.g., Ref. [131]).

a. Uniqueness of solution. The algorithm produces a prioritized list of solutions that all match a provided XRD, ordered from the lowest predicted formation energy up to a cutoff value related to the accuracy of the energy predictions. In some cases, the result is a single solution, but there may be more, especially if the experimental data is of a nature that does not allow the requirements on the R -value match to be very stringent. In such cases, reducing the results into a single true solution requires further in-depth theoretical or experimental analysis, which may not be easy to automate and can require significant effort. Strictly speaking, the same kind of deeper characterization is required also to confirm that the true solution has been found even if there is a single well-fitting identification when there are limits on how exhaustively all possibilities have been screened (e.g., see discussion on complexity below). Nevertheless, the XRD pattern inversion problem has long been known not to be uniquely solvable, and our algorithm arguably handles the ambiguity as well as is possible under the circumstances by finding all solutions matching the provided information. This differs from other tools that are deliberately designed to be biased towards those solutions that share features of previously identified structures.

b. Bounds on structural complexity. Our systematic enumeration of protostructures enables an efficient coarse-grained screening of candidates. However, the complexity of the brute force exploration of Wyckoff positions in a given space group is similar to the subset sum problem, which is identified to be an NP-hard problem. The combinatorics are especially challenging for compositions with many atoms in space groups with a large number of Wyckoff positions with at least one degree of freedom (e.g., 47 and 123). For example, this class includes many molecular crystals, which end up having many atoms assigned to general Wyckoff positions in space groups representing fairly low symmetry. Hence, in practice, the enumeration has to stop at some level of complexity set by the number of occupied Wyckoff positions and the number of atoms in the unit cell. In this work we have set those limits at 16 assigned Wyckoff positions and 64 atoms per unit cell, which are chosen because structures beyond these are not well represented in the Wren training data [16], and we therefore suggest not going beyond these limits in HHTK-SYMGEM. These bounds preclude finding structures of

arbitrary complexity; however, the user can choose the limits based on available computational resources.

c. Need for initial indexing, space group identification. The method as presented relies on a prior step of accurate indexing and space group identification. A reasonably small deviation in lattice shape significantly affects the cost function defined in Eq. (6). Also, XRD patterns may have been obtained under high pressure, which affects subsequent analysis of candidates, e.g., by DFT relaxations using zero external stress.

d. Presence of multiple phases, disorder, and background contributions. Our present implementation assumes the provided powder XRD is for a single ordered phase with no significant background contribution, possibly by first taking it through some form of preprocessing. As discussed in the Introduction, much prior literature (many involving ML models) deals with the complex problem of separating contributions from multiple phases. However, this problem has been argued to remain a major challenge for automated XRD analysis [132]. There is nothing inherent to our algorithm preventing a future extension to consider possible mixtures of enumerated reference structures, and such an extension could help further address these challenges. However, such an extension would significantly increase the computational effort and has not yet been implemented.

VI. TESTING GENERALIZABILITY OF ENERGY PREDICTIONS

Predictive ML models are known to be accurate primarily for data similar to the training dataset, which often is discussed in terms of interpolative vs extrapolative use of models. In this section, we investigate the out-of-dataset predictive power of Wren in the context of this work by creating two new retrained versions.

The first model is trained only on a subset of the original dataset where we have removed an entire anonymized composition AB_2C_6 from the original dataset, which results in 314 554 materials. A technicality of this test is that it would be expensive to train the model to the same level of convergence as the original pretrained Wren model provided with [16], given that it will be used only for this test. Hence, we run the training only for 50 epochs, where we observe that when used with 20 ensembles, the overall performance is on par with Wren. We then create a second model trained on the entire dataset used for the original Wren in the same way as the first model (50 epochs, 20 ensembles).

We can now compare these two models for protostructures with the composition removed from the training of the first model, AB_2C_6 . In this domain, the first model error comes out to 88 meV/atom (see Fig. 2), whereas the second model gives an error similar to the original Wren, 37 meV/atom. The conclusion is that the restricted model can generalize well into the space of an entire composition that is not part of its training data without uncontrollably large errors. Rather, the error for out-of-dataset predictions is thus still below the reported mean absolute error (MAE) of DFT calculations on general chemistries and structures (~ 100 meV/atom [133,134]). Hence, the remaining accuracy is sufficient for the enumeration-based screening workflow of this work.

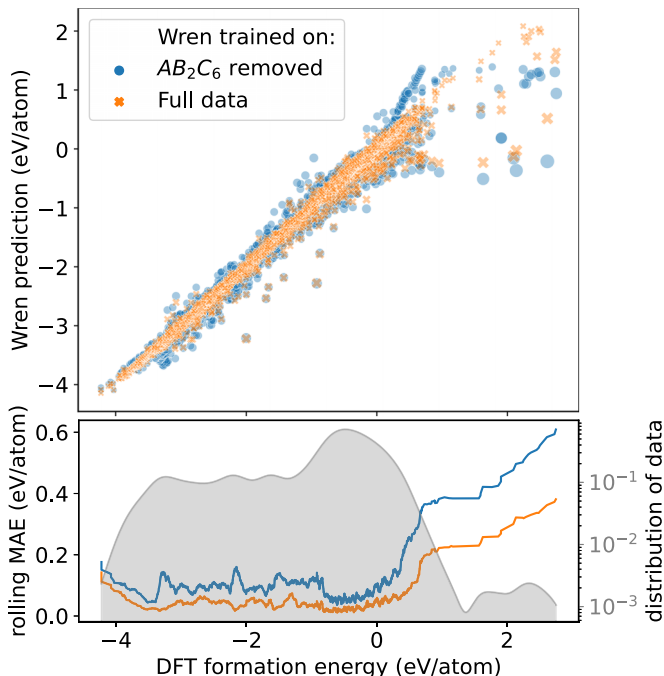


FIG. 2. Top: A scatter plot comparing the predictive accuracy of Wren models trained on two different datasets: one on the complete WBM + MP dataset and the other omitting all data points with the composition AB_2C_6 . The size of each point represents the level of predictive accuracy, with larger points indicating higher errors. Bottom: The rolling MAE of the two models as a function of the DFT formation energy and, in the background, a gray kernel density estimate plot of the data distribution. The model trained on the complete dataset has an error of 37 meV/atom, while the model trained on the dataset with the excluded composition reports an error of 88 meV/atom. Wren displays higher accuracy for protostructures with lower formation energy, and its accuracy generally goes down as the formation energy increases.

