JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

POMFinder: identifying polyoxometallate cluster structures from pair distribution function data using explainable machine learning

Andy S. Anker,[a] Emil T. S. Kjær,[a] Mikkel Juelsholt[b] and Kirsten M. Ø. Jensen[a]*

[a]Department of Chemistry and Nano-Science Center, University of Copenhagen, 2100 Copenhagen Ø, Denmark, and [b]Department of Materials, University of Oxford, Parks Road, Oxford, Oxfordshire OX1 3PH, United Kingdom. *Correspondence e-mail: kirsten@chem.ku.dk

Received 27 October 2023
Accepted 16 November 2023

Edited by G. J. McIntyre, Australian Nuclear Science and Technology Organisation, Lucas Heights, Australia

Keywords: computational modelling; machine learning; polyoxometallate clusters; *POMFinder*.

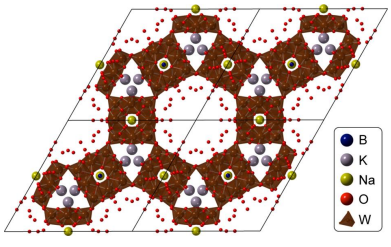
Supporting information: this article has supporting information at journals.iucr.org/j

Characterization of a material structure with pair distribution function (PDF) analysis typically involves refining a structure model against an experimental data set, but finding or constructing a suitable atomic model for PDF modelling can be an extremely labour-intensive task, requiring carefully browsing through large numbers of possible models. Presented here is *POMFinder*, a machine learning (ML) classifier that rapidly screens a database of structures, here polyoxometallate (POM) clusters, to identify candidate structures for PDF data modelling. The approach is shown to identify suitable POMs from experimental data, including *in situ* data collected with fast acquisition times. This automated approach has significant potential for identifying suitable models for structure refinement to extract quantitative structural parameters in materials chemistry research. *POMFinder* is open source and user friendly, making it accessible to those without prior ML knowledge. It is also demonstrated that *POMFinder* offers a promising modelling framework for combined modelling of multiple scattering techniques.

1. Introduction

The continued development of increasingly bright synchrotron and neutron facilities means that scattering and spectroscopy data can now be measured at impressive speeds (Wang *et al.*, 2018; Dong *et al.*, 2021; Pacchioni, 2019). Hundreds of gigabytes or even terabytes of data are now commonly collected in each experiment, each data set containing thousands or millions of individual measurements. With this amount of data, it is an enormous challenge to work through each data set manually, and the development of automated methods for data analysis is thus becoming more and more necessary (Dong *et al.*, 2021; Chen *et al.*, 2021; Choudhary *et al.*, 2022).

For many X-ray- and neutron-based scattering techniques such as small-angle scattering, powder diffraction and total scattering with pair distribution function (PDF) analysis, data analysis is often done through least-squares optimization (Pedersen, 1997; Rietveld, 1969; Chepkemboi *et al.*, 2022). Here, structure models found in *e.g.* structure databases are used to simulate data, which are then refined against experimental data. This allows the extraction of quantitative structural parameters. This approach can, in principle, be automated by *e.g.* testing entire databases of structures against experimental data sets (Banerjee *et al.*, 2020; Yang *et al.*, 2020; Aimi & Fujimoto, 2020; Christiansen *et al.*, 2020*b*). However, least-squares fitting algorithms are computationally expensive, which makes them unsuited for automatically identifying and



refining structures from experiments with many data sets (Wang *et al.*, 2018). Consequently, identifying structural models is currently a bottleneck for modelling large quantities of scattering data.

In this study, we present a tree-based machine learning (ML) classifier that identifies a chemical structure from a PDF in less than one second, enabling high-throughput database screening. The PDF here refers to the reduced pair distribution function $G(r)$, which represents a histogram of real-space interatomic distances and can be used to identify atomic arrangements in materials. $G(r)$ is obtained by Fourier transforming the total scattering structure function $S(Q)$, which is the set of corrected and normalized total scattering data (Egami & Billinge, 2012),

$$G(r) = (2/\pi) \int_{Q_{\min}}^{Q_{\max}} Q[S(Q) - 1] \sin(Qr)\, \mathrm{d}Q. \qquad (1)$$

Here, Q is the magnitude of the scattering vector [$Q = (4\pi/\lambda) \times \sin(\theta/2)$, where θ is the scattering angle and λ is the wavelength of the incident radiation], while r is the interatomic distance. The Q range used for modern total scattering experiments ranges from $Q_{\min}$ = 0.1–1 Å^{-1} to $Q_{\max}$ = 15–30 Å^{-1}.

In recent years PDF analysis has been shown to be a powerful technique for characterization of disordered materials (Christiansen *et al.*, 2020*b*; Yang *et al.*, 2013; Billinge & Kanatzidis, 2004; Keen & Goodwin, 2015), amorphous materials (Christiansen *et al.*, 2020*a*; Juelsholt *et al.*, 2021; Bennett & Cheetham, 2014), clusters in solution (Anker *et al.*, 2021; Jensen *et al.*, 2016; Szczerba *et al.*, 2021) and nanomaterials (Billinge & levin, 2007; Cooper *et al.*, 2020) where conventional crystallographic approaches are challenged (Billinge & Kanatzidis, 2004; Keen & Goodwin, 2015).

PDFs are usually analysed by fitting a reasonable starting model to the experimental PDF using dedicated software such as *PDFgui* (Farrow *et al.*, 2007), *DiffPy-CMI* (Juhás *et al.*, 2015), *DISCUS* (Proffen & Neder, 1997, 1999) or *TOPAS* (Coelho, 2018). In some cases, for example for well characterized crystalline materials, identifying a starting model for structural refinements is easily done. In other cases, finding or constructing a good initial atomic model for modelling the PDF can be an extremely labour-intensive task, requiring carefully browsing through large numbers of possible starting models. However, we and others have shown that ML methods such as neural networks and tree-based ML have much potential to improve the speed of PDF analysis (Anker *et al.*, 2020, 2022, 2023; Liu *et al.*, 2019; Kjær *et al.*, 2023; Kløve *et al.*, 2023; Skjaervø *et al.*, 2023; Magnard *et al.*, 2022). ML has, for example, been used to identify crystallographic space groups from PDFs (Liu *et al.*, 2019), to extract structural motifs (Anker *et al.*, 2022; Skjaervø *et al.*, 2023; Magnard *et al.*, 2022) and to determine the structure of small metallic nanoparticles (Anker *et al.*, 2020; Kjær *et al.*, 2023).

We here use a tree-based ML classifier to identify the structure of polyoxometallate (POM) clusters in solution on the basis of a PDF. POM clusters are a family of large polyanion clusters mostly constructed of [MO_6] octahedra, where M is often Mo, W, V or Nb (Gumerova & Rompel, 2018, 2020; Miras *et al.*, 2012; Long *et al.*, 2010). POMs have been extensively studied due to both their rich chemistry and their many applications, *e.g.* in molecular magnets, as catalysts for water splitting, as conductors or in medicine (Gumerova & Rompel, 2018, 2020; Miras *et al.*, 2012; Long *et al.*, 2010). Furthermore, it has been shown that the formation of metal oxide crystals can be dependent on the structure of the POM cluster, which has a huge impact on the formation mechanism (Christiansen *et al.*, 2020*b*; Juelsholt *et al.*, 2019). While POMs have so far been mainly studied in the crystalline form, PDF analysis allows POM structure studies in solution, which paves the way for a new understanding of their chemistry.

