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

Andy S. Anker[1], Emil T. S. Kjær[1], Mikkel Juelsholt[2], Troels Lindahl Christiansen[1], Susanne Linn Skjærvø[1], Mads Ry Vogel Jørgensen[3,4], Innokenty Kantor[4,5], Daniel Risskov Sørensen[3,4], Simon J. L. Billinge,[6,7] Raghavendra Selvan,[8,9] and Kirsten M. Ø. Jensen[1]*

*Correspondence to kirsten@chem.ku.dk (KMØJ)

1: Department of Chemistry and Nano-Science Center, University of Copenhagen, 2100 Copenhagen Ø, Denmark
2: Department of Materials, University of Oxford, Parks Road, Oxford, UK
3: Department of Chemistry & iNANO, Aarhus University, 8000 Aarhus C, Denmark
4: MAX IV Laboratory, Lund University, 224 84 Lund, Sweden
5: Department of Physics, The Technical University of Denmark, 2880 Lyngby, Denmark
6: Department of Applied Physics and Applied Mathematics Science, Columbia University, New York, NY 10027, USA
7: Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, NY 11973, USA
8: Department of Computer Science, University of Copenhagen, 2100 Copenhagen Ø, Denmark
9: Department of Neuroscience, University of Copenhagen, 2200, Copenhagen N

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 **M**achine **L**earning based **Mot**if **Ex**tractor (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.

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 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, we introduced a new approach, where we automatically generated a large number of MoO_x cluster structure models 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] We 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, we cut out thousands of different cluster structure models of different sizes to build a 'catalogue' of structure candidates. These models were all tested against our data to identify the best fitting structural motif. We recently used a similar approach 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 structure that are tested. For example, our brute force approach for cluster identification above generate 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 explainable ML inform which features in the model are important for the agreement factor.[19-24] Our **M**achine **L**earning based **Mot**if **Ex**tractor (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 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, as discussed in detail below.

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 (~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 ~24 days for starting models with 24 atoms, ~$3*10^6$ years for starting models with 48 atoms and ~$6*10^{13}$ 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 completely 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 in 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.

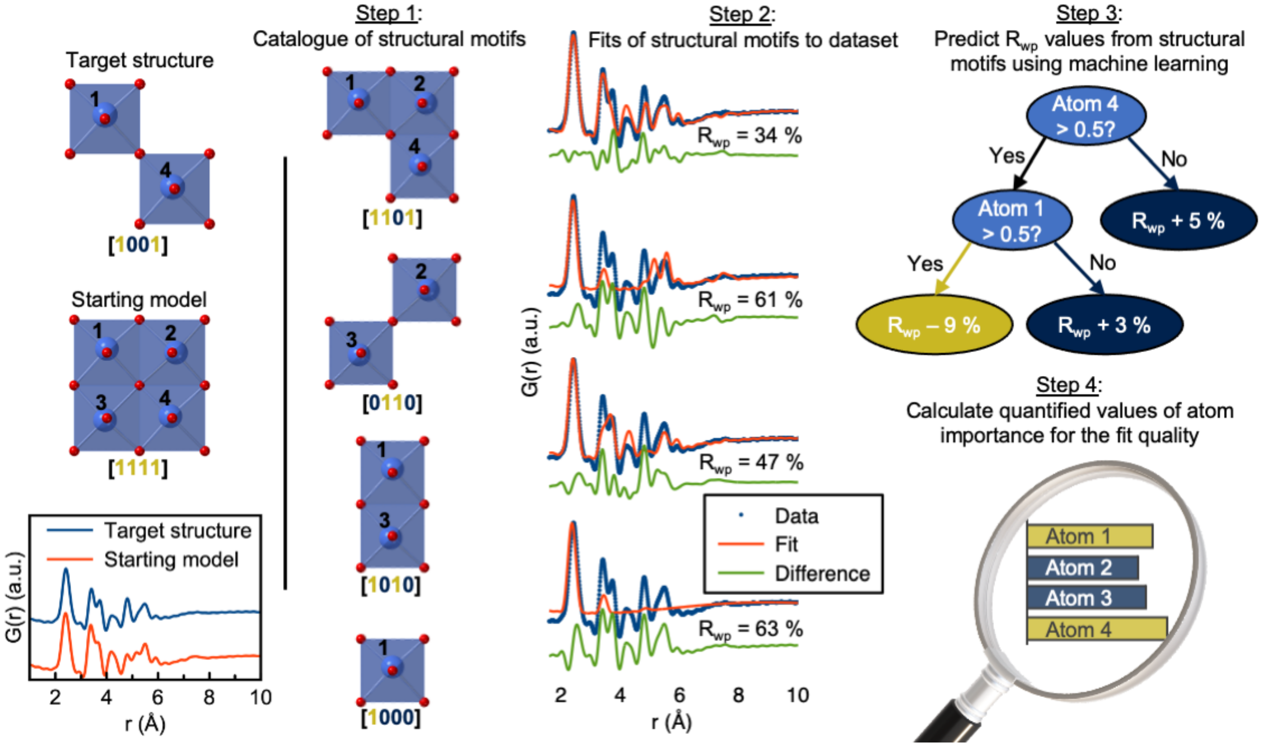
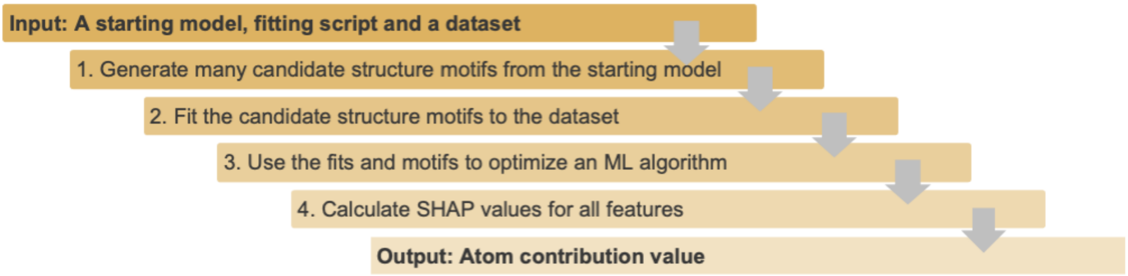