VII. TESTS OF SYNTHETIC XRD INVERSION

To evaluate the performance of the complete workflow in identifying the atomic positions belonging to a previously unseen prototype from an XRD pattern, we randomly chose a few materials from the materials project dataset. These materials were part of the AB_2C_6 space that is excluded from the training data in the retrained models discussed in the previous section. We chose structures that are evaluated to be on the convex hull of thermodynamical stability. Then, we simulated the XRD patterns for these structures. We were able to reproduce the original crystal structure from which the XRD was generated by starting only from the simulated XRD and the composition. The two examples are discussed in more detail below, with a summary presented in Table I (more data is available in the Supplemental Material, Ref. [117]).

In the first test, we start from a synthetic XRD for Li_2MnF_6 (mp-752936) and the knowledge that it is in space group 150. Using `AflowStringGenerator`, we find that this composition and space group allow 7150 different enumerated protostructures. Ranking the Wyckoff protostructure IDs using Wren and sorting them based on the aleatoric and epistemic uncertainty (explained in workflow) and using a

TABLE I. Summary of the highlighted structures. “enum” represents the total number of enumerated candidates and “cutoff” represents the number of candidates left after applying a cutoff of 40 meV above the lowest prediction. “rank” shows at what position the structure is ranked based on predicted formation energy.

mp-id	Composition	spgp	enum	Cutoff	Rank
mp-752936	Li_2MnF_6	150	7150	180	1
mp-9126	Sb_2SrO_6	162	41	9	3

cutoff of 0.04 eV/atom from the lowest predicted energy gives a shortlisted number of 180 candidates. The original protostructure ID is `A6B2C_hP27_150_3g_ef_ad:F-Li-Mn`, and it was predicted as the most likely structure (at the top of the sorted list) in the Wren ranking. Hence, the number of protostructures to take to the next step of matching the R value using `StructureSolver` is brought down by the screening with Wren from 7150 to 180, i.e., in this case, the screening of formation energies reduces the computational effort of the following step with a factor of 40.

The remaining list of protostructures, sorted by the energy predicted by Wren, is sequentially processed using `StructureSolver`. We initialize 512 random structures for a given protostructure and optimize them to match the XRD profile. For each protostructure, multiple instantiations among the 512 structures relax into the same structure. Duplicate structures are trivially excluded by comparing the values of the degrees of freedom and picking the structure corresponding to the best R value. For this particular case, the `StructureSolver` screening results in three protostructures `3g_ef_ad`, `3g_2e_ad`, and `3g_2e_bd` reproduced by the original XRD. It may be sufficient to use DFT to discern the most stable of the structures.

In the second test, we use a synthetic XRD for Sb_2SrO_6 (mp-9126) and the knowledge that it is in space group 162. Enumeration gives in this case that the composition and space group only allow 42 protostructures. Wren is used to predict formation energies, and the cutoff criterion is applied to yield nine shortlisted Wyckoff protostructures. The original structure has a protostructure ID `A6B2C_hP9_162_k_c_b:0-Sb-Sr` with two degrees of freedom appearing as the third-ranked option. In this case, solving the structure requires optimizing only two degrees of freedom, which exactly reproduces the synthetic XRD. The next best structure has an R value greater than 0.15.

Our tests have been run on a NVIDIA A40 GPU, and the Wren formation energy predictions are found to generally, on average, take approximately 0.002 s per protostructure. The processing with `StructureSolver` takes around 30 s per protostructure. The computational load can be scaled up more or less perfectly in parallel onto multiple GPUs. This is helpful for potentially solving a large number of XRD peaks that are indexed to sufficient accuracy. Analysis of an additional 25 structures is provided in the Supplemental Material, Ref. [117].

VIII. TESTS OF EXPERIMENTAL XRD INVERSION

In this section, we use the proposed framework to identify crystal structures from experimental XRD. We first validate