The ML model, which we refer to as *POMFinder*, has been trained on simulated PDFs from 443 POM clusters, cut out of crystal structures containing POMs obtained from the Crystallography Open Database (COD; Gražulis *et al.*, 2018) and the Inorganic Crystal Structure Database (ICSD; Allen *et al.*, 1987). *POMFinder* allows identification of POM structures from PDFs and has an accuracy of 94.0% on simulated data within the first prediction. It also shows good performance on experimental PDF data. We use SHapley Additive exPlanations (SHAP; Lundberg & Lee, 2017; Lundberg *et al.*, 2020) analysis to understand the predictions of *POMFinder*. With SHAP analysis, we can calculate the contribution of each input feature in the ML model to its predictions. Using SHAP analysis on simulated PDFs from X-rays (xPDF), neutrons (nPDF) and electrons (ePDF), we show that *POMFinder* learns trends corresponding to the scattering power of the different elements in the POMs and uses this information in its predictions. Finally, we show that the method can be extended to use data jointly from multiple scattering techniques instead of analysing the data separately, comparable to the 'complex modelling' approach (Billinge & Levin, 2007). We use simulated xPDF, small-angle X-ray scattering (SAXS) data, nPDF and ePDF, as well as combinations of the above data sets. A common problem of complex modelling is to weight the data sets (Anker *et al.*, 2021; Juhás *et al.*, 2015; Krayzman *et al.*, 2008), but this is not necessary when using ML to identify the structural model.

2. Construction of the POM database and training of the *POMFinder* model

We aim to create an ML model that can quickly and efficiently match an atomic POM structure to experimental data. We have chosen to focus on X-ray PDF data as the structural characterization technique and on POMs as the structures of interest. In principle, however, the data set can be any information that can be modelled using an atomistic model and represented in a tabular data format. The goal is *not* to have an algorithm that can output the perfect model to all experimental data every time with no user input. Instead, the successful ML model can filter out all bad models and give the user a handful of models which can be used for further analysis.

How to create a POM database and train *POMFinder*

Input : Structural database (COD + ICSD) & chemical restraints (if any)

1. Build a database of polyoxometallate clusters
   a. Remove crystals which do not fulfil chemical restraints
   b. Cut out clusters from crystal structures in the database
   c. Remove clusters which do not fulfil chemical restraints
   d. Remove all equivalent clusters based on a similarity score
2. Simulate N data sets per structure (N is normally set to 100) (data set can be any data set such as PDF or SAXS etc.)
3. Train a GBDT algorithm to classify to which structure a data set matches

Output : A tool that can output a cluster structure based on a data set

Figure 1
Pseudo-code describing how to create POM clusters from a CIF database and how to train *POMFinder*. A POM database is built from the ICSD and COD by cutting out clusters from all crystal structures with chemical compositions similar to POM clusters in solution (step 1) (Gumerova & Rompel, 2020). A number of PDFs are simulated for each POM cluster using various parameters (step 2). These PDFs are then used to train a GBDT model for classifying the corresponding structure from a PDF (step 3).

A pseudo-code of how to create the POM database and train *POMFinder* can be seen in Fig. 1. The structural database of POM clusters is built from crystallographic information files (CIFs) obtained from both the COD and ICSD using chemical restraints appropriate for POM clusters, as discussed below. Afterwards, a number (N) of PDFs are simulated for each POM structure with varying simulation parameters ($Q_{\rm min}$, $Q_{\rm max}$, an instrumental damping parameter $Q_{\rm damp}$ and an isotropic atomic displacement parameter ADP). The Q range used in the PDF [equation (1)] affects the r resolution of the PDF, and the limited Q range ($Q_{\rm min}$–$Q_{\rm max}$) creates termination ripples in the PDF (Egami & Billinge, 2012). Therefore, automated PDF data analysis must be applicable across various Q ranges. The parameters are varied using Latin hypercube sampling (Bouhlel *et al.*, 2019), and a gradient-boosting decision tree (GBDT) model (Chen & Guestrin, 2016; https://xgboost.readthedocs.io/en/stable/index.html#) is trained to classify which POM structure best matches the input data. In the following sections we will elaborate on this process.

2.1. Building a database of polyoxometallate clusters

The COD and ICSD databases contain hundreds of thousands of CIFs. When building our database, we first screened for CIFs with the same metal–oxygen ratios as described in a comprehensive review of POM clusters in solution by Gumerova & Rompel (2020). This restrained the database to 56 different metal–oxygen ratios, yielding 1281 CIFs. Clusters were then cut out of the CIFs by creating a 2 × 2 × 2 unit cell of the crystal and extracting all clusters of atoms not bonded to other atoms in the structure. Some clusters span more than a single unit cell, so to capture the complete POM cluster, a 2 × 2 × 2 unit cell was needed. Next, all isolated clusters that did not fulfil the chemical restraints (the 56 different metal–oxygen ratios) were removed. Fig. 2 illustrates an example of a cluster that was cut out of a crystal built from Keggin polyoxoanions, $K_2NaH_2[BW_{12}O_{40}]\cdot 12H_2O$ (Han *et al.*, 2012).

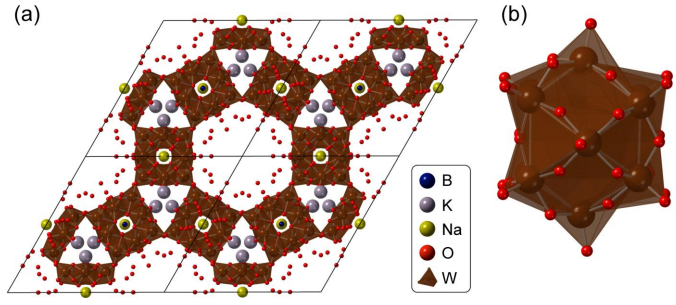


Figure 2
A POM cluster cut out from a crystal structure. (*a*) The crystal structure of $K_2NaH_2[BW_{12}O_{40}]\cdot 12H_2O$ (Han *et al.*, 2012) and (*b*) the corresponding POM cluster. W is shown in brown, O in red, Na in yellow, B in blue and K in grey. H is omitted for clarity.

This procedure yielded 969 potential polyoxometallate clusters. The simulated PDFs' Pearson correlation coefficients (PCCs) were used to remove similar structures from the database (Kjær *et al.*, 2022). The PDFs were compared iteratively by simulating a PDF of the first and second clusters with the parameters given in Section A in the supporting information and comparing their absolute PCCs. The PCC is a measure from −1 to 1 of how linearly correlated two continuous data sets are, where −1 represents inverse data sets and 1 represents identical data sets. If the absolute PCC was higher than 0.99, the second cluster was not included in the database. The third cluster was then compared with the first and second clusters by the same procedure and so on. The value of 0.99 was defined by manually inspecting the structures, their corresponding PDFs and the PCCs. Examples of three structures, their corresponding simulated PDFs and the PCCs can be seen in Section A in the supporting information. This process was performed with all 969 structures, yielding 443 unique structures. We note here that it is not guaranteed that the clusters are perfectly cut out of the crystal structure, which makes it important for the user to inspect the results of *POMFinder* and establish whether they make chemical sense.

2.2. Simulation of PDFs from the POM structures and training process of *POMFinder*

For each structure, a number (N) of PDFs were simulated with a broad range of instrumental parameters sampled using Latin hypercube sampling (Bouhlel *et al.*, 2019.) The simulations were done using *DiffPy-CMI* (Juhás *et al.*, 2015). The parameters are $Q_{\rm min}$, $Q_{\rm max}$, $Q_{\rm damp}$ and the ADPs. Section B in the supporting information gives the range of simulation parameters for the PDF data. The PDFs are normalized to have $G(r)_{\rm max} = 1$, and all intensities up to $r = 1$ Å are set to 0 since the POM clusters are unlikely to have atomic distances that contribute to the signal in this range of the PDF. While this normalization is vital for aligning the training set with experimental PDFs, we use the PDFs in their unmodified form

How to use *POMFinder*

Input : A data set

1. Give the data set as input to the algorithm

Output : An ordered list of suitable POM clusters

Figure 3
Pseudo-code describing how to use *POMFinder*. The data set is simply given as input to *POMFinder*, which outputs an ordered list of suitable POM clusters from which a few can be fitted to the data set.

for fitting procedures and for all visual representations within this paper.