Figure 1: Illustration of the ML-MotEx process. Firstly, a starting model is provided. Using this starting model, a structure catalogue is generated, and the structures in the catalogue are fitted to the experimental data in question. An ML algorithm is then trained to predict R_{wp} values and finally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

Results

ML-MotEx algorithm

ML-MotEx consists of four steps, which are all fully automated. These four steps are shown in Figure 1 and the simplified pseudo-code of the algorithm in Figure 2. 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. A further definition and description of the individual steps of the algorithm is given in the Methods section.



Input: A starting model, fitting script and a dataset

1. Generate many candidate structure motifs from the starting model
2. Fit the candidate structure motifs to the dataset
3. Use the fits and motifs to optimize an ML algorithm
4. Calculate SHAP values for all features

Output: Atom contribution value

Figure 2: Pseudo-code describing the four steps of ML-MotEx. A starting model, fitting script, and dataset are given as input. Firstly, a catalogue of candidate structure motifs is generated (step 1), which are fitted to the dataset (step 2). The output from step 1 and 2 is then given to an ML algorithm, which learns to predict goodness-of-fit (R_{wp}) values based on the structure motif (step 3). Lastly, SHAP values are calculated for each feature (step 4) which can be converted to atom contribution values.

Example 1: Proof-of-concept: Identification of the C_{60} bucky ball

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 3A). The aim is to identify the structural motif, the C_{60} bucky ball, 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 3B, where one whole C_{60} structure (Figure 3A) 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 3C.

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 is 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 ~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} bucky ball 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) where refined, as detailed in section H of the SI. To get an overview of the results from these fits, we plot the R_{wp} value versus the number of atoms in the structure. 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 3E, where transparent grey atoms represent atoms deleted from the models. The fits of these structures to the dataset are presented in Figure 3E, 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.

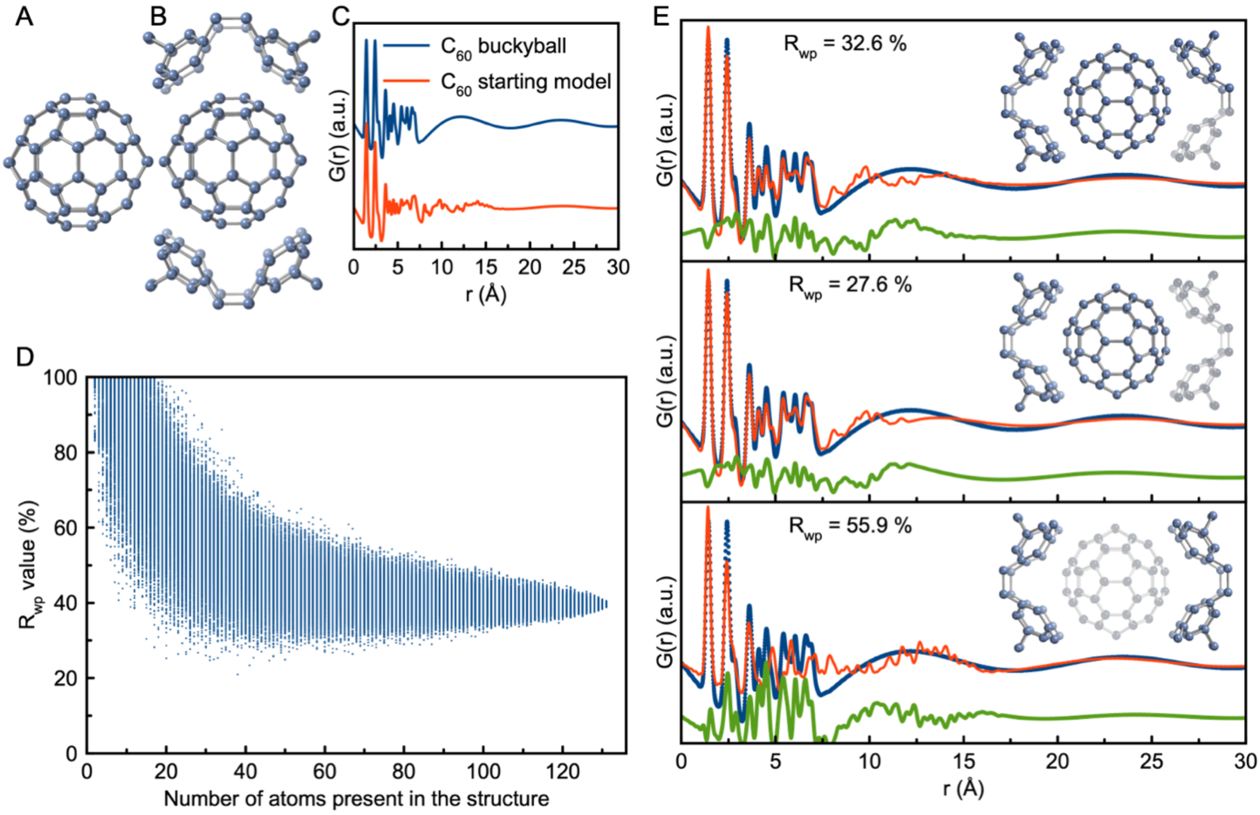


Figure 3: Analysis of a simulated PDF from a C_{60} buckyball. A) C_{60} buckyball, B) single C_{60} unit cell,[29] treated as a discrete structure with 132 atoms and C) their simulated PDFs. The simulation parameters mimic typical values of a PDF dataset and are presented in section B in the Supplementary Information. D) R_{wp} values obtained in the fits using the C_{60} structure catalogue, plotted as function of number of atoms in the structure motifs. Note that the model with a single C_{60} bucky ball is not included in the set of 384,260 structures tested. This would result in a perfect fit with R_w = 0 %. E) Examples of candidate structure motifs with their corresponding fits to the simulated C_{60} buckyball data. Grey, semitransparant atoms are removed from the starting model.

