Extracting Structural Motifs from Pair Distribution Function Data of Nanostructures using Explainable Machine Learning

Andy S. Anker
University of Copenhagen
andy@chem.ku.dk

Emil T. S. Kjær
University of Copenhagen

Mikkel Juelsholt
University of Oxford

Troels Lindahl Christiansen
University of Copenhagen

Susanne Linn Skjærvø
University of Copenhagen

Mads Ry Vogel Jørgensen
Aarhus University
Lund University

Innokenty Kantor
Lund University
Technical University of Denmark

Daniel Risskov Sørensen
Aarhus University
Lund University

Simon J. L. Billinge
Columbia University
Brookhaven National Laboratory

Raghavendra Selvan
University of Copenhagen

Kirsten M. Ø. Jensen
University of Copenhagen

Abstract

Characterization of material structure with X-ray or neutron scattering using e.g. Pair Distribution Function (PDF) analysis most often rely on refining a structure model against an experimental dataset. However, identifying a suitable model is often a bottleneck. Recently, new automated approaches have made it possible to test thousands of models for each dataset, but these methods are computationally expensive, and analysing the output, i.e., extracting structural information from the resulting fits in a meaningful way is challenging. Our Machine Learning based Motif Extractor (ML-MotEx) trains an ML algorithm on thousands of fits, and uses SHAP (SHapley Additive exPlanation) values to identify which model features are important for the fit quality. We use the method for 4 different chemical systems including disordered nanomaterials and clusters. ML-MotEx opens for a new type of modelling where each feature in a model is assigned an importance value for the fit quality based on explainable ML.[1]

1 Introduction

The development of advanced, functional materials builds on an understanding of the intricate relationship between material structure and properties, and over the past century, crystallographic methods using scattering and diffraction have thus been essential for materials science. Crystallography allows ab initio determination of crystal structures from diffraction data, and has provided us with the vast knowledge of crystal chemistry that is now used in design of functional materials. However, in the case of nanomaterials with limited long-range order, crystallographic methods are challenged, and *ab*

[1] A complete version of this manuscript is accepted at npj Computational Materials [https://doi.org/10.1038/s41524-022-00896-3].

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

initio structure determination, or structure solution, is not currently possible. Over the past decades, total scattering with Pair Distribution Function (PDF) analysis has become an essential tool for characterisation of nanomaterial structure.[1, 2] The PDF is the Fourier transform of normalized and corrected X-ray, neutron, or electron scattering intensities, and is a function in real space representing a histogram of interatomic distances in the sample. Compared to crystallographic methods relying on long-range order, PDF analysis can be applied for nanomaterials,[3-5] disordered[1, 6, 7] or amorphous materials.[3, 5, 8] However, structure solution from the PDF is not possible except in a very few simple cases,[9] using either the Reverse Monte Carlo method[10] or the LIGA algorithm.[11, 12] In the absence of broadly applicable *ab initio* nanostructure determination methods, it is therefore necessary to propose reasonable starting models and to then 'refine' the model parameters against the data using local minimization methods. The step of finding a starting model can be a major challenge and is thus a bottleneck in complex material characterization. In the case of PDF analysis of nanomaterials, such models are often guessed at by considering related bulk materials, however these are often not good starting models for very small clusters and nanoparticles, where significant structural changes may take place.[3, 5, 13, 14] A way of building plausible starting models is thus needed, where structure models can be built capturing local bonding topologies suggested by known chemistries.

Recently, automated methods such as 'structure mining' and 'cluster mining' have appeared in the literature to help overcome this challenge.[15-17] In a study of the structure of metallic nanoparticles, Banerjee et al. automatically generated thousands of discrete metal nanocluster structures and fitted PDFs from each of them to experimental data to identify the best model in an automated manner.[17] In a recent study of molybdenum oxide nanomaterials, a new approach were introduced, where a large number of MoO_x cluster structure models were automatically generated and compared their PDFs to experimental data in order to identify dominating structural motifs in the sample, i.e. arrangements of atoms that dominate the material structure on the local scale.[7] The authors hypothesised that the structural motifs present in amorphous molybdenum oxides can also be found in crystalline structures, and therefore used crystal structures of molybdenum oxides as starting models. From these models, they cut out thousands of different cluster structure models of different sizes to build a 'catalogue' of structure candidates. These models were all tested against the experimental PDFs to identify the best fitting structural motif. In another study, a similar approach were used for identification of a bismuth oxido cluster intermediate structure in a study of cluster growth.[18]

While these approaches can extend the structural space searched when identifying models for structure refinement, new challenges arise. Firstly, the refinement processes can be computationally heavy, which can limit the number of catalogue structures that are tested. For example, our brute force approach for cluster identification above generates $2^N - 1$ structures for starting model sizes with N atoms. Each structure must have its PDF computed and then refined against the target measured PDF, so that its fit quality can be evaluated. This process is computationally costly and does not scale well with number of structure candidates. Furthermore, for disordered, amorphous, and nanostructured systems many hundred models may provide similar fit qualities, and if only reporting a few of them, it is difficult to assess which structural features of these models are important. We therefore need effective and unbiased methods to compare many fits to extract structural information. Here, we introduce a completely new approach that uses an explainable Machine Learning (ML) model that, after training, will predict the agreement factor for a test cluster with a given dataset. Furthermore, the use of explainable ML informs which features in the model are important for the agreement factor.[19-24] Our Machine Learning based Motif Extractor (ML-MotEx) model is illustrated in Figure 1. Firstly, it builds a large catalogue of thousands of candidate structural motifs, which are 'cut outs' from a chosen bulk structure[7, 18] (step 1). The PDF is then computed from each one, and each model is fit to the target dataset (step 2). The structures and R_{wp} values (explained in the Methods section) from each fit are handed to an ML algorithm applying gradient boosting decision trees (GBDTs),[25] which learns to predict R_{wp} values for new fits based on an atomic structure model (step 3). The ML-MotEx algorithm then outputs quantified values of how important each atom or feature in the starting structure is for the fit to yield a low R_{wp} value with the given fitting-algorithm (step 4). This is done by using SHAP (Shapley Additive exPlanation)[26, 27] values, which is a known method for explaining tree-based ML models. The amplitude of the SHAP value reflects how important a structural feature is for the fit quality, while the sign of the SHAP value reflects whether the feature affects the R_{wp} value of the fit towards 1 (poor fit) or 0 (perfect fit), in other words why it is important.

