Acta Cryst A FOUNDATIONS ADVANCES

ISSN 2053-2733

ClusterFinder: a fast tool to find cluster structures from pair distribution function data

Andy S. Anker,[a] Ulrik Friis-Jensen,[a,b] Frederik L. Johansen,[a,b] Simon J. L Billinge[c]* and Kirsten M. Ø. Jensen[a]*

[a]Department of Chemistry and Nano-Science Center, University of Copenhagen, 2100 Copenhagen Ø, Denmark, [b]Department of Computer Science, University of Copenhagen, 2100 Copenhagen Ø, Denmark, and [c]Department of Applied Physics and Applied Mathematics Science, Columbia University, New York, NY 10027, USA. *Correspondence e-mail: sb2896@columbia.edu, kirsten@chem.ku.dk

Received 22 November 2023
Accepted 1 February 2024

Edited by T. E. Gorelik, Helmholtz Centre for Infection Research, Germany

Keywords: pair distribution function analysis; nanoclusters; nanomaterials; screening.

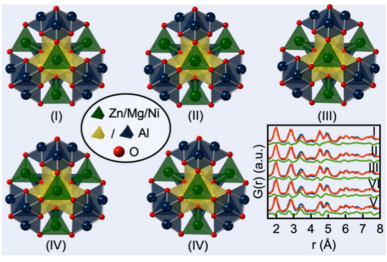
Supporting information: this article has supporting information at journals.iucr.org/a

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 nanoclusters or small nanoparticles. The reported *ClusterFinder* algorithm can screen 10^4 to 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 has performed well for simulated and measured PDFs of metal–oxido clusters such as Keggin clusters. This is therefore a powerful approach to finding structural cluster candidates in a modelling campaign for PDFs of nanoparticles and nanoclusters.

1. 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). Total scattering with pair distribution function (PDF) analysis has shown promise for characterizing such nanomaterials (Billinge & Levin, 2007; Christiansen *et al.*, 2020). The PDF, derived from the Fourier transform of normalized and corrected X-ray, neutron or electron scattering intensities, offers a real-space representation of interatomic distances in the sample. As the data used in the Fourier transform include both Bragg and diffuse scattering, PDF analysis can be used to characterize the structure of materials with or without long-range atomic order (Egami & Billinge, 2012; Christiansen *et al.*, 2020).

The challenge of *ab initio* structure solution from PDFs has long been pursued (Juhás *et al.*, 2006, 2008, 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 simple inorganic crystals, 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. For crystalline or nanocrystalline materials, such starting models can, in many cases, easily be identified from crystallographic databases. However, this task becomes exceptionally difficult for small clusters and nanomaterials with significant disorder. Recent methods such as *ClusterMining* (Banerjee *et al.*, 2020), *StructureMining* (Yang *et al.*,



2020) and *POMFinder* (Anker *et al.*, 2024) 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.

We recently presented a hybrid approach, *ML-MotEx* (Anker *et al.*, 2022), where the user initially selects candidate crystal structures from a crystallographic database. Explainable machine learning is then used to find sub-clusters from the candidate structure that are consistent with the data, which can then be used for further structure refinement and analysis. The approach works well but is slow, taking several minutes for each starting structure. This limits its application to cases where the candidate parent crystal structures are 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 from minutes 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.

2. Method

The basic strategy for finding clusters from crystalline fragments was described by 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 atomic arrangement in highly disordered materials can also sometimes be described using cluster structures (Du *et al.*, 2012; Castillo-Blas *et al.*, 2020; Christiansen *et al.*, 2020). The resulting measured PDF has a small number of peaks confined to the low-r region, indicating the presence of unknown atomic clusters of small size (see Fig. 1).