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 4 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 squared error of 2.0 %.

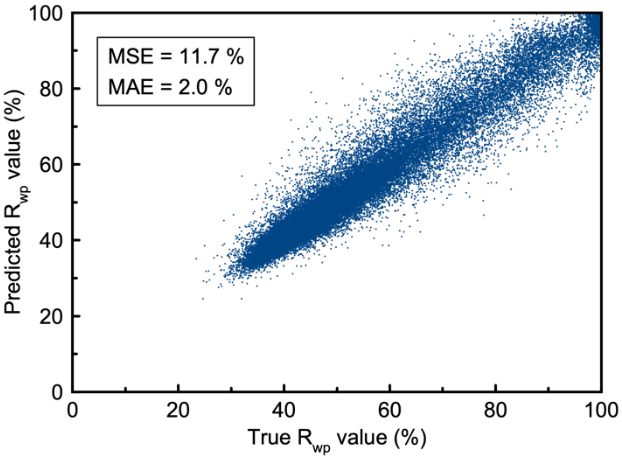


Figure 4: Predicted R_{wp} values versus true R_{wp} values. R_{wp} values from the fits of the catalogue structures to the simulated C_{60} dataset, plotted versus the predicted R_{wp} values from the GBDT model from the same structures. The mean squared error (MSE) and the mean absolute error (MAE) are based on all 76,852 predictions in the test set, which are structures the model has not been trained on.

We now use explainable ML to explain R_{wp} values with the use of the feature importance tool SHAP values.[27] As described in details in the Methods section, a SHAP value can 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 5A 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 5A. 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 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 5B. 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 in the atom. 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 5B 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 based on their impact on the R_{wp} value. The result can be seen in Figure 5B where the 60 atoms with the lowest atom contribution values are coloured yellow. The results are also shown in section C in the Supplementary Information, 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. 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.

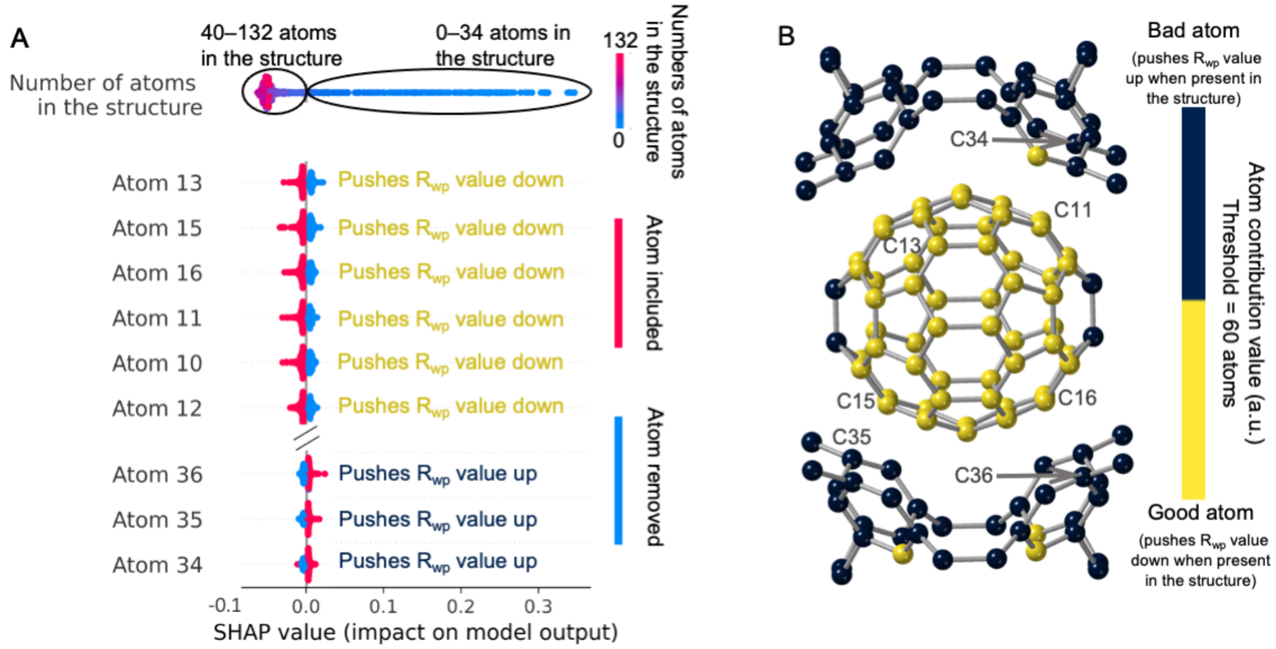


Figure 5: Summary of the ML-MotEx analysis of C_{60} PDF. A) Plot of the SHAP values obtained in the C_{60} analysis, showing if atoms in the starting model are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown 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 Supplementary Information shows a similar representation but where the atom contribution values are directly shown from a continuous colour bar.