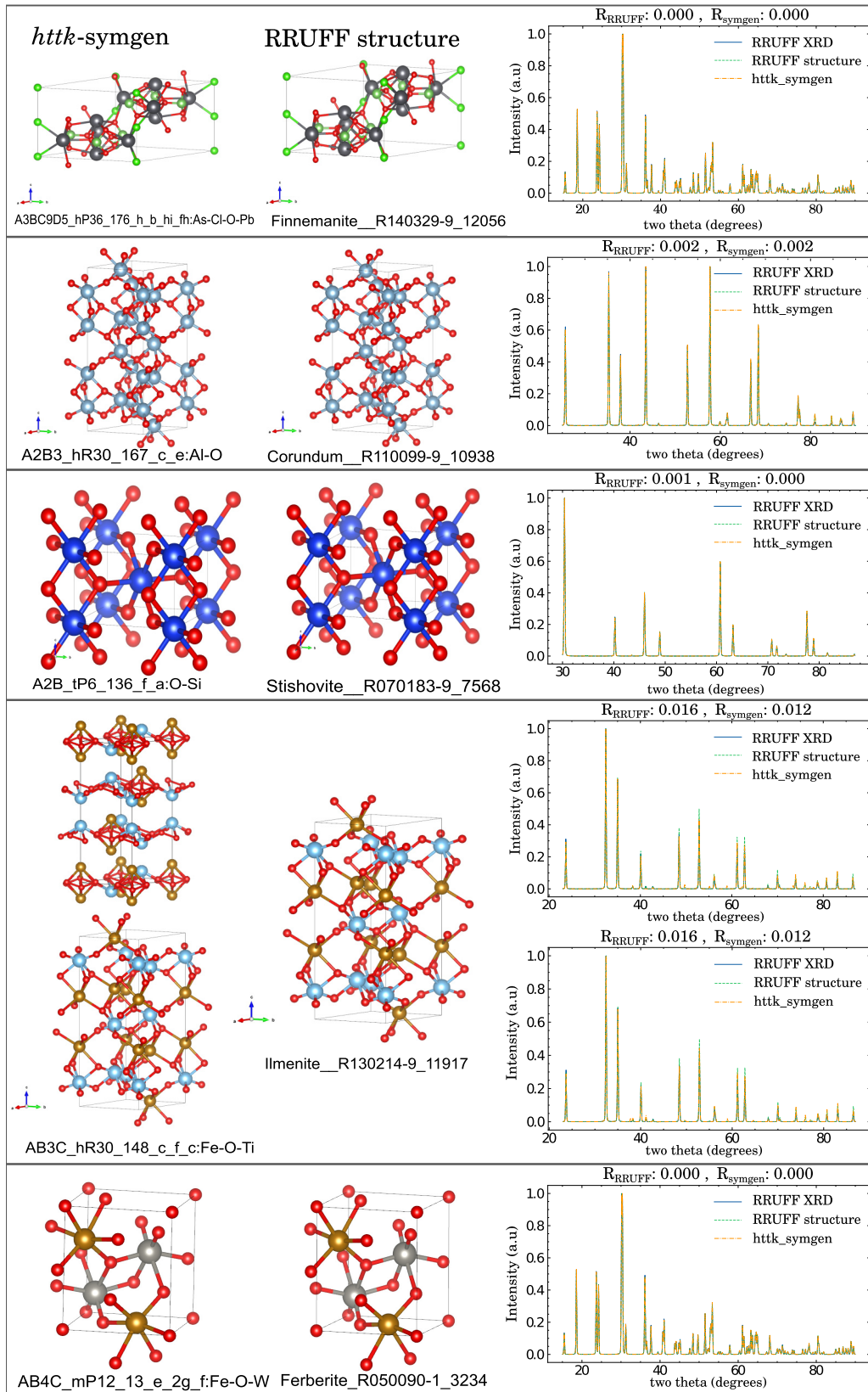


FIG. 3. A few examples of structures identified from experimental XRD patterns in the RRUFF dataset. Each row corresponds to an entry in the RRUFF dataset, showing (first column) all the candidate structures identified by our algorithm; (second column) the crystal structure in the RRUFF dataset, regarded here as the true one; (third column) comparison of the synthetic XRD from our solutions and the experimental XRD in RRUFF. For the fourth row, our algorithm found two solutions with the same good R value of 0.012. For the fifth row, we found five solutions and only show the one with the lowest R value here (the others are included in the Supplemental Material [117]).

TABLE II. Summary of the crystal structures identified from experimental XRD data. The “DFT hull dist” in the table is the positive or negative distance to the known convex hull of thermodynamic stability from Materials Project; i.e., when negative, it indicates a structure that is stable with respect to the competing phases, and thus that the known hull needs to be updated with the listed structure. The “known” column means that the prototype is present in at least one of the databases used to train Wren, i.e., MP or WBM.

ICDD-ID	Composition	Protostructure ID	Known	DFT hull dist. (meV/atom)	<i>R</i> value
00-044-1073	LiSbFe ₂ O ₆	A2BC6D_oP20_34_2b_a_3c_a:Fe-Li-O-Sb	No	9	0.090
00-024-1293	TiFeCl ₃	A3BC_hP10_186_c_a_b:Cl-Fe-Tl	Yes	-5	0.065
00-060-0238	SnTe ₂ Ni ₃	A3BC2_hP12_194_ae_c_f:Ni-Sn-Te	No	7	0.059
00-040-1408	In ₂ Se ₃	A2B3_hP10_194_f_af:In-Se	Yes	55	0.230

the framework by reproducing already identified structures in the RRUFF project database [135] and then apply the same methods on previously unidentified powder diffraction data in the International Centre for Diffraction Data (ICDD) PDF-2 2012 database [136,137]. The closed nature of the latter imposes technical limitations on large-scale access to the atomic coordinates of identified structures. Hence, we were not able to devise a test of the ability of our framework to recover already identified structures in this database. This highlights the important contribution of open datasets, such as the one provided by RRUFF.

A. Test of recovery of structures in RRUFF

We start from the RRUFF project database retrieved on Jan 24, 2024, with 2902 XRD patterns. The dataset is filtered to remove disordered structures (1992), structures that are difficult to automatically process (e.g., internal inconsistencies, unusual symmetry settings, or where the RRUFF data indicates low confidence; 447), and structures of complexity beyond what our present implementation can handle (163) (see Sec. V). This leaves 300 entries. We used DFT calculations using the PBE functional [129] to relax all these structures and kept only the ones we could confirm to have reasonable stability at zero temperature (i.e., less than 100 meV/atom above the Materials Project [9] convex hull of thermodynamical stability). The end result is a data set of 189 structures suitable for testing our algorithm. (More details on the filtering are available in the Supplemental Material, Ref. [117]).

Our implementation enumerates and identifies crystal structure candidates for the remaining 189 XRD patterns. We count the inversion as a success if the known structure in RRUFF appears anywhere in our final remaining candidates after applying the energy screening cutoff of 40 meV/atom above the energy of the structure with the lowest predicted energy and has an *R*-value match better than a chosen cutoff of less than 0.1. The result is that 186 out of the 189 structures are recovered. Figure 3 shows some examples of these recovered structures. However, we stress that this test only demonstrates the ability of our algorithm to reproduce structures from experimental XRD that are similar to those available in the training data set (which is very likely to hold for all the RRUFF structures). Also, the workflow yields multiple candidate structures for a given XRD pattern (e.g., Ilmenite_R130214-9_11917 in Fig. 3), which required additional DFT evaluations. More details on this test are available in the Supplemental Material, Ref. [117].

B. Recovery of unidentified ICDD structures

In this section, we investigate previously unidentified experimental XRD data from the (ICDD) database. The edition we use omits the precise atomic coordinates for all entries and does not clearly specify which structures the coordinates are known for. Hence, our selection of relevant XRD data to investigate has used the patterns marked as unidentified in the work presented by Griesemer *et al.* [98]. We detail four cases here (summarized in Table II), and results for additional ICDD XRD patterns are provided in the Supplemental Material, Ref. [117].

LiSbFe₂O₆ (ICDD ID: 00-044-1073). the ICDD database provides the diffraction pattern, unit cell parameters, and space group *Pnn2*, along with the reduced formula unit. The space group 34 has only three Wyckoff positions. Upon running the workflow, we identified that there were 128 unique Wyckoff protostructures that the given composition can accommodate. We identified that there were two formula units per unit cell, and the protostructure with ID A2BC6D_oP20_34_2b_a_3c_a:Fe-Li-O-Sb was determined to be the solution. This poststructure has 13 degrees of freedom, and our workflow yielded a crystal structure with an *R* value of 0.09 and an energy of 9 meV/atom above the convex hull of thermodynamic stability from the Materials Project. The entire enumeration workflow was completed in 23 min.

