ClusterFinder:
A non-machine learning approach to find cluster structures from pair distribution function data

Andy S. Anker[*1], Frederik L. Johansen*[†1,2], Ulrik Friis-Jensen*[†1,2], Simon J. L. Billinge*[*3], Kirsten M. Ø. Jensen*[*1]

*Correspondence to andy@chem.ku.dk (ASA), sb2896@columbia.edu (SJLB), kirsten@chem.ku.dk (KMØJ)

†Both authors contributed equally to this work

1: Department of Chemistry and Nano-Science Center, University of Copenhagen, 2100 Copenhagen Ø, Denmark

2: Department of Computer Science, University of Copenhagen, 2100 Copenhagen Ø, Denmark

3: Department of Applied Physics and Applied Mathematics Science, Columbia University, New York, NY 10027, USA

Abstract

A novel, automated, high throughput screening approach, ClusterFinder, is reported for finding candidate structures for atomic pair distribution function (PDF) structural refinements. Finding starting models for PDF refinements is notoriously difficult when the PDF originates from small chemical clusters. The reported ClusterFinder algorithm is able to screen 10^4 – 10^5 candidate structures from structural databases such as the inorganic crystal structure database (ICSD) in minutes, using the crystal structures as templates in which it looks for atomic clusters that result in a PDF similar to the target measured PDF. The algorithm returns a rank

ordered list of clusters for further assessment by the user. The algorithm performed well for simulated and measured PDFs of metal oxido clusters such as Keggin clusters. The approach is therefore a powerful approach to finding structural cluster candidates in a modelling campaign for PDFs of nanoparticles and nanoclusters.

Introduction

Throughout the last century, crystallographic methods have played a crucial role in advancing materials science. Yet, they often struggle when examining nanomaterials with limited long-range order (Billinge & Levin, 2007). Lately, total scattering with PDF analysis has shown promise for characterizing such nanomaterials (Billinge & Levin, 2007, Juelsholt *et al.*, 2021, Christiansen *et al.*, 2020), including polyoxometalate (POM) clusters (Juelsholt *et al.*, 2019, Benseghir *et al.*, 2020), and ionic clusters (Szczerba *et al.*, 2021, Anker *et al.*, 2021, Van den Eynden *et al.*, 2023). The PDF, derived from the Fourier transform of normalized and corrected X-ray, neutron, or electron scattering intensities, offers a real-space representation of inter-atomic distances in the sample (Egami & Billinge, 2012, Christiansen *et al.*, 2020).

Researchers have long pursued the challenge of deriving *ab initio* structure solutions from PDFs (Juhás *et al.*, 2006, Juhás *et al.*, 2008, Juhas *et al.*, 2010, Cliffe *et al.*, 2010, Cliffe & Goodwin, 2013, Anker *et al.*, 2020, Kjær *et al.*, 2023, Kløve *et al.*, 2023). However, success remains limited to rather simple chemical systems like the C_{60} buckyball and mono-metallic nanoparticles. In the absence of broadly applicable *ab initio* structure solution methods, suitable starting models are necessary to refine the PDFs. Known crystal structures are often used for crystalline materials. However, this task becomes exceptionally difficult for small clusters and nanoparticles. Recent methods such as clusterMining (Banerjee *et al.*, 2020) and structureMining have taken the approach of screening large numbers of structures that are pulled from databases or algorithmically generated. Nonetheless, they are all restrained to the presence of a suitable database of structures or an algorithmic structure generator.

A hybrid approach, ML-MotEx (Anker *et al.*, 2022), was recently demonstrated that used chemical knowledge to select candidate crystal structures from a crystallographic database, then explainable ML to find sub-clusters from the candidate structure that were consistent with the data. The approach worked well, but was slow, taking several minutes for each starting structure, which limited its application to cases where the candid parent crystal structures were few and obvious to the user. Here, we propose a novel algorithm, ClusterFinder, that follows the same approach of sampling sub-clusters from larger structural candidates, but it uses a non-machine learning direct scoring approach for identifying high performing sub-clusters. This speeds up the selection procedure to seconds, allowing for an automated search for sub clusters over large numbers of candidate parent structures that can be selected in an automated way from structural databases.

Method

The basic strategy was described in (Anker et al 2022). We summarize it here. The starting point is an atomic PDF experiment of a sample that contains small clusters, for example, a soluble reagent or nanoparticles suspended in a solvent. The resulting PDF has a small number of peaks in it confined to the low-*r* region, indicating the presence of unknown atomic clusters of small size (e.g., see Fig. 1).

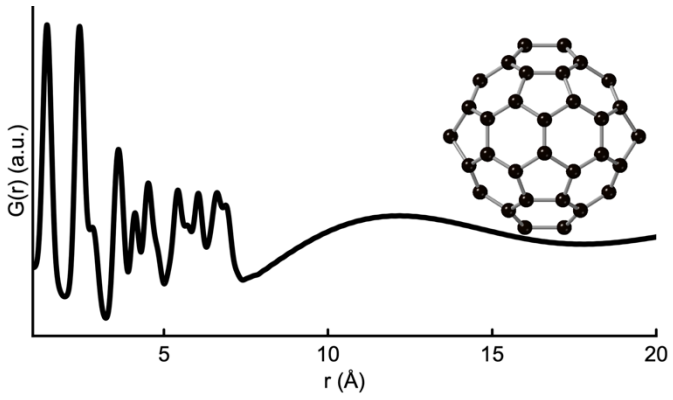

Fig. 1 | Simulated PDF from a C_{60} buckyball from a single unit cell of a C_{60} crystal structure (Chen & Yamanaka, 2002); The simulation parameters mimic typical PDF dataset values and can be seen in section A in the SI.