Example 2: Identification of structural motifs in disordered molybdenum oxides

As discussed above, we have recently used the brute-force automated modelling method to identify structural motifs in disordered molybdenum oxides from PDF analysis.[7] Here we show that by reassessing the data with ML-MotEx, we can reproduce the results from Christiansen et al.[7] but in a fully automated way that allows analysis of the resulting structure model using SHAP values. Figure 6A shows the difference-PDF (d-PDF) obtained from amorphous molybdenum oxide supported on γ-Al_2O_3 nanoparticles (15 w% Mo), where the signal from the γ-Al_2O_3 nanoparticles has been subtracted. The d-PDF thus only reflects the structure of the supported material. The aim is to develop a structural model for the amorphous MoO_x. In our previous paper, different starting models were tested, which were all based on structures of molybdenum-based polyoxometalates (POMs) built from $[MoO_4]$ tetrahedra and $[MoO_6]$ octahedra. The analysis showed that the best fitting models did not contain tetrahedral motifs. Instead, the brute-force automated modelling approach hinted to a unit of three edge-sharing $[MoO_6]$ octahedra, or a 'triad', to be present in the structure. However, the use of the computationally expensive brute-force method limited the number of atoms that could be included in the starting model. This meant that a range of different smaller starting models were used to test different structure hypotheses. With ML-MotEx, we can instead test much larger systems and thereby include several different structural motifs at the same time in one starting model, as well as a quantitatively analyse the results using SHAP values. We therefore use a larger POM as starting model, namely the entire $Mo_{36}O_{128}$ cluster cut out of the $K_8(Mo_{36}O_{112}(H_2O)_{16})\cdot(H_2O)_{37}$ crystal structure.[31] Figure 6A shows the simulated data from the $Mo_{36}O_{128}$ cluster and Figure 6B shows the structure of the $Mo_{36}O_{128}$ cluster.

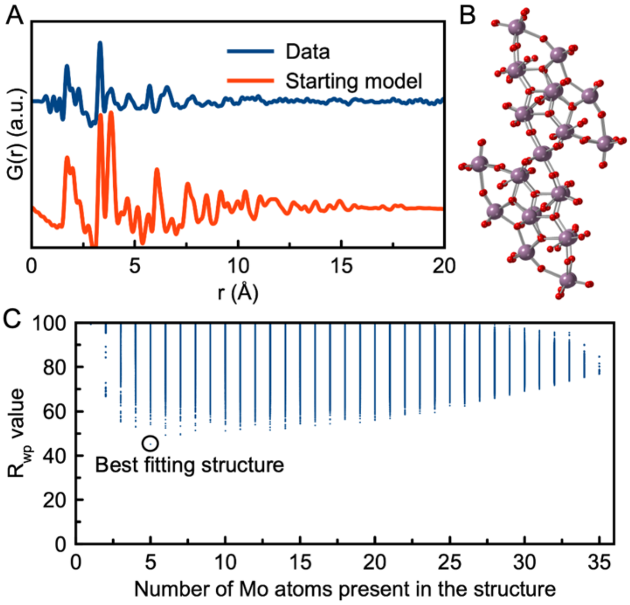


Figure 6: Analysis of experimental PDF from disordered molybdenum oxide. A) Comparison of experimental PDF from a disordered molybenum oxide,[7] and simulated data from $Mo_{36}O_{128}$ cluster, used as starting model. The simulation parameters mimic typical values of a PDF dataset and can be seen in section B in the Supplementary Information. B) Structure of the $Mo_{36}O_{128}$ cluster. C) R_{wp} values obtained in the fits using the $Mo_{36}O_{128}$ structure catalogue, plotted as function of number of atoms in the structure motifs.

We apply ML-MotEx to the molybdenum oxide system in the same manner as we did to the C_{60} buckyball. First, we used the starting model to make a catalogue of candidate structure motifs, as described in detail in the methods section. These are all fit to the experimental PDF, and the results are used to train the GBDT model. The fits are made with the same fitting algorithm as used in the paper from Christiansen et al.[7] Figure 6C illustrates the R_{wp} values of the fits, plotted as a function of the number of molybdenum atoms present in the structural motif. The best fitting models contain 5–7 molybdenum atoms. The model that fits the data with the lowest R_{wp} value (45 %) can be identified as a Mo_5O_{24} structure as shown in section D in the Supporting Information. However, it is difficult to justify that this structural model is unique representing the structure in the sample, purely based on the R_{wp} value.

We therefore use step 3 and 4 of ML-MotEx to analyse the results of the ensemble of fits. The resulting SHAP values are shown in Figure 7A. The plot should be interpreted in the same way as for the C_{60} example: Each atom is assigned a SHAP value in each of the fits in the training set. 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. When considering the amplitudes of the SHAP values, we see that the atoms labelled with 14, 15, 19, and 20 are marked as very important by ML-MotEx. When these atoms are present in the structure (red), they all have large negative SHAP values, indicating that their presence in the model pushes the R_{wp} down. When they are not present in the structure (blue), they all have large positive SHAP values, also indicating that they should be present in the structure to obtain a good fit. Atom 22 and 23 are examples of atoms that ML-MotEx do not suggest keeping in the structure. As seen from the SHAP values, its presence pushes up the R_{wp} value.