A unique Wyckoff protostructure ID does not absolutely ensure that there is no other way to reach the same structure through the already known prototypes, for example, via sub- and supergroup relationships and by realizing the crystal structure by placing atoms within the degrees of freedom just slightly on or off symmetry in a different prototype. Hence, we calculated formation energies with DFT of all the relevant substitutions into alternative candidate Wyckoff prototypes from both the materials project and WBM to ensure that no other structure reproduced a similar or lower formation energy. We conclude that the found prototype is a new prototype that is not present in those databases. However, we find a very similar prototype ABC2D6_oP20_34_a_b_ab_3c (Materials Project entry: mp-8673), with a comparable *R* value of 0.102, which from DFT calculations comes out as slightly less stable at 36 meV/atom above the convex hull. A deeper look into the atomic placements reveals a layered nature of these structures, and it appears the two prototypes represent different stackings between Sb and Fe atoms. (These differences are elaborated on in the Supplemental Material, Ref. [117], including a radial distribution plot.)

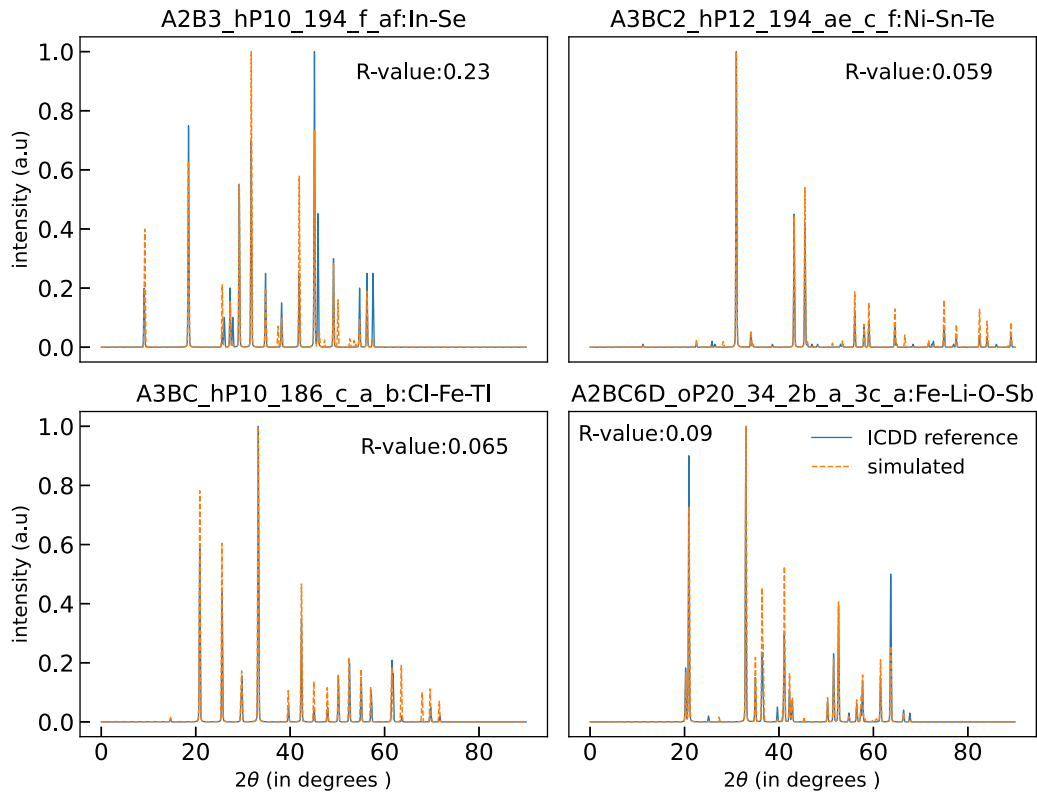


FIG. 4. Experimental XRD resolved using HTTK-SYMGEM. Four structures are shown to be solved using our workflow. The blue line shows the experimentally obtained peaks fitted over a pseudo-Voigt function. Yellow-dashed lines show the XRD profile generated using structures solved by HTTK-SYMGEM. R value is calculated only in the region where ICDD data is available; regions outside are removed from the plot.

SnTe_2Ni_3 (ICDD ID: 00-060-0238). The ICDD entry specifies space group 194 ($P63/mmc$). Utilizing the `AflowStringGenerator` method, we identified a total of 74 unique Wyckoff prototypes. The protostructure with ID `A3BC2_hP12_194_ae_c_f:Ni-Sn-Te` was determined to be the solution. The calculated crystal structure based on this protostructure is 7 meV/atom above the convex hull of thermodynamical stability from the Materials Project, with an R value of 0.059. None of the prototypes present in these datasets met either the R value or the formation energy criteria. The entire computational workflow was completed in a span of 17 min.

There are no entries with the same protostructure ID `AB2C3_hP12_194_c_f_ae` in the Materials Project or the WBM dataset (but we did not do the same exhaustive analysis via DFT calculations of all possible alternatives as for $\text{LiSbFe}_2\text{O}_6$).

In_2Se_3 (ICDD ID: 00-040-1408). The pattern files were provided and were identified to be in space group $P63/mmc$. Using our algorithm, we identify that there are two formula units per unit cell, and the protostructure with ID `A2B3_hP10_194_f_af:In-Se` was determined to be the solution, with two degrees of freedom along the z direction at the Wyckoff position f , which was occupied by In (f) and Se (a,f). There were only 16 different combinations of Wyckoff positions that satisfied the composition. The R value is 0.23, with an energy above the hull of 55 meV/atom. The corresponding Wyckoff prototype (`A2B3_hP10_194_f_af`) is previously reported, as seen in mp-1094034 and mp-1246153.

TiFeCl_3 (ICDD ID: 00-024-1293). The structure was reported to be in space group 186 ($P63mc$). Upon enumeration, there were a total of 1874 possible Wyckoff protostructures, from which we identified `A3BC_hP10_186_c_a_b:Cl-Fe-Tl` as the correct protostructure label for the structure with an R value of 0.065, and an energy of 5 meV/atom below the convex hull. The prototype corresponding to this structure turns out to already exist in the Materials Project (mp-1096852, mp-19241, etc.) and OQMD. Hence, it was also identified by Griesemer *et al.* [98] with a formation energy insignificantly higher (3 meV/atom) than our structure. (There is a slight complication here in that while they also report their match to be in space group $P63mc$, their final atomic coordinates differ slightly from ours in a way that makes us identify their structure to be in a supergroup, space group 194 ($P63/mmc$). Hence, it is possible that the structure actually is in this supergroup rather than the space group reported in ICDD.) A comparison of simulated and experimental patterns for these identifications is shown in Fig. 4.