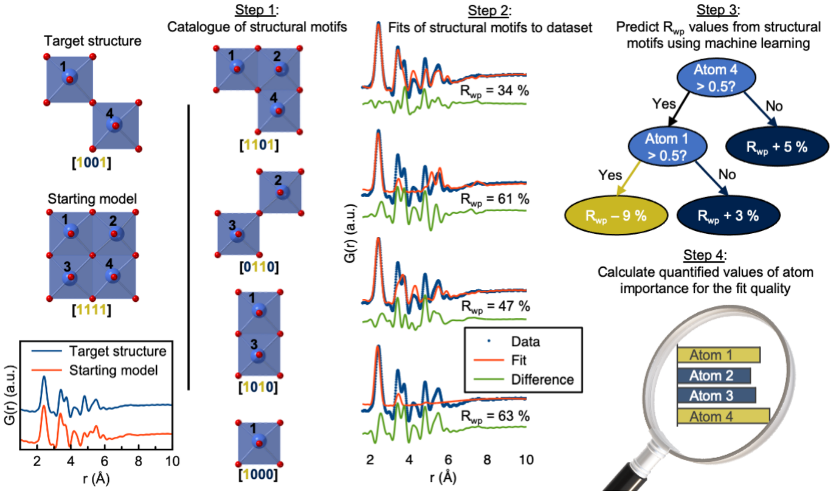

Figure 1: Illustration of the ML-MotEx process. Firstly, a starting model is provided. Using this starting model, a structure catalogue is generated, and the structures in the catalogue are fitted to the experimental data in question. An ML algorithm is then trained to predict R$_{wp}$ values and finally calculating quantified values of feature importance for the fit quality.

Compared to the automated, brute-force methods previously introduced for PDF analysis,[7, 15-17] we can much faster screen a larger number of structures. Our method only needs to screen a sub-sample ($\approx$ 10.000) of the much larger number of motifs that can be generated from a bulk material to learn how to predict which structures provide a good agreement with the data. The analysis done for the examples presented below would take $\approx$ 24 days for starting models with 24 atoms, $\approx 3 \cdot 10^6$ years for starting models with 48 atoms and $\approx 6 \cdot 10^{13}$ years for starting models with 72 atoms using a brute-force approach (section A in the SI), while ML-MotEx analysis is done in minutes or hours. Furthermore, the use of explainable ML provides a way to better analyse the output of the screening: instead of just identifying the model that provides the lowest R$_{wp}$ value, we are able to output a measure of how important each atom or feature (e.g. size or shape) in the starting model is for the fit to yield a low R$_{wp}$ value (step 4). This procedure is automated, can be done in quasi-real experimental time and without human bias.

We illustrate the use of ML-MotEx using 4 different examples. We first show the principles of the method using a simple model system based on simulated X-ray PDF data from a C_{60} buckyball. We further demonstrate the use of ML-MotEx on experimental X-ray PDF data from amorphous, disordered molybdenum oxides[7] and tungstate α-Keggin clusters in solution,[28] where it allows identifying the main structural motifs present in the samples using different starting models. Lastly, we extend the method to use a ‘cookie-cutter’ strategy to generate structures for the catalogue of candidate motifs. Here, the algorithm is used to identify a bismuth oxido cluster by using a cut-out of the β-Bi_2O_3 structure as starting model. The examples illustrate that it is possible to obtain knowledge of dominating structural motifs from PDF in an automated manner using ML.

2 Results

2.1 ML-MotEx algorithm

ML-MotEx consists of four steps. These four steps are shown in Figure 1. In the first step, a starting structure model is used to generate a catalogue of candidate structure motifs. As detailed in the Methods section, the structures are generated by removing different numbers of atoms from the original starting structure which results in thousands of smaller, candidate structure motifs. In the second step, a fitting script is used to fit the generated candidate structures to the dataset. In the third step, the fitting results are handed to the explainable ML algorithm which is optimised and trained. By using this information, SHAP values of the atoms or structural features in the starting model are calculated in the fourth step. The output of the algorithm is thus the starting model along with SHAP values, indicating the importance of each individual atom in the structure for the fit quality, or in other words; how much each individual atom or feature affects the R$_{wp}$ value either positively or negatively. We refer to this value as the "atom contribution value". We furthermore define the ratio between the atom contribution value and its uncertainty as the "confidence factor". Further definitions and descriptions of the individual steps of the algorithm are given in the Methods section.

2.2 Example 1: Proof-of-concept: Identification of the C_{60} buckyball

We first show the use of ML-MotEx with a simple, proof-of-concept example, using a calculated PDF from an ideal C_{60} buckyball (Figure 2A). The aim is to identify the structural motif, the C_{60} buckyball, from the data. We first need a starting structure that contains the motifs we are looking for. In this simplified example, we use a single unit cell of the crystal structure of C_{60}.[29] However, we discarded all symmetry and generated a discrete structure model corresponding to the 132 atoms in one unit cell. This model is shown in Figure 2B, where one whole C_{60} structure (Figure 2A) is seen along with fragments of the neighbouring C_{60} buckyballs. The simulated PDF of the C_{60} buckyball and the starting model are shown in Figure 2C. We can now use this starting model to generate a catalogue of structures, which are all fitted to the data. The structures are created by removing different numbers of atoms from the original starting structure, which results in thousands of smaller, candidate structure motifs. This model generation and fitting steps are identical to our previously reported brute-force approach, where we simply compare the R$_{wp}$ values of all the fits to identify the best structure motif. We first consider this simple approach. One of the limitations of the brute-force method is that the possible candidate structures is exponential in N, the number of atoms in the model. Since each atom in the starting model can be present or absent, the number of possible sub-clusters is equal to $2^N - 1$. For large models such as the C_{60} starting model containing 132 atoms, this is $\approx 10^{40}$, a gigantic number, making it impossible to investigate all candidate structures. For this example, we used 384,260 structures to train ML-MotEx, which is only a very small fraction of the $2^{132} - 1$ possible candidate structures. Note that the model with a single C_{60} buckyball was not in the generated structure catalogue. All these 384,260 structures were fitted to the PDF calculated from the C_{60} cluster. Only a scale factor, an isotropic expansion/contraction factor, and isotropic Atomic Displacement Parameters (ADPs) were refined, as detailed in the Methods section. We note that refinement of the atom positions can be added to the fitting procedure to expand the chemical space that is investigated. However, this would be computationally expensive and it would allow deviations from the chemical topologies set up in the starting model.