Based on the SHAP analysis, atom contribution values were calculated. The results are visually illustrated in Figure 7B, where the molybdenum atoms in the structure are coloured yellow if the atom contributed to a better fit quality, otherwise it is coloured black. Figure 7B clearly shows a specific motif that ML-MotEx wants to keep in the model. The yellow molybdenum atoms are all part of a 'triad' structure, where three $[MoO_6]$ octahedra share edges, and all oxygen atoms that bond to 3 or 4 Mo atoms are connected to yellow molybdenum atoms. This is further illustrated in section D in the Supplementary Information. Specifically, the resulting structural unit that ML-MotEx wants to keep is similar to heptamolybdate $[Mo_7O_{24}]^{6-}$, which can be described as several triads connected through edge-sharing. These results indicate that a motif of connected edge-sharing triads as shown in Figure 7C are important in order to describe the data of the disordered molybdenum oxides, which was also found by Christiansen et al.[7] We note here that when fitting this model to the PDF itself, we cannot describe the medium-range order present in the PDF. The ML-MotEx rather allows identifying the main local motifs in the data.

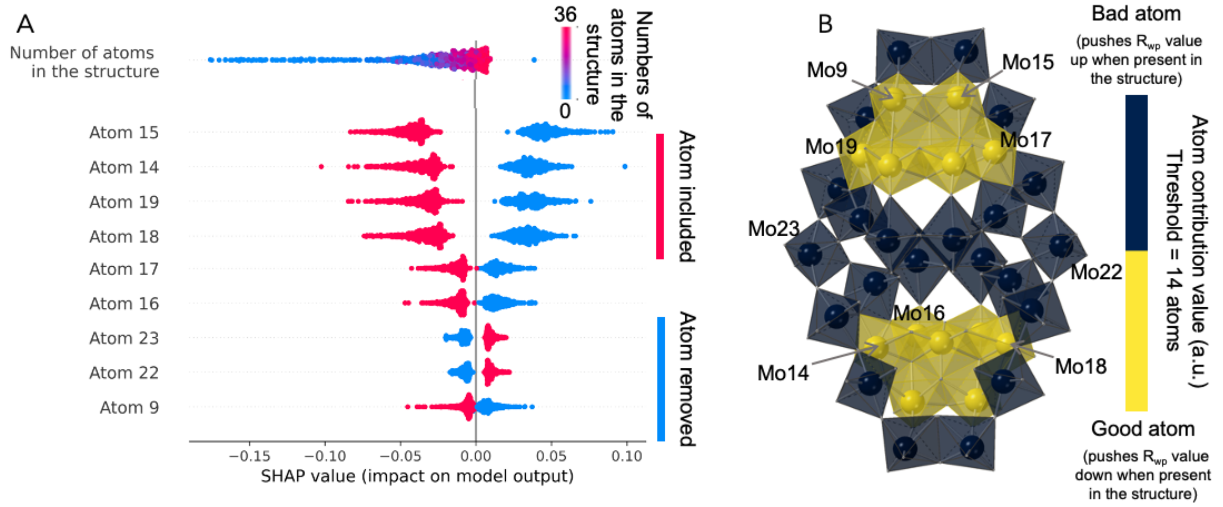


Figure 7: Summary of the ML-MotEx analysis of experimental PDF from disordered molybdenum oxide. A) Plot of the SHAP values obtained in the molybdenum oxide analysis, showing if atoms in the starting model $Mo_{36}O_{128}$ are favourable for the fit quality. For the models where the atom is not present 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 (yellow) and removed (black) atoms. Section D in the Supplementary Information shows a similar representation but where the atom contribution values are directly shown from a continuous colour bar.

Example 3: 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 E in the Supporting Information.

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]**,[32] $(CH_3)_4N)_4SiW_{12}O_{40}$ **[2]**,[33] $(((CH_3)_2NH_2)_6(Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$[34] **[3],** and $((CH_3)_2NH_2)_3(PW_{12}O_{40})$ **[4]**.[35] 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 8A shows the experimental dataset with simulated PDFs from the 4 different starting models. Figure 8B illustrates a $W_{12}O_{40}$ α-Keggin structure.

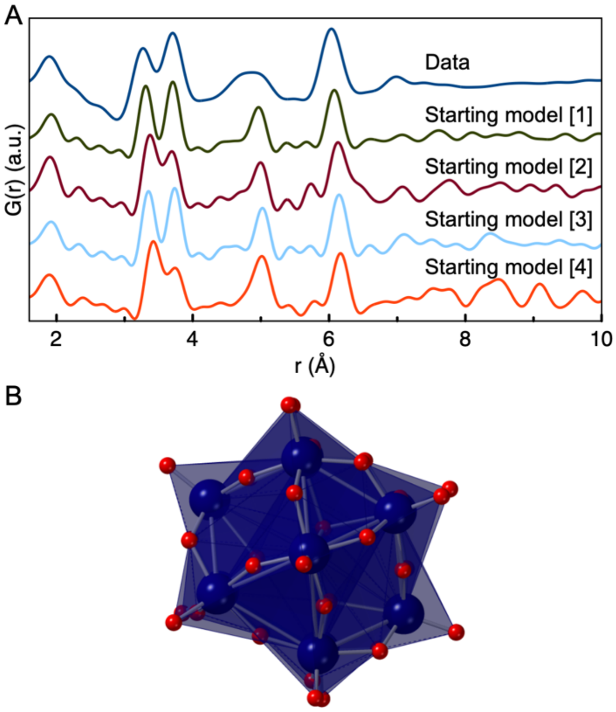


Figure 8: Experimental PDF from Keggin clusters in solution. A) Comparison of experimental data from a 0.05 M ammonium metatungstate hydrate solution, and simulated PDFs from the four different starting models **[1]**–**[4]**. The simulation parameters mimic typical values of a PDF dataset and can be seen in section B in the Supplementary Information. B) The $W_{12}O_{40}$ 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 10^4 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 E in the Supporting Information. While ML-MotEx takes about 100 seconds on an AMD Ryzen Threadripper 3990X with 64-core 2.9/4.3GHz using 10^4 fits on a structure with 48 atoms, it would take about ~$3*10^6$ years (section A in the Supplementary Information) to make fits of all the 2^{48}-1 possible structures using the brute-force approach. Section F in the Supplementary Information shows the exact computer time of the fits on a MacBook Pro and a Threadripper, which clearly demonstrates the scalability of ML-MotEx.