IX. DISCUSSION

The ability in tests on synthetic patterns to identify prototypes in the structural space omitted in training, along with the identification from previously unidentified experimental XRD patterns from the ICDD database of two crystal structures with Wyckoff prototypes not present in existing databases, substantiates the capability of our method to identify new, previously

unseen, prototypes. The success in applying the method to experimental patterns also demonstrates robustness against noise and other imperfections absent in simulated patterns. Therefore, our approach shows promise not only in theoretical simulations but also in its applicability to real-world experimental data.

Our test on XRD patterns from the ICDD database have been successful in finding new structures with better R -value matches, and higher stability based on DFT formation energy calculations, compared to solutions using known structures and prototypes. Since the DFT calculations place them below, or just slightly above, the convex hull of thermodynamic stability, well within the theoretical accuracy of DFT with semilocal exchange-correlation functionals of approximately 100 meV/atom [133,134] they represent new relevant crystal structures that belong in the materials databases. However, given the nonuniqueness of XRD inversion, further in-depth experimental and theoretical characterization would formally be required to completely confirm these as the true solutions for the respective ICDD entries and dismiss the possibility of even better matches (cf. the discussion on limitations in Sec. V).

The identified prototypes could arguably have been found by a type of extended screening suggested in some prior works. For example, the FPASS method by Ward *et al.* [96] describes how to stochastically pivot into a different combination of Wyckoff sites within the same space group during its evolutionary exploration. However, the difference with the workflow presented in this paper is that it describes how to *systematically* explore the entire space of these previously unseen prototypes.

In Sec. V, we have extensively discussed various limitations and challenges in applying our algorithm, including its combinatorial complexity. Nevertheless, our tests demonstrate that our implementation in practice has a computational efficiency sufficient to explore relevant levels of structural complexity.

X. CONCLUSIONS

In this work, we have presented a scheme that can resolve experimental XRD patterns without relying on a database of previously resolved structural prototypes. The use of a coarse-grained descriptor in the Wren ML model allows exploring candidate structures at a low computational effort,

which enables prioritizing among large sets of candidates, including in the unexplored parts of the structural space.

We have demonstrated the ability of our model to invert both simulated and experimental XRD patterns. Our tests show how crystal structures with new Wyckoff prototypes can be obtained, i.e., that symmetric arrangements of atoms that have never been observed before can be identified. We are unaware of other equally automated and systematic techniques with this capability. Our work represents a significant advancement in the field, offering a highly automated, efficient, and versatile tool for the identification of new crystal structures.

The source code is being prepared for public release and is available on request. It will be available as a standalone package and bundled with a future release of HTTK.

ACKNOWLEDGMENTS

The authors acknowledge insightful discussions with Alpha A. Lee in the early stages of this project, as it spawned out of the development of the Wren machine learning model. A.S.P. and R.A. acknowledge useful discussions related to the work with Florian Trybel. We acknowledge support from the Swedish Research Council (VR) Grant No. 2020-05402 and the Swedish e-Science Centre (SeRC). The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at NSC and the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at C3SE, partially funded by the Swedish Research Council through Grant No. 2018-05973. Development of the project and calculations were conducted using resource allocations in Berzelius (Berzelius-2022-192) and Alvis (NAISS 2023/22-559).

A.S.P. contributed to the methodology, software, investigation, visualization, and writing the first draft. R.E.A.G. contributed to the conceptualization and methodology. F.A.F. contributed to the conceptualization, supervision, and writing review and editing. R.A. contributed to the conceptualization, supervision, and writing review and editing.

F.A.F. is employed by AstraZeneca at the time of publication; however, none of the work presented in this paper was conducted at or influenced by this affiliation.

-
- [1] W. Massa and R. O. Gould, *Crystal Structure Determination* (Springer, New York, 2004).
 - [2] B. D. Cullity, *Elements of X-ray Diffraction* (Addison-Wesley Publishing, Reading, MA, 1956).
 - [3] R. Černý, *Crystals* 7, 142 (2017).
 - [4] A. L. Patterson, *Phys. Rev.* **65**, 195 (1944).
 - [5] R. W. Harrison, *J. Opt. Soc. Am. A* **10**, 1046 (1993).
 - [6] A. Gindhart, T. Blanton, J. Blanton, and S. Gates-Rector, *Microsc. Microanal.* **24**, 1154 (2018).
 - [7] B. Meredig and C. Wolverton, *Nat. Mater.* **12**, 123 (2013).
 - [8] K. D. Harris, M. Tremayne, and B. M. Kariuki, *Angew. Chem., Int. Ed.* **40**, 1626 (2001).
 - [9] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, *APL Mater.* **1**, 011002 (2013).
 - [10] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, *JOM* **65**, 1501 (2013).
 - [11] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, *Comput. Mater. Sci.* **58**, 227 (2012).
 - [12] M. Hellenbrandt, *Crystallogr. Rev.* **10**, 17 (2004).
 - [13] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P.