To get an overview of the results from these fits, we plot the R$_{wp}$ value versus the number of atoms in the structure, Figure 2D. To further investigate the results, one must visually inspect the fits of the catalogue of candidate structure motifs and their R$_{wp}$ value. Some of the candidate structure motifs are shown as inserts in Figure 2E, where transparent grey atoms represent atoms deleted from the models. The fits of these structures to the dataset are presented in Figure 2E, along with the R$_{wp}$ values. The R$_{wp}$ value appears to drop when the 'outer' atoms are removed, while it increases when the atoms that are part of the center C_{60} buckyball are removed. From investigating these few, but manually selected, structures and their corresponding fitted R$_{wp}$ value, one can hypothesize that the structure giving the best fit should be the C_{60} buckyball. However, this method can be biased by human interaction, and it is time-consuming and difficult to go through the many fits to extract structural information. We therefore move on to the ML-MotEx method. Using the catalogue of candidate structure motifs and the corresponding R$_{wp}$ values obtained above, we train a GBDT model on the training set to predict the R$_{wp}$ value of the candidate structure motifs. Figure 2F shows the predicted R$_{wp}$ values of the ML algorithm versus the R$_{wp}$ value of the structures when they are fitted to the simulated C_{60} dataset in DiffPy-CMI.[30] For the structures used in the test set, the GBDT model predicts the R$_{wp}$ value with a mean absolute error of 2.0 %. We now use explainable ML to explain R$_{wp}$ values with the use of the feature importance tool SHAP values.[27] As described in detail in the Methods section, a SHAP value is calculated for each structural feature (here each atom and the cluster size) for each candidate structure motif that is fitted to the PDF during the training process. The amplitude of the SHAP value reflects how important a structural feature is for the fit quality, while the sign of the SHAP value reflects whether the feature affects the R$_{wp}$ value of the fit towards 1 (poor fit) or 0 (perfect fit), in other words why it is important.

Figure 3A shows the most important results from the SHAP value analysis. The first feature we consider is the number of atoms, with SHAP values shown in the top part of Figure 3A. The plot represents SHAP values for the cluster size feature with the size shown on a colour scale, going from small (blue) to large clusters (red). From the large amplitude of some of the SHAP values observed

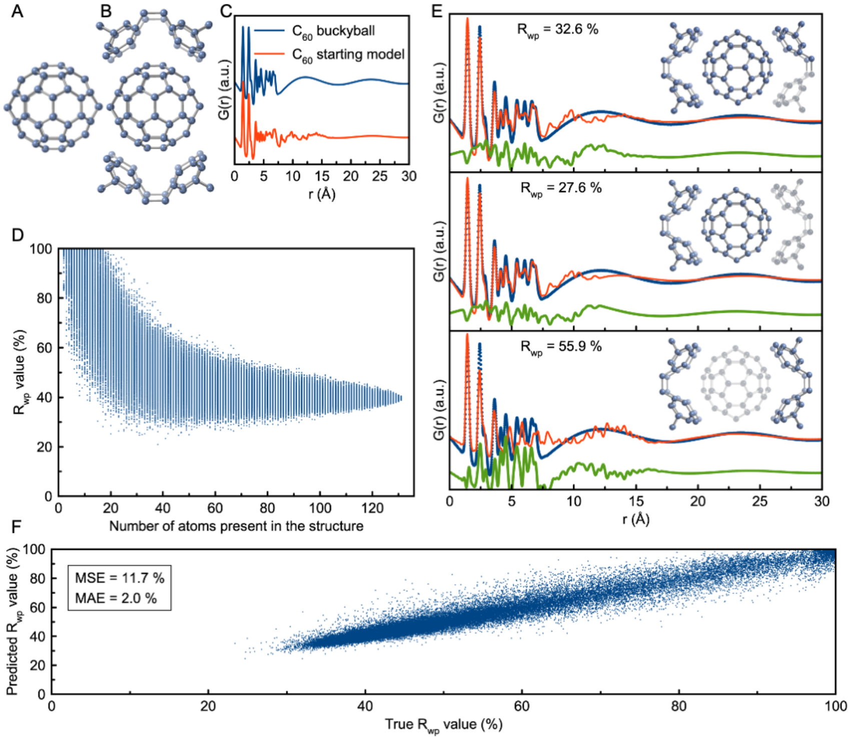

Figure 2: A) C_{60} buckyball, B) single C_{60} unit cell,[29] treated as a discrete structure with 132 atoms and C) their simulated PDFs. The simulation parameters (presented in section B in the SI) mimic typical values of a PDF dataset. D) R_{wp} values obtained in the fits using the C_{60} structure catalogue, plotted as a function of number of atoms in the structure motifs. Note that the model with a single C_{60} buckyball is not included in the set of 384,260 structures tested. This would result in a perfect fit with R_{wp} = 0 %. E) Examples of candidate structure motifs with their corresponding fits to the simulated C_{60} buckyball data. Grey, semitransparent atoms are removed from the starting model. F) Predicted R_{wp} values versus true R_{wp} values. R_{wp} values from the fits of the catalogue structures to the simulated C_{60} dataset, plotted versus the predicted R_{wp} values from the GBDT model from the same structures. The mean squared error (MSE) and the mean absolute error (MAE) are based on all 76,852 predictions in the test set, which are structures the model has not been trained on.