In principle, the data can be fit using programs such as PDFgui (Farrow *et al.*, 2007), Topas (Coelho, 2018), or DiffPy-CMI (Juhás *et al.*, 2015) to understand the structure of the clusters, but this process requires a good initial candidate structure to be given. The main challenge is to find good starting models for the fit. ClusterFinder addresses this need. It reuses the approach taken by ML-MotEx where a set of chemically reasonable crystal structures is identified. Large-enough candidate template clusters are then cut out from that crystal structure. Assuming for now that the cluster present in the experimental data, the target cluster, is contained within the template, the principal goal is to find the subset of occupied sites in the template that corresponds to the target cluster. A search over all possible permutations of present vs. absent atoms is impossible because of the combinatorics with 2^N-1 possibilities for a template of N sites. ML-MotEx (Anker *et al.*, 2022) used an explainable machine learning approach to optimize this problem by learning probabilities that each atom might be present in the target cluster after iterating over a small subset of all the possible permutations. This placed atom-sites in a rank ordered list and made it easy for the user to select a cut-off for which sites were occupied and which not to generate the target cluster configuration. Of course, the target cluster may not be present in the template and in general there is a further outer-loop that needs to be iterated over of all possible candidate crystal structures and templates. The ML-MotEx algorithm is too slow to do this over a large number of template candidates and the success of the approach relies on a strong chemical intuition suggesting a small number of candidate structures.

At the heart of the algorithm is the calculation to generate an ordered list of sites based on the probability that they are present in the target cluster. The Liga algorithm (Juhás *et al.*, 2006, Juhás *et al.*, 2008), also scores atoms in a cluster as part of its backtracking cluster reduction step, where poor performing clusters are reduced in size

by preferentially removing atoms that are contributing more error to the agreement with the data. The ranking was done using the commonly used PDF weighted profile agreement factor.

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

where G_{obs} and G_{calc} are the observed and calculated PDF intensities for the set, P, of model refinement parameters. The sum is over the *n* points in the PDF.

Taking inspiration from the Liga algorithm (Juhás *et al.*, 2006, Juhás *et al.*, 2008), we attempt an approach of computing the contribution to the fitting error for each atom site in the cluster. We call this the atom-removal error, and denote it for the i^{th} atom by ΔR^i_{wp}. It is computed by evaluating R_{wp} for the full set of atoms, then recomputing R_{wp} for the cluster with the i^{th} atom removed and taking the difference. This allows us to identify which atoms contribute the most error to the fit allowing us to target them for removal. For each atom, a scale factor and an isotropic expansion/contraction factor are allowed to refine to give the best agreement before computing R_{wp}. This procedure is extremely rapid and results in a list of atomic sites ranked by ΔR^i_{wp}.

To visualise the results, we plot the templates with each atom-site colour-coded based on its ΔR^i_{wp}. Atom sites with negative (good) ΔR^i_{wp} are coloured yellow and those with positive (bad) ΔR_{wp} are coloured blue. The approach is illustrated schematically for a trivial example of a binary molecule in Fig. 2. Note that ClusterFinder only ranks the atoms in the template and a human input is still needed to determine which atoms to remove when finding the best cluster candidates. In Fig. 2, it is trivial to remove atom 3 and 4 from the ClusterFinder output but this task might not always be trivial and may include chemical intuition of the user. However, it is still extremely valuable because, due to its speed, it can be used to screen large numbers of structures to find the best cluster candidates, removing the need for chemical intuition in the template structure selection part of the task.

To test the ClusterFinder approach, the results were compared to a known "ground-truth" and to the already published results of ML-MotEx. We found that ClusterFinder provided comparable results to ML-MotEx in quality but orders of magnitude quicker.

The speed-up was sufficient to allow us to screen large databases of starting models for the right starting template in minutes. To demonstrate the power of this, we provide five examples where we screen the entire ICSD (Zagorac *et al.*, 2019), containing 188,631 structure entries, for a suitable starting model in a timeframe ranging from 3 to 42 minutes. We expect this to make ClusterFinder highly valuable since if the target cluster exists anywhere in any known crystal structure it will automatically be found without any user input at this stage.

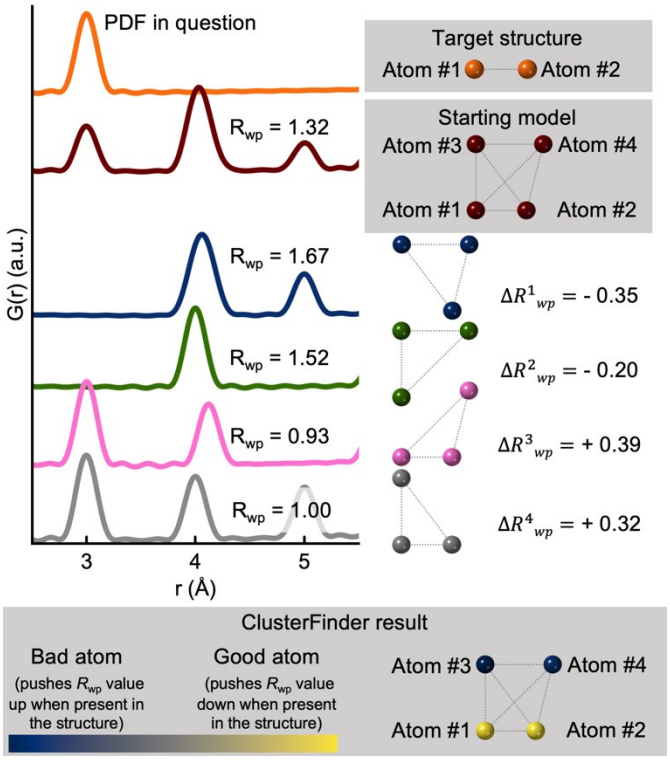

Fig. 2 | Illustration of the ClusterFinder process. A starting model is provided as input and the R_{wp} value is calculated by structure refinement. Atoms are iteratively removed from the starting model and fitted to the experimental PDF. The atom-removal error, ΔR^{i}_{wp}, is calculated by taking the difference between the R_{wp} value of the full starting model and when the atoms are removed. Atoms are colour-coded based on atom-removal error

– yellow indicates a negative ΔR^{i}_{wp} value (improved fit) while blue signifies a positive ΔR^{i}_{wp} value (worsened fit).