- Moeck, and A. Le Bail, *J. Appl. Crystallogr.* **42**, 726 (2009).
- [14] T. Xie and J. C. Grossman, *Phys. Rev. Lett.* **120**, 145301 (2018).
- [15] C. Chen and S. P. Ong, *Nat. Comput. Sci.* **2**, 718 (2022).
- [16] R. E. A. Goodall, A. S. Parackal, F. A. Faber, R. Armiento, and A. A. Lee, *Sci. Adv.* **8**, eabn4117 (2022).
- [17] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, *Nat. Machine Intelligence* **5**, 1031 (2023).
- [18] K. D. M. Harris and M. Tremayne, *Chem. Mater.* **8**, 2554 (1996).
- [19] J. W. Visser, *J. Appl. Crystallogr.* **2**, 89 (1969).
- [20] R. Shirley, National Bureau of Standards Special Publication **567**, 361 (1980).
- [21] P.-E. Werner, L. Eriksson, and M. Westdahl, *J. Appl. Crystallogr.* **18**, 367 (1985).
- [22] C. J. Gilmore, G. Barr, and J. Paisley, *J. Appl. Crystallogr.* **37**, 231 (2004).
- [23] A. Altomare, C. Cuocci, A. Moliterni, and R. Rizzi, Indexing a powder diffraction pattern, in *International Tables for Crystallography* (International Union of Crystallography, 2007), pp. 270–281.
- [24] S. J. L. Billinge and T. Proffen, *Acta Crystallogr., Sect. A: Found. Adv.* **80**, 139 (2024).
- [25] R. LeBras, T. Damoulas, J. M. Gregoire, A. Sabharwal, C. P. Gomes, and R. B. van Dover, in *Principles and Practice of Constraint Programming—CP 2011*, edited by J. Lee (Springer, Berlin, Heidelberg, 2011), pp. 508–522.
- [26] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, and K.-S. Sohn, *IUCrJ* **4**, 486 (2017).
- [27] Z. Liu, H. Sharma, J.-S. Park, P. Kenesei, A. Miceli, J. Almer, R. Kettimuthu, and I. Foster, *IUCrJ* **9**, 104 (2022).
- [28] D. M. de Oca Zapiain, T. Ao, B. Donohoe, C. Martinez, D. Morgan, M. A. Rodriguez, M. D. Knudson, and J. M. D. Lane, *AIP Conf. Proc.* **2844**, 280004 (2023).
- [29] H. Yanxon, J. Weng, H. Parraga, W. Xu, U. Ruett, and N. Schwarz, *J. Synchrotron Radiat.* **30**, 137 (2023).
- [30] L. Banko, P. M. Maffettone, D. Naujoks, D. Olds, and A. Ludwig, *npj Comput. Mater.* **7**, 104 (2021).
- [31] B. Sullivan, R. Archibald, V. Vandavasi, P. Langan, L. Coates, and V. Lynch, in *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)* (IEEE, Piscataway, NJ, 2019).
- [32] J. K. Bunn, S. Han, Y. Zhang, Y. Tong, J. Hu, and J. R. Hattrick-Simpers, *J. Mater. Res.* **30**, 879 (2015).
- [33] A. G. Kusne, D. Keller, A. Anderson, A. Zaban, and I. Takeuchi, *Nanotechnology* **26**, 444002 (2015).
- [34] S. Ermon, R. L. Bras, S. Suram, J. Gregoire, C. Gomes, B. Selman, and R. v. Dover, in *Proceedings of the AAAI Conference on Artificial Intelligence* (PKP, 2015), Vol. 29.
- [35] J. K. Bunn, J. Hu, and J. R. Hattrick-Simpers, *JOM* **68**, 2116 (2016).
- [36] Y. Iwasaki, A. G. Kusne, and I. Takeuchi, *npj Comput. Mater.* **3**, 4 (2017).
- [37] Y. Xue, J. Bai, R. L. Bras, R. Bernstein, J. Bjorck, L. Longpre, S. Suram, R. v. Dover, J. Gregoire, and C. Gomes, in *Proceedings of the AAAI Conference on Artificial Intelligence* (PKP, 2017), Vol. 31, pp. 4635–4642.
- [38] Z. Xiong, Y. He, J. R. Hattrick-Simpers, and J. Hu, *ACS Comb. Sci.* **19**, 137 (2017).
- [39] S. K. Suram, Y. Xue, J. Bai, R. Le Bras, B. Rappazzo, R. Bernstein, J. Bjorck, L. Zhou, R. B. van Dover, C. P. Gomes, and J. M. Gregoire, *ACS Comb. Sci.* **19**, 37 (2017).
- [40] V. Stanev, V. V. Vesselinov, A. G. Kusne, G. Antoszewski, I. Takeuchi, and B. S. Alexandrov, *npj Comput. Mater.* **4**, 43 (2018).
- [41] H. Wang, Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin, and J. Lin, *J. Chem. Inf. Model.* **60**, 2004 (2020).
- [42] J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh, and K.-S. Sohn, *Nat. Commun.* **11**, 86 (2020).
- [43] J.-W. Lee, W. B. Park, M. Kim, S. P. Singh, M. Pyo, and K.-S. Sohn, *Inorg. Chem. Front.* **8**, 2492 (2021).
- [44] H. Dong, K. T. Butler, D. Matras, S. W. T. Price, Y. Odarchenko, R. Khatry, A. Thompson, V. Middelkoop, S. D. M. Jacques, A. M. Beale, and A. Vamvakeros, *npj Comput. Mater.* **7**, 74 (2021).
- [45] P. M. Maffettone, L. Banko, P. Cui, Y. Lysogorskiy, M. A. Little, D. Olds, A. Ludwig, and A. I. Cooper, *Nat. Comput. Sci.* **1**, 290 (2021).
- [46] N. J. Szymanski, C. J. Bartel, Y. Zeng, Q. Tu, and G. Ceder, *Chem. Mater.* **33**, 4204 (2021).
- [47] D. Chen, Y. Bai, S. Ament, W. Zhao, D. Guevarra, L. Zhou, B. Selman, R. B. van Dover, J. M. Gregoire, and C. P. Gomes, *Nat. Mach. Intell.* **3**, 812 (2021).
- [48] Y. Suzuki, *Synchrotron Radiat. News* **35**, 9 (2022).
- [49] M. Zhdanov and A. Zhdanov, Machine learning-assisted close-set x-ray diffraction phase identification of transition metals, in *Workshop on Machine Learning for Materials, ICLR, 2023* (ICLR, 2023).
- [50] N. Q. Le, M. Pekala, A. New, E. B. Gienger, C. Chung, T. J. Montalbano, E. A. Pogue, J. Domenico, and C. D. Stiles, *J. Phys. Chem. C* **127**, 21758 (2023).
- [51] H. Fang, E. Hovad, Y. Zhang, and D. Juul Jensen, *Mater. Charact.* **201**, 112983 (2023).
- [52] N. J. Szymanski, C. J. Bartel, Y. Zeng, M. Diallo, H. Kim, and G. Ceder, *npj Comput. Mater.* **9**, 31 (2023).
- [53] R. Drapeau, Improving XRD analysis with machine learning, Master’s thesis, Brigham Young University, 2023, <https://scholarsarchive.byu.edu/etd/10078>.
- [54] W. Yue, P. K. Tripathi, G. Ponon, Z. Ualikhankyzy, D. W. Brown, B. Clausen, M. Strantza, D. C. Pagan, M. A. Willard, F. Ernst, E. Ayday, V. Chaudhary, and R. H. French, *Integr. Mater. Manuf. Innovation* **13**, 36 (2024).
- [55] J. Oppliger, M. M. Denner, J. Küspert, R. Frison, Q. Wang, A. Morawietz, O. Ivashko, A.-C. Dippel, M. v. Zimmermann, I. Biało, L. Martinelli, B. Fauqué, J. Choi, M. Garcia-Fernandez, K.-J. Zhou, N. B. Christensen, T. Kurosawa, N. Momono, M. Oda, and F. D. Natterer, *Nat. Mach. Intell.* **6**, 180 (2024).
- [56] J. Schuetzke, A. Benedix, R. Mikut, and M. Reischl, *IUCrJ* **8**, 408 (2021).
- [57] B. Zhao, J. A. Greenberg, and S. Wolter, in *Anomaly Detection and Imaging with X-Rays (ADIX) III*, edited by A. Ashok, M. A. Neifeld, M. E. Gehm, and J. A. Greenberg (SPIE, 2018), pp. 20–25, Vol. 10632.
- [58] L. C. O. Tiong, J. Kim, S. S. Han, and D. Kim, *npj Comput. Mater.* **6**, 196 (2020).
- [59] J. E. Salgado, S. Lerman, Z. Du, C. Xu, and N. Abdolrahim, *npj Comput. Mater.* **9**, 214 (2023).