An example of an experimental PDF before and after normalization is shown in Section B in the supporting information.

The simulated data sets and their corresponding instrumental parameters (Q_{min}, Q_{max}, Q_{damp} and ADPs) are input in a GBDT model. The GBDT algorithm used is *XGBoost* with default parameters, except for the learning rate, which was set to 0.3, and the early stop criterion of five rounds without improvement (Chen & Guestrin, 2016; https://xgboost.readthedocs.io/en/stable/index.html#). The problem is a 443 class classification problem with an input of 443 × N simulated PDFs. For each structure, two of the 100 simulated PDFs are randomly chosen and set aside during the training of the model and later used as validation and test sets. The validation set is used to validate when the GBDT model has converged.

The loss curve (multiclass log loss; https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html) is plotted in Section C in the supporting information, which shows that the model can predict the training data with 100% accuracy, while the validation data are predicted with a small loss. The concluding accuracy of the model can be determined on the test set, which are data on which the model has not been trained or validated, *i.e.* comparable to how *POMFinder* can be used for experimental data. When *POMFinder* is trained on 100 PDFs for each structure, the accuracy on the test set is 94.0% according to test set predictions.

3. Use of *POMFinder*

POMFinder is a simple tool to use since everything is fully automated. As seen in Fig. 3, one simply provides a data set as input to *POMFinder*, and it will return a list of likely structures as output. The input here is a PDF but it can, in principle, be any data that can be modelled using an atomistic model and represented in a tabular data format. As *POMFinder* is designed for predicting single-phase POMs from their corresponding PDFs, it will be challenged if confronted with PDFs obtained from multi-phase cluster systems or from crystalline structures. The output will be given in the *XYZ* format providing the elements and coordinates of all the atoms in the structure.

4. Results and discussion

4.1. Identification of POM structures from experimental PDFs

We start by demonstrating the power of *POMFinder* on an experimental PDF from a 0.05 *M* aqueous solution of ammonium metatungstate hydrate, $(NH_4)_6[H_2W_{12}O_{40}]\cdot xH_2O$, which is known to yield $[H_2W_{12}O_{40}]^{6-}$ ions with the α-Keggin structure (Juelsholt *et al.*, 2019). The data were collected on

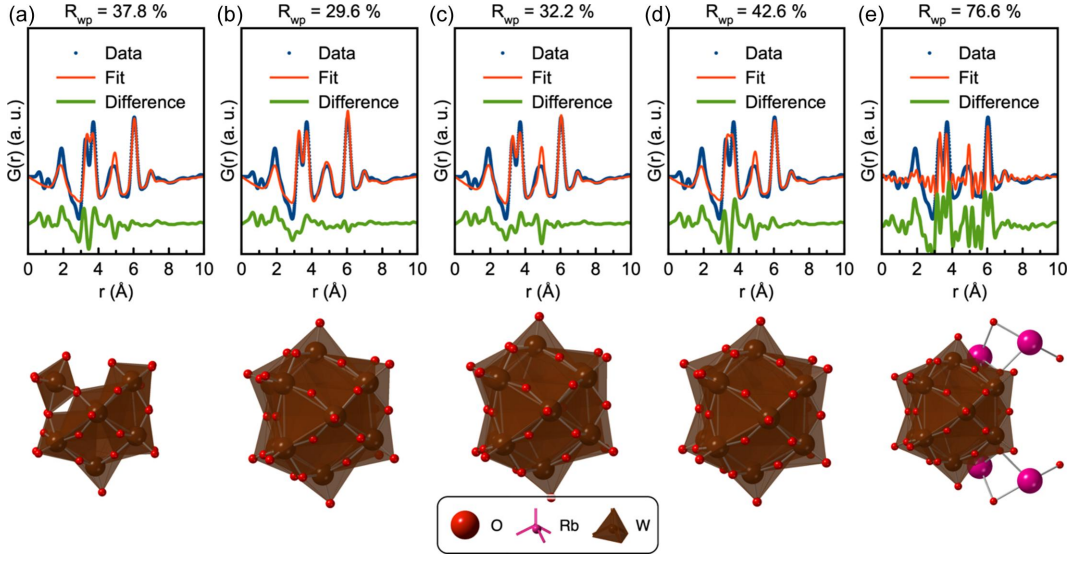


Figure 4
POMFinder's top five predictions on experimental high-quality X-ray PDF data. Comparisons of the PDF obtained from the 0.05 *M* ammonium metatungstate solution and the fitted PDF of (*a*) a $W_{11}O_{35}$ Keggin-based fragment from the dimeric $K_{5.5}Na_7Nd[SiW_{11}O_{39}(H_2O)]_2(CH_3COO)_2(H_2O)_{10}$ complex (Saini *et al.*, 2014), (*b*) a $W_{12}O_{36}$ fragment from the $K_5H(CoW_{12}O_{40})(H_2O)_{15}$ crystal (Glass *et al.*, 2014), (*c*) a $W_{12}O_{40}$ fragment from an ionic crystal structure of $[Al_{13}O_4(OH)_{24}(H_2O)_{12}](H_2W_{12}O_{40})(OH)(H_2O)_{23.12}$ (Son *et al.*, 2003), (*d*) a $W_{12}O_{36}$ fragment from the porous inorganic structure of the formula $K_2NaH_2(BW_{12}O_{40})(H_2O)_{12}$ (Han *et al.*, 2012) and (*e*) a $W_{12}Rb_4BO_{43}$ fragment from another ionic crystal, $Rb_4[Cr_3O(OOCH)_6(H_2O)_3(BW_{12}O_{40})](H_2O)_{16}$ (Uchida *et al.*, 2006). W is shown in brown, O in red and Rb in pink. Refinement parameters are reported in Section D in the supporting information.

the DanMAX beamline (MAX IV, Lund, Sweden) using a wavelength of $\lambda = 0.3542$ Å, achieving a Q_{max} of 20 Å^{-1}. The acquisition time for the total scattering data set was 15 min. Keggin structures [Fig. 2(*b*)] have the chemical composition $[XM_{12}O_{40}]^{n-}$, where X is a tetrahedrally coordinated cationic central atom in the middle of the cluster or one to three H^+ ions, M is the metal atom of the cluster, and n is the negative charge of the cluster. Keggin clusters are divided into five rotational isomers with increasing degrees of edge sharing, namely α, β, γ, δ and ε (Gumerova & Rompel, 2020; Jeannin, 1998; Sartzi *et al.*, 2015), although the δ isomer is not present in our POM database.

When the experimental PDF is given as input to *POMFinder*, the output is an ordered list of how probable it is that the PDF originates from each of the 443 POM structures in the POM database. The first five entries of the list are given in Section D in the supporting information, along with the probabilities assigned by *POMFinder*. The list clearly shows a dominance of structures with $W_{11-12}O_{35-43}$ composition which correspond to Keggin fragments. The five structures with the highest probability assigned by *POMFinder* are shown in Fig. 4, along with the fits to the experimental PDF. The first four candidate structures fit the PDF reasonably well. The best candidate, Fig. 4(*b*), with an R_{wp} value of 29.6%, is an α-Keggin structure. All the other structures are also α-Keggin structures or fragments.