from this feature, we see that the number of atoms in the structure motif is the most important feature for the R_{wp} value. All small clusters (0–34 atoms, plotted in blue colours) show a large positive SHAP value, which implies that the R_{wp} value of the fit to the PDF data is high, i.e. the fit quality is low. All small clusters can thereby be discarded as structural models for satisfyingly describing the data. Next, we can investigate the SHAP values obtained for the individual atoms in the structure. We first consider atom 13, as labelled in the structure drawing in Figure 3B. The SHAP values obtained from this atom for each of the fits in the training set are all plotted on the SHAP axis. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. If first considering the cases where the atom is kept in the model, the atom 13 SHAP values are generally negative, which means that the presence of this atom pushes the R_{wp} value towards 0. We interpret this as ML-MotEx wants to keep the atom in the model. The SHAP values obtained for the fits without the atom present are positive, which confirms that if removing the atom, the fit quality becomes worse. Based on the SHAP values obtained for the atom in each fit, we calculate an atom contribution value. The atom contribution value is defined in the Methods section, and is calculated as the difference between the average SHAP values obtained for the atom when kept in the model, and when removed from the model. A negative atom contribution value means that the atom pushes the R_{wp} value down if kept in the structure. The atom contribution value obtained for atom 13 is negative, and we therefore colour it yellow in the structural representation in Figure 3B to indicate that it should be kept in the model. We use this strategy to automatically go through all the atoms in the starting model and colour them yellow/black

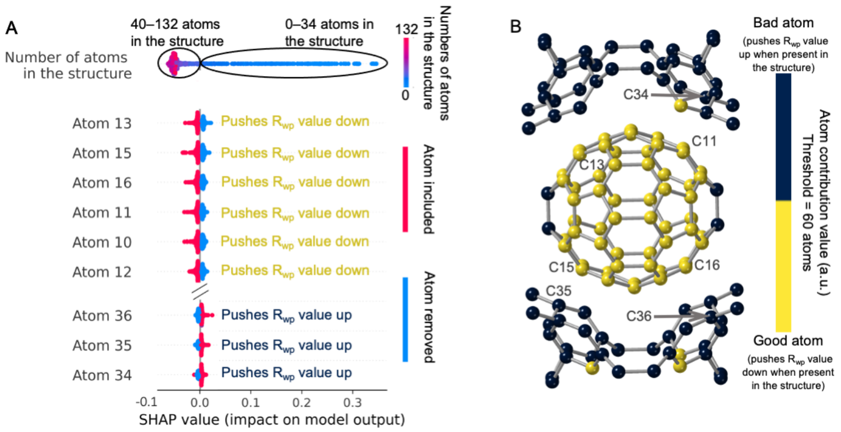

Figure 3: A) Plot of the SHAP values obtained in the C_{60} analysis, showing if atoms in the starting model are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. The SHAP values are plotted as a violin plot. B) Structural visualisation of kept and removed atoms. The atoms with the 60 lowest atoms contribution values have been coloured yellow, while the rest are coloured black. Section C in the Supporting Information (SI) shows a similar representation using a continuous colorbar for the atom contribution values.

based on their impact on the R_{wp} value. The result can be seen in Figure 3B where the 60 atoms with the lowest atom contribution values are coloured yellow. The results are also shown in section C in the SI, where the atom contribution values are plotted using a continuous colour bar. The results show that ML-MotEx mainly favours the atoms comprising the central buckyball. While the average confidence factor (as defined in the Methods section) is 1.26 for all of the atoms in the starting model, we observe that the average confidence factor of the mislabelled atoms is 0.37, meaning that ML-MotEx is less confident about the atom contribution values of those.

The ML-MotEx algorithm thus provides an unbiased method to extract important motifs from PDF data, without any inputs other than a starting model and a fitting script. We emphasize that the structural motifs extracted with ML-MotEx are based on the R_{wp} value of the fits and are thereby not necessarily physically reasonable. It is therefore important to still critically consider the extracted motif with chemical knowledge, in the same manner as for conventional PDF refinements. In this process, one could refine additional parameters such as atom positions. Consequently, in Figure 3B, the user should identify the full C_{60} buckyball as the structural motif rather than just choosing the motif of the yellow atoms. Another approach to avoid unphysically results from ML-MotEx would may be to include e.g., density function theory (DFT) calculations in the goodness-of-fit value.

2.3 Example 2: Identification of the ionic cluster structure from PDFs

To investigate the reproducibility of the ML-MotEx method, we investigate if similar results are achieved with different starting models, all containing the correct structure motif. We here model a PDF obtained from a solution of 0.05 M ammonium metatungstate hydrate, $(NH_4)_6[H_2W_{12}O_{40}]\cdot H_2O$ in water, which dissolves to form monodisperse α-Keggin clusters.[28] Experimental details are provided in section D in the SI. To test the ML-MotEx method we use four different starting models of tungstate oxide crystals, all including the α-Keggin cluster motif with varying complexity. Unit cells from the 4 following crystal structures were used as starting models: $[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py=pyridine) (1),[31] $(CH_3)_4N)_4SIW_{12}O_{40}$ (2),[32] $(((CH_3)_2NH_2)_6(Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$ [33] (3), and $(CH_3)_2NH_2)_3(PW_{12}O_{40})$ (4).[34] Again, we discarded all symmetry and generated discrete structure model corresponding to the atoms in one single unit cell. All other atoms than tungsten and oxygen were furthermore removed from the structures before catalogue structures were created. Figure 4A shows the experimental dataset with simulated PDFs from the 4 different starting models. Figure 4B illustrates a $W_{12}O_{40}$ α-Keggin structure.

