Equivariant Diffusion Solution for Inorganic Crystal Structure Determination from Powder X-ray Diffraction Data

Dongfang Yu[a,b], Zhewen Zhu[a,b], Fucheng Leng[c], Yizhou Zhu[b]

[a] School of Materials Science and Engineering, Zhejiang University, Hangzhou 310058, China

[b] Department of Materials Science and Engineering, Westlake University, Hangzhou 310030, China

[c] School of Engineering and Architecture, University College Cork, Cork, Ireland

*Corresponding authors: Yizhou Zhu

E-mail: zhuyizhou@westlake.edu.cn.

Abstract

Determining the crystal structures of inorganic crystalline materials is crucial as the structures encode essential information about their physical, chemical, and mechanical properties. Powder X-ray diffraction is one of the most widely used structural characterization techniques. However, determining crystal structure directly from experimental powder X-ray diffraction patterns can be challenging and requires significant crystallographic knowledge, which still heavily relies on manual inspection by human experts. Even the state-of-the-art databases contain thousands of entries with incomplete or implausible crystal structure information. In this work, we trained a diffusion model based on equivariant graph neural networks that can infer atomic coordinates from powder X-ray diffraction patterns. Starting from a random guess, our model iteratively refines atom coordinates until it reaches a chemically reasonable structure that matches the target diffraction pattern. Our approach is both efficient and accurate. It takes on average 0.6 seconds to solve the atomic positions per crystal structure, which is several orders of magnitude faster than previous approaches. The success rate reaches 82.3% and 81.6% on the simulated and experimental diffraction datasets, respectively. We revisited energetically unfavorable crystal structures in the database and demonstrated that our model can propose more plausible structure solutions for 39 entries. We also suggested 912 complete crystal structure models for entries in the database lacking all or partial atomic positions, including entries that contain light elements, are natural minerals, or exhibit chemical disorder lattice sites. We demonstrated that conditional equivariant generative model can tackle the structure determination problem and provide high-quality structure models for inorganic crystalline materials, paving the way for automated structural analysis of diffraction patterns in autonomous materials development loops.

Keywords: Crystal structure; X-ray diffraction; Equivariant graph neural network; Generative diffusion model

X-ray diffraction (XRD) is one of the most widely used structural characterization techniques for inorganic crystalline materials. Since Bragg's law was first proposed in 1912, X-ray crystallography has become a fundamental tool that has revolutionized many scientific fields, including mineralogy, geology, chemistry, biology, and materials science[1]. As a core structural characterization technique, X-ray crystallography plays a pivotal role in studying the structure-property relationship of materials. The earliest batch of crystal structures solved by XRD used single-crystal samples[2,3], which are not always readily accessible. Soon after, powder XRD (PXRD) techniques were developed to analyze polycrystalline samples, significantly reducing sample preparation requirements. Beyond phase identification, PXRD is also used to extract information on texture, lattice parameters, grain size, and crystallinity. These advantages make PXRD one of the most widely used structural characterization techniques for inorganic crystalline materials.

However, determining crystal structure directly from PXRD patterns is challenging. When diffraction patterns are collected from polycrystalline samples, the three-dimensional atomic spatial information is compressed into one-dimensional pattern, resulting in a significant loss of information. In most scenarios, PXRD is used jointly with large diffraction or structure databases, such as the International Centre for Diffraction Data (ICDD) and the Inorganic Crystal Structure Database (ICSD). If the acquired PXRD patterns match any existing patterns within the database, the phase identity is revealed. If not, it becomes necessary to infer the crystal structure directly from the PXRD pattern, a difficult task requiring domain knowledge of crystallography and solid-state chemistry.

A reasonably inferred crystal structure model should meet two requirements: chemical reasonableness and matching the target PXRD patterns. Chemical reasonableness requires that the crystal structures adhere to explicit and implicit rules of chemical bonding and atomic arrangement. Pattern matching involves determining whether the reconstructed pattern matches the target pattern while considering various factors, such as texture, background noise, sample size, and X-ray polarization. Therefore, ab initio crystal structure determination from PXRD patterns has long been a manual task that heavily relies on the expertise of human experts, during which determining atomic coordinates is one of the most crucial and challenging steps. The traditional approach generally involves first determining the lattice parameters and possible space groups, then exploring reasonable combinations of atom coordinates, and finally performing Rietveld refinement. Specifically, determining atomic coordinates usually relies on prior knowledge, prerequisite information, and the expertise of expert crystallographers. As a result, for complex structures or those lacking prior information, even experienced crystallographers may require multiple rounds of tedious trial-and-error attempts to arrive at the correct structure. Besides, most researchers who perform PXRD analysis focus on specific materials of their interests rather than crystallography. In these cases, structural analysis based on PXRD could be challenging, potentially resulting in incomplete, controversial, or even incorrect solutions. For example, the ICSD database, which is the largest database for completely identified inorganic crystal structures, contains about 300,000 entries. However, a previous study suspected that among the unique entries with ordered structures, about one-fifth may be spurious[4]. The ICDD database, one of the most widely used diffraction databases, contains more than 480,000 sets of inorganic diffraction data[5]. Nevertheless, it also includes tens of thousands of entries without atomic coordinates in their associated structures (Supplementary Table S1). Given that these two databases already represent cutting-edge resources for structural data on inorganic materials, developing efficient and accurate approaches to solve crystal structures from PXRD is urgently important.

Apart from the traditional method solely relying on the experience and intuition of crystallographers, several alternative approaches have been proposed to tackle the challenge of structure determination from PXRD patterns, including structure optimization algorithms[6-9] and database-based prototype search methods[10]. Density functional theory (DFT) calculations are commonly used to select plausible solutions among candidates. FPASS employed a DFT-supported genetic algorithm for automatic crystal structure solutions based on given PXRD patterns, stoichiometry, and space group[6]. Evolve&Morph combined a DFT-supported evolutionary algorithm with crystal morphing to optimize structure solutions from initial random guesses[8].

However, these approaches typically require hundreds to thousands of DFT calculations to solve a single structure, which is computationally expensive. Besides, they require predetermined space group information, which may be unavailable or unreliable[11]. For database-based methods, Griesemer utilized prototypes as initial guesses to solve about 500 previously unsolved compounds from experimental PXRD patterns in ICDD[10]. Nevertheless, the effectiveness of this approach is limited by the availability and reliability of known structural prototypes.

Recently, machine learning approaches, particularly generative diffusion models, have provided a new paradigm for materials structure generation[12-15]. Xie et al. built the Crystal Diffusion Variational Autoencoder (CDVAE), which employs score matching to generate realistic crystal structures by learning from the data distribution of known stable materials[12]. Without any predefined prototypes, CDVAE shows its ability to generate valid and diverse crystal structures that captures the physical inductive bias of material stability. Equivariant diffusion can effectively preserve reflection, translation, rotation, and mirror symmetries[12-16], enabling generation of physically reasonable crystal structures. Such equivariant denoising models were employed to generate inorganic crystal structures[12,14,15] and molecules[13,16], and showed remarkable performance. Besides generating crystal structures without constraints, conditioning diffusion models allow crystal generation with desired chemistry, mechanical, electronic, or magnetic properties, paving the way towards generative AI-assisted materials design paradigms. Recently, Lipson et al. proposed a generative diffusion model, PXRDnet, to determine the crystal structures from nanocrystalline powder diffraction patterns[17]. This approach addresses the challenge of peak broadening effects due to nanosize crystallites, and solved 200 structures of varying symmetry and complexity from simulated nanocrystalline PXRD patterns. Consequently, the conditional equivariant diffusion model holds tremendous potential for determining atomic coordinates under PXRD pattern conditions.

In this work, we introduce XRDSol, an equivariant diffusion model that can propose crystal structure solutions based on given PXRD patterns. Besides the diffraction patterns, XRDSol requires only stoichiometry and unit cell as inputs, and can reasonably infer atomic coordinates without human intervention. By taking advantage of the equivariance inherently implemented in the model, the reflection, translation, rotation, and mirror symmetries in crystal structures are guaranteed in solutions. XRDSol takes, on average, about 0.6 seconds to provide a crystal structure solution based on a given PXRD pattern, making it approximately 10^4–10^5 times faster than previous works[6-8]. The success rate of XRDSol is 82.3% on the simulated PXRD test dataset (MP-20), and 81.6% on the experimental PXRD dataset (ICDD-20). We revisited hundreds of previously solved (mostly by human experts) but energetically unfavorable structures, and XRDSol successfully provides more plausible structure solutions for at least 39 entries from ICSD. Moreover, for 912 ICDD entries without documented atomic coordinates, XRDSol is able to derive structure solutions from PXRD patterns, and we verified that these solutions are highly plausible with both DFT calculations and manual inspections. The performance of XRDSol is robust across different chemistry systems, including those involving light-element materials, natural minerals, and structures with substitutional disorder. Such performance demonstrates that successfully trained equivariant diffusion models can implicitly learn crystallographic knowledge, and can be deployed to provide crystal structure solutions based on PXRD patterns, paving the way for automated, efficient, and accurate structural analysis for inorganic materials.