4.2. Using *POMFinder* on fast acquisition data sets with a lower Q_{max}

Having established that *POMFinder* can identify a POM structure from a high-quality experimental PDF, we are interested in examining the use of *POMFinder* for data acquired with a fast time resolution, as is the case for *in situ* data. X-ray total scattering with PDF analysis is a powerful technique to study the formation of *e.g.* oxides, and it has previously been shown that POM structures can play an important role in their formation (Skjaervø *et al.*, 2023; Juelsholt *et al.*, 2019; Bøjesen *et al.*, 2016; Saha *et al.*, 2014). Therefore, we tested *POMFinder* on fast-acquisition experimental PDFs with 2 s time resolution from the 0.05 *M* solution of ammonium metatungstate. The data quality for this data set only allows a Q_{max} of 16 Å^{-1}. The data are the same as reported by Juelsholt *et al.* (2019) on the formation of tungsten oxide. The experimental PDF of the ammonium metatungstate solution (Fig. 5) shows a small structure with PDF peaks up to about 7 Å. When inputting the PDF to *POMFinder*, we again obtain an ordered list of possible structures, with the best five listed in Section D in the supporting information. Fig. 5 shows the fit of the five best predictions on the experimental PDF. The best fitting POM fragments, Figs. 5(*a*) and 5(*d*), are lacunary α-Keggin structures with three out of four triads. The second structure, Fig. 5(*b*), also reasonably fits the experimental PDF with an α-Keggin structure. In contrast, the structures in both Figs. 5(*c*) and 5(*e*) are too large to describe the experimental PDF well. Nevertheless, using *POMFinder* we can identify main motifs and thus determine a good model from fast-acquisition PDFs.

As discussed by Juelsholt *et al.* (2019), another cluster appears when heating the 0.05 *M* solution of ammonium metatungstate in oleylamine to 200°C. The experimental PDF after *ca* 4 min of heating is shown in Fig. 6. When inputting the PDF to *POMFinder*, we again obtain an ordered list of

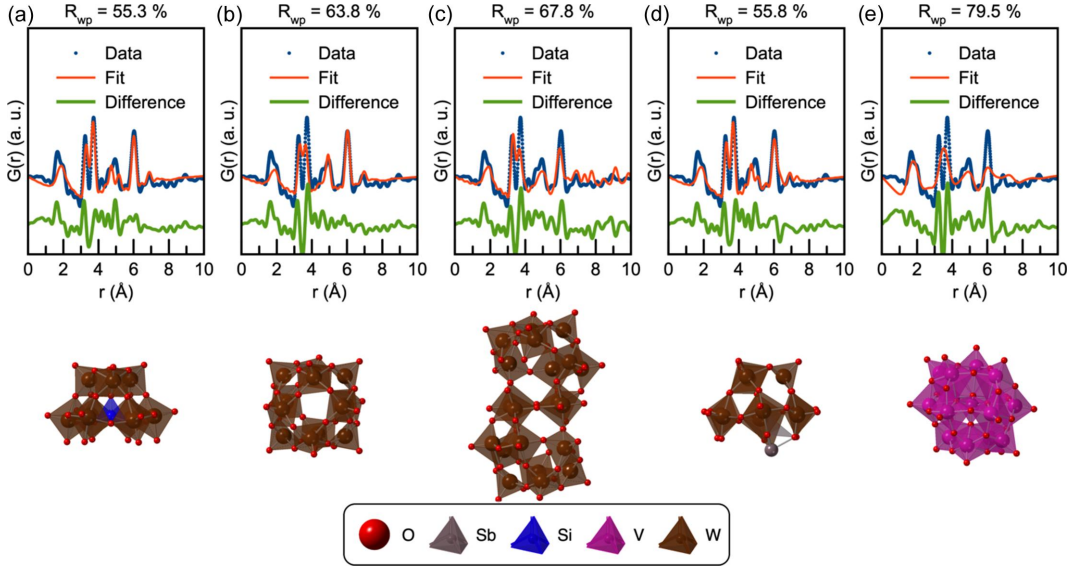


Figure 5
POMFinder's top five predictions on experimental fast-acquisition X-ray PDF data. Comparisons of the PDF from a 0.05 *M* solution of ammonium metatungstate in oleylamine with (*a*) a W_9SiO_{34} fragment from a Keggin-based $Na_2[C(NH_2)_3]_2[[(CH_3)_2Sn(H_2O)]_3(A\text{-}\alpha\text{-}SiW_9O_{34})]\cdot 10H_2O$ crystal (Piedra-Garza *et al.*, 2009), (*b*) a $W_{12}O_{36}$ fragment from the crystal structure of a porous framework based on Keggin polyoxoanions, $K_2NaH_2[BW_{12}O_{40}]\cdot 12H_2O$ (Han *et al.*, 2012), (*c*) a $W_{20}O_{64}$ fragment from a pseudo-Keggin-based crystal with chemical composition $H_{2-x}Bi_2W_{20}O_{70}(HWO_3)$ (Patrut *et al.*, 2010), (*d*) an SbW_9O_{30} fragment from a $K_{11}[Sb_3(SiW_9O_{34})_2]\cdot 31H_2O$ crystal structure (Assran *et al.*, 2012) and (*e*) a $V_{15}O_{42}$ fragment from the bicapped Keggin structure $(TMA)_3H_6V^V{}_{15}0_{42}\cdot 2.5H_20$ (TMA = tetramethylammonium) (Hou *et al.*, 1993). W is shown in brown, Sb in grey, O in red, Si in blue and V in pink. Refinement parameters are reported in Section D in the supporting information.

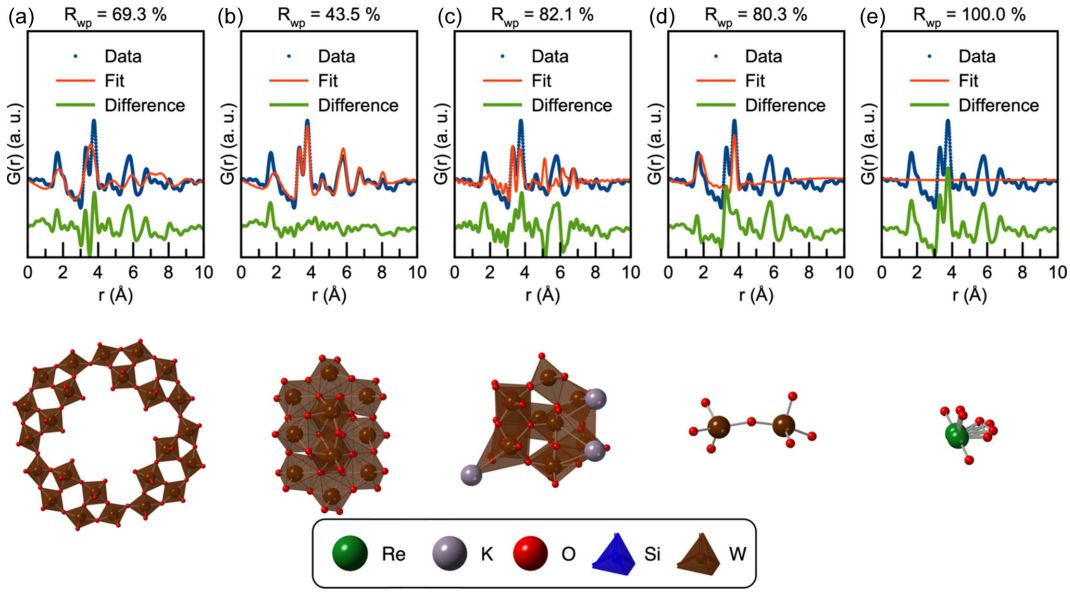


Figure 6
POMFinder's top five predictions on experimental fast-acquisition X-ray PDF data. Comparisons of the PDF from a 0.05 *M* solution of ammonium metatungstate in oleylamine heated to 200°C for 4 min and the calculated PDF of (*a*) a $W_{48}O_{152}$ fragment from the polyanion $K_{26.5}Li_{9.5}[H_4As_8W_{48}O_{184}]\cdot 90H_2O$ (Mbomekallé *et al.*, 2014), (*b*) a $W_{12}O_{42}$ fragment from the acidic sodium polytungstate $Na_5[H_7W_{12}O_{42}]\cdot 20H_2O$ (Redrup & Weller, 2009), (*c*) a $W_{11}K_3O_{38}$ fragment from the crystal structure $K_6H_4W_{11}O_{38}\cdot H_2O$ (Lehmann & Fuchs, 1988), (*d*) a W_2O_7 fragment from the crystal structure of $Bi_2W_2O_9$ (Champarnaud-Mesjard *et al.*, 1999) and (*e*) an Re_2O_8 fragment from the crystal structure $Bi_{28}Re_2O_{49}$ (Crumpton *et al.*, 2005). W is shown in brown, K in grey, O in red, Si in blue and Re in green. Refinement parameters are reported in Section D in the supporting information.