In principle, the data can be fitted using the Debye scattering equation in programs such as *DISCUS* (Proffen & Neder, 1997, 1999) 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 a set of good starting models for the fit. *ClusterFinder* addresses this need. It reuses the approach taken by *ML-MotEx* (Anker *et al.*, 2022) where a set of chemically reasonable crystal structures is first identified. From the crystal structures, which are represented using crystallographic information files (CIFs), candidate templates are then cut out. The candidate templates are represented in *xyz* format (a list of atomic identities and their respective Cartesian coordinates x, y and z). Assuming for now that the cluster present in the experimental data, the target cluster, is contained within the candidate template, the principal goal is to find the subset of occupied atom sites within that template that corresponds to the target cluster. A search over all possible permutations of present versus absent atoms is impossible because of the combinatorics, with $2^N - 1$ possibilities for a template of N sites. *ML-MotEx* uses an explainable machine learning approach to optimize this problem by learning the probabilities that each atom might be present in the target cluster after iterating over a small subset of all the possible permutations. This places the atom sites in a rank-ordered list and makes it easy for the user to select a cut-off for which sites are occupied 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 all possible candidate crystal structures and templates. The *ML-MotEx* algorithm is too slow to do this over many template candidates and the success of the approach relies on a strong chemical intuition suggesting a small number of candidate structures.

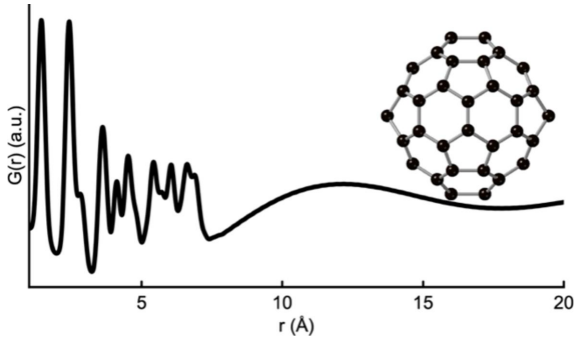


Figure 1
A simulated PDF for a C_{60} buckyball from a single unit cell of a C_{60} crystal structure (Chen & Yamanaka, 2002). The simulation parameters mimic typical PDF dataset values and can be seen in Section A in the supporting information.

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, 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,

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

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

Taking inspiration from the *LIGA* algorithm, 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 ith atom by $\Delta R^i_{\rm wp}$. It is computed by evaluating $R_{\rm wp}$ for the full set of atoms, then recomputing $R_{\rm wp}$ for the cluster with the ith 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 computation of $R_{\rm wp}$, a scale factor and an isotropic expansion/contraction factor are allowed to be

refined to give the best agreement. Atomic displacement parameters (ADPs) were fixed to 0.3 Å^2 for the metallic atoms and 0.4 Å^2 for the oxygen atoms. This procedure is extremely rapid and results in a list of atomic sites ranked by ΔR^i_{wp}.

The candidate structure must be large enough to encapsulate the target cluster, but the computational cost scales linearly with the number of atoms of the template structure and so the cluster size chosen can thus be a compromise between time and the cluster structures screened.

To visualize 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^i_{wp} are coloured blue. The colour coding is further explained in Section B in the supporting information. The approach is illustrated schematically for a trivial example of a small cluster consisting of two atoms 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 in the subsequent task of finding the best cluster candidates. In Fig. 2, it is trivial to remove atoms 3 and 4 from the *ClusterFinder* output, but this task might not always be trivial and may rely on the 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.