Results and discussion

Conditional Equivariant Diffusion for Crystal Structure Solutions

Solving crystal structures from PXRD patterns typically begins with determining lattice parameters and identifying space group, both of which are relatively well-established tasks[18-20] if the pattern has reasonable resolution and especially when the unit cell is not large. The information on the number and type of atoms within the determined unit cell can also be obtained using well-established techniques. However, determining atomic coordinates could be challenging due to the vast number of possibilities in the configurational space. Our XRDSol is specifically designed to address this challenging step of determining atomic coordinates.

The training of XRDSol includes a forward noising process and a reverse denoising process (Fig. 1a and Fig. S1). In the forward noising process, we disrupt the atomic arrangement of a known crystal structure S_0 within its unit cell by applying random displacements with gradually increasing noise level over 1000 steps ($t = 0 – 1000$). To account for the periodic invariance of the crystalline materials, the random displacements were set to follow a wrapped normal distribution in all three lattice dimensions during the forward process. At the noise limit, atoms in structure S_T acquire random atomic coordinates. In the reverse denoising process, the model is trained to restore the atomic coordinates from S_T to S_0 with the guidance of target PXRD patterns. The PXRD patterns are compressed into a feature space using a neural network architecture, and then mapped to nodes feature distribution in an equivariant graph neural network, which predicts the atomic coordinates. Since our training set (MP-20) consists of thermodynamically stable structures, the model implicitly learns crystallographic knowledge throughout the denoising process. Consequently, it can provide structures that are both chemically reasonable and match the target PXRD patterns, thus enabling crystal structure determination from PXRD patterns.

After successful training, the model is then capable of using given PXRD patterns to infer the correct crystal structures from random initial guesses through the denoising process. We illustrate this process using Eu_2CrSbO_6 (mp-1516449) from the MP-20 dataset as an example (Fig. 1b). 10 atoms (2 Eu, 1 Cr, 1 Sb, and 6 O atoms) with random atomic coordinates were initialized in a primitive cell ($a = b = c = 5.67$ Å, $\alpha = \beta = \gamma = 60°$). We use cosine similarity $R_{\cos}$ to quantify the consistency between the target and reconstructed PXRD patterns (more details in SI). The randomly initialized structure differs significantly from the ground truth structure, reflected by a low $R_{\cos}$ of 0.315. Under the corresponding PXRD pattern condition, the atomic positions are iteratively denoised, and the simulated PXRD pattern gradually aligns with the target pattern. At $t = 300$, the intermediate reconstructed structure still shows unreasonable atomic coordinates and a dissimilar PXRD pattern ($R_{\cos} = 0.488$). By $t = 600$, the $R_{\cos}$ increases sharply to 0.981, indicating closer alignment with the target structure. Finally, at t = 1000, the resulting crystal structure has an extremely high R_{cos} of 0.999, demonstrating excellent reconstruction performance guided by the target PXRD pattern. The reconstructed crystal structure also closely resembles the target structure (Fig. S2). In this way, a successfully trained model can reconstruct a reasonable crystal structure based on a given PXRD pattern.

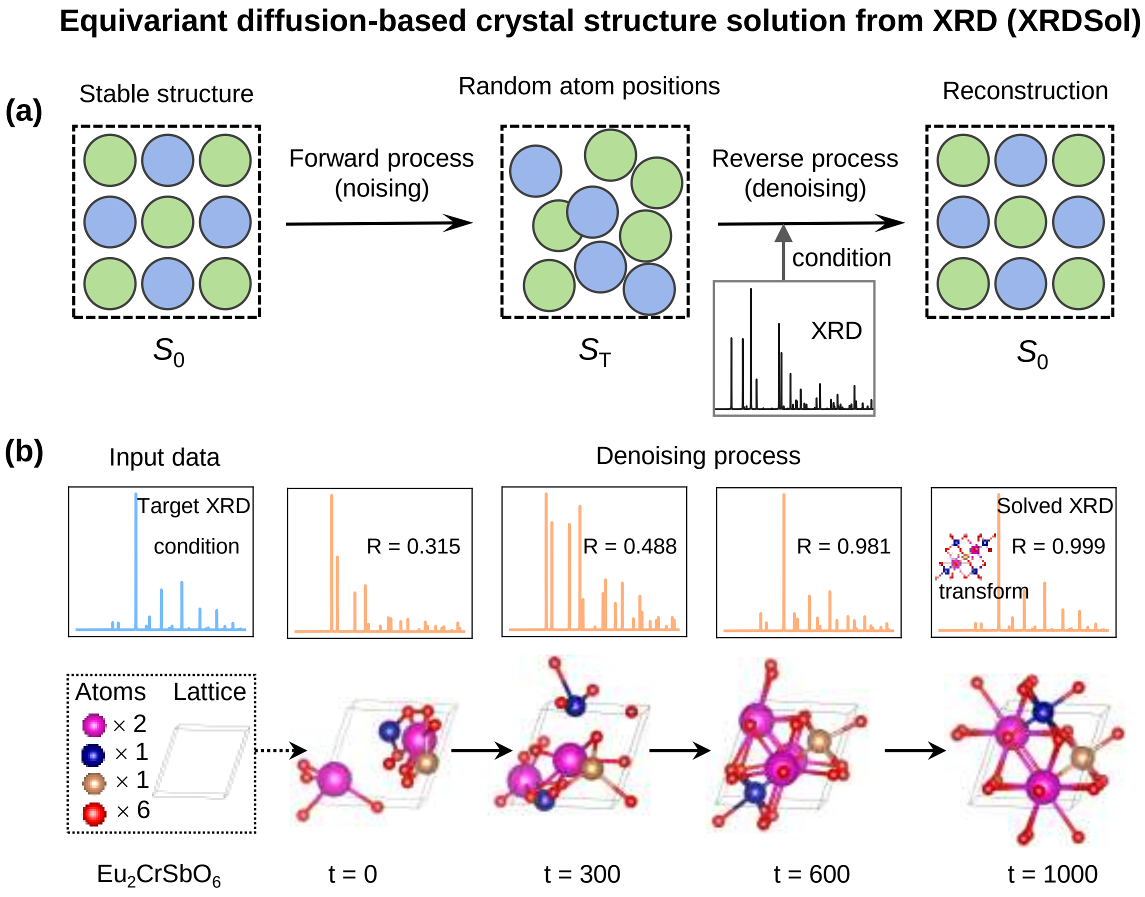

Figure 1. Equivariant diffusion-based crystal structure solution from PXRD patterns (XRDSol). (a) The training process of XRDSol consists of a forward noising process and a reverse denoising process. In the forward process, we disrupt the atomic arrangement within the unit cell of a stable crystal structure by applying random displacements with a gradually increasing noise level, resulting in completely random atomic positions at the noise limit. In the reverse process, the model is trained to recover the atomic positions by removing the random displacements we applied.

During this process, the target PXRD pattern is used as a condition, and the model aims to recover the crystal structure while also matching the target PXRD pattern. (b) Eu_2CrSbO_6 (mp-1516449) as an example to illustrate the evolution of crystal structure from random initialization to the final solution during the denoising process. The initial structure consists of 10 atoms in a given unit cell (2 Eu, 1 Cr, 1 Sb, and 6 O atoms) with random atomic coordinates ($a = b = c = 5.67$ Å, $\alpha = \beta = \gamma = 60°$). During the solution process, the atomic coordinates are iteratively updated over 1000 steps, progressively leading to the correct crystal structure while matching the target PXRD patterns as well. Two intermediate reconstructed structures at $t = 300$ ($R_{\mathrm{cos}} = 0.488$) and $t = 600$ ($R_{\mathrm{cos}} = 0.981$) are shown to illustrate the solving process. Finally, at $t = 1000$, the reconstructed Eu_2CrSbO_6 crystal structure matches the ground truth, and the reconstructed PXRD patterns closely resemble the target patterns ($R_{\mathrm{cos}} = 0.999$).

Ultrafast and High-Quality Crystalline Materials Reconstruction

We evaluated the performance of XRDSol in terms of efficiency and accuracy. The test dataset contains 9046 set-aside unseen crystal structures in the MP-20 dataset. Starting from a randomly initialized structure, XRDSol reconstructs the crystal structure with a simulated PXRD pattern as a condition. The average solution time is only 0.6 seconds per crystal structure with a batch size of 128 on a single NVIDIA V100 GPU (Fig. S3). In contrast, previous approaches that start from randomly initialized structures but are based on evolution algorithms or particle swarm optimization[8,21], typically use a runtime of 0.5 to 1 day and require multiple DFT calculations[8] to solve a single structure. Our XRDSol is about 10^4 to 10^5 times faster than the previous works[6-8] without using any database or first principles calculation (Fig. 2a). Furthermore, we evaluated the accuracy of both the reconstructed crystal structures and their reconstructed PXRD patterns in the test dataset. Compared to the ground truth, 91.5% of the reconstructed PXRD patterns exhibit a high similarity to the target patterns, with an R_{cos} exceeding 0.9, and the average R_{cos} reaches 0.974 (Fig. 2b).