Figure 9 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 9C–D, 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 F in the Supporting Information, 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, as discussed in the Supporting Information. This illustrates reproducibility of the method.

We have also used the ML-MotEx method for a larger ionic cluster, namely [$Bi_{38}O_{45}$]. Here, we use the β-Bi_2O_3 structure as starting model, and used a 'cookier-cutter' strategy to generate structures for the motif catalogue. This example, and the 'cookie-cutter' approach, is described further in Section G of the SI.

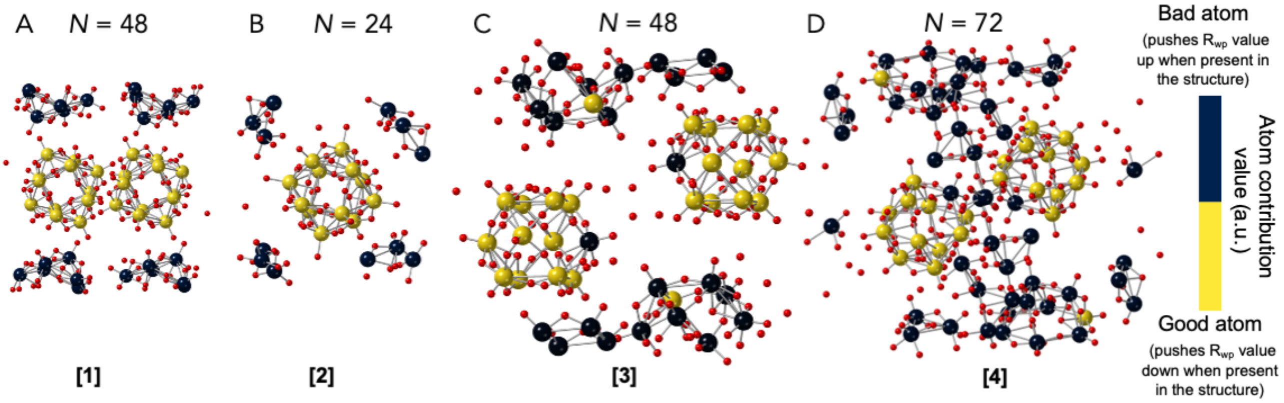

A N = 48
B N = 24
C N = 48
D N = 72
[1]
[2]
[3]
[4]
Bad atom
(pushes R_{wp} value up when present in the structure)
Atom contribution value (a.u.)
Good atom
(pushes R_{wp} value down when present in the structure)

Figure 9: Summary of the ML-MotEx analysis of experimental PDFs from Keggin clusters in solution. Results from the ML-MotEx method on a PDF from a solution of ammonium metatungstate hydrate, using four different starting models: A) $[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine),[32] B) $(CH_3)_4N)_4SiW_{12}O_{40,}$[33] C) $(((CH_3)_2NH_2)_6$ $(Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_{2,}$[34] D) $((CH_3)_2NH_2)_3(PW_{12}O_{40})$.[35] 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 (model A), 12 atoms (model B), 24 atoms (model C), and 24 atoms (model D) with the lowest atom contribution values. In section E in the Supplementary Information a similar representation is shown, but where the atom contribution values are directly shown using a continuous colorbar.

Discussion

In the four 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 quantitate 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 cluster (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 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. With this in mind we plan to deploy ML-MotEx as an application on the PDFitc.org web server.[36]

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. It therefore 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. However, 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.[15,36] 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. 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.

DATA AVAILABILITY

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

References

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

2 Keen, D. A. & Goodwin, A. L. The crystallography of correlated disorder. *Nature* **521**, 303-309 (2015).

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

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

5 Juelsholt, M., Anker, A. S., Christiansen, T. L., Jørgensen, M. R. V., Kantor, I., Sørensen, D. R. & Jensen, K. M. Ø. Size-induced amorphous structure in tungsten oxide nanoparticles. *Nanoscale* **13**, 20144-20156 (2021).

6 Yang, X., Masadeh, A. S., McBride, J. R., Božin, E. S., Rosenthal, S. J. & Billinge, S. J. L. Confirmation of disordered structure of ultrasmall CdSe nanoparticles from X-ray atomic pair distribution function analysis. *Phys. Chem. Chem. Phys.* **15**, 8480-8486 (2013).

7 Christiansen, T. L., Kjær, E. T. S., Kovyakh, A., Röderen, M. L., Høj, M., Vosch, T. & Jensen, K. M. Ø. Structure analysis of supported disordered molybdenum oxides using pair distribution function analysis and automated cluster modelling. *J. Appl. Crystallogr.* **53**, 148-158 (2020).

8 Bennett, T. D. & Cheetham, A. K. Amorphous Metal–Organic Frameworks. *Acc. Chem. Res.* **47**, 1555-1562 (2014).

9 Kjær, E. S. T., Anker, A. S., Weng, M. N., Billinge, S. J. L., Selvan, R. & Jensen, K. M. Ø. DeepStruc: Towards structure solution from pair distribution function data using deep generative models. *ChemRxiv* doi:10.26434/chemrxiv-2022-0zrdl (2022).

10 Cliffe, M. J., Dove, M. T., Drabold, D. & Goodwin, A. L. Structure determination of disordered materials from diffraction data. *Phys. Rev. Lett.* **104**, 125501 (2010).