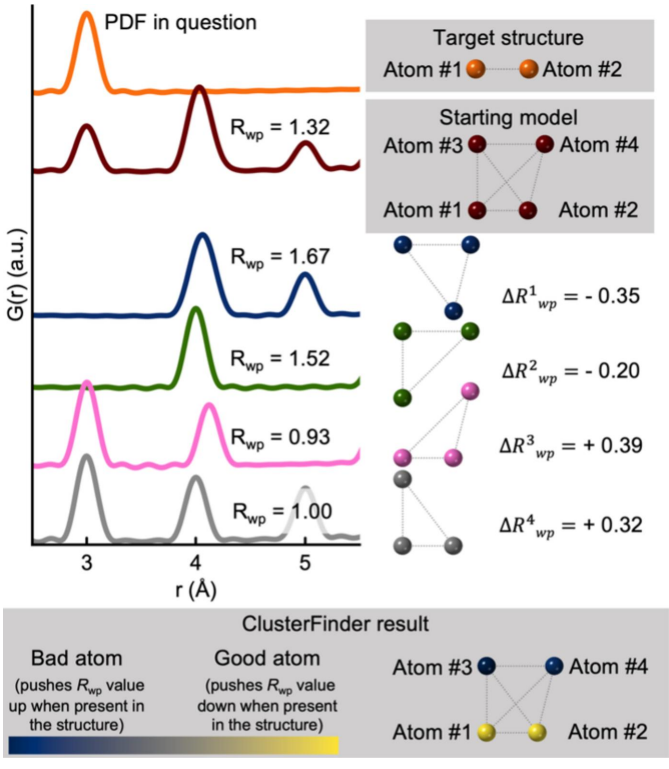


Figure 2
An illustration of the *ClusterFinder* process. A starting model is provided as input and the R_{wp} value is calculated by structure refinement. Atoms are iteratively removed from the starting model and the revised model is fitted to the experimental PDF. The atom-removal error ΔR^i_{wp} is calculated by taking the difference between the R_{wp} values of the full starting model and when the atoms are removed. Atoms are colour coded based on the atom-removal error – yellow indicates a negative ΔR^i_{wp} value (improved fit) while blue signifies a positive ΔR^i_{wp} value (worsened fit).

To test the *ClusterFinder* approach, we here use it on simulated and experimental PDF data. *ClusterFinder* provides comparable results to *ML-MotEx* in quality but orders of magnitude more quickly. The acceleration is sufficient to allow screening of large databases of starting models in minutes. To demonstrate the power of this, we provide five examples where we screen the Inorganic Crystal Structure Database (ICSD, https://icsd.fiz-karlsruhe.de/index.xhtml; Zagorac *et al.*, 2019), containing 188 631 structure entries, for a suitable starting model. This is done in a time frame ranging from 3 to 42 min. 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.

3. Results and discussion

3.1. Applying *ClusterFinder* to extract cluster motifs from simulated PDFs

We first demonstrate *ClusterFinder*'s ability to extract cluster motifs from simulated PDFs. Fig. 3 shows three simulated PDFs, each corresponding to a distinct structure: a decatungstate polyoxometallate cluster from an $Na_5(H_7W_{12}O_{42})(H_2O)_{20}$ crystal structure (Redrup & Weller, 2009), coloured in blue; a C_{60} buckyball from the C_{60} crystal structure (Chen & Yamanaka, 2002), coloured in green; and a paratungstate polyoxometallate 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 are listed in Section A in the supporting information. Figs. 3(*b*)–3(*d*) show the structural templates used by *ClusterFinder*.

In these tests, the structural templates were constructed using the crystal structures containing each of the cluster structures, and including the minimum number of unit cells needed to include the full cluster (Section C in the supporting information). *ClusterFinder* outputs a list of atomic sites ranked by the ΔR^i_{wp} value, and we visualize atom sites with negative ΔR^i_{wp} as yellow and those with positive ΔR^i_{wp} as blue. The ranking is here done on the metal atoms, while oxygen atoms are removed if they are beyond a distance threshold of 2.6 Å from any other atom. The resulting visualizations are shown in Figs. 3(*b*)–3(*d*), where the atoms with the lowest ΔR^i_{wp} values have been coloured yellow, while the rest are coloured blue. Section C in the supporting information shows a similar representation but where the atom-removal values are directly shown using a continuous colour bar. Oxygen atoms are coloured red and polyhedra are coloured according to their metal atom centre.

ClusterFinder correctly extracted all three cluster structures from their starting model in under a minute using a standard laptop (Intel Core i7-8665U CPU at 1.9/2.11 GHz), demonstrating a significant speed advantage over the *ML-MotEx* algorithm (Anker *et al.*, 2022), which takes approximately an