Structural reasonableness is an important metric to assess the quality of crystal structure solutions from PXRD patterns. To evaluate the quality, we compared the reconstructed crystal structures to the ground truth solutions from the test dataset. A reconstruction is considered successful only if site-to-site matching is achieved and the maximum distance between all paired atomic sites (MaxDist) falls below a specified threshold. For all successfully reconstructed solutions, we also computed the scaled root mean square distance (sRMS) between the original and reconstructed crystal structures to quantify the accuracy of the reconstruction. All these structure analysis were performed with pymatgen and a typical threshold for MaxDist is chosen as 0.5[22]. XRDSol successfully restored the atomic coordinates of 82.3% of structures in the MP-20 test dataset (Fig. 2c). Furthermore, the reconstructed crystal structures from XRDSol generally exhibit low sRMS values. The average and median of the sRMS are 0.024 and 0.006, respectively (Fig. 2c). Furthermore, we also calculated the root mean square deviation (RMSD) of atom positions to reflect the reconstruction quality of the crystal structures (Fig. S4). 72.15% of the structures achieved an RMSD less than 0.1 Å and 78.13% achieved an RMSD less than 0.3 Å. Among the structures with RMSD < 0.1 Å, the mean RMSD is 0.0215 Å, the median RMSD is 0.0129 Å, the first quartile is 0.0024 Å, and the third quartile is 0.0325 Å. These results confirm that most of the reconstructed crystal structures align well with the original ones. Such performance indicates that XRDSol is capable of reconstructing high-quality structures based on the given PXRD patterns. We showcase six materials from different chemical systems (K_2NaYCl_6, Li_5SbS, $CaEuNbFeO_6$, CuNi, $CsYbBr_3$, and $InGa(AgTe_2)_2$) as examples (Fig. 2d). These examples contain 2 to 5 different elements in the unit cell, covering halides, sulfide, oxide, intermetallic, and tellurite. XRDSol successfully inferred the correct crystal structure for all of them, highlighting its robust and excellent performance across various chemical systems. Furthermore, quantitative evaluation reveals that XRDSol's strong dependence on structural symmetry, with higher success rates for high-symmetry crystals (e.g., 97.7% success rate for cubic) compared to low-symmetry ones (e.g., 48.5% success rate for triclinic). This performance trend across all seven crystal systems and individual space groups is detailed in Fig.S5 and S6.

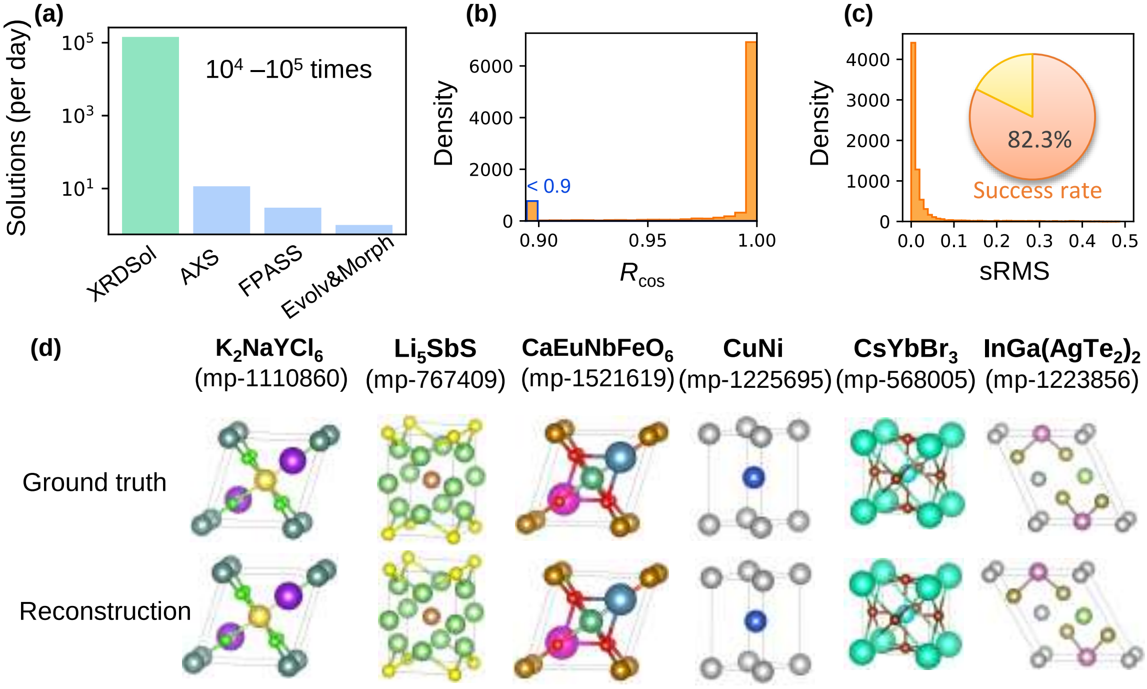

Figure 2. Speed and accuracy performance of XRDSol on MP-20 test dataset. (a) Comparison of solving speed (solutions per day) between XRDSol and three other baseline models, including AXS[7], FPASS[6], and Evolv&Morph[8]. (b) Distribution of R_{cos} for 9046 structures solved by XRDSol in the MP-20 test dataset, reflecting the reconstruction quality of the PXRD patterns (c) Distribution of sRMS between the original structures and those successfully solved by XRDSol, reflecting the crystal structure reconstruction quality. (d) Visual verification of six reconstructed crystal structures compared to their original counterparts.

Revisiting energetically unfavorable structures in the ICSD database

Thermodynamic stability is a crucial property for inorganic crystalline materials, and convex hull analysis is commonly used for thermodynamic assessment[4]. Although first-principles calculations confirm that most experimentally reported structures have energy close to the convex hull, Sun *et al.* suspected that about 20% of the structures in the ICSD database exhibit poor thermodynamic stability with implausibly high energy above the convex hull (E_{hull})[4]. Manual inspection suggests that some of these structures, previously solved and reported by human experts, are probably incorrect or incomplete, due to unreasonable or partially missing atomic positions. We revisited these high-energy structures sourced from ICSD, all of which contain no more than 20 atoms in the primitive cell and their E_{hull} are higher than 0.1 eV/atom (Table S2). The workflow of revisiting high-energy structures from ICSD is shown in Fig. S7. Calculated PXRD patterns were used as inputs to see if XRDSol can propose more reasonable or complete structures that are energetically favorable. At least 39 structures proposed by XRDSol have a reasonable pattern match and E_{hull} lower than 0.1 eV/atom, suggesting that these structures are highly plausible (Supplementary Table S3–S4, Fig. S8–S16). We manually inspected these structure solutions with a few representative examples analyzed below.

HoGeAg is an intermetallic compound with Ho ions forming a distorted Kagome lattice, leading to the frustrated spin ice state[23,24]. An earlier report based on neutron diffraction claimed that silver occupies two distinct Wyckoff positions, while 3 germanium atoms are symmetrically equivalent[24]. However, the E_{hull} of this originally reported structure is 0.17 eV/atom, extremely high for an intermetallic compound. XRDSol proposed a different structure from the original one, in which Ho and Ag atoms arrange in alternating layers and occupy 3f and 3g Wyckoff positions, respectively. Ge occupies two distinct Wyckoff positions, 1a site within the Ho plane, and 2d sites within the Ag plane. Our proposed structure agrees with a recent literature report[23] (RMSD = 0.04 Å) (Fig. S17), and it is predicted as a stable compound based on first principles calculations (E_{hull} = 0 eV/atom), suggesting the solution is highly plausible.