Again, we first build structure catalogues based on the starting models (step 1) and fit them to the experimental PDF (step 2). In this case, we extract 104 structures from each starting model, which is just a small fraction of all possible structures that can be made from the starting models that have 24 (2), 48 (1) and (3), and 72 (4) atoms that are permuted. Again, a GBDT model was trained to predict the R_{wp} values of the structures (step 3), and SHAP values were obtained to calculate atom contribution values (step 4). The resulting SHAP value plots can be seen in section D in the SI. While ML-MotEx takes about 100 seconds on an AMD Ryzen Threadripper 3990X with 64-core 2.9/4.3GHz using 104 fits on a structure with 48 atoms, it would take about $\approx 3 \cdot 10^6$ years (section A in the SI) to make fits of all the $2^{48} - 1$ possible structures using the brute-force approach. Table S1 in the SI shows the exact computer time of the fits on a MacBook Pro and a Threadripper, which clearly demonstrates the scalability of ML-MotEx.

Figure 4C-F shows the results of applying ML-MotEx to the 4 different starting models. For structures (1), (3), and (4), the 24 atoms most preferred by ML-MotEx were coloured yellow, while the rest were coloured black. For structure (2), 12 atoms were coloured yellow. In all 4 examples, the yellow atoms have a motif of a α-Keggin cluster, however, in Figure 4E–F, we see a few mislabelled atoms (2 of 24 atoms in the worst case). The mislabelled atoms are found in the starting models containing most atoms, i.e. with the highest permutation value N. To achieve a better prediction, we could have built larger catalogues of candidate structure motifs and thus performed more fits. We therefore conclude that the ML-MotEx method is not completely insensitive to the starting model, but that it yields very similar results for all the tested starting models if it contains similar motifs. Furthermore, the example shows that ML-MotEx can be used to investigate PDF data from clusters in solution, whose structure also is part of known crystal structures. As described in section E in the SI, we performed an identical analysis of a different dataset also obtained from a second solution of 0.05 M ammonium metatungstate hydrate. This analysis provided highly comparable results. This illustrates the reproducibility of the method. In section F in the SI, we discuss what happens if a poor staring model is used, and how one can identify if the starting model does contain the right motif using the confidence factor. In the SI, we describe two other examples where we have used ML-MotEx. Firstly, we have used the ML-MotEx method to identify the main structural motifs present in an amorphous, disordered molybdenum oxide[7] from its experimental X-ray PDF. This example is described in Section G of the SI. Secondly, we have identified a larger ionic cluster, namely [$Bi_{38}O_{45}$], from an experimental PDF. Here, we use the β-Bi_2O_3 structure as starting model, and used a 'cookie-cutter' strategy to generate structures for the motif catalogue. This example, and the 'cookie-cutter' approach, are described further in Section I of the SI.

3 Discussion

In the 4 examples presented above, we have shown how explainable ML can aid in identifying structural motifs in nanostructured materials and presented a new approach to structure characterization. Traditional PDF analysis investigates how an entire structure model agrees with an experimental PDF, rather than identifying how different features in the model affect the fit quality. Instead, ML-MotEx provides a quantitative measure of how each atom or feature contributes to the fit. The use of ML furthermore allows screening of a large number of models in an automated and fast manner. In the examples described here, ML-MotEx has been used with various starting models with up to 256 metal atoms, however, the algorithm can handle larger systems, as it is highly scalable. In comparison, a full brute-force approach is computationally restricted to systems with up to 15–30 atoms. For the type of systems described here, it is possible to use the method in quasi-experimental time which could, for example, be useful for analysis of time-resolved scattering data, where the structural motifs present might change with time, which would be revealed by changing SHAP values.

ML-MotEx shares some similarities with the cluster build-up algorithm LIGA,[11, 12] which automatically builds clusters of different sizes based on information that is contained in inter-atomic distance lists extracted from the PDF. LIGA has shown to be successful at automatically reconstructing clusters (up to 150 atoms) with no user input except the interatomic distance list, extracted from an experimental PDF, and at low computational cost. However, its use has not caught on because extracting the distance list from the data presents significant practical difficulties, and is not unique. As with ML-MotEx it uses the error each atom in a cluster contributes to the fit to weight the decision about which atom to include in the model. Presumably, part of the success of LIGA and ML-MotEx is its use of this atom contribution for rapidly finding good candidate motifs. Unlike LIGA, ML-MotEx

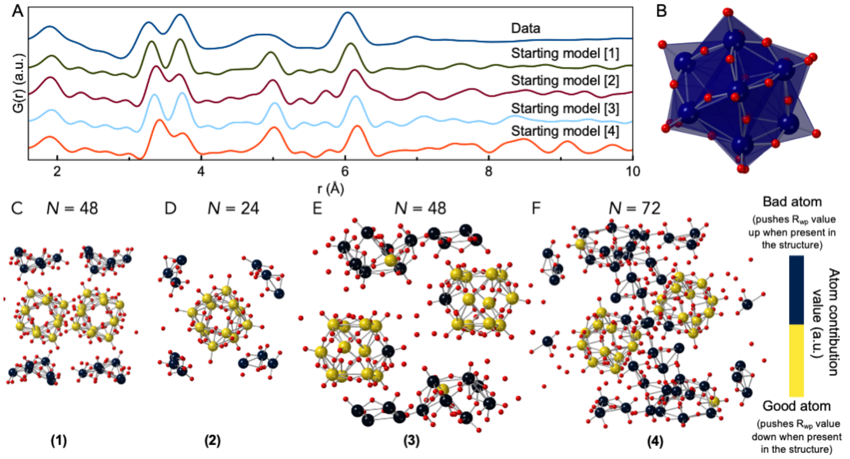

Figure 4: A) Comparison of experimental data from a 0.05 M ammonium metatungstate hydrate solution, and simulated PDFs from the four different starting models (1)–(4). The simulation parameters mimic typical values of a PDF dataset and can be seen in section B in the SI. B) The $W_{12}O_{40}$ α-Keggin structure. C-F) Results from ML-MotEx on a PDF from a solution of ammonium metatungstate hydrate, using four different starting models: C) $[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py=pyridine),[31] D) $(CH_3)_4N)_4SIW_{12}O_{40}$,[32] E) $(((CH_3)_2NH_2)_6(Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$,[33] F) $(CH_3)_2NH_2)_3(PW_{12}O_{40})$.[34] Atoms kept by ML-MotEx are shown in yellow while removed atoms are shown in black. The kept atoms were chosen as the 24 atoms (1), 12 atoms (2), 24 atoms (3), and 24 atoms (4) with the lowest atom contribution values. In section D in the SI a similar representation is shown using a continuous colorbar for the atom contribution values.