Results & Discussion

Applying ClusterFinder to Extract Cluster Motifs from Simulated PDFs

We first demonstrate ClusterFinder's ability to extract cluster motifs from simulated PDFs. Figure 3 shows three simulated PDFs, each corresponding to a distinct structure: a decatungstate polyoxometalate cluster from a $Na_5(H_7W_{12}O_{42})(H_2O)_{20}$ crystal structure (Redrup & Weller, 2009), coloured in blue; a C_{60} buckyball from a single unit cell of the C_{60} crystal structure (Chen & Yamanaka, 2002), coloured in green; and a paratungstate polyoxometalate cluster originated from a $(Ba(H_2O)_2(H(N(CH_3)_2)CO)_3)_2(W_{10}O_{32})(H(N(CH_3)_2)CO)_2$ crystalline model (Poimanova *et al.*, 2015), coloured in red. The values of the simulation parameters used to mimic typical PDF dataset values are listed in Section A in the Supplementary Information (SI). Figure 3B–D show the structural templates used by ClusterFinder. In these tests, structural templates were manually constructed with the minimum unit cells needed to include the full cluster. ClusterFinder outputs a list of atomic sites ranked by the ΔR^{i}_{wp} value, and again, we visualise atom sites with negative ΔR^{i}_{wp} as yellow and those with positive ΔR^{i}_{wp} as blue. The resulting visualisations are shown in Figure 3B-D.

ClusterFinder correctly extracted all three cluster structures from their starting model in under a minute using a standard laptop, demonstrating a significant speed advantage over the ML-MotEx algorithm (Anker *et al.*, 2022), which takes approximately an hour on a standard laptop. Although ClusterFinder accurately extracts the decatungstate polyoxometalate cluster (blue) and the paratungstate polyoxometalate cluster (red), it does not completely recover the C_{60} buckyball (green), incorrectly labelling two atoms. The ML-MotEx algorithm also exhibited similar limitations in extracting this structure.

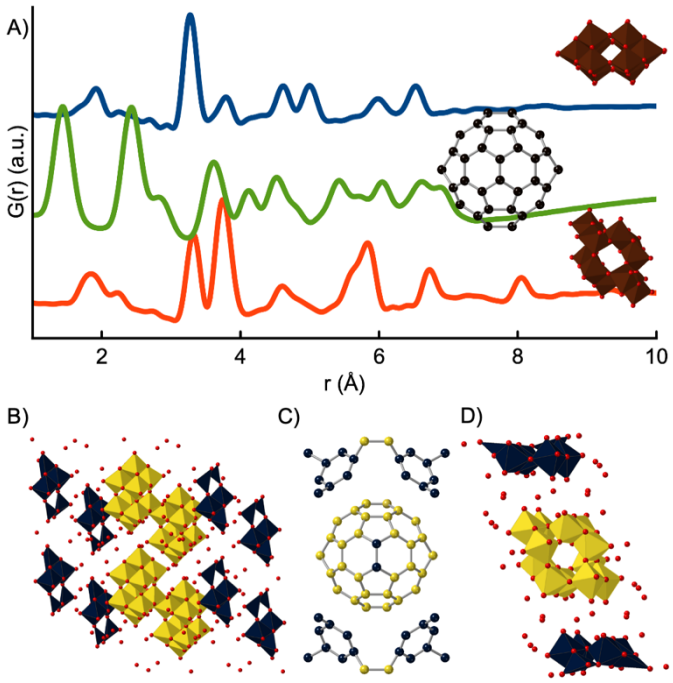

Fig. 3 | Analysis of simulated PDFs of well-known cluster structures. A) Simulated PDFs of (blue) a decatungstate polyoxometalate cluster from the $Na_5(H_7W_{12}O_{42})(H_2O)_{20}$ crystal structure (Redrup & Weller, 2009); (green) a C_{60} buckyball from a single unit cell of a C_{60} crystal structure (Chen & Yamanaka, 2002); and (red) a paratungstate polyoxometalate cluster obtained from the $(Ba(H_2O)_2(H(N(CH_3)_2)CO)_3)_2(W_{10}O_{32})(H(N(CH_3)_2)CO)_2$ crystalline model (Poimanova *et al.*, 2015). Simulation parameters chosen to mimic typical measured PDF datasets and are reproduced in section A in the SI. B-D) Results of using ClusterFinder on the three simulated PDFs where the atoms with the B) 40, C) 60 and D) 12 lowest ΔR^i_{wp} values have been coloured yellow, while the rest are coloured blue. Section C in the SI shows a similar representation but where the atom-removal values are directly visualised using a continuous colour bar. Oxygens are coloured red and polyhedra are coloured according to their metal atom center.

Applying ClusterFinder to Extract Cluster Motifs from Experimental PDFs

While ClusterFinder's potential to extract cluster motifs from various crystalline supercell structures has been demonstrated with simulated PDFs, it is essential that ClusterFinder possesses similar abilities on an experimental PDF. Here we benchmark the performance of ClusterFinder against that of the previously published ML-MotEx algorithm by comparing its performance on the same set of experimental PDFs and clusters.

The experimental PDF was 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 (Juelsholt *et al.*, 2019). Experimental details can be found in the ML-MotEx paper (Anker *et al.*, 2022). We employed four different crystallographic models to extract templates for ClusterFinder/ML-MotEx, as listed in Table 1.

Starting models	Crystal Composition	Reference
I	$[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine)	(Niu *et al.*, 2004)
II	$(CH_3)_4N)_4SiW_{12}O_{40}$	*(Joachim et al., 1981)*
III	$(((CH_3)_2NH_2)_6(Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$	(Niu *et al.*, 2003)
IV	$((CH_3)_2NH_2)_3(PW_{12}O_{40})$	*(Busbongthong & Ozeki, 2009)*

Table 1 | Four starting models containing the α-Keggin clusters used with ClusterFinder to extract an α-Keggin cluster.