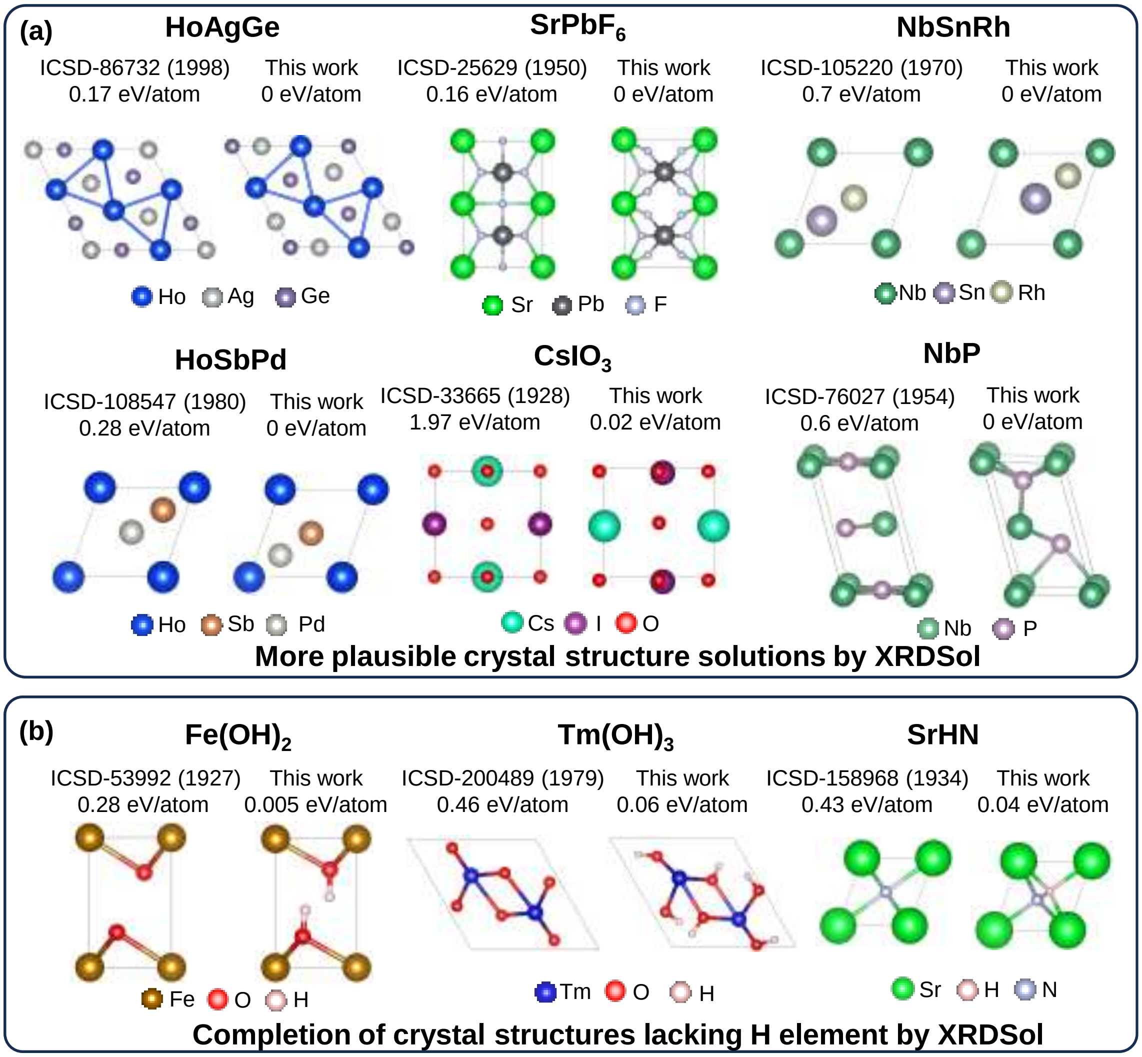
(a)
HoAgGe
ICSD-86732 (1998) 0.17 eV/atom
This work 0 eV/atom
Ho Ag Ge
SrPbF6
ICSD-25629 (1950) 0.16 eV/atom
This work 0 eV/atom
Sr Pb F
NbSnRh
ICSD-105220 (1970) 0.7 eV/atom
This work 0 eV/atom
Nb Sn Rh
HoSbPd
ICSD-108547 (1980) 0.28 eV/atom
This work 0 eV/atom
Ho Sb Pd
CsIO3
ICSD-33665 (1928) 1.97 eV/atom
This work 0.02 eV/atom
Cs I O
NbP
ICSD-76027 (1954) 0.6 eV/atom
This work 0 eV/atom
Nb P
More plausible crystal structure solutions by XRDSol
(b)
Fe(OH)2
ICSD-53992 (1927) 0.28 eV/atom
This work 0.005 eV/atom
Fe O H
Tm(OH)3
ICSD-200489 (1979) 0.46 eV/atom
This work 0.06 eV/atom
Tm O H
SrHN
ICSD-158968 (1934) 0.43 eV/atom
This work 0.04 eV/atom
Sr H N
Completion of crystal structures lacking H element by XRDSol

Figure 3. Examples of revisited entries in the ICSD database that originally associate with energetically unfavorable crystal structures, with original crystal structures on the left, and the reconstructed crystal structures on the right. Examples include (a) entries that likely have incorrect atomic arrangements in the unit cell, and (b) entries associated with incomplete crystal structures missing light elements, such as hydrogen.

$SrPbF_6$, a hexafluorideplumbate, was first synthesized and characterized by Hoppe in 1958[25]. It was originally proposed to crystallize in space group $P4_2/mmc$. However, the original report noted the presence of several unindexed peaks in the PXRD pattern, which were attributed to another unknown phase. First principles calculations reveal that the original structure has a high

E_{hull} of 0.16 eV/atom. It was not until half a century later, in 2018, that this hexafluorideplumbate was experimentally revisited with its crystal structure corrected[26]. This revisited structure perfectly matches the PXRD profile obtained from the experiment with no unindexed peaks, suggesting its correctness. The crystal structure proposed by XRDSol is identical to the later corrected one with RMSD of 0.08 Å (Fig. S18), and is predicted to be a stable compound (E_{hull} = 0 eV/atom). The corrected structure belongs to space group $P4_2/mcm$, in which Sr is eight-coordinated instead of ten-coordinated by F. This example demonstrates the capability of XRDSol to propose structures with more chemically reasonable coordination environments.

Half-Heusler is a group of ternary intermetallic compounds, among which many are promising thermoelectric materials. Three equal molar elements form three interpenetrating face-centered cubic sublattices, and only one element is eightfold coordinated. There are two types of half-Heusler materials, with anion (MgAgAs-type) and cation (MgCuSb-type) occupying the eight-coordinated site, respectively. Correct assignment of lattice sites is crucial to understand their structure-property relationships[27]. Among the ICSD entries we investigated, XRDSol proposed different lattice site assignments for 10 half-Heusler compounds, including NbRhSn, TiNiSn, ZrPdSn, HfPtSn, HfRhSb, NiTiSb, HoPdSb, TmSbPd, MgCuSn, and MgAgAs. The E_{hull} of the original structures range from 0.170 eV/atom to 0.704 eV/atom. In contrast, structures proposed by XRDSol exhibit much lower E_{hull}, ranging from 0 eV/atom to 0.079 eV/atom, suggesting solutions from XRDSol are more plausible.

$CsIO_3$, a covalent perovskite material[28], was first reported in 1928 with a cubic perovskite structure in the $Pm\bar{3}m$ space group, where Cs occupies the six-coordinate B-site[29]. The calculated E_{hull} of this original structure is extremely high at 1.97 eV/atom. In contrast, our solution from XRDSol exhibits a much reasonably lower E_{hull} of 0.02 eV/atom. The larger Cs cations are moved to twelve-coordinate A-sites. Our structure solution is very close to the recently synthesized and experimentally characterized structure from a single crystal with RMSD of 0.28 Å (Fig. S19)[30], validating its plausibility.

NbP was first characterized by PXRD in 1954 and reported as centrosymmetric with $I4_1/amd$ space group. It contains a very uncommon four-fold square planar coordination for Nb and P. DFT calculation reveals an extremely high E_{hull} of 0.6 eV/atom, further raising doubts about the validity of this structure. In contrast, XRDSol suggests a non-centrosymmetric structure with a different space group $I4_1md$, where Nb is in a trigonal prismatic coordination environment with six P atoms. DFT calculation confirms this proposed structure is stable (E_{hull} = 0 eV/atom), and it agrees with a few other literature reports[31-33]. In particular, our solution matches that reported by Zhao *et al.* with RMSD of 0.07 Å (Fig. S20)[33].

Locating hydrogen atomic coordinates is an intricate task due to their weak scattering property, heavily dependent on the crystallographer's expertise. Typically, people choose to combine other characterization techniques, such as neutron scattering, with PXRD to determine the positions of hydrogen atoms, as it is very challenging to determine hydrogen atom positions solely based on the PXRD pattern. For hydrogen-containing materials, XRDSol can determine the positions of other heavy atoms while suggesting the most reasonable and likely positions for hydrogen atoms based on its embedded crystallographic knowledge. Here, we demonstrate such capability of XRDSol using three examples: $Fe(OH)_2$, $Tm(OH)_3$, and SrNH, which are three compounds that had unassigned hydrogen atoms structures in early reports (Fig. 3b)[34-36]. These incomplete structures violate charge neutrality, resulting in significantly high E_{hull} of 0.28, 0.46, and 0.43 eV/atom, respectively. XRDSol can infer reasonable hydrogen positions from its implicitly learned crystallographic knowledge. For $Fe(OH)_2$, it places H at 2d Wyckoff position (1/3, 2/3, 0.453), in good agreement with neutron diffraction results[37], with a much lower E_{hull} of 0.005 eV/atom. Similarly, in $Tm(OH)_3$, XRDSol restored hydrogen at 6h Wyckoff sites (0.721, 0.863, 3/4), reducing E_{hull} to 0.06 eV/atom. While we failed to find experimental reports on H positions in $Tm(OH)_3$, our solved structure is isomorphous to other lanthanide trihydroxide, such as $Er(OH)_3$, $Ho(OH)_3$, and $Dy(OH)_3$[38]. For SrNH, early PXRD reports showed Sr and N forming a rocksalt-type lattice and H coordinates were missing. Structure proposed by XRDSol agrees with the rocksalt framework. Complex rotational disorder of the imide ion may lead to multiple partially occupied H sites[39], which XRDSol is incapable to handle. This issue will be further elaborated in the next section. Instead, it places H along the <111> direction, consistent with literature reports suggesting this is the most energetically favorable site with minimized electrostatic energy[39]. The SrNH with a complete crystal structure has a reduced E_{hull} of 0.04 eV/atom.