11 Juhás, P., Cherba, D. M., Duxbury, P. M., Punch, W. F. & Billinge, S. J. L. Ab initio determination of solid-state nanostructure. *Nature* **440**, 655-658 (2006).

12 Juhás, P., Granlund, L., Duxbury, P. M., Punch, W. F. & Billinge, S. J. L. The Liga algorithm for ab initio determination of nanostructure. *Acta Cryst. A* **64**, 631-640 (2008).

13 Christiansen, T. L., Bøjesen, E. D., Juelsholt, M., Etheridge, J. & Jensen, K. M. Ø. Size Induced Structural Changes in Molybdenum Oxide Nanoparticles. *ACS Nano* **13**, 8725-8735 (2019).

14 Aalling-Frederiksen, O., Juelsholt, M., Anker, A. S. & Jensen, K. M. Ø. Formation and growth mechanism for niobium oxide nanoparticles: atomistic insight from in situ X-ray total scattering. *Nanoscale* **13**, 8087-8097 (2021).

15 Yang, L., Juhas, P., Terban, M. W., Tucker, M. G. & Billinge, S. J. L. Structure-mining: screening structure models by automated fitting to the atomic pair distribution function over large numbers of models. *Acta Crystallogr. A* **76**, 395-409 (2020).

16 Anker, A. S., Kjær, E. T. S., Dam, E. B., Billinge, S. J. L., Jensen, K. M. Ø. & Selvan, R. *Characterising the Atomic Structure of Mono-Metallic Nanoparticles from X-Ray Scattering Data Using Conditional Generative Models*. (Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG), 2020).

17 Banerjee, S., Liu, C.-H., Jensen, K. M. O., Juhas, P., Lee, J. D., Tofanelli, M., Ackerson, C. J., Murray, C. B. & Billinge, S. J. L. Cluster-mining: an approach for determining core structures of metallic nanoparticles from atomic pair distribution function data. *Acta Crystallogr. A* **76**, 24-31 (2020).

18 Anker, A. S., Christiansen, T. L., Weber, M., Schmiele, M., Brok, E., Kjær, E. T. S., Juhás, P., Thomas, R., Mehring, M. & Jensen, K. M. Ø. Structural Changes during the Growth of Atomically Precise Metal Oxido Nanoclusters from Combined Pair Distribution Function and Small-Angle X-ray Scattering Analysis. *Angew. Chem. Int. Ed.* **60**, 2-12 (2021).

19 Butler, K. T., Le, M. D., Thiyagalingam, J. & Perring, T. G. Interpretable, calibrated neural networks for analysis and understanding of inelastic neutron scattering data. *J. Phys.: Condens. Matter* **33**, 194006 (2021).

20 Suzuki, Y., Hino, H., Hawai, T., Saito, K., Kotsugi, M. & Ono, K. Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach. *Sci. Rep.* **10**, 21790 (2020).

21 Torrisi, S. B., Carbone, M. R., Rohr, B. A., Montoya, J. H., Ha, Y., Yano, J., Suram, S. K. & Hung, L. Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. *npj Comput. Mater.* **6**, 109 (2020).

22 Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. Interpretable and Explainable Machine Learning for Materials Science and Chemistry. *arXiv preprint arXiv:2111.01037* (2021).

23 Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 83 (2019).

24 Lee, K., Ayyasamy, M. V., Delsa, P., Hartnett, T. Q. & Balachandran, P. V. Phase classification of multi-principal element alloys via interpretable machine learning. *npj Comput. Mater.* **8**, 25 (2022).

25 Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).

26 Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. & Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56-67 (2020).

27 Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4765-4774 (2017).

28 Juelsholt, M., Lindahl Christiansen, T. & Jensen, K. M. Ø. Mechanisms for Tungsten Oxide Nanoparticle Formation in Solvothermal Synthesis: From Polyoxometalates to Crystalline Materials. *J. Phys. Chem. C* **123**, 5110-5119 (2019).

29 Chen, X. & Yamanaka, S. Single-crystal X-ray structural refinement of the `tetragonal' C_{60} polymer. *Chem. Phys. Lett.* **360**, 501-508 (2002).

30 Juhás, P., Farrow, C. L., Yang, X., Knox, K. R. & Billinge, S. J. L. Complex modeling: a strategy and software program for combining multiple information sources to solve ill posed structure and nanostructure inverse problems. *Acta Cryst. A* **71**, 562-568 (2015).

31 Krebs, B. & Paulat-Böschen, I. The structure of the potassium isopolymolybdate $K_8[Mo_{36}O_{12}(H_2O)_{16}]\cdot nH_2O$ (n = 36···40). *Acta Cryst.* **38**, 1710-1718 (1982).

32 Niu, J., Zhao, J., Wang, J. & Bo, Y. Syntheses, spectroscopic characterization, thermal behavior, electrochemistry and crystal structures of two novel pyridine metatungstates. *J. Coord. Chem.* **57**, 935-946 (2004).

33 Joachim, F., Axel, T. & Rosemarie, P. Strukturen und Schwingungsspektren des Tetramethylammonium-α-dodekawolframatosilikats und des Tetrabutylammonium-β-dodekawolframatosilikats: Structures and Vibrational Spectra of Tetramethylammonium α-Dodecatungstosilicate and Tetrabutylammonium β-Dodecatungstosilicate. *Z. Naturforsch.* **36**, 161-171 (1981).

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

35 Busbongthong, S. & Ozeki, T. Structural Relationships among Methyl-, Dimethyl-, and Trimethylammonium Phosphododecatungstates. *Bull. Chem. Soc. Jpn.* **82**, 1393-1397 (2009).