Again, only a scale factor and an isotropic expansion/contraction factor were refined during the ClusterFinder process. Both ClusterFinder and ML-MotEx successfully extracted the α-Keggin clusters with few mislabelled atoms for all four starting models. Although, ClusterFinder has slightly more mislabelled atoms compared to ML-MotEx, it is orders of magnitude faster, making it an ideal choice for screening larger databases.

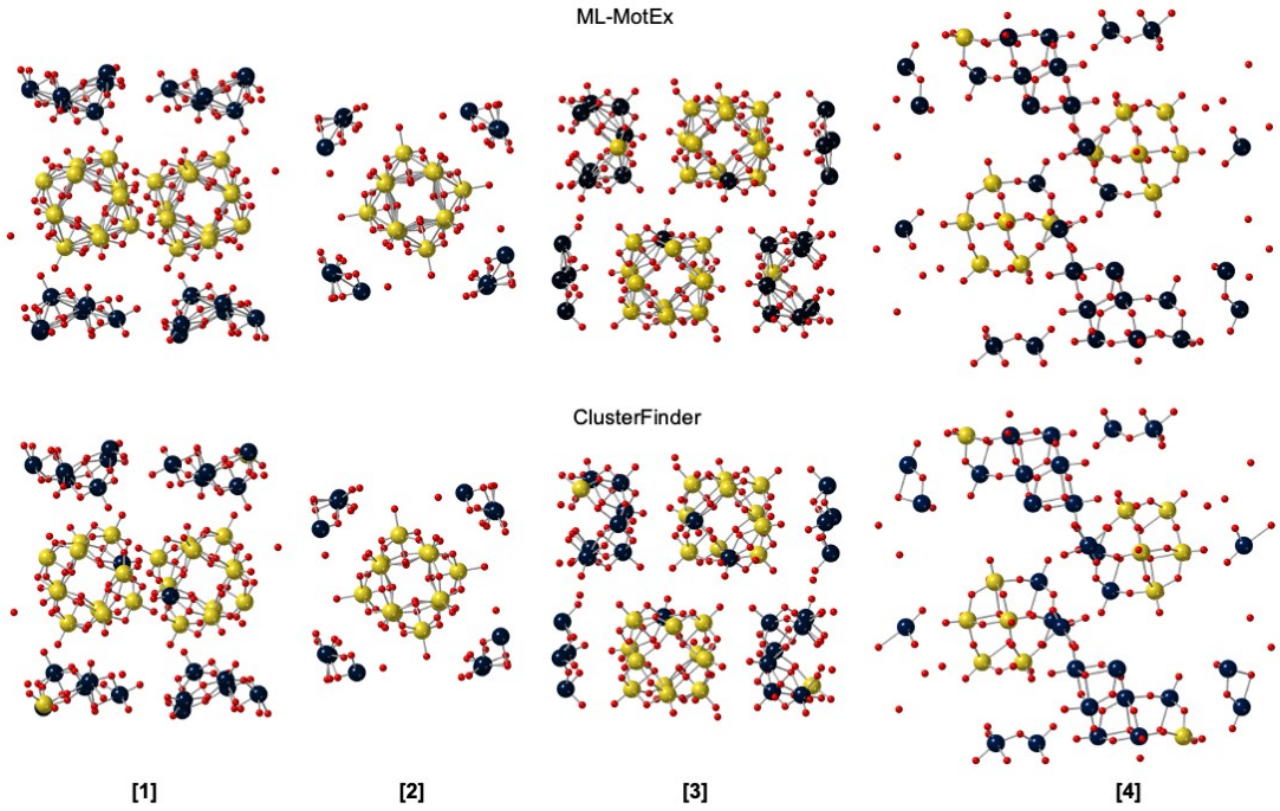

Fig. 4 | Comparison of the ML-MotEx- and ClusterFinder analysis of an experimental PDF obtained from Keggin clusters in solution. Results from the ML-MotEx- and ClusterFinder methods on a PDF obtained from a solution of ammonium metatungstate hydrate, using four different starting models: I) $[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine) *(Niu et al., 2004),* II) $(CH_3)_4N)_4SiW_{12}O_{40}$ *(Joachim et al., 1981),* III) $(((CH_3)_2NH_2)_6(Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$ *(Niu et al., 2003),* IV) $((CH_3)_2NH_2)_3(PW_{12}O_{40})$ *(Busbongthong & Ozeki, 2009).* The 24 ([1]+[3]+[4]) and 12 ([2]) atoms with the lowest atom-removal values have been coloured yellow, while the rest are coloured blue.

Screening the ICSD for a Suitable Starting Model with ClusterFinder

We now use ClusterFinder to scan the ICSD for the best-fitting structure models for the experimental PDF obtained from α-Keggin clusters in solution. ClusterFinder iteratively uses a supercell containing a single unit cell of each crystalline structure (188,631 structures) in the ICSD as the starting template. To accelerate the

ClusterFinder process, only the scale factor was refined, and structures without W, Fe and Mo atoms (158,399 structures), or supercells with over 1000 atoms (0 structures) were excluded. This left 29,070 candidate structures.