possible structures, with the best five listed in Section D in the supporting information. Figs. 6(*a*)–6(*e*) show the first five structures suggested by *POMFinder* and their fits to the PDF. The second prediction, Fig. 6(*b*), is the only POM fragment that reasonably fits the experimental PDF. This is a paratungstate POM, which agrees with the conclusion reached by Juelsholt *et al.* (2019). *POMFinder* is thus very well suited for analysis of *in situ* data where small structural changes in the cluster structure are observed. We attempted to analyse the entire *in situ* data set from Juelsholt *et al.* (2019), comprising 1022 PDFs. This analysis was completed in 66.5 s using a standard laptop equipped with an Intel Core i7-8665U CPU at 1.9/2.11 GHz. *POMFinder* performs well for the stages in the *in situ* data set where only one cluster is present. However, many of the PDFs in the time-resolved data set contain signals from multiple cluster species. Here, *POMFinder* is challenged, as these types of data go beyond the training set used. At this point, *POMFinder* thus cannot be used for identifying suitable POM clusters for PDFs obtained from multiple coexisting POMs. This challenge could possibly be overcome by combining the use of *POMFinder* with *e.g.* principal component analysis or negative matrix factorization, which potentially could separate the signals from each POM in the PDF.

4.3. Rationalizing *POMFinder*'s predictions using SHAP values

The above results have established that *POMFinder* can identify the POM structure present in solutions from experimental PDFs, yet it is not clear on what *POMFinder* bases its predictions. To obtain this understanding, we use SHAP analysis. SHAP is a feature importance measure which yields information about how the ML model exploits the individual features in the input data to make its predictions. Here, the features are Q_{min}, Q_{max}, Q_{damp} and $G(r)$ values for r values between 0.0 and 10.0 Å with a step size of 0.1 Å. A SHAP value is calculated for each feature for each PDF in the training set. The amplitude of the calculated SHAP value for a given feature provides information about how important the feature is, while the sign of the SHAP value tells whether the feature is confirming or disqualifying the specific structure as a match to the data set. Figs. 7(*a*) and 7(*c*) show a SHAP analysis of the two fast-acquisition PDFs discussed above, predicting the α-Keggin and the paratungstate cluster, respectively. The top of the figure shows the SHAP values of the most important features, *i.e.* those that give the highest amplitude of SHAP values. The value of the features, in this case the $G(r)$ intensity, is indicated by colour: high PDF intensities [$G(r)$ values] in the PDFs in the training set are represented in red, while low $G(r)$ values are coloured blue. For the α-Keggin cluster, the SHAP analysis shows that the two most important features are the $G(r)$ values at r = 6.0 Å and r = 3.6 Å. When inspecting the PDF and POM structures, these r values correspond to two W—W distances, as indicated in the structure drawing in Fig. 7(*b*). This means that *POMFinder* bases its predictions strongly on PDF peaks arising from W—W distances. W has a higher X-ray scattering power compared with O (W has 74 electrons, whereas O has eight), and W—W peaks are thus much more prominent in X-ray PDFs compared with W—O or O—O peaks (Prince, 2004). For paratungstate, we observe the same trend [Figs. 7(*c*) and 7(*d*)]. We conclude that *POMFinder* predominantly bases its predictions on the intensities of the

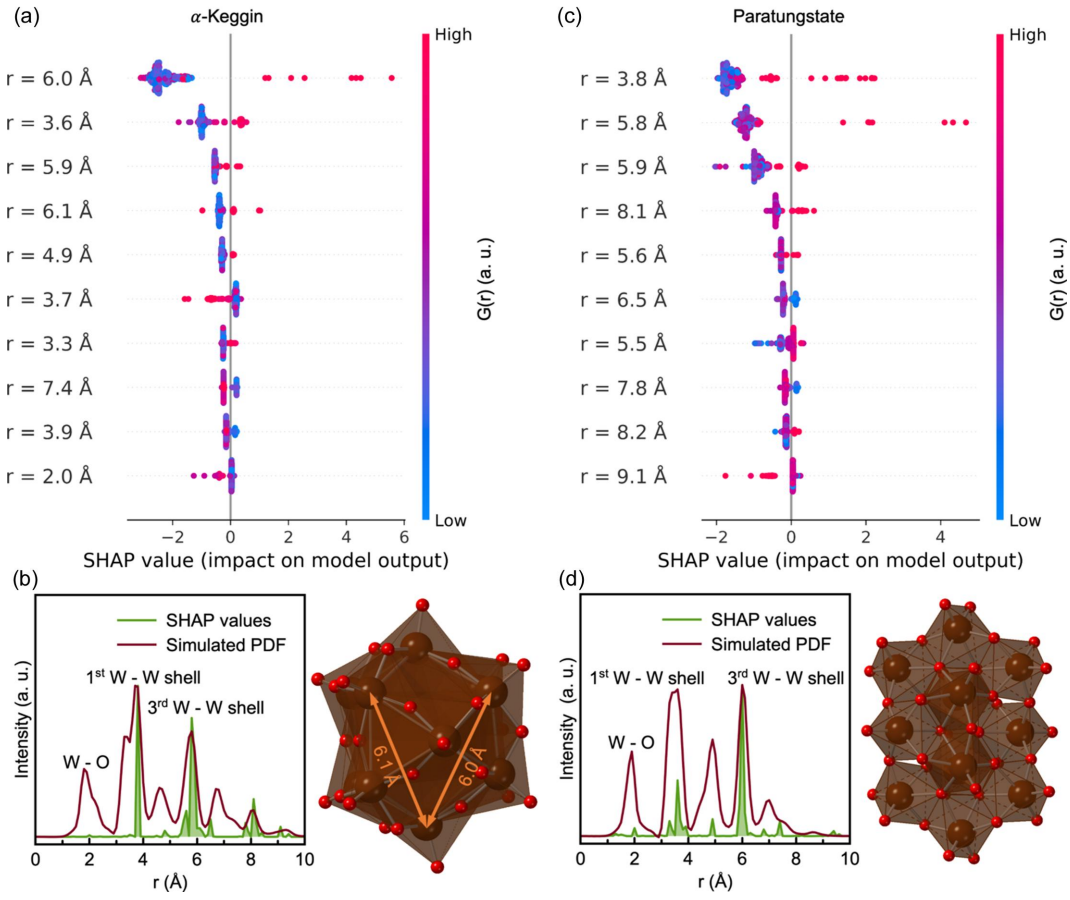


Figure 7
SHAP analysis of *POMFinder* on experimental PDFs. (*a*) and (*c*) For every PDF in the test set, SHAP values are calculated for all PDF intensities [$G(r)$ values], indicated with red for peaks and blue for low intensities. The r values of the PDF intensity are shown as labels. In panel (*a*), only the impact of predicting the α-Keggin cluster is shown, while (*c*) shows the impact of predicting the paratungstate cluster. (*b*) and (*d*) Histograms of absolute SHAP values for each PDF intensity plotted versus the r values on top of the PDFs of (*b*) the α-Keggin cluster and (*d*) the paratungstate cluster. W is shown in brown and O in red.

PDF peaks describing the first and third metal–metal shells for these two structures.