requires a starting model that contains the target structural motif, and it leverages ML to rapidly compute the atom contributions. It can therefore be positioned between traditional refinement (where the complete starting model is needed) and LIGA (which is *ab initio*) as it finds structural motifs from within a larger model as a starting model for a subsequent refinement. However, it has the significant advantage over LIGA that it works directly on the measured PDF and does not require the inter-atomic distance list to be extracted from the PDF data and we expect it to be of great practical value. It may be considered as a significant drawback that ML-MotEx requires as an input a structure fragment that contains the target motif within it in order to work. We provide a confidence factor for the starting model but ML-MotEx still requires significant chemical/structural knowledge and intuition to be of use. We first note that such intuition is widespread in the chemistry community and is unlikely to be a significant drawback in practice. For example, ML-MotEx has recently been used to identify the structure of intermediates in the formation of transition metal tungstates from polyoxometalate ions using in situ PDF data, and for identifying stacking fault domain sizes in manganese oxides from PDF and PXRD.[36, 37] We also note that the method is sufficiently fast that it would be possible to combine it with structural screening applications such as structureMining@PDFitc.[15, 37] Given chemical information about elements that are present, structureMining searches structural databases for candidate structures. These are then refined to a target dataset and a rank ordered list returned to the user. If the PDF represents a signal from a short-range ordered structural motif, we could insert ML-MotEx between the database mining and refinement steps to search over sets of plausible structures to look for structural sub-motifs. It may be possible to first use structure mining to identify starting models, which could then be used for ML-MotEx analysis. The models could then be further evaluated using both the resulting R_{wp} values and confidence factor. The ML-MotEx method is currently limited to PDF analysis in the fitting procedure of the algorithm (step 2), however, the rest of ML-MotEx (step 1+3+4) is ready to use with data from other techniques. We are confident that a similar approach, taking advantage of explainable ML and SHAP values can be broadly useful for enhancing and developing how models for data analysis are identified and constructed.

4 Methods

4.1 Step 1: Creation of a catalogue of candidate structure motifs

The first step in ML-MotEx is to use a starting structure model to generate a catalogue of candidate structure motifs, which are all fitted to the data. The structures are generated by removing different numbers of atoms from the original starting structure resulting in thousands of smaller, candidate structure motifs. This process, which we refer to as 'structure permutation', is illustrated in Figure 1, step 1. Here, the starting model contains 4 metal atoms, which are each bonded to 6 oxygen atoms. Before candidate structure motifs are generated, we select which atom type should be included in the permutation process. For the project discussed here, this selection is based on the X-ray scattering power of the atoms (i.e., heavier atoms scatter X-rays strongly, while lighter ones do not), and we therefore choose to permute over the 4 metal atoms in the structure rather than oxygen atoms. The total number of atoms that are selected for permutation (here 4) is referred to as the permutation number, N. Note that we do not take symmetry into account in this process. The selected atoms are removed or kept in the model by randomly associating them with zeros and ones, where 0 means that we remove the atom and 1 means we keep it. This is repeated multiple times to generate a large catalogue of candidate structure motifs. The total number of possible motifs from the permutations is equal to 2^N-1, but only a small fraction of these needs to be produced for ML-MotEx to provide satisfactory results. In section J in the SI, we discuss how large a catalogue of candidate structure motifs ML-MotEx needs as training data to output reasonable results. This is likely to be highly system dependent and especially dependent on N and structure symmetry. For the examples presented in the paper, we use $\approx$ 140–3000 structure motifs per N. The atoms which were not chosen for permutation, in this case oxygen, are removed if they are not within a distance threshold from any other atom. The threshold is user-defined and can be set according to PDF peaks and/or chemically valid distances (i.e., bond lengths) for the expected compounds.

4.2 Step 2: Fitting the catalogue of candidate structure motifs to the data

We fit each of the candidate structures in the catalogue to the experimental PDF using the Python-based program DiffPy-CMI[30, 38-40]. We apply the Debye equation for calculation of scattering intensities and PDFs from the structures. The fitting strategies and parameters for all 4 examples are listed in section K in the SI including a description of the fit quality measure, R_{wp}.

4.3 Step 3: Predicting R_{wp} values using Gradient Boosting Decision Trees

GBDTs[25] are a tool that can do classification or regression using decision trees. In this work, we are using XGBoost[25] as the GBDT algorithm to do the regression task of predicting the fit quality (step 2) based on the structural input given as zeros or ones (step 1) and the number of atoms in the structure. Section L in the SI demonstrates how the structure can be given as a input to the GBDT model. The optimisation is done by making trees of 'yes' and 'no' questions on whether to keep an atom in the structure or not, based on the resulting R_{wp} value. A hypothetical example of a simple tree can be seen in Figure 1, step 3. When atom 4 is present in the structure, the GBDT model will predict a R_{wp} value which is 5 % lower than if atom 4 is not present in the structure. In the same way, it will predict an R_{wp} value which is 12 % lower if atom 1 is present in the structure. In the decision tree, the algorithm will therefore say 'yes' to keep both atoms 1 and 4 in the structure. In this project, the GBDT model predicts the R_{wp} value using a weighted average of 100 trees. The GBDT model performance is improved with a large amount of training data, which in this tool is provided by creating a larger catalogue of candidate structure motifs and fitting them to the data. The GBDT model is trained on 80 % of the data, which is referred to as the training set. XGBoost[25] were used with default parameters except for learning rate and max depth, which were optimised with the use of Bayesian optimization using 50 iterations and cross-validation split on 3.[41, 42] While this procedure automates the hyperparameter tuning, we demonstrate in section M in the SI that similar results are achieved across various hyperparameters. The last 20 % of the data is used to evaluate the performance of the algorithm and is referred to as test set.