Afterwards, the template structures from crystals in the ICSD were ranked according to their average ΔR^{i}_{wp} value during the ClusterFinder process. The complete computation took ~17.5 min (1,046 seconds) on an AMD Ryzen Threadripper 3990X with 64-core 2.9/4.3GHz or 10 hrs. (34,882 s) on an Intel(R) Core™ i7-8665U CPU @ 1.9/2.11 GHz. Figure 5 demonstrates that all top five crystal structures contained the α-Keggin cluster. This demonstrates ClusterFinder's ability to effectively scan large structural databases, such as ICSD, for appropriate cluster structure. The five α-Keggin cluster structures are extracted from:

Ranked structure	Crystal composition	Reference
I)	$((CH_3)_4N)_6(Cu_{0.5}(H_2)_{0.5}O_4W_{12}O_{36})(H_2O)_{10}$	(Lunk *et al.*, 1993)
II)	$Cs_5(Cr_3O(OOCH)_6(H_2O)_3)(CoW_{12}O_{40})(H_2O)_2$	*(Uchida et al., 2006)*
III)	$(CH_3)_4N)_6(H_2W_{12}O_{40})(H_2O)_9$	(Asami *et al.*, 1984)
IV)	$Al_{13}O_4(OH)_{24}(H_2O)_{12})(H_2W_{12}O_{40})(OH)(H_2O)_{23.12}$	*(Son et al., 2003)*
V)	$K_2(H_2O)_4Eu\ (H_2O)_7(Eu(H_2O)_3HAlW_{11}O_{39})(H_2O)_7$	(Niu *et al.*, 2013)

Table 1 | Crystal composition of the top five candidate crystal structures ranked by ClusterFinder for the PDF obtained from α-Keggin clusters in solution.

ClusterFinder ranks supercells containing only essential cluster structures (in which no atoms need removal) over supercells containing both essential clusters and additional atoms. Consequently, the supercell generation influences the ranking of crystals in the ICSD. In instances where only essential clusters are present, the colour-coding still reflects the internal atomic ranking, even if all atoms are good and none requires removal. Figure 5 demonstrates this phenomenon; for instance, supercell (IV) contains only four essential α-Keggin clusters, with

no atoms needing removal. However, some atoms are coloured blue, as the colour bar merely signifies the internal atomic ranking. In the case of a supercell containing essential clusters with additional atoms, as seen in Figure 5, ClusterFinder indicates which atoms require removal.

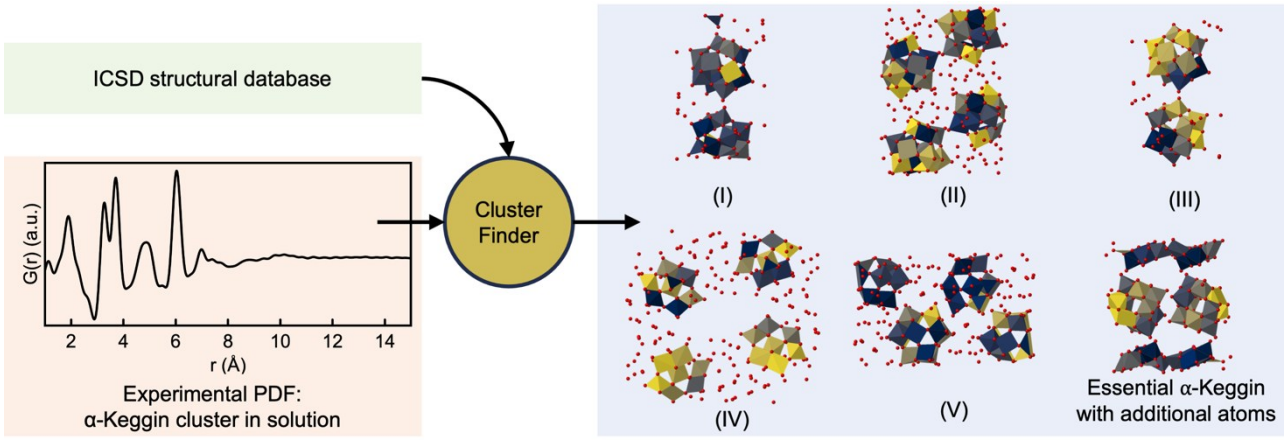

Fig. 5 | Illustration of how ClusterFinder is used to screen ICSD for the correct starting model for an experimental PDF obtained from α-Keggin clusters in solution. For each structure in the ICSD, the ClusterFinder procedure is performed, and the atoms are colour-coded based on their impact on fit quality using a continuous colour bar. Afterwards, the ICSD structures are sorted according to their average ΔR^{i}_{wp} value during the ClusterFinder process. The five candidates with the lowest average R_{wp} value are highlighted, along with an example of an essential α-Keggin structure with additional atoms.

ClusterFinder can also extract a cluster structure from a crystalline metal oxide structure. The ε-Keggin cluster serves as an excellent example of a cluster structure that can be directly cut out from a spinel structure. A PDF of an $Al_{12}O_{40}$ ε-Keggin cluster from the spinel $MgAl_2O_4$ crystal structure (Ji *et al.*, 2020) was calculated with parameters that mimic typical PDF dataset values, as seen in section A in the SI. Again, ClusterFinder was used iteratively to scan all ICSD structures. This time, crystals without W, Fe, Mo and Al atoms (143,956 structures) or supercells with more than 1,000 atoms (704 structures) were excluded. After evaluation, 42,809 structures

were ranked based on their average ΔR^{i}_{wp} value during the ClusterFinder process. The entire procedure takes ~42 min (2,495 seconds) on an AMD Ryzen Threadripper 3990X with 64-core 2.9/4.3GHz or ~23 hrs. (82,100 s) on an Intel(R) Core™ i7-8665U CPU @ 1.9/2.11 GHz. Figure 6 shows that the top five structures are all spinel structures:

Ranked structure	Crystal composition	Reference
I)	Al_2NiO_4	(Videnskaps-Akademi, 1925)
II)	Al_2MgO_4	(Zorina & Kvitka, 1968)
III)	$ZnAl_2O_4$	(Holgersson, 1927)
IV)	Al_2ZnO_4	(Videnskaps-Akademi, 1925)
V)	$ZnAl_2O_4$	*(Strukturuntersuchungen im System Al_2O_3–Cr_2O_3, 1964)*

Table 2 | Crystal composition of the top five candidate crystal structures ranked by ClusterFinder for the simulated PDF from the $Al_{12}O_{40}$ ε-Keggin cluster cut out from the spinel $MgAl_2O_4$ crystal structure.