- [60] A. Chakraborty and R. Sharma, *Vis. Comput.* **38**, 1275 (2022).
- [61] L. Chen, B. Wang, W. Zhang, S. Zheng, Z. Chen, M. Zhang, C. Dong, F. Pan, and S. Li, *J. Am. Chem. Soc.* **146**, 8098 (2024).
- [62] B. D. Lee, J.-W. Lee, W. B. Park, J. Park, M.-Y. Cho, S. Pal Singh, M. Pyo, and K.-S. Sohn, *Adv. Intell. Syst.* **4**, 2200042 (2022).
- [63] A. Hamza, U. Hayat, W. Hussain, and A. Mumtaz, in *2023 3rd International Conference on Artificial Intelligence (ICAI)* (IEEE, Piscataway, NJ, 2023), pp. 64–69.
- [64] N. Corriero, R. Rizzi, G. Settembre, N. D. Buono, and D. Diacono, *J. Appl. Crystallogr.* **56**, 409 (2023).
- [65] A. N. Zaloga, V. V. Stanovov, O. E. Bezrukova, P. S. Dubinin, and I. S. Yakimov, *Mater. Today Commun.* **25**, 101662 (2020).
- [66] Y. Suzuki, H. Hino, T. Hawaii, K. Saito, M. Kotsugi, and K. Ono, *Sci. Rep.* **10**, 21790 (2020).
- [67] Y. Li, R. Dong, W. Yang, and J. Hu, *Comput. Mater. Sci.* **198**, 110686 (2021).
- [68] P. M. Vecsei, K. Choo, J. Chang, and T. Neupert, *Phys. Rev. B* **99**, 245120 (2019).
- [69] C.-H. Liu, Y. Tao, D. Hsu, Q. Du, and S. J. L. Billinge, *Acta Crystallogr., Sect. A: Found. Adv.* **75**, 633 (2019).
- [70] J. A. Aguiar, M. L. Gong, and T. Tasdizen, *Comput. Mater. Sci.* **173**, 109409 (2020).
- [71] Y. Suzuki, H. Hino, Y. Takeichi, T. Hawaii, M. Kotsugi, and K. Ono, *Microsc. Microanal.* **24**, 142 (2018).
- [72] B. D. Lee, J.-W. Lee, J. Ahn, S. Kim, W. B. Park, and K.-S. Sohn, *Adv. Intell. Syst.* **5**, 2300140 (2023).
- [73] J. A. Aguiar, M. L. Gong, R. R. Unocic, T. Tasdizen, and B. D. Miller, *Sci. Adv.* **5**, eaaw1949 (2019).
- [74] V.-A. Surdu and R. György, *Appl. Sci.* **13**, 9992 (2023).
- [75] N. J. Szymanski, Y. Zeng, H. Huo, C. J. Bartel, H. Kim, and G. Ceder, *Mater. Horiz.* **8**, 2169 (2021).
- [76] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. L. Billinge, E. Holm, S. P. Ong, and C. Wolverton, *npj Comput. Mater.* **8**, 59 (2022).
- [77] A. Hinderhofer, A. Greco, V. Starostin, V. Munteanu, L. Pithan, A. Gerlach, and F. Schreiber, *J. Appl. Crystallogr.* **56**, 3 (2023).
- [78] P. G. Jones, *Chem. Soc. Rev.* **13**, 157 (1984).
- [79] A. T. Brünger, *Nature (London)* **355**, 472 (1992).
- [80] J. P. Glusker, M. Lewis, and M. Rossi, *Crystal Structure Analysis for Chemists and Biologists* (Wiley, New York, 1996), Vol. 16.
- [81] H. D. Flack and G. Bernardinelli, *Chirality* **20**, 681 (2008).
- [82] A. Altomare, C. Giacovazzo, A. Guagliardi, A. G. G. Moliterni, and R. Rizzi, *J. Appl. Crystallogr.* **32**, 963 (1999).
- [83] F. Massuyeau, T. Broux, F. Coulet, A. Demessence, A. Mesbah, and R. Gautier, *Adv. Mater.* **34**, 2203879 (2022).
- [84] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. Gilad Kusne, and T. Buonassisi, *npj Comput. Mater.* **5**, 60 (2019).
- [85] M. F. C. Ladd, R. A. Palmer, and R. A. Palmer, *Structure Determination by X-ray Crystallography* (Springer, New York, 1977), Vol. 233.
- [86] C. Giacovazzo, C. Giacovazzo, H. L. Monaco, G. Artioli, D. Viterbo, M. Milanesio, G. Gilli, P. Gilli, G. Zanotti, G. Ferraris, and M. Catti, *Fundamentals of Crystallography* (Oxford University Press, New York, 2011).
- [87] P. Hohenberg and W. Kohn, *Phys. Rev.* **136**, B864 (1964).
- [88] W. Kohn and L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).
- [89] D. C. Lonie and E. Zurek, *Comput. Phys. Commun.* **182**, 372 (2011).
- [90] L. Ward, K. Michel, and C. Wolverton, *Acta Crystallogr., Sect. A: Found. Adv.* **71**, 542 (2015).
- [91] H. Putz, J. C. Schön, and M. Jansen, *J. Appl. Crystallogr.* **32**, 864 (1999).
- [92] T. S. Bush, C. R. A. Catlow, and P. D. Battle, *J. Mater. Chem.* **5**, 1269 (1995).
- [93] A. Altomare, R. Caliendo, M. Camalli, C. Cuocci, C. Giacovazzo, A. G. G. Moliterni, and R. Rizzi, *J. Appl. Crystallogr.* **37**, 1025 (2004).
- [94] G. M. Sheldrick, *Acta Crystallogr., Sect. A: Found. Adv.* **71**, 3 (2015).
- [95] R. L. McGreevy, *J. Phys.: Condens. Matter* **13**, R877 (2001).
- [96] L. Ward, K. Michel, and C. Wolverton, *Phys. Rev. Mater.* **1**, 063802 (2017).
- [97] U. Müller, in *International Tables for Crystallography Volume A1: Symmetry Relations between Space Groups*, edited by H. Wondratschek and U. Müller (Springer Netherlands, Dordrecht, 2004), pp. 24–26.
- [98] S. D. Griesemer, L. Ward, and C. Wolverton, *Phys. Rev. Mater.* **5**, 105003 (2021).
- [99] L. Yang, P. Juhás, M. W. Terban, M. G. Tucker, and S. J. L. Billinge, *Acta Crystallogr., Sect. A: Found. Adv.* **76**, 395 (2020).
- [100] J. Schmidt, N. Hoffmann, H.-C. Wang, P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira, S. Botti, and M. A. L. Marques, *arXiv:2210.00579*.
- [101] V. V. Gusev, D. Adamson, A. Deligkas, D. Antypov, C. M. Collins, P. Krysta, I. Potapov, G. R. Darling, M. S. Dyer, P. Spirakis, and M. J. Rosseinsky, *Nature (London)* **619**, 68 (2023).
- [102] R. Shirley, in *Accuracy in Powder Diffraction*, NBS Special Publication No. 567, edited by S. Block and C. R. Hubbard (National Bureau of Standards, Gaithersburg, Maryland, 1980), pp. 361–382.
- [103] M. J. Mehl, D. Hicks, C. Toher, O. Levy, R. M. Hanson, G. Hart, and S. Curtarolo, *Comput. Mater. Sci.* **136**, S1 (2017).
- [104] M. I. Aroyo, J. M. Perez-Mato, C. Capillas, E. Kroumova, S. Ivantchev, G. Madariaga, A. Kirov, and H. Wondratschek, *Z. Für Krist. - Cryst. Mater.* **221**, 15 (2006).
- [105] M. De Graef and M. E. McHenry, *Structure of Materials, an Introduction to Crystallography, Diffraction and Symmetry* (Cambridge University Press, Cambridge, UK, 2007).
- [106] L. R. B. Elton and D. F. Jackson, *Am. J. Phys.* **34**, 1036 (1966).
- [107] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, *Comput. Mater. Sci.* **68**, 314 (2013).
- [108] R. L. Harlow, *J. Res. Natl. Inst. Stand. Technol.* **101**, 327 (1996).
- [109] A. L. Spek, *Inorg. Chim. Acta* **470**, 232 (2018).
- [110] D. K. Smith, G. G. Johnson, A. Scheible, A. M. Wims, J. L. Johnson, and G. Ullmann, *Powder Diffr.* **2**, 73 (1987).
- [111] B. H. Toby, *Powder Diffr.* **21**, 67 (2006).
- [112] H. C. Wang, S. Botti, and M. A. Marques, *npj Comput. Mater.* **7**, 12 (2021).
- [113] F. Gurski, D. Komander, and C. Rehs, *Math. Methods Oper. Res.* **92**, 401 (2020).