4.4 Step 4: Quantifying the contribution of each atom using SHAP values

SHAP values are used to analyse the R$_{wp}$ values resulting from the process described above. For each fit (step 2), each atom in the starting model is assigned a SHAP value. The amplitude of the SHAP value reflects how important a structural feature is for the fit quality, while the sign of the SHAP value reflects whether the feature affects the R$_{wp}$ value of the fit towards 1 (poor fit) or 0 (perfect fit), in other words why it is important. Each atom in the starting model will thus get F number of SHAP values, where F corresponds to the number of fits made in step 2 of the algorithm. We divide the F number of SHAP values into two categories; firstly the ones where the atom was kept in the structure motif (kept atom SHAP value list) and secondly the ones where the atom was removed to create the structure motif (removed atom SHAP value list). From each of the two lists, an average SHAP value for the atoms can be calculated, defined as SHAP$_{average-kept}$ and SHAP$_{average-removed}$. We then define an atom contribution value, which is calculated as the difference between two average SHAP values, i.e. atom contribution value = SHAP$_{average-kept}$ - SHAP$_{average-removed}$. We also define the uncertainty on this value as: atom contribution value RMS = (SHAP$^2_{average-kept}$ − SHAP$^2_{average-removed})^2$. We define a confidence factor for each atom that describes how confident we can be about including/excluding that atom in a structural motif; Confidence factor = atom contribution value / atom contribution value RMS. ML-MotEx outputs a VESTA[43] and CrystalMaker[44] file where all the atoms are coloured with regard to their atom contribution value.

5 Data and Code Availability

The authors declare that the data and code supporting this study are available within the paper, its Supplementary Information files and the associated Github to the paper: https://github.com/AndySAnker/ML-MotEx. Additional data that support the findings of this study are available from the corresponding author upon request.

Acknowledgments and Disclosure of Funding

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 804066). We are grateful to the Villum Foundation for financial support through a Villum Young Investigator grant (VKR00015416). Funding from the Danish Ministry of Higher Education and Science through the SMART Lighthouse is gratefully acknowledged We furthermore thank DANSCATT (supported by the Danish Agency for Science and Higher Education) for support. A. S. A and M. J. acknowledge the Siemens Foundation for support for their thesis projects. We acknowledge DESY (Hamburg, Germany), a member of the Helmholtz Association HGF, for the provision of experimental facilities. Parts of this research were carried out at beamline P02.1 and P07 at Petra III and we would like to thank Martin Etter, Jozef Bednarcik and Ann-Christin Dippel for assistance in using the beamlines. We acknowledge SOLEIL for provision of synchrotron radiation facilities and we would like to thank Thomas Bizien for assistance in using beamline SWING. This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract No. DE-AC02-06CH11357. We acknowledge MAX IV Laboratory for time on Beamline DanMAX under Proposal 20200731. Research conducted at MAX IV is supported by the Swedish Research council under contract 2018-07152, the Swedish Governmental Agency for Innovation Systems under contract 2018-04969, and Formas under contract 2019-02496. DanMAX is funded by the NUFI grant no. 4059-00009B. S. J. L. B. was supported by the U.S. National Science Foundation through grant DMREF-1922234

6 References

1. Billinge, S.J.L. and M.G. Kanatzidis, Beyond crystallography: the study of disorder, nanocrystallinity and crystallographically challenged materials with pair distribution functions. Chemical Communications, 2004(7): p. 749-760.

2. Keen, D.A. and A.L. Goodwin, The crystallography of correlated disorder. Nature, 2015. 521(7552): p. 303-309.

3. Christiansen, T.L., S.R. Cooper, and K.M.Ø. Jensen, There's no place like real-space: elucidating size-dependent atomic structure of nanomaterials using pair distribution function analysis. Nanoscale Advances, 2020. 2(6): p. 2234-2254.

4. Billinge, S.J.L. and I. Levin, The Problem with Determining Atomic Structure at the Nanoscale. Science, 2007. 316(5824): p. 561-565.

5. Juelsholt, M., et al., Size-induced amorphous structure in tungsten oxide nanoparticles. Nanoscale, 2021. 13(47): p. 20144-20156.

6. Yang, X., et al., Confirmation of disordered structure of ultrasmall CdSe nanoparticles from X-ray atomic pair distribution function analysis. Physical Chemistry Chemical Physics, 2013. 15(22): p. 8480-8486.

7. Christiansen, T.L., et al., Structure analysis of supported disordered molybdenum oxides using pair distribution function analysis and automated cluster modelling. Journal of Applied Crystallography, 2020. 53(1): p. 148-158.

8. Bennett, T.D. and A.K. Cheetham, Amorphous Metal–Organic Frameworks. Accounts of Chemical Research, 2014. 47(5): p. 1555-1562.

9. Kjær, E.S.T., et al., DeepStruc: Towards structure solution from pair distribution function data using deep generative models. ChemRxiv, 2022. doi:10.26434/chemrxiv-2022-0zrdl.

10. Cliffe, M.J., et al., Structure determination of disordered materials from diffraction data. Physical review letters, 2010. 104(12): p. 125501.

11. Juhás, P., et al., *ab initio* determination of solid-state nanostructure. Nature, 2006. 440(7084): p. 655-658.

12. Juhás, P., et al., The Liga algorithm for *ab initio* determination of nanostructure. Acta Crystallogr A, 2008. 64(Pt 6): p. 631-40.

13. Christiansen, T.L., et al., Size Induced Structural Changes in Molybdenum Oxide Nanoparticles. ACS Nano, 2019. 13(8): p. 8725-8735.

14. Aalling-Frederiksen, O., et al., Formation and growth mechanism for niobium oxide nanoparticles: atomistic insight from in situ X-ray total scattering. Nanoscale, 2021. 13(17): p. 8087-8097.

15. Yang, L., et al., Structure-mining: screening structure models by automated fitting to the atomic pair distribution function over large numbers of models. Acta Crystallographica Section A, 2020. 76(3): p. 395-409.

16. Anker, A.S., et al., Characterising the Atomic Structure of Mono-Metallic Nanoparticles from X-Ray Scattering Data Using Conditional Generative Models. 2020: Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG).