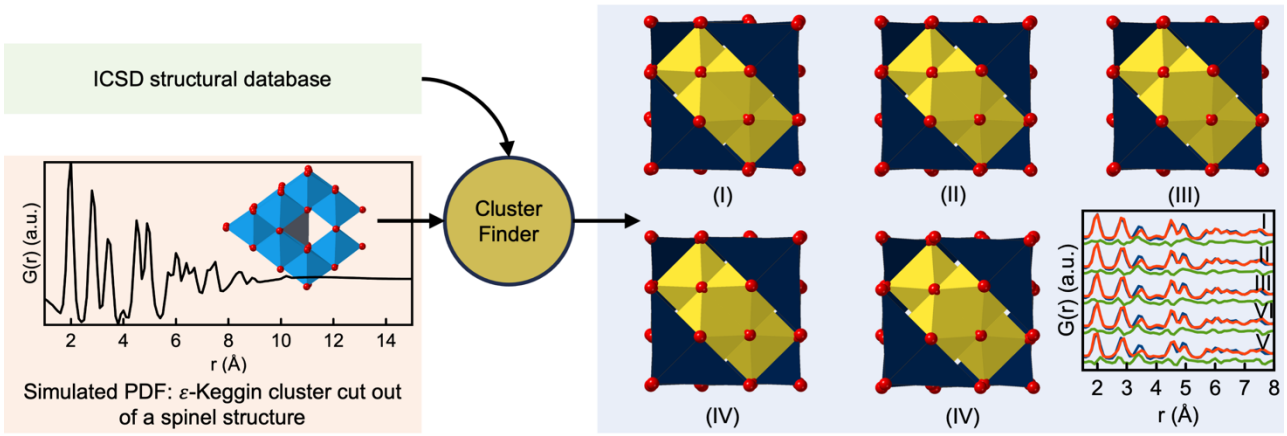

Fig. 6 | Illustration of how ClusterFinder is used to screen ICSD for the correct starting model of a simulated PDF obtained from a ε-Keggin cluster cut out of a spinel crystal. For each structure in the ICSD, the ClusterFinder procedure is performed, and the atoms are colour-coded based on their impact on the fit quality using a continuous colour bar. Afterwards, the ICSD structures are sorted according to their average ΔR^{i}_{wp} value

during the ClusterFinder process. The five candidates with the lowest R_{wp} value are highlighted. More extensive views of the PDF fits, including the calculated R_{wp} values, can be seen in section D in the SI.

We now proceed to apply ClusterFinder to a simulated PDF calculated from the ε-Keggin cluster cut out from a ε-Keggin crystal instead of a spinel crystal. As a result, the ε-Keggin cluster is less ordered. Specifically, we simulate a PDF of an $Al_{12}O_{40}$ ε-Keggin cluster cut out from a $(Al_{13}O_4(OH)_{24}(H_2O)_{12})_2(V_2W_4O_{19})_3(OH)_2(H_2O)_{27}$ crystal (Son & Kwon, 2004) with parameters mimicking typical values of an experimental PDF dataset, as seen in section A in the SI. The disorder can both be seen in the structures and their PDFs, where the PDF simulated from the ε-Keggin cluster cut out of the spinel structure exhibits more intense peaks than the PDF simulated from the ε-Keggin cluster cut out of the $(Al_{13}O_4(OH)_{24}(H_2O)_{12})_2(V_2W_4O_{19})_3(OH)_2(H_2O)_{27}$ crystal (Son & Kwon, 2004).

Again, we use ClusterFinder iteratively on all ICSD structures containing W, Fe, Mo and Al atoms and rank the structure based on their average ΔR^i_{wp} value during the ClusterFinder process. Figure 7 shows that the top five structures are mainly ε-Keggin clusters or crystal variants of the spinel structure (structure III and V):

Ranked structure	Crystal composition	Reference
I)	$(Al_{13}O_4(OH)_{24}(H_2O)_{12})(H_2W_{12}O_{40})(OH)(H_2O)_{23.12}$	(Son *et al.*, 2003)
II)	$(Al_{13}O_4(OH)_{24}(H_2O)_{12})(CoW_{12}O_{40})(OH)(H_2O)_{20}$	(Son *et al.*, 2003)
III)	$Ca_2Mg_2Fe_2(Al_{14}O_{31}(OH))(Al_2O)(Al)(Al(OH))$	(Rastsvetaeva *et al.*, 2010)
IV)	$((GeO_4)Al_{12}(OH)_{24}(H_2O)_{12})(SeO_4)_4(H_2O)_{14}$	(Lee *et al.*, 2001)
V)	$(Al_2O_3)_{13}(SO_3)_6(H_2O)_{79}$	(Nordstrom, 1982)

Table 3 | Crystal composition of the top five candidate crystal structures calculated by ClusterFinder for the simulated PDF from the ε-Keggin cluster cut out of the $Al_{12}O_{40}$ $(Al_{13}O_4(OH)_{24}(H_2O)_{12})_2(V_2W_4O_{19})_3(OH)_2(H_2O)_{27}$ crystal(Son & Kwon, 2004).

ClusterFinder's sensitivity to minor changes in the PDF suggests that it can potentially be used to estimate similarities between ionic cluster structures. In conclusion, two distinct variations of ε-Keggin can be observed; The ε-Keggin cluster in its crystallised form as ε-Keggin crystals, or the ε-Keggin cluster extracted from related metal oxide structures, such as the spinel structure. The main differentiator is the degree of disorder present in each crystal representation. ClusterFinder can discern between the more ordered spinel-obtained motifs and the more distorted Keggin crystal structure. It highlights the level of detailed description attained in this modelling approach. Additionally, the supercell structure (I) and (II) in Figure 7 demonstrate that ClusterFinder can differentiate between α-Keggin (blue) and ε-Keggin clusters (yellow).