- [114] I. Borosh and L. B. Treybig, *Proc. Am. Math. Soc.* **55**, 299 (1976).
- [115] P. D. P. G. Reinhard and P. D. E. Suraud, Gross properties of atoms and solids, in *Introduction to Cluster Dynamics* (John Wiley and Sons, Ltd, 2003), Chap. Appendix B, pp. 259–266.
- [116] R. Armiesto, in *Machine Learning Meets Quantum Physics*, Lecture Notes in Physics (Springer International Publishing, Cham, 2020), pp. 377–395.
- [117] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevMaterials.8.103801> for additional test cases, more information about the tests, and further details on the enumeration algorithm.
- [118] S. Fredericks, K. Parrish, D. Sayre, and Q. Zhu, *Comput. Phys. Commun.* **261**, 107810 (2021).
- [119] W.-L. Loh, *Ann. Stat.* **24**, 2058 (1996).
- [120] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, and C. J. Carey, *Nat. Methods* **17**, 261 (2020).
- [121] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX: Composable transformations of PYTHON+NumPy programs (2018), <http://github.com/google/jax>.
- [122] I. Babuschkin, K. Baumli, A. Bell, S. Bhupatiraju, J. Bruce, P. Buchlovsky, D. Budden, T. Cai, A. Clark, I. Danihelka, A. Dedieu, C. Fantacci, J. Godwin, C. Jones, R. Hemsley, T. Hennigan, M. Hessel, S. Hou, S. Kapturowski, T. Keck *et al.*, The DeepMind JAX Ecosystem (2020), <http://github.com/deepmind>.
- [123] S. Ruder, [arXiv:1609.04747](https://arxiv.org/abs/1609.04747).
- [124] D. P. Kingma and J. Ba, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [125] G. Kresse and J. Furthmüller, *Comput. Mater. Sci.* **6**, 15 (1996).
- [126] G. Kresse and J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).
- [127] G. Kresse and D. Joubert, *Phys. Rev. B* **59**, 1758 (1999).
- [128] P. E. Blöchl, *Phys. Rev. B* **50**, 17953 (1994).
- [129] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [130] A. Jain, G. Hautier, S. P. Ong, C. J. Moore, C. C. Fischer, K. A. Persson, and G. Ceder, *Phys. Rev. B* **84**, 045115 (2011).
- [131] N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng, and G. Ceder, *Nature (London)* **624**, 86 (2023).
- [132] J. Leeman, Y. Liu, J. Stiles, S. B. Lee, P. Bhatt, L. M. Schoop, and R. G. Palgrave, *PRX Energy* **3**, 011002 (2024).
- [133] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, *npj Comput. Mater.* **1**, 15010 (2015).
- [134] M. Aykol, S. S. Dwaraknath, W. Sun, and K. A. Persson, *Sci. Adv.* **4**, eaaq0148 (2018).
- [135] B. Lafuente, R. T. Downs, H. Yang, and N. Stone, 1. The power of databases: The RRUFF project, in *Highlights in Mineralogical Crystallography*, edited by T. Armbruster and R. M. Danisi (De Gruyter (O), Berlin, 2016), pp. 1–30.
- [136] S. Gates-Rector and T. Blanton, *Powder Diffr.* **34**, 352 (2019).
- [137] ICDD, <https://www.icdd.com/pdf-2/>.