Instead of using SHAP to explain how *POMFinder* makes its predictions on individual PDFs, it is possible to get a global explanation by calculating an average of all the absolute SHAP values from the 443 POM structures in the POM structure database (shown in Section E in the supporting information). This analysis shows that the average absolute SHAP value for the Q_{min}, Q_{max} and Q_{damp} values is insignificant, meaning that *POMFinder* is not sensitive to the provided user input of Q_{min}, Q_{max} and Q_{damp} in the ranges used for training *POMFinder*. The average absolute SHAP value for $G(r)$ values in the r = 0–1 Å range is 0 since they are fixed to $G(r < 1\ \text{Å}) = 0$. However, the rest of the $G(r)$ values all have some contribution to the prediction of *POMFinder*. In particular, the $G(r)$ values corresponding to PDF peaks for M—O distances (∼2.0 Å) and the first and third metal–metal distances (∼3.3 and ∼6.2 Å, respectively) are important for *POMFinder*'s predictions.

To confirm that the predictions from *POMFinder* relate to the scattering power of the elements, we conducted the same SHAP analysis on simulated nPDF and ePDF data. From the same POM database, we first simulated 100 nPDFs and ePDFs from each POM structure in our database with different Q_{min}, Q_{max}, Q_{damp} and atomic displacement parameters, trained a GBDT model using a 98:1:1 training:validation:test set split, and then applied SHAP analysis to investigate the results. Fig. 8 shows the SHAP values for each feature in *POMFinder* when trained on the simulated xPDF, nPDF and ePDF data compared with their relative simulated PDFs.

When *POMFinder* is trained on nPDFs, the SHAP value is high for features corresponding to O—O peaks and W—O peaks. The neutron scattering lengths of W and O are 4.9 and 5.8 fm, respectively (Prince, 2004). This means that *POMFinder* bases its predictions on the O—O and W—O peaks when trained on nPDFs, rather than on the W—W peaks as seen for xPDFs as discussed above. This is probably due to the comparable neutron scattering lengths of W and O in contrast to the scattering contrast between W and O in xPDF and ePDF experiments. As expected, *POMFinder* primarily bases its predictions on the W—W distances when trained on ePDFs (electron scattering factors: W 12.5 Å, O 2.0 Å; Prince, 2004), but it gives higher weights to the O atoms than for xPDF. We thus see a clear trend between the scattering power of the element and the reasoning of *POMFinder*. A similar global analysis of all structures in our POM database

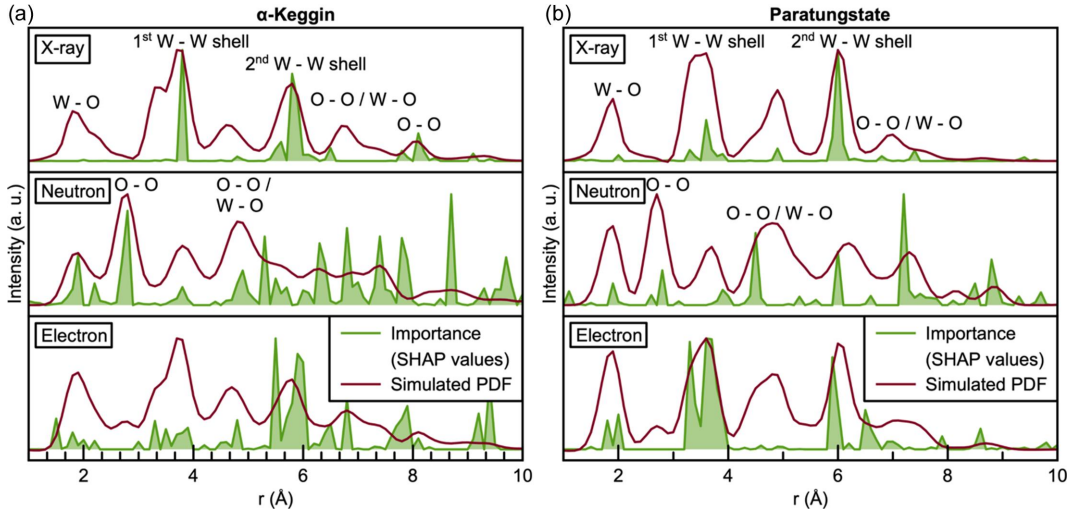


Figure 8
Analysis of the influence of the scattering probe on *POMFinder*'s predictions. (Top) The X-ray, (middle) the neutron and (bottom) the electron PDFs are plotted on top of a measure (SHAP values) of how important each data point in the PDF is for *POMFinder* to make its prediction on (*a*) the α-Keggin cluster and (*b*) the paratungstate cluster.

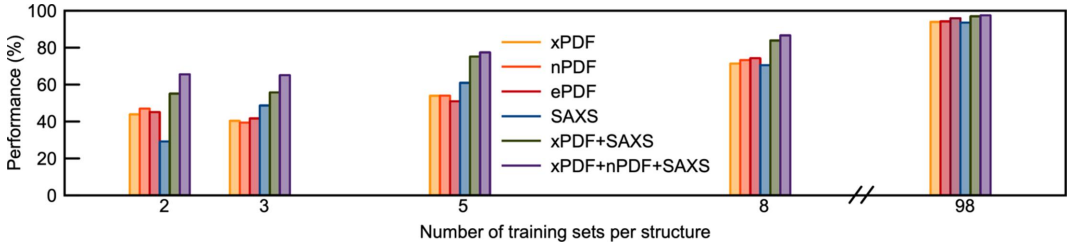


Figure 9
The performance of the model trained with various simulated data sets and different numbers of data sets per structure. Section F in the supporting information lists the mean and standard deviation based on five iterations where the model was trained on different simulated PDFs and predictions were made on the same test set.

is shown in Section E in the supporting information, which provides comparable results. We therefore hypothesize that *POMFinder* learns about the scattering contrast of different elements when predicting which POM fragment a PDF matches.

Fig. 9 shows the performance of *POMFinder* on the test set when *POMFinder* is trained using splits of 2:1:1, 3:1:1, 5:1:1, 8:1:1 and 98:1:1 PDFs per POM structure. Unsurprisingly, the performance (defined as the accuracy of the model on the test set) of *POMFinder* increases when trained on more data. Generally, *POMFinder* performs comparably when trained on xPDF, nPDF and ePDF data, as seen in Fig. 9.

4.4. Combination with data from other techniques

It has previously been shown that combined modelling of data from multiple scattering techniques can provide more robust results than separately modelling data from the individual scattering techniques (Anker *et al.*, 2021; Juhás *et al.*, 2015; Farrow *et al.*, 2014; Farrow & Billinge, 2009; Tucker *et al.*, 2007; Krayzman *et al.*, 2008). However, it is a cumbersome process to do combined modelling of data from multiple scattering techniques using a least-squares approach, and it can be challenging to weight the contribution from each data set (Anker *et al.*, 2021; Juhás *et al.*, 2015; Krayzman *et al.*, 2008; Terban & Billinge, 2022). We hypothesize that this problem can be overcome with ML methods, and here we take the first steps to extend *POMFinder* to combined data sets. Specifically, we train *POMFinder* on a combination of xPDF/SAXS and a combination of xPDF/SAXS/nPDF data. We do not weight the data sets. The SAXS simulations provide information on the size and shape of the POM clusters and are thus highly complementary to the PDF data discussed above. Details of the SAXS simulations are given in Section B in the supporting information.

The results on performance are given in Fig. 9, where we observe that when combining information from PDF and SAXS experiments, the performance increases, especially when using small training sets where *POMFinder* is challenged when using data from only one technique. This example demonstrates that *POMFinder* can easily be extended to identify a structure from combined data sets and that combining information from various data sets provides a higher performance on a test set.