36 Yang, L., Culbertson, E. A., Thomas, N. K., Vuong, H. T., Kjaer, E. T. S., Jensen, K. M. O., Tucker, M. G. & Billinge, S. J. L. A cloud platform for atomic pair distribution function analysis: PDFitc. *Acta Cryst. A* **77**, 2-6 (2021).

37 Proffen, T. & Neder, R. B. DISCUS, a program for diffuse scattering and defect structure simulations – update. *J. Appl. Cryst.* **32**, 838-839 (1999).

38 Proffen, T. & Neder, R. B. DISCUS: a program for diffuse scattering and defect-structure simulation. *J. Appl. Cryst.* **30**, 171-175 (1997).

39 Coelho, A. A. TOPAS and TOPAS-Academic: an optimization program integrating computer algebra and crystallographic objects written in C++. *J. Appl. Crystallogr.* **51**, 210-218 (2018).

40 Putatunda, S. & Rama, K. 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*, 6-10 (2018).

41 Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **44**, 1272-1276 (2011).

42 Palmer, D. C. Visualization and analysis of crystal structures using CrystalMaker software. *Z. Kristallogr.* **230**, 559-572 (2015).

Methods

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 10. 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. We have not studied exactly 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 ~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.

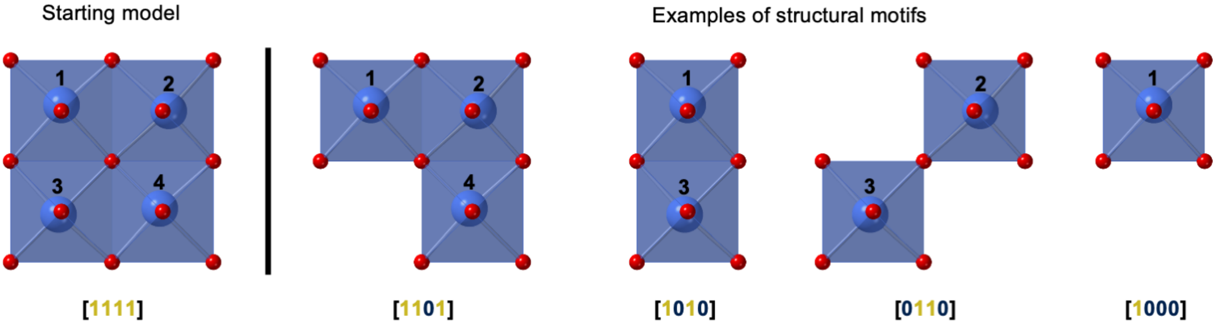


Figure 10: Example of how structure motifs can be extracted from a starting model with 4 metal atoms coordinated to oxygen. The metal atoms are permuted randomly by creating an array of zeros and ones, where 0 refers to a deleted atom and 1 refers to an atom that is kept in the structure. Oxygen atoms are removed if they do not bond to any metal atoms within a distance threshold that is set by the user. Note that the metal atoms (blue) are slightly distorted from the centre of the octahedra.

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

In the next step, we fit each of the candidate structures in the catalogue to the experimental PDF. We here use the Python-based program DiffPy-CMI[30] for PDF fitting[37-39] and apply the Debye equation for calculation of scattering intensities and PDFs from the structures. The fitting strategies and parameters for each of the examples presented below are listed in section H in the Supplementary Information. The output of the fit is a R_{wp} value reflecting the quality of the fit:

$$R_{wp} = \sqrt{\frac{\sum_{i=1}^{n}[G_{obs}(r_i) - G_{calc}(r_i, P)]^2}{\sum_{i=1}^{n} G_{obs}(r_i)^2}} \cdot 100\ \%$$

Here, G_{obs} and G_{calc} are the observed and calculated PDFs, and P is the refinement parameters in the model.

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

Gradient Boosting Decision Trees (GBDTs)[25] are a tool that can do classification or regression using decision trees. In this work, we are using 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).

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 atom 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. The GBDT parameters were chosen using Bayesian optimisation and a five-fold cross-validation meaning that 20 % of the training set were left out of the training process sequently to avoid overfitting.[40] The last 20 % of the data is used to evaluate the performance of the algorithm and is referred to as test set.

A further description of the GBDT method used for ML-MotEx are given in section I in the Supporting Information.

Step 4: Quantifying the effect of structural features using SHAP values, assigning atomic contribution 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 $SHAP_{average\text{-}kept}$ and $SHAP_{average\text{-}removed}$. We then define an *atom contribution value,* which is calculated as the difference between two average SHAP values, i.e. *atom contribution value* = $SHAP_{average\text{-}kept} - SHAP_{average\text{-}removed}$. The results can be visually inspected as the atoms in the starting model are coloured according to their atom contribution value using yellow for low atom contribution value (tendency to keep atom, pushing R_{wp} down) and black for high atom contribution value (tendency to remove atom, pushing R_{wp} up). Any 3D visualisation tools for structural models, which can visualise XYZ files, can be used such as VESTA[41] or CrystalMaker.[42]

Acknowledgements

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. 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

AUTHOR CONTRIBUTIONS

A.S.A., E.T.S.K., and K.M.Ø.J. conceptualized the project. A.S.A., E.T.S.K., M.J., T.L.C, S.L.S, S. B. J. L, and R.S. designed the methodology and A.S.A. and E.T.S.K. wrote the code. A.S.A, M.J., T.L.C, M.R.V.J, I.K. and D.R.S measured the data. K.M.Ø.J procured funding and supervised the project. All authors were involved with the writing of the paper.

COMPETING INTERESTS

The authors declare no competing interests.