17. Banerjee, S., et al., Cluster-mining: an approach for determining core structures of metallic nanoparticles from atomic pair distribution function data. Acta Crystallographica Section A, 2020. 76(1): p. 24-31.

18. Anker, A.S., et al., Structural Changes during the Growth of Atomically Precise Metal Oxido Nanoclusters from Combined Pair Distribution Function and Small-Angle X-ray Scattering Analysis. Angewandte Chemie International Edition, 2021. 60: p. 2-12.

19. Butler, K.T., et al., Interpretable, calibrated neural networks for analysis and understanding of inelastic neutron scattering data. Journal of Physics: Condensed Matter, 2021. 33(19): p. 194006.

20. Suzuki, Y., et al., Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach. Scientific Reports, 2020. 10(1): p. 21790.

21. Torrisi, S.B., et al., Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. npj Computational Materials, 2020. 6(1): p. 109.

22. Oviedo, F., et al., Interpretable and Explainable Machine Learning for Materials Science and Chemistry. arXiv preprint arXiv:2111.01037, 2021.

23. Schmidt, J., et al., Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 2019. 5(1): p. 83.

24. Lee, K., et al., Phase classification of multi-principal element alloys via interpretable machine learning. npj Computational Materials, 2022. 8(1): p. 25.

25. Chen, T. and C. Guestrin, XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22^{nd} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 785–794.

26. Lundberg, S.M., et al., From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020. 2(1): p. 56-67.

27. Lundberg, S.M. and S.-I. Lee, A Unified Approach to Interpreting Model Predictions. Proceedings of the 31^{st} International Conference on Neural Information Processing Systems, 2017: p. 4765-4774.

28. Juelsholt, M., T. Lindahl Christiansen, and K.M.Ø. Jensen, Mechanisms for Tungsten Oxide Nanoparticle Formation in Solvothermal Synthesis: From Polyoxometalates to Crystalline Materials. The Journal of Physical Chemistry C, 2019. 123(8): p. 5110-5119.

29. Chen, X. and S. Yamanaka, Single-crystal X-ray structural refinement of the 'tetragonal' C_{60} polymer. Chemical Physics Letters, 2002. 360(5): p. 501-508.

30. Juhás, P., et al., Complex modeling: a strategy and software program for combining multiple information sources to solve ill posed structure and nanostructure inverse problems. Acta Crystallogr A Found Adv, 2015. 71(Pt 6): p. 562-568.

31. Niu, J., et al., Syntheses, spectroscopic characterization, thermal behavior, electrochemistry and crystal structures of two novel pyridine metatungstates. Journal of Coordination Chemistry, 2004. 57(11): p. 935-946.

32. Joachim, F., T. Axel, and P. Rosemarie, Strukturen und Schwingungsspektren des Tetramethylammonium-α-dodekawolframatosilikats und des Tetrabutylammonium-β-dodekawolframatosilikats: Structures and Vibrational Spectra of Tetramethylammonium α-Dodecatungstosilicate and Tetrabutylammonium β-Dodecatungstosilicate. Zeitschrift für Naturforschung B, 1981. 36(2): p. 161-171.

33. Niu, J.-Y., Q.-X. Han, and J.-P. Wang, A Novel Keggin Units-Supported Complex: Synthesis, Characterization and Crystal Structure of $[(CH_3)_2NH_2]_6[Cu(DMF)_4(GeW_{12}O_{40})_2]\cdot 2DMF$. Journal of Coordination Chemistry, 2003. 56(6): p. 523-530.

34. Busbongthong, S. and T. Ozeki, Structural Relationships among Methyl-, Dimethyl-, and Trimethylammonium Phosphododecatungstates. Bulletin of the Chemical Society of Japan, 2009. 82(11): p. 1393-1397.

35. Skjærvø, S.L., et al., Atomic structural changes in the formation of transition metal tungstates: The role of polyoxometalate structures in material crystallization. 2022.

36. Magnard, N., et al., Characterisation of intergrowth in metal oxide materials using structure-mining: the case of γ-MnO_2. 2022. 37. Yang, L., et al., A cloud platform for atomic pair distribution function analysis: PDFitc. Acta Crystallographica Section A, 2021. 77(1): p. 2-6.

38. Proffen, T. and R.B. Neder, DISCUS, a program for diffuse scattering and defect structure simulations – update. Journal of Applied Crystallography, 1999. 32(4): p. 838-839.

39. Proffen, T. and R.B. Neder, DISCUS: a program for diffuse scattering and defect-structure simulation. Journal of Applied Crystallography, 1997. 30(2): p. 171-175.

40. Coelho, A.A., TOPAS and TOPAS-Academic: an optimization program integrating computer algebra and crystallographic objects written in C++. Journal of Applied Crystallography, 2018. 51(1): p. 210-218.

41. Nogueira, F., Bayesian Optimization: Open source constrained global optimization tool for Python. 2014.

42. Putatunda, S. and K. Rama, A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. Proceedings of the 2018 International Conference on Signal Processing and Machine Learning, 2018: p. 6-10.

43. Momma, K. and F. Izumi, VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. Journal of Applied Crystallography, 2011. 44(6): p. 1272-1276.

44. Palmer, D.C., Visualization and analysis of crystal structures using CrystalMaker software. Zeitschrift für Kristallographie - Crystalline Materials, 2015. 230(9-10): p. 559-572.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Line 287 - 294.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We have not identified any potential negative societal impacts of this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] We do not include theoretical results
   (b) Did you include complete proofs of all theoretical results? We do not include theoretical results
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code, data and instructions needed to reproduce the results are uploaded to Github
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Not in a traditional way but we have reported the results using many different seeds, starting models and hyperparameters and shown that it nearly provide identical results.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] The assets we have used are standard open-source software as XGBoost, scikit learn ect. Our software is also open-source.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] Consent was obtained from all people whose data is used. They are all part of the author list.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We have not identified any offensive or identifiable information in our X-ray scattering data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable?
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable?
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?