5. Conclusions

We have demonstrated how our tree-based ML classifier, *POMFinder*, can screen a POM structure database to identify

structural candidates for the modelling of PDF data. Instead of using the traditional approach in scattering data analysis, where PDFs from all POM clusters in the database are fitted to the data through a least-squares refinement, we have shown that *POMFinder* can be used first to narrow down the field of candidate structures very quickly to five POM clusters, which can then be analysed further.

The POM database was made by cutting out clusters from the COD and ICSD databases following appropriate chemical restraints for POM structures. A GBDT model, *XGBoost*, was trained on simulated X-ray PDF data to classify the POM clusters with an accuracy of 94.0% on simulated PDFs. *POMFinder* also performs well on experimental data, including *in situ* data collected with a fast acquisition time. This ultrafast method allows *e.g.* visualizing the structural model in three dimensions while collecting data.

Using SHAP analysis, we have shown that *POMFinder* bases its predictions on trends comparable to the scattering contrast of the elements in the clusters.

Finally, we have shown that, in contrast to conventional complex modelling refinement methods, ML offers a promising and more flexible modelling framework for structure identification from combined data sets as it is not necessary to weight the data contributions (Anker *et al.*, 2021; Juhás *et al.*, 2015; Krayzman *et al.*, 2008).

POMFinder is open source, and the method can be directly applied by users without prior ML knowledge to characterize POM clusters.

POMFinder can be extended to include more types of chemical systems by extending the structural database used to generate the training data. In this project, we have focused on screening a database of POM fragments. However, the ultimate goal is to include any cluster fragment from the databases of known crystal structures, such as COD and ICSD which have more than 600 000 entries between them. The approach used for *POMFinder* can also be extended to analyse data from other scattering and spectroscopy techniques. We thus see *POMFinder* as a proof of concept, showing how a database of known structures can quickly be screened for analysis of *e.g.* scattering data using simple explainable ML methods.

6. Data availability

The database of POM clusters and the code used to train *POMFinder* is available at https://zenodo.org/records/10055030. *POMFinder* is available on GitHub at https://github.com/AndySAnker/POMFinder/. A web app to use *POMFinder* is available at https://huggingface.co/spaces/AndySAnker/POMFinder.

Acknowledgements

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 on beamline P02.1 at PETRAIII, and we would like to thank Martin Etter for his assistance in using the beamline.

Funding information

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 also thank DANSCATT (supported by the Danish Agency for Science and Higher Education) for support. A. S. Anker and M. Juelsholt acknowledge the Siemens Foundation for support for their thesis projects. We acknowledge the 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 NUFI (grant No. 4059-00009B).