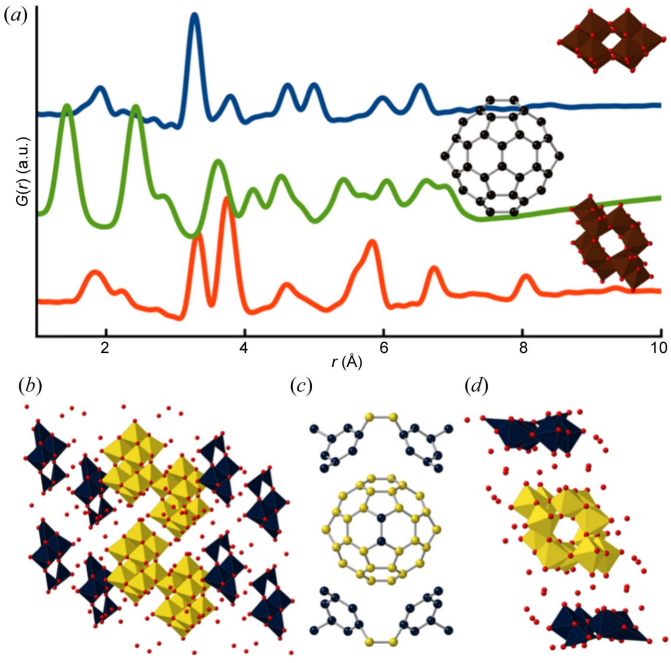


Figure 3
Analysis of simulated PDFs of well known cluster structures. (*a*) Simulated PDFs of a decatungstate polyoxometallate cluster from the $Na_5(H_7W_{12}O_{42})(H_2O)_{20}$ crystal structure (blue) (Redrup & Weller, 2009), a C_{60} buckyball from a single unit cell of a C_{60} crystal structure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

hour on the same computer. Although *ClusterFinder* accurately extracts the decatungstate polyoxometallate cluster (blue) and the paratungstate polyoxometallate 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. Note that while *ClusterFinder* is faster than *ML-MotEx*, the latter algorithm is more versatile and has, for example, also been used to determine stacking fault size domain distributions from experimental powder diffraction and PDF data from γ-MnO_2 nanoparticles (Magnard *et al.*, 2022).

3.2. 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 must also work on experimental data. 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.

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

Starting model	Crystal composition	Reference
1	$[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine)	Niu *et al.* (2004)
2	$[(CH_3)_4N]_4SiW_{12}O_{40}$	Joachim *et al.* (1981)
3	$([(CH_3)_2NH_2]_6\{Cu[HCON(CH_3)_2]_4\}\text{-}(GeW_{12}O_{40})_2)[HCON(CH_3)_2]_2$	Niu *et al.* (2003)
4	$[(CH_3)_2NH_2]_3(PW_{12}O_{40})$	Busbongthong & Ozeki (2009)

An 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.

Again, only a scale factor and an isotropic expansion/contraction factor were refined during the *ClusterFinder* process. As seen in Fig. 4, both *ClusterFinder* and *ML-MotEx* successfully extracted the α-Keggin clusters with few mislabelled atoms for all four starting models. *ClusterFinder* has slightly more mislabelled atoms than *ML-MotEx*, but it is orders of magnitude faster, making it an ideal choice for screening larger databases.

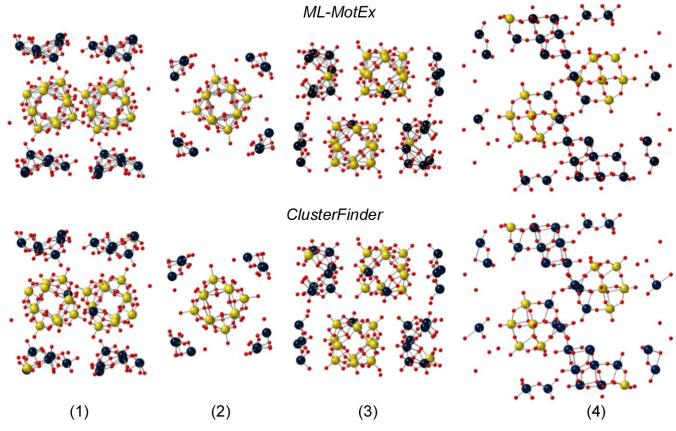


Figure 4
Comparisons of the *ML-MotEx* and *ClusterFinder* analyses of an experimental PDF obtained from Keggin clusters in solution. Results are given from the *ML-MotEx* and *ClusterFinder* methods on a PDF obtained from a solution of ammonium metatungstate hydrate using four different starting models,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

3.3. Screening the ICSD for a suitable starting model with *ClusterFinder*

We now use *ClusterFinder* to scan the whole ICSD for the best-fitting structure models for the experimental PDF obtained from α-Keggin clusters in solution. *ClusterFinder* uses a single unit cell of each crystal structure (188 631 structures, although we removed unreadable CIFs making it 187 469 structures) in the ICSD as the starting template. To accelerate the *ClusterFinder* process, only the scale factor was refined, and structures without W, Fe or Mo atoms (158 399 structures), or starting templates with over 1000 atoms (zero structures) were excluded. This left 29 070 candidate structures. For database screening, an isotropic contraction/expansion factor was not refined. 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 (1046 s) on an AMD Ryzen Threadripper 3990X with 64 cores at 2.9/4.3 GHz, or 10 h (34 882 s) on a standard laptop (Intel Core i7-8665U CPU at 1.9/2.11 GHz). Fig. 5 demonstrates that all of the top five crystal structures (Table 2) contained the α-Keggin cluster. This shows *ClusterFinder*'s ability to scan large structural databases effectively, such as the ICSD, for appropriate cluster structures.