Solving crystal structures without atomic coordinates information

Experimental PXRD patterns differ from simulated PXRD profiles, especially in peak intensities, due to various factors, including texture, polarization, background noise, and temperature effects. Therefore, it is necessary to benchmark the performance of XRDSol on experimental diffraction data. We assembled the ICDD-20 dataset, which contains 1000 sets of experimental PXRD data and their ground truth crystal structures (up to 20 atoms per primitive cell). XRDSol achieved a remarkable 81.6% success rate on this dataset, demonstrating robustness against intensity variations in experimental PXRD profiles. R_{cos} is expectedly lower compared to the performance on the MP-20 test dataset. The average R_{cos} is 0.67, with 71.1% of the samples exceeding 0.6 (Fig.4a). The distribution of R_{cos} on ICDD-20 dataset provides a gauge of the fitting quality by using simulated PXRD patterns to fit experimental PXRD patterns. Since structural reasonableness is more important than achieving better R factors[40], arriving at the correct structure is the ultimate goal of structure solutions.

We then deployed XRDSol to solve crystal structures from ICDD entries missing atomic coordinates. Obviously, no ground truth is available for these entries. XRDSol inferred missing atomic coordinates based on its training experience (Fig. 4b). The input for solving these entries includes known target PXRD patterns, lattice parameters, number and type of atoms. Among 912 entries we investigated, the average R_{cos} reaches 0.708, with 79.6% exceeding 0.6, indicating high pattern reconstruction quality. Note that the distribution of R_{cos} is quite similar to that on ICDD-20 dataset, indicating that such a distribution is typical for experimental PXRD patterns.

Thermodynamic assessment is another important metric to evaluate the plausibility of structure solutions. We performed first principles calculations to evaluate E_{hull} of these solved structures (Fig. 4c). All 912 structures exhibit E_{hull} below 0.1 eV/atom, which is a typical threshold for determining structure plausibility, with 81.3% exhibiting E_{hull} below 0.05 eV/atom. To the best of our knowledge, this is the largest dataset containing the most structures solved from experimental PXRD patterns using an automated algorithm-based solver (Fig. S21)[6-8,10]. The reconstructed structures and their PXRD patterns of 912 entries are available in the supporting dataset. The remarkable performance demonstrates the capability of XRDSol to infer crystal structures from experimental PXRD patterns.

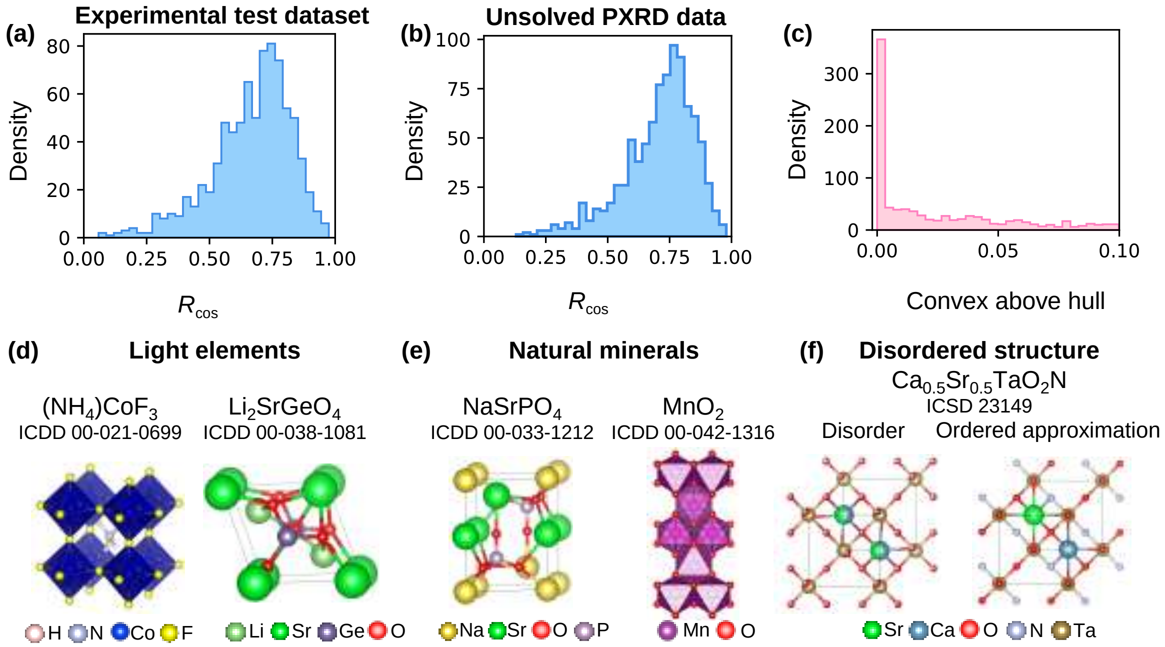

Figure 4. Performance and examples of XRDSol on solving crystal structure from experimental PXRD patterns. PXRD pattern reconstruction quality (evaluated by R_{cos}) for (a) the 816 crystal structures successfully solved by XRDSol in experimental test dataset ICDD-20. (b) the 912 crystal structures solved by XRDSol based on ICDD entries without documented atomic coordinates in the database. (c) Thermodynamic assessment of the 912 crystal structures solved by XRDSol based on ICDD entries without documented atomic coordinates in the database, evaluated by the energy

above the convex hull E_{hull}. Visual verification of six selected crystal structures solved by XRDSol that (d) include light elements, such as hydrogen and lithium, (e) are natural minerals that may contain associated impurity phases, and (f) exhibit chemical disorder at certain lattice sites.

Light elements such as hydrogen and lithium exhibit weak X-ray scattering due to their low atomic numbers, making it challenging to determine their lattice sites directly only based on PXRD patterns. $(NH_4)CoF_3$ is an ammonium cobalt trifluoride, and was documented in 1970 without atomic coordinates[41]. XRDSol solved its structure as a cubic perovskite, with cobalt at the B-site and ammonium at the A-site. Nitrogen is tetrahedrally coordinated with hydrogen, forming the correct ammonium structure. The structure has a low E_{hull} of 0.03 eV/atom and a good diffraction pattern match with the original one (R_{cos} = 0.846) (Fig. S22). Moreover, it aligns well with a previous report with RMSD of 0.28 Å (Fig. S23)[42]. Li_2SrGeO_4 is a candidate host material for luminescent phosphors. It was synthesized in 1986 and characterized as a tetragonal phase with *I*-42*m* space group, but its atomic coordinates were not identified. The structure XRDSol proposed is a Sr-analog of Li_2CaGeO_4 phase. It is thermodynamically stable (E_{hull} = 0 eV/atom), and its diffraction pattern matches well with the original pattern (R_{cos} = 0.767) (Fig. S24). Given the chemical similarity between Ca and Sr, the proposed structure is highly plausible and consistent with literature[10,43]. In particular, our solution agrees with the recently reported structure by Griesemer *et al.* with RMSD of 0.08 Å (Fig. S25)[10].

The presence of impurities represents a significant factor that can adversely impact the success rate of structure determination from PXRD patterns. At least 15 structures of natural minerals proposed by XRDSol have the E_{hull} lower than 0.1 eV/atom, suggesting that these structures are highly plausible (Supplementary Table S5). $NaSrPO_4$ is a natural olgite mineral discovered in 1980. Early XRD experiments revealed that it belongs to the hexagonal lattice system and was originally assumed to have the *P*3 space group[44]. Lattice parameters were extracted from diffraction patterns but without atomic coordinates. XRDSol provided a solution with a different space group *P*3*m*1. Despite that the reconstructed pattern matches most of the major peaks, strong intensity deviation was observed, leading to a relatively low $R_{\cos}$ of 0.313 (Fig. S26). This is probably due to the strong texture of the sample and potentially associated impurity phases[45]. DFT calculation confirms that the structure is highly plausible with a low E_{hull} 0.01 eV/atom. We confirmed that our proposed structure agrees with a recent report on olgite crystal structure, except that one lattice site might be co-occupied by Na and Sr[45]. Natural ramsdellite MnO_2 was documented by ICDD Grant-in-Aid in 1989. However, the atomic coordinates for this entry were missing, possibly due to the absence of indexed peaks[46,47]. The structure solved by XRDSol agrees with literature reports[46,47], and it has a low E_{hull} of 0.009 eV/atom. In particular, our solution agrees with the recently reported structure by Post *et al.* with RMSD value of 0.1 Å (Fig. S27)[46]. We believe the proposed structure is highly plausible. Compared to the simulated pattern of proposed structure, a few low-intensity peaks were missing in the experimental XRD pattern (Fig. S28), which is probably the cause of preferred orientation. In contrast, XRDSol shows robust performance in these natural mineral samples.