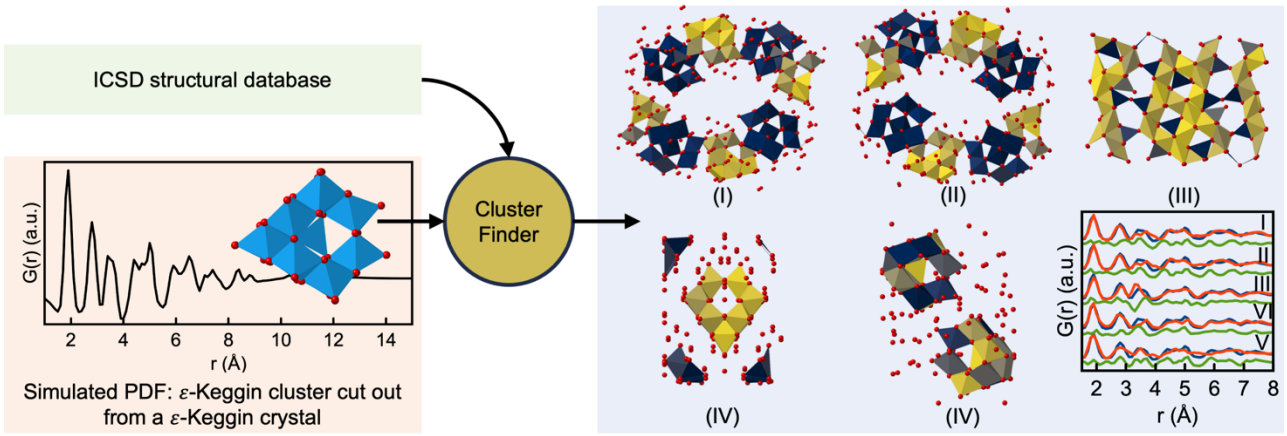

Fig. 7 | Illustration of how ClusterFinder is used to screen ICSD for the correct starting model of a simulated PDF obtained from a ε-Keggin cluster cut out of an ε-Keggin crystal. For each structure in the ICSD, the ClusterFinder procedure is performed, and the atoms are colour-coded based on their impact on the fit quality using a continuous colour bar. Afterwards, the ICSD structures are sorted according to their average

ΔR^{i}_{wp} value during the ClusterFinder process. The five candidates with the lowest R_{wp} value are highlighted. More extensive views of the PDF fits, including the calculated R_{wp} values, can be seen in section E in the SI.

In section F and G in the SI, we present two similar examples in which we rank the ICSD structures according to experimental datasets obtained from ionic [$Bi_{38}O_{45}$] clusters and ceria (CeO_2) nanoparticles. We find that the highest ranked structures from the [$Bi_{38}O_{45}$] cluster example are δ-Bi_2O_3 crystal structures, as previously observed by Weber et al. (Weber *et al.*, 2017). For the ceria nanoparticles, the highest ranked structures correspond to the fluorite structure.

Conclusions

We have introduced a new automated structure selection approach called ClusterFinder for extracting cluster motifs from PDF data and identifying suitable starting models for refining PDFs of nanoclusters. We have demonstrated the effectiveness of ClusterFinder on simulated and experimental PDFs obtained from POM and ionic clusters. ClusterFinder is inspired by our previously developed algorithms, LIGA and ML-MotEx, but is significantly faster, facilitating the screening of large databases in minutes. Our study demonstrates ClusterFinder's efficacy as a robust tool for extracting appropriate starting models from extensive structural databases like the ICSD for experimental PDF analysis. By applying ClusterFinder to diverse scenarios, such as α-Keggin clusters, ε-Keggin clusters, ionic [$Bi_{38}O_{45}$] clusters, and ceria nanoparticles, we showcase its ability to effectively rank and select the most relevant structures based on fitting quality.

Our findings reveal ClusterFinder's sensitivity to subtle variations in PDFs, indicating its potential use in estimating similarities among ionic cluster structures. It can also differentiate between varying degrees of disorder in crystal structures, as illustrated by the contrast between more ordered spinel-derived motifs and more distorted Keggin crystal structures.

References

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. (2021). *Angew. Chem. Int. Ed.* **60**, 2-12.

Anker, A. S., Kjær, E. T. S., Dam, E. B., Billinge, S. J. L., Jensen, K. M. Ø. & Selvan, R. (2020). *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).

Anker, A. S., Kjær, E. T. S., Juelsholt, M., Christiansen, T. L., Skjærvø, 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.

Asami, M., Ichida, H. & Sasaki, Y. (1984). *Acta Crystallogr. C* **40**, 35-37.

Banerjee, S., Liu, C.-H., Jensen, K. M. Ø., Juhas, P., Lee, J. D., Tofanelli, M., Ackerson, C. J., Murray, C. B. & Billinge, S. J. L. (2020). *Acta Crystallogr. A* **76**, 24-31.

Benseghir, Y., Lemarchand, A., Duguet, M., Mialane, P., Gomez-Mingot, M., Roch-Marchal, C., Pino, T., Ha-Thi, M.-H., Haouas, M., Fontecave, M., Dolbecq, A., Sassoye, C. & Mellot-Draznieks, C. (2020). *J. Am. Chem. Soc.* **142**, 9428-9438.

Billinge, S. J. L. & Levin, I. (2007). *Science* **316**, 561-565.

Busbongthong, S. & Ozeki, T. (2009). *Bull. Chem. Soc. Jpn.* **82**, 1393-1397.

Chen, X. & Yamanaka, S. (2002). *Chem. Phys. Lett.* **360**, 501-508.

Christiansen, T. L., Cooper, S. R. & Jensen, K. M. Ø. (2020). *Nanoscale Adv.* **2**, 2234-2254.