References

Aimi, A. & Fujimoto, K. (2020). *ACS Comb. Sci.* **22**, 35–41.
Allen, F. H., Bergerhoff, G. & Sievers, R. (1987). *Crystallographic Databases.* Chester: International Union of Crystallography.
Anker, A. S., Butler, K. T., Selvan, R. & Jensen, K. M. Ø. (2023). *Chem. Sci.* **14**, 14003–14019.
Anker, A. S., Christiansen, T. L., Weber, M., Schmiele, M., Brok, E., Kjaer, E. T. S., Juhás, P., Thomas, R., Mehring, M. & Jensen, K. M. Ø. (2021). *Angew. Chem. Int. Ed.* **60**, 20407–20416.
Anker, A. S., Kjaer, E. T. S., Dam, E. B., Billinge, S. J. L., Jensen, K. M. Ø. & Selvan, R. (2020). *Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG)*, 24 August 2020, San Diego, California, USA, Abstract No. 22. New York: Association for Computing Machinery.
Anker, A. S., Kjaer, E. T. S., Juelsholt, M., Christiansen, T. L., Skjaervø, S. L., Jørgensen, M. R. V., Kantor, I., Sørensen, D. R., Billinge, S. J. L., Selvan, R. & Jensen, K. M. Ø. (2022). *npj Comput. Mater.* **8**, 213.
Assran, A. S., Izarova, N. V., Banerjee, A., Rabie, U. M., Abou-El-Wafa, M. H. M. & Kortz, U. (2012). *Dalton Trans.* **41**, 9914–9921.
Banerjee, S., Liu, C.-H., Jensen, K. M. Ø., Juhás, P., Lee, J. D., Tofanelli, M., Ackerson, C. J., Murray, C. B. & Billinge, S. J. L. (2020). *Acta Cryst.* A**76**, 24–31.
Bennett, T. D. & Cheetham, A. K. (2014). *Acc. Chem. Res.* **47**, 1555–1562.
Billinge, S. J. L. & Kanatzidis, M. G. (2004). *Chem. Commun.* pp. 749–760.
Billinge, S. J. L. & Levin, I. (2007). *Science*, **316**, 561–565.
Bøjesen, E. D., Jensen, K. M. Ø., Tyrsted, C., Mamakhel, A. H., Andersen, H. L., Reardon, H., Chevalier, J., Dippel, A.-C. & Iversen, B. B. (2016). *Chem. Sci.* **7**, 6394–6406.
Bouhlel, M. A., Hwang, J. T., Bartoli, N., Lafage, R., Morlier, J. & Martins, J. R. R. A. A. (2019). *Adv. Eng. Softw.* **135**, 102662.
Champarnaud-Mesjard, J.-C., Frit, B. & Watanabe, A. (1999). *J. Mater. Chem.* **9**, 1319–1322.
Chen, T. & Guestrin, C. (2016). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining, 13–17 August 2016, San Francisco, California, USA, pp. 785–794. New York: Association for Computing Machinery.
Chen, Z., Andrejevic, N., Drucker, N. C., Nguyen, T., Xian, R. P., Smidt, T., Wang, Y., Ernstorfer, R., Tennant, D. A., Chan, M. & Li, M. (2021). *Chem. Phys. Rev.* **2**, 031301.
Chepkemboi, C., Jorgensen, K., Sato, J. & Laurita, G. (2022). *ACS Omega*, **7**, 14402–14411.
Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., Park, C. W., Choudhary, A., Agrawal, A., Billinge, S. J., Holm, E., Ong, S. P. & Wolverton, C. (2022). *npj Comput. Mater.* **8**, 59.
Christiansen, T. L., Cooper, S. R. & Jensen, K. M. Ø. (2020*a*). *Nanoscale Adv.* **2**, 2234–2254.
Christiansen, T. L., Kjær, E. T. S., Kovyakh, A., Röderen, M. L., Høj, M., Vosch, T. & Jensen, K. M. Ø. (2020*b*). *J. Appl. Cryst.* **53**, 148–158.
Coelho, A. A. (2018). *J. Appl. Cryst.* **51**, 210–218.
Cooper, S. R., Candler, R. O., Cosby, A. G., Johnson, D. W., Jensen, K. M. Ø. & Hutchison, J. E. (2020). *ACS Nano*, **14**, 5480–5490.
Crumpton, T. E., Mosselmans, J. F. W. & Greaves, C. (2005). *J. Mater. Chem.* **15**, 164.
Dong, H., Butler, K. T., Matras, D., Price, S. W. T., Odarchenko, Y., Khatry, R., Thompson, A., Middelkoop, V., Jacques, S. D. M., Beale, A. M. & Vamvakeros, A. A. (2021). *npj Comput. Mater.* **7**, 74.
Egami, T. & Billinge, S. J. (2012). *Underneath the Bragg Peaks: Structural Analysis of Complex Materials*, 2nd ed. Oxford: Pergamon.
Farrow, C., Shi, C., Juhás, P., Peng, X. & Billinge, S. J. L. (2014). *J. Appl. Cryst.* **47**, 561–565.
Farrow, C. L. & Billinge, S. J. L. (2009). *Acta Cryst.* A**65**, 232–239.
Farrow, C. L., Juhas, P., Liu, J. W., Bryndin, D., Božin, E. S., Bloch, J., Proffen, T. & Billinge, S. J. L. (2007). *J. Phys. Condens. Matter*, **19**, 335219.
Glass, E. N., Fielden, J., Kaledin, A. L., Musaev, D. G., Lian, T. & Hill, C. L. (2014). *Chem. Eur. J.* **20**, 4297–4307.
Gražulis, S., Merkys, A. & Vaitkus, A. (2018). *Handbook of Materials Modeling – Methods: Theory and Modeling*, edited by W. Andreoni & S. Yip, pp. 1–19. Cham: Springer International Publishing.
Gumerova, N. I. & Rompel, A. (2018). *Nat. Rev. Chem.* **2**, 0112.
Gumerova, N. I. & Rompel, A. (2020). *Chem. Soc. Rev.* **49**, 7568–7601.
Han, X.-B., Zhang, Z.-M., Wang, Z.-S., Zhang, H., Duan, H. & Wang, E.-B. A. (2012). *Inorg. Chem. Commun.* **18**, 47–49.
Hou, D., Hagen, K. S. & Hill, C. L. (1993). *J. Chem. Soc. Chem. Commun.* pp. 426–428.
Jeannin, Y. P. (1998). *Chem. Rev.* **98**, 51–76.
Jensen, K. M. Ø., Juhas, P., Tofanelli, M. A., Heinecke, C. L., Vaughan, G., Ackerson, C. J. & Billinge, S. J. L. (2016). *Nat. Commun.* **7**, 11859.
Juelsholt, M., Anker, A. S., Christiansen, T. L., Jørgensen, M. R. V., Kantor, I., Sørensen, D. R. & Jensen, K. M. Ø. (2021). *Nanoscale*, **13**, 20144–20156.
Juelsholt, M., Lindahl Christiansen, T. & Jensen, K. M. Ø. (2019). *J. Phys. Chem. C*, **123**, 5110–5119.
Juhás, P., Farrow, C., Yang, X., Knox, K. & Billinge, S. (2015). *Acta Cryst.* A**71**, 562–568.
Keen, D. A. & Goodwin, A. L. (2015). *Nature*, **521**, 303–309.
Kjaer, E. T. S., Aalling-Frederiksen, O., Yang, L., Thomas, N. K., Juelsholt, M., Billinge, S. J. L. & Jensen, K. M. Ø. (2022). *Chem. Methods*, **2**, e202200034.
Kjaer, E. T. S., Anker, A. S., Weng, M. N., Billinge, S. J. L., Selvan, R. & Jensen, K. M. Ø. (2023). *Digit. Discov.* **2**, 69–80.
Kløve, M., Sommer, S., Iversen, B. B., Hammer, B. & Dononelli, W. A. (2023). *Adv. Mater.* **35**, 2208220.
Krayzman, V., Levin, I. & Tucker, M. G. (2008). *J. Appl. Cryst.* **41**, 705–714.
Lehmann, T. & Fuchs, J. (1988). *Z. Naturforsch. Teil B*, **43**, 89–93.
Liu, C.-H., Tao, Y., Hsu, D., Du, Q. & Billinge, S. J. L. (2019). *Acta Cryst.* A**75**, 633–643.
Long, D.-L., Tsunashima, R. & Cronin, L. (2010). *Angew. Chem. Int. Ed.* **49**, 1736–1758.
Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. & Lee, S.-I. (2020). *Nat. Mach. Intell.* **2**, 56–67.
Lundberg, S. M. & Lee, S.-I. (2017). *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4–9 December 2017, Long Beach, California, USA, pp. 4765–4774. Red Hook: Curran Associates.
Magnard, N. P. L., Anker, A. S., Aalling-Frederiksen, O., Kirsch, A. & Jensen, K. M. Ø. (2022). *Dalton Trans.* **51**, 17150–17161.
Mbomekallé, I.-M., Bassil, B. S., Suchopar, A., Keita, B., Nadjo, L., Ammam, M., Haouas, M., Taulelle, F. & Kortz, U. (2014). *J. Clust. Sci.* **25**, 277–285.
Miras, H. N., Yan, J., Long, D.-L. & Cronin, L. (2012). *Chem. Soc. Rev.* **41**, 7403–7430.
Pacchioni, G. (2019). *Nat. Rev. Phys.* **1**, 100–101.
Patrut, A., Bögge, H., Forizs, E., Rusu, D., Lowy, D. A., Margineany, D. & Naumesco, A. (2010). *Rev. Roum. Chim.* **55**, 865–870.
Pedersen, J. S. (1997). *Adv. Colloid Interface Sci.* **70**, 171–210.
Piedra-Garza, L. F., Reinoso, S., Dickman, M. H., Sanguineti, M. M., Kortz, U. (2009). *Dalton Trans.* pp. 6231–6234.
Prince, E. (2004). *International Tables for Crystallography*, Vol. C, *Mathematical, Physical and Chemical Tables*, 3rd ed., ch. 6. Dordrecht: Kluwer Academic Publishers.
Proffen, Th. & Neder, R. B. (1997). *J. Appl. Cryst.* **30**, 171–175.
Proffen, Th. & Neder, R. B. (1999). *J. Appl. Cryst.* **32**, 838–839.
Redrup, K. V. & Weller, M. T. (2009). *Dalton Trans.* pp. 4468–4472.
Rietveld, H. M. (1969). *J. Appl. Cryst.* **2**, 65–71.
Saha, D., Jensen, K. M., Tyrsted, C., Bøjesen, E. D., Mamakhel, A. H., Dippel, A. C., Christensen, M. & Iversen, B. B. (2014). *Angew. Chem.* **126**, 3741–3744.
Saini, M. K., Gupta, R., Parbhakar, S., Kumar Mishra, A., Mathur, R. & Hussain, F. (2014). *RSC Adv.* **4**, 25357–25364.
Sartzi, H., Miras, H. N., Vilà–Nadal, L., Long, D.-L. & Cronin, L. (2015). *Angew. Chem. Int. Ed.* **54**, 15488–15492.
Skjaervø, S. L., Anker, A. S., Wied, M. C., Kjaer, E. T. S., Juelsholt, M., Christiansen, T. L. & Ø. Jensen, K. M. (2023). *Chem. Sci.* **14**, 4806–4816.
Son, J. H., Kwon, Y.-U. & Han, O. H. (2003). *Inorg. Chem.* **42**, 4153–4159.
Szczerba, D., Tan, D., Do, J. L., Titi, H. M., Mouhtadi, S., Chaumont, D., Del Carmen Marco de Lucas, M., Geoffroy, N., Meyer, M., Rousselin, Y., Hudspeth, J. M., Schwanen, V., Spoerk-Erdely, P., Dippel, A. C., Ivashko, O., Gutowski, O., Glaevecke, P., Bazhenov, V., Arhangelskis, M., Halasz, I., Friščić, T. & Kimber, S. A. J. (2021). *J. Am. Chem. Soc.* **143**, 16332–16336.
Terban, M. W. & Billinge, S. J. L. (2022). *Chem. Rev.* **122**, 1208–1272.
Tucker, M. G., Keen, D. A., Dove, M. T., Goodwin, A. L. & Hui, Q. (2007). *J. Phys. Condens. Matter*, **19**, 335218.
Uchida, S., Kawamoto, R. & Mizuno, N. (2006). *Inorg. Chem.* **45**, 5136–5144.
Wang, C., Steiner, U. & Sepe, A. (2018). *Small*, **14**, e1802291.
Yang, L., Juhás, P., Terban, M. W., Tucker, M. G. & Billinge, S. J. L. (2020). *Acta Cryst.* A**76**, 395–409.
Yang, X., Masadeh, A. S., McBride, J. R., Božin, E. S., Rosenthal, S. J. & Billinge, S. J. L. (2013). *Phys. Chem. Chem. Phys.* **15**, 8480–8486.