Chemical disorder is a common phenomenon in inorganic crystalline materials, where different types of atoms are randomly distributed within a crystal lattice. Currently, XRDSol cannot propose crystal structures with chemical disorder, since all atoms have pre-assigned element types, similar to the limitation of DFT calculations. Nevertheless, we challenged XRDSol with two PXRD patterns from disordered materials: $Sr_{0.5}Ca_{0.5}TaO_2N$[48] and $Eu_{0.5}Dy_{0.5}TiO_3$[49]. Surprisingly, we all reached solutions that are "ordered approximations" of the ground truth structures that contain chemical disorder. $Sr_{0.5}Ca_{0.5}TaO_2N$ is an oxynitride perovskite material, in which complex anion order behavior may exist. The A-site should be co-occupied by Sr and Ca. In contrast, in the structure proposed by XRDSol, two A-sites in a supercell are occupied by Sr and Ca, respectively. For the anion-sites, XRDSol proposed an ordered *trans*-type anion configuration, in which nitrogen occupies axial sites while oxygen occupies the equatorial sites. Remarkably, this is in excellent agreement with neutron diffraction results[50]. The ordered structure that we solved reaches $R_{\cos}$ of 0.879 (Fig. S29). Similarly, for perovskite $Eu_{0.5}Dy_{0.5}TiO_3$ with Eu/Dy chemical disordering, XRDSol successfully proposed a fully ordered structure model, which shows a good match between the reconstructed and the original PXRD patterns (Fig. S30). Atoms at all Wyckoff positions are correctly assigned except for the presence of chemical disordering[49] (Fig. S31). Such performance demonstrates that for materials with chemical disorder, XRDSol is capable of proposing reasonably ordered structure models, which is crucial for further structural refinement tasks.

Discussion

Determining crystal structures of inorganic crystalline materials plays a crucial role in physical science. In this work, we propose using an equivariant diffusion-based model to infer atomic coordinates from PXRD patterns. We demonstrate that after training, our model implicitly learns crystallographic knowledge, enabling it to propose chemically reasonable structure solutions that match the target PXRD patterns. It shows efficient and robust performance across various chemistry systems and on both simulated and experimental PXRD patterns. Our approach is several orders of magnitude faster than several previous methods that require first-principles calculations or evolution algorithms. We demonstrated its capability by correcting crystal structures of implausible entries and completing structures for entries lacking partial or all atomic positions in the databases. We envision three practical application scenarios for XRDSol: AI-assisted manual PXRD analysis, quality control of large-scale structural databases for inorganic materials, and integration into autonomous materials discovery platforms. For experienced crystallographers, XRDSol can provide high-quality initial models that can serve as a starting point for the following Rietveld refinement, saving time and effort from human experts. For researchers who focus on specific materials rather than crystallography, XRDSol can provide a baseline model with good quality that can potentially satisfy their need for routine structural characterization in their research. For large-scale structural databases, XRDSol can be valuable for database maintenance and quality control, by providing an initial screening of suspicious entries, identifying potential errors in existing records, supplementing missing structural parameters, and offering cross-validation for ambiguous cases. Besides, the automated nature of our algorithm also makes it suitable for potential integration into autonomous materials discovery platforms, since it requires minimal human intervention and has a rapid processing capability.

Our current approach could still be improved in several ways. The inherent stochastic nature of the diffusion model can introduce uncertainty in crystal structure solutions and sometimes requires repeated runs. Moreover, the success rate decreases significantly as the number of atoms in the unit cell exceeds twenty. Imposing additional constraints, such as space groups, will significantly reduce the degrees of freedom when the model searches the configuration space, making it possible to tackle the solution of crystal structures with larger unit cells or more atoms. Furthermore, the low-symmetry structures still pose great challenges. In the future, we will focus on overcoming the current limitations in handling large-unit-cell crystals and low-symmetry systems. In addition, the lattice parameters can in principle be inferred from the PXRD patterns as well. Integrating more domain-specific crystallographic knowledge will be a future research direction, ultimately aiming to enable automated structural analysis of PXRD patterns. We believe that such automated analysis tools have the potential to be integrated into autonomous materials discovery platforms and paving the way for the automated design, development, and optimization of novel functional materials.

Acknowledgments

We thank Dr. Thomas Blanton from International Center for Diffraction data, Prof. Shifeng Jin from institute of Physics, Chinese Academy of Sciences, and Dr. Yinghui Zhang from Chinese Academy of Geological Sciences for helpful discussions. We acknowledge the funding support from the Research Center for Industries of the Future (RCIF) from Westlake University and the computation resource from the High-Performance Computing Center (HPC) at Westlake University.

Author Contribution

Y.Z. conceived and supervised the project. Y.Z. and D.Y. designed the algorithm and experiments. D.Y. wrote the code and performed computations and analyses. D.Y. and Z.Z. conducted first-principles calculations. Y.Z. and D.Y. wrote the manuscript. All authors contributed to discussions and manuscript revisions.

Competing Interests

The authors declare no competing interests.

I. Methods

A. Conditional equivariant diffusion

Graph representation of crystal structure. For each given crystal structure, its primitive reduced cell was chosen as its unique represetation of the infinite 3D periodic crystal structure. The 3D periodic crystal structures were represented as a graph $\boldsymbol{G}$ = ($\boldsymbol{L}$, $\boldsymbol{A}$, $\boldsymbol{P}$, $\boldsymbol{E}$). $\boldsymbol{L}$ is a 3 × 3 matrix that represents a set of lattice vectors of the crystal structure, which is equivalent to the conventional lattice parameters representation (a, b, c, α, β, and γ). $\boldsymbol{A}$ is the atom type matrix $\boldsymbol{A}$ = [a_1, a_2, …, a_n]^T$\in\mathbb{R}^{n\times d_a}$, $P=[\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_n]^T\in\mathbb{R}^{n\times 3}$, where $\boldsymbol{p}_i=(x_i,y_i,z_i)\in\mathbb{R}^3$denotes the 3D position of $\boldsymbol{a}_i$, and $E\in\mathbb{R}^{n\times n}$ is the adjacency matrix and denotes the edge features.

Conditional equivariant diffusion model. The diffusion model is a generative model inspired by non-equilibrium thermodynamics. It includes a forward noising process and a reverse denoising process (Fig. S22). During the forward process, random normal distribution noise is gradually added to the input data x_0 via T steps, thereby establishing a Markov chain $x_0\rightarrow x_1\rightarrow\cdots\rightarrow x_t\rightarrow\cdots\rightarrow x_T$. This noising process can be described by the transition distribution as $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})=$

$\mathcal{N}(\boldsymbol{x}_t|\alpha_{t|t-1}\boldsymbol{x}_{t-1}, \sigma_{t|t-1}^2\mathbf{I})$, where the parameter α_t controls the amount of x_0 retained and σ_t controls the amount of noise added. For XRDSol, we perform diffusion on the atomic coordinates of the given structure $\boldsymbol{S}$, which defines a Markov chain $\boldsymbol{S}_0 \rightarrow \boldsymbol{S}_1 \rightarrow \cdots \rightarrow \boldsymbol{S}_t \rightarrow \cdots \rightarrow \boldsymbol{S}_T$ [12,14]. Due to the periodicity of crystal structures, we added wrapped normal distribution noise $\mathcal{N}_W$ on the atomic coordinates to account for the intrinsic periodicity. We define our forward process via transition distribution as $q(\boldsymbol{S}_t|\boldsymbol{S}_{t-1}) = \mathcal{N}_W(\boldsymbol{S}_t|\alpha_{t|t-1}\boldsymbol{S}_{t-1}, \sigma_{t|t-1}^2\mathbf{I})$. During the reverse denoising process, the equivariant graph neural network (EGNN) ϕ predicts the noise to construct the desired samples. EGNN is a type of Graph Neural Network that satisfies the equivariance constraint[51]. In this work, the graph is initialized by a fully connected graph $\boldsymbol{G}$ based on atomic positions and lattice information. Each node is endowed with coordinates as well as features. The edge construction follows a fully connected graph $\boldsymbol{G}$, where every atom in the unit cell is connected to every other atom, forming a fully connected graph. The edge features are defined as fractional coordinate differences between connected atoms. In this setting, EGNN consists of 3 EGNN blocks, each of which composed of equivariant convolutional layers. The noisy structure S_t is passed through the EGNN to extract structure features. And the structure features are concatenated with the PXRD features that are derived from the target PXRD pattern by a neural network. Next, the concatenated features are fed into a neural network to predict the atomic coordinates of $S_{t\text{-}1}$. This process is repeated for 1000 steps and gradually restores the atomic coordinates from S_T to S_0 with the guidance of target PXRD patterns. We define the reverse denoising process via transition distribution as $p(\boldsymbol{S}_{t-1}|\boldsymbol{S}_t) = \mathcal{N}_W(\boldsymbol{S}_{t-1}|\phi(\boldsymbol{S}_t), \sigma_{t\rightarrow t-1}^2\mathbf{I})$. To solve crystal structures conditioned on given PXRD patterns C, we input the PXRD condition along with atomic coordinates into the neural network to guide the correct placement of atoms. The final denoising process was defined as $p(\boldsymbol{S}_{t-1}|\boldsymbol{S}_t,\ C) = \mathcal{N}_W(\boldsymbol{S}_{t-1}|\phi(\boldsymbol{S}_t,\ C), \sigma^2_{t\rightarrow t-1}\mathbf{I})$. The pseudo-code of XRDSol is provided to illustrate nosing and denoising steps (Table S6).