Cliffe, M. J., Dove, M. T., Drabold, D. & Goodwin, A. L. (2010). *Phys. Rev. Lett.* **104**, 125501.

Cliffe, M. J. & Goodwin, A. L. (2013). *J. Phys.: Condens. Matter* **25**, 454218.

Coelho, A. A. (2018). *J. Appl. Cryst.* **51**, 210-218.

Egami, T. & Billinge, S. J. L. (2012). *Underneath the Bragg Peaks, Pergamon*.

Farrow, C. L., Juhas, P., Liu, J. W., Bryndin, D., Božin, E. S., Bloch, J., Th, P. & Billinge, S. J. L. (2007). *J. Phys.: Condens. Matter* **19**, 335219.

Holgersson, S. (1927). *Lunds universitets årsskrift. NF Avd* **2**, 1-9.

Ji, H., Hou, X., Molokeev, M. S., Ueda, J., Tanabe, S., Brik, M. G., Zhang, Z., Wang, Y. & Chen, D. (2020). *Dalton Trans.* **49**, 5711-5721.

Joachim, F., Axel, T. & Rosemarie, P. (1981). *Z. Naturforsch.* **36**, 161-171.

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., Cherba, D. M., Duxbury, P. M., Punch, W. F. & Billinge, S. J. L. (2006). *Nature* **440**, 655-658.

Juhás, P., Farrow, C. L., Yang, X., Knox, K. R. & Billinge, S. J. L. (2015). *Acta Crystallogr. A* **71**, 562-568.

Juhás, P., Granlund, L., Duxbury, P. M., Punch, W. F. & Billinge, S. J. L. (2008). *Acta Crystallogr. A* **64**, 631-640.

Juhas, P., Granlund, L., Gujarathi, S. R., Duxbury, P. M. & Billinge, S. J. L. (2010). *J. Appl. Cryst.* **43**, 623-629.

Kjær, E. T. S., Anker, A. S., Weng, M. N., Billinge, S. J. L., Selvan, R. & Jensen, K. M. Ø. (2023). *Digital Discovery* **2**, 69-80.

Kløve, M., Sommer, S., Iversen, B. B., Hammer, B. & Dononelli, W. (2023). *Adv. Mater.* **35**, 2208220.

Lee, A. P., Phillips, B. L., Olmstead, M. M. & Casey, W. H. (2001). *Inorg. Chem.* **40**, 4485-4487.

Lunk, H.-J., Giese, S., Fuchs, J. & Stösser, R. (1993). *Zeitschrift für anorganische und allgemeine Chemie* **619**, 961-968.

Niu, J., Zhao, J., Wang, J. & Bo, Y. (2004). *J. Coord. Chem.* **57**, 935-946.

Niu, J.-Y., Han, Q.-X. & Wang, J.-P. (2003). *J. Coord. Chem.* **56**, 523-530.

Niu, L., Li, Z., Xu, Y., Sun, J., Hong, W., Liu, X., Wang, J. & Yang, S. (2013). *ACS Appl. Mater. Inter.* **5**, 8044-8052.

Nordstrom, D. K. (1982). *Geochim. Cosmochim. Acta* **46**, 681-692.

Poimanova, O. Y., Radio, S. V., Bilousova, K. Y., Baumer, V. N. & Rozantsev, G. M. (2015). *J. Coord. Chem.* **68**, 1-17.

Rastsvetaeva, R., Aksenov, S. & Verin, I. (2010). *Crystallogr. Rep.* **55**, 563-568.

Redrup, K. V. & Weller, M. T. (2009). *Dalton Trans.*, 4468-4472.

Son, J.-H. & Kwon, Y.-U. (2004). *Inorg. Chem.* **43**, 1929-1932.

Son, J. H., Kwon, Y.-U. & Han, O. H. (2003). *Inorg. Chem.* **42**, 4153-4159.

Strukturuntersuchungen im System Al_2O_3–Cr_2O_3 (1964). **120**, 342-348.

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.*

Uchida, S., Kawamoto, R. & Mizuno, N. (2006). *Inorg. Chem.* **45**, 5136-5144.

Van den Eynden, D., Pokratath, R., Mathew, J. P., Goossens, E., De Buysser, K. & De Roo, J. (2023). *Chem. Sci.* **14**, 573-585.

Videnskaps-Akademi, O. D. N. (1925). *Avhandlinger/Norske Videnskaps-Akademi, Matematisk-Naturvidenskapelig Klasse*. Dybwad [in Komm.].

Weber, M., Schlesinger, M., Walther, M., Zahn, D., Schalley, C. A. & Mehring, M. (2017). *Zeitschrift für Kristallographie-Crystalline Materials* **232**, 185-207.

Yang, L., Juhás, P., Terban, M. W., Tucker, M. G. & Billinge, S. J. L. (2020). *Acta Crystallogr. A* **76**, 395-409.

Zagorac, D., Müller, H., Ruehl, S., Zagorac, J. & Rehme, S. (2019). *J. Appl. Cryst.* **52**, 918-925.

Zorina, N. & Kvitka, S. (1968). *Kristallografiya* **13**, 703-705.

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/AndySAnker/ClusterFinder. Additional data that support the findings of this study are available from the corresponding authors upon request.

Code availability

The authors declare that the code supporting this study are available on the associated Github to the paper: https://github.com/AndySAnker/ClusterFinder. Additional code that supports the findings of this study are available from the corresponding authors upon request.

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). Work in the Billinge group was supported by the U.S. National Science Foundation through grant DMREF-1922234. 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 acknowledge MAX IV Laboratory for time on Beamline DanMAX under Proposal 20200731. 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 at Petra III, and we thank Martin Etter, Jozef Bednarcik for assistance in using the beamline

Author contributions

ASA contributed to all aspects of the paper. ASA, UFJ and FLJ wrote the code. KMØJ procured funding. SJLB and KMØJ supervised the project. All authors contributed to the writing of the manuscript.

Competing interests

The authors declare no competing interests.