ClusterFinder prioritizes starting templates exclusively comprising the essential cluster structure, *i.e.* clusters in which no atoms need removal and that thereby inherently match their target cluster, over those that contain additional atoms. Consequently, the starting template generation influences the ranking of crystal structures in the ICSD. In instances where exclusively essential clusters are present, the colour coding still reflects the internal atomic ranking, even if all atoms are good and none requires removal. Fig. 5 demonstrates this phenomenon; for instance, starting template (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 starting template containing essential clusters with additional atoms, as seen in Fig. 5, *ClusterFinder* indicates which atoms require removal.

Table 2
Crystal composition of the top five candidate crystal structures ranked by *ClusterFinder* for the PDF obtained from α-Keggin clusters in solution.

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)

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 simulated with parameters that mimic typical PDF dataset values, as seen in Section A in the supporting information. The PDF and structure are illustrated in Fig. 6. Again, *ClusterFinder* was used to scan all structures in the ICSD. This time, crystal structures without W, Fe, Mo or Al atoms (143 956 structures) or starting templates with more than 1000 atoms (704 structures) were excluded. After evaluation, 42 809 structures were ranked based on their average ΔR^i_{wp} value found during the *ClusterFinder* process. The entire procedure takes 42 min (2495 s) on an AMD Ryzen

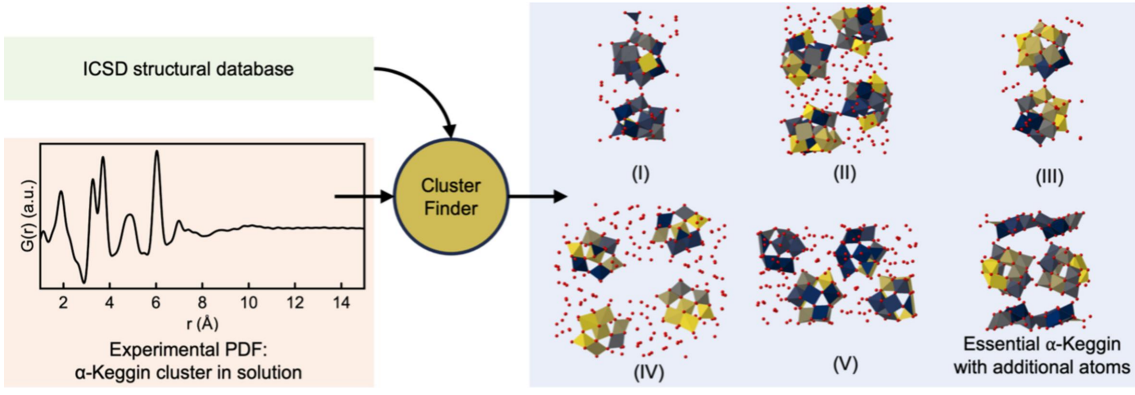


Figure 5
An illustration of how *ClusterFinder* is used to screen the ICSD for the correct starting model for an experimental PDF obtained from α-Keggin clusters in solution. For each structure in the ICSD, the *ClusterFinder* procedure is performed, and the atoms are colour coded based on their impact on fit quality using a continuous colour bar. Afterwards, the ICSD structures are sorted according to their average ΔR^i_{wp} values. The five candidate ICSD structures with the lowest average R_{wp} value are highlighted. The top five candidates are all starting templates exclusively comprising essential cluster structures – clusters in which no atoms need removal and that thereby inherently match their target cluster. An example of an essential α-Keggin structure with additional atoms (non-essential structure) is shown to exemplify that *ClusterFinder* provides meaningful atomic rankings of non-essential structures. Oxygen atoms are coloured red. Atoms different from W, Fe, Mo or O are omitted for clarity.

Table 3
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.

Ranked structure	Crystal composition	Reference
(I)	$NiAl_2O_4$	Vegard & Borlaug (1943)
(II)	$MgAl_2O_4$	Zorina & Kvitka (1968)