B. Experimental details

Datasets. (1) **MP-20 dataset (simulated PXRD patterns with ground truth structures):** We employed the MP-20 dataset (45231 structures) to train and evaluate our XRDSol model. The dataset consists of materials sourced from ICSD database. Their energy above the hull and the formation energy are constrained to be less than 0.08 and 2 eV/atom, respectively. These thermodynamic stability constraints ensure the experimental validity of the materials under consideration. (2) **ICDD-20 dataset (experimental PXRD patterns with ground truth structures):** We assembled the ICDD-20 dataset (1000 structures) with at most 20 atoms in the primitive cell and collected their corresponding experimental PXRD patterns. As an experimental dataset, ICDD-20 serves as a crucial benchmark for validating the performance of XRDSol under real-world conditions. (3) **ICSD-20-HE dataset (high-energy structures with simulated PXRD patterns):** We have created the ICSD-20-HE dataset (4157 structures) to revisit energetically unfavorable structures in the ICSD. Their E_{hull} are constrained to be higher than 0.1 eV/atom, respectively. This dataset is used to explore the model's ability to handle the probably incorrect or incomplete structures. (4) **Unsolved experimental XRD pattern dataset:** We use the trained XRDSol model to solve the entries from the ICDD database, of which the atomic coordinates are missing but lattice parameters are known. We start with the entries whose diffraction quality is listed as “star”, “good”, or “indexed”, indicating that the diffraction pattern represents a single-phase crystal with minimal impurities. The lattice and number of formula units per unit cell are already known. The number of atoms should be less than 20.

Together, these four datasets enable a comprehensive evaluation of the model across both simulated and experimental XRD patterns, with or without ground truth structural data. The target PXRD patterns, lattice parameters, number and type of atoms are already known and used as the inputs of XRDSol.

Training XRDSol model. We provide detailed configuration parameters for the training and sampling process within the conditional XRDSol model. The conditional XRDSol model was trained for 1000 steps with a batch size of 256 across four V100 GPUs. The Adam optimizer was employed. The initial learning rate was set to 0.001 and was subjected to decay using the ReduceLROnPlateau scheduler, with a decay factor of 0.6, patience of 30, and a minimum learning rate of 10^{-4}. The loss function includes L_{XRD} and $L_{coordinates}$, which is defined as $L = \lambda_1 \times L_{XRD} + \lambda_2 \times L_{coordinates}$, where λ_1= 1000 and λ_2 = 1. The L_{XRD} is given a much higher weight, emphasizing the dominant role of XRD data in guiding the crystal structure solution. For the denoising process, we discretized the reverse diffusion process over the continuous time interval [0, 1] into T = 1000 steps. For each time step, we sample (X_{t-1}) given (X_t) using the EGNN.

Evaluating the crystal structure solutions. We used the E_{hull}, R_{cos}, and sRMS to evaluate the quality of resolved structures. (1) **Energy assessment**. We used DFT to compute the formation energy of all our crystal structure solutions. All DFT calculations were performed using the Vienna Ab-Initio Simulation Package (VASP) using the PBE exchange-correlation functional, and potentials supplied by VASP with the projector augmented-wave method. The parameters of DFT calculations, such as the plane-wave energy cutoff and *k*-points density, were consistent with the parameters used for the Materials Project[52]. The energy correction schemes for oxides and transition metals were implemented as the Materials Project setup. To measure the energetic stability of a candidate structure, we compute the convex hull of formation energies in the relevant phase space and compute the difference between the candidate structure's formation energy and the convex hull energy E_{hull} at the target composition. (2) **XRD match**. The simulated diffraction data were calculated by XRDCalculator module in *pymatgen*[22]. We evaluated the difference between the target and solved XRD diffraction data using R_{cos}. R_{cos} is based on cosine similarity, which measures the similarity of two vectors[8]. In this work, the two vectors are the original XRD pattern and reconstructed XRD patterns, respectively. The cosine similarity is calculated as the cosine of the angle between the vectors. The value ranges from -1 to 1, where -1 means completely dissimilar, 1 means completely similar, and 0 means no correlation. (3) **Structural match**. We used the get_rms_dist method in *pymatgen* package to calculate the sRMS[22], which is normalized by the average free length per atom. Furthermore, we used StructureMatcher in *pymatgen*[22] to evaluate the success rate by calculating the maximum root-mean-square displacement (MaxDist) between the structures. The fit function in StructureMatcher compares the normalized MaxDist against the specified site tolerance (stol). If the MaxDist is less than stol (stol = 0.5), the fit function returns True, indicating a successful match. We also calculated the RMSD of atom positions between the original and reconstructed crystal structures to quantitatively measure the structure similarity.

Solving details. For each PXRD pattern, we ran XRDSol 25 times, generating 25 crystal structures independently. First, we rank the 25 potential solutions using R_{cos} (top-25). The solution with the highest R_{cos} is chosen as the final crystal structure solution of the corresponding given XRD pattern. We observed a systematically improving performance for the XRDSol as the number of total independent runs increased (Fig. S32 and 33).

C. Data availability

The MP-20 dataset used for training the XRDSol model has been deposited to XRDSol repository: https://github.com/ai4mat-zhu/XRDSol/tree/main/data, and can be also found in the previous work. The Powder Diffraction File used in this work can be requested from the ICDD PDF-4+ database, Their Powder Diffraction File numbers and the solved structures have been deposited to https://github.com/ai4mat-zhu/XRDSol/tree/main/data. The high-energy structures used in this work can be requested from the ICSD database or the Materials Project, and their ICSD collection codes have been provided in the Supplementary Information.

D. Code availability

The full code has been deposited to XRDSol repository at the following link: https://github.com/ai4mat-zhu /XRDSol.

II. References

1 Bragg, W. H. X-Rays and Crystals. *Nature* **90**, 360-361, doi:10.1038/090360d0 (1912).

2 Bragg, W. L. & Bragg, W. H. The structure of some crystals as indicated by their diffraction of X-rays. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **89**, 248-277, doi:10.1098/rspa.1913.0083 (1913).

3 Bragg, W. H. & Bragg, W. L. The structure of the diamond. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **89**, 277-291, doi:10.1098/rspa.1913.0084 (1913).

4 Sun, W. *et al.* The thermodynamic scale of inorganic crystalline metastability. *Science Advances* **2**, e1600225, doi:10.1126/sciadv.1600225 (2016).

5 Gindhart, A., Blanton, T., Blanton, J. & Gates-Rector, S. The Power of Electron Diffraction Phase Analysis and Pattern Simulations Using the ICDD® Powder Diffraction File™ (PDF-4+). *Microscopy and Microanalysis* **24**, 1154-1155, doi:10.1017/S1431927618006256 (2018).

6 Meredig, B. & Wolverton, C. A hybrid computational–experimental approach for automated crystal structure solution. *Nature Materials* **12**, 123-127, doi:10.1038/nmat3490 (2013).

7 Ling, H., Montoya, J., Hung, L. & Aykol, M. Solving inorganic crystal structures from X-ray powder diffraction using a generative first-principles framework. *Computational Materials Science* **214**, 111687, doi:10.1016/j.commatsci.2022.111687 (2022).

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

9 Ozaki, Y. *et al.* Automated crystal structure analysis based on blackbox optimisation. *npj Computational Materials* **6**, doi:10.1038/s41524-020-0330-9 (2020).

10 Griesemer, S. D., Ward, L. & Wolverton, C. High-throughput crystal structure solution using prototypes. *Physical Review Materials* **5**, 105003, doi:10.1103/PhysRevMaterials.5.105003 (2021).

11 Müller, U., Ivlev, S., Schulz, S. & Wölper, C. Automated Crystal Structure Determination Has its Pitfalls: Correction to the Crystal Structures of Iodine Azide. *Angewandte Chemie International Edition* **60**, 17452-17454, doi:10.1002/anie.202105666 (2021).

12 Xie, T., Fu, X., Ganea, O. E., Barzilay, R. & Jaakkola, T. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. *arXiv pre-print server*, doi:arxiv:2110.06197 (2021).

13 Hoogeboom, E., Satorras, V. c. G., Vignac, C. & Welling, M. in *Proceedings of the 39th International Conference on Machine Learning* Vol. 162 8867-8887 (PMLR, 2022).

14 Jiao, R. *et al.* in *Thirty-seventh Conference on Neural Information Processing Systems* (NeurIPS, 2023).

15 Zeni, C. *et al.* Mattergen: a generative model for inorganic materials design. *arXiv preprint* doi:10.48550/arXiv.2312.03687 (2023).

16 Igashov, I. *et al.* Equivariant 3D-conditional diffusion model for molecular linker design. *Nature Machine Intelligence* **6**, 417-427, doi:10.1038/s42256-024-00815-9 (2024).

17 Guo, G. *et al.* Ab Initio Structure Solutions from Nanocrystalline Powder Diffraction Data. *arXiv preprint arXiv:2406.10796* (2024).

18 Favre-Nicolin, V. & Cerny, R. FOX, `free objects for crystallography': a modular approach to ab initio structure determination from powder diffraction. *Journal of Applied Crystallography* **35**, 734-743, doi:10.1107/S0021889802015236 (2002).

19 Altomare, A. *et al.* Advances in powder diffraction pattern indexing: N-TREOR09. *Journal of Applied Crystallography* **42**, 768-775, doi:10.1107/S0021889809025503 (2009).

20 Coelho, A. A. TOPAS and TOPAS-Academic: an optimization program integrating computer algebra and crystallographic objects written in C++. *Journal of Applied Crystallography* **51**, 210-218, doi:10.1107/S1600576718000183 (2018).

21 Gao, P., Tong, Q., Lv, J., Wang, Y. & Ma, Y. X-ray diffraction data-assisted structure searches. *Computer Physics Communications* **213**, 40-45, doi:10.1016/j.cpc.2016.11.007 (2017).

22 Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314-319, doi:10.1016/j.commatsci.2012.10.028 (2013).

23 Zhao, K. *et al.* Realization of the kagome spin ice state in a frustrated intermetallic compound. *Science* **367**, 1218-1223, doi:10.1126/science.aaw1666 (2020).

24 Baran, S. *et al.* Magnetic order in RAgGe (R=Gd–Er) intermetallic compounds. *Journal of Alloys and Compounds* **281**, 92-98, doi:10.1016/S0925-8388(98)00721-X (1998).

25 Hoppe, R. & Blinne, K. Erdalkalihexafluoroplumbate(IV). *Zeitschrift für anorganische und allgemeine Chemie* **293**, 251-263, doi:10.1002/zaac.19582930504 (1958).

26 Bandemehr, J., Deubner, H. L., Sachs, M. & Kraus, F. Li_2PbF_6 and $SrPbF_6$ Revisited. *Zeitschrift für anorganische und allgemeine Chemie* **644**, 1721-1726, doi:10.1002/zaac.201800299 (2018).

27 Graf, T., Felser, C. & Parkin, S. S. P. Simple rules for the understanding of Heusler compounds. *Progress in Solid State Chemistry* **39**, 1-50, doi:10.1016/j.progsolidstchem.2011.02.001 (2011).

28 Gu, G. H., Jang, J., Noh, J., Walsh, A. & Jung, Y. Perovskite synthesizability using graph neural networks. *npj Computational Materials* **8**, 71, doi:10.1038/s41524-022-00757-z (2022).

29 Zachariasen, W. H. Untersuchungen über die Kristallstruktur von Sesquioxyden und Verbindungen ABO_3. **42**, 797-797, doi:10.1002/ange.19290423015 (1928).

30 Zhang, M., Hu, C., Abudouwufu, T., Yang, Z. & Pan, S. Functional Materials Design via Structural Regulation Originated from Ions Introduction: A Study Case in Cesium Iodate System. *Chemistry of Materials* **30**, 1136-1145, doi:10.1021/acs.chemmater.7b05252 (2018).

31 Rundqvist, S. New Metal-rich Phosphides of Niobium, Tantalum and Tungsten. *Nature* **211**, 847-848, doi:10.1038/211847a0 (1966).

32 Willerström, J. O. Stacking disorder in NbP, TaP, NbAs and TaAs. *Journal of the Less Common Metals* **99**, 273-283, doi:10.1016/0022-5088(84)90225-X (1984).

33 Xu, J. *et al.* Crystal Structure, Electrical Transport, and Magnetic Properties of Niobium Monophosphide. *Inorganic Chemistry* **35**, 845-849, doi:10.1021/ic950826f (1996).

34 Natta, G. & Casazza, E. Crystal and atomic structure of ferrohydrate. *Atti della Accademia Nazionale dei Lincei, Classe di Scienze Fisiche, Matematiche e Naturali, Rendiconti, Serie 7* (1927).

35 Hartmann, H., Fröhlich, H. J. & Ebert, F. Über ein neues Pernitrid des Strontiums und Calciums und über die Imide der Erdalkalimetalle. *Zeitschrift für anorganische und allgemeine Chemie* **218**, 181-189, doi:10.1002/zaac.19342180209 (1934).

36 Mullica, D. F., Milligan, W. O. & Beall, G. W. Crystal structures of $Pr(OH)_3$, $Eu(OH)_3$ and $Tm(OH)_3$. *Journal of Inorganic and Nuclear Chemistry* **41**, 525-532, doi:10.1016/0022-1902(79)80438-8 (1979).

37 Lutz, H. D., Möller, H. & Schmidt, M. Lattice vibration spectra. Part LXXXII. Brucite-type hydroxides $M(OH)_2$ (M = Ca, Mn, Co, Fe, Cd) — IR and Raman spectra, neutron diffraction of $Fe(OH)_2$. *Journal of Molecular Structure* **328**, 121-132, doi:10.1016/0022-2860(94)08355-X (1994).

38 Beall, G. W., Milligan, W. O. & Wolcott, H. A. Structural trends in the lanthanide trihydroxides. *Journal of Inorganic and Nuclear Chemistry* **39**, 65-70, doi:10.1016/0022-1902(77)80434-X (1977).

39 Brese, N. E., O'Keeffe, M. & Von Dreele, R. B. Synthesis and crystal structure of SrD_2 and SrND and bond valence parameters for hydrides. *Journal of Solid State Chemistry* **88**, 571-576, doi:10.1016/0022-4596(90)90255-V (1990).

40 Toby, B. H. R factors in Rietveld analysis: How good is good enough? *Powder Diffraction* **21**, 67-70, doi:10.1154/1.2179804 (2006).

41 Swanson, H. E., McMurdie, H. F., Morris, M. C. & Evans, E. H. in *Standard X-ray Diffraction Powder Patterns: Section 8* (1970).

42 Siebeneichler, S., Dorn, K. V., Smetana, V., Valldor, M. & Mudring, A.-V. A soft chemistry approach to the synthesis of single crystalline and highly pure $(NH_4)CoF_3$ for optical and magnetic investigations. *The Journal of Chemical Physics* **153**, 104501, doi:10.1063/5.0023343 (2020).

43 Huang, S. & Li, G. Photoluminescence properties of Li_2SrGeO_4:RE^{3+} (RE=Ce/Tb/Dy) phosphors and enhanced luminescence through energy transfer between Ce^{3+} and Tb^{3+}/Dy^{3+}. *Optical Materials* **36**, 1555-1560, doi:10.1016/j.optmat.2014.04.024 (2014).

44 Sokolova, E., EGOROVTISMENKO, Y. K., Yamnova, N. & Simonov, M. The Crystal-structure of Olgite Na (Sr0.52Ba0.48)(Sr0.58Na0.42)(Na0.81Sr0. 19)[PO3. 40][P0.76O3. 88] *Kristallografiya* **29**, 1079-1083 (1984).

45 Sokolova, E. V., Hawthorne, F. C. & Khomyakov, A. P. REFINEMENT OF THE CRYSTAL STRUCTURE AND REVISION OF THE CHEMICAL FORMULA OF OLGITE: (Ba,Sr) $(Na,Sr,REE)_2Na[PO_4]_2$. *Canadian Mineralogist* **43**, 1521-1526, doi:10.2113/gscanmin.43.5.1521 (2005).

46 Post, J. E. & Heaney, P. J. Neutron and synchrotron X-ray diffraction study of the structures and dehydration behaviors of ramsdellite and “groutellite”. **89**, 969-975, doi:10.2138/am-2004-0706 (2004).

47 Dent Glasser, L. S. & Ingram, L. Refinement of the crystal structure of groutite - MnOOH. *Acta Crystallographica Section B* **24**, 1233-1236, doi:10.1107/S0567740868004036 (1968).

48 Seol, J. W., Kim, Y.-I., Pham, T. L. & Lee, J.-S. Syntheses and characterizations of complex perovskite oxynitrides (Ca, Sr, Ba)TaO_2N. *Journal of the Korean Ceramic Society* **57**, 432-439, doi:10.1007/s43207-020-00041-0 (2020).

49 Yoshii, K., Mizumaki, M., Nakamura, A. & Abe, H. Structure and magnetism of $Eu_{1-x}DyxTiO_3$. *Journal of Solid State Chemistry* **171**, 345-348, doi:10.1016/S0022-4596(02)00231-1 (2003).

50 Oka, D. *et al.* Strain engineering for anion arrangement in perovskite oxynitrides. *ACS Nano* **11**, 3860-3866, doi:10.1021/acsnano.7b00144 (2017).

51 Satorras, V. c. G., Hoogeboom, E. & Welling, M. in *Proceedings of the 38th International Conference on Machine Learning* Vol. 139 (eds Meila Marina & Zhang Tong) 9323-9332 (PMLR, 2021).

52 Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002, doi:10.1063/1.4812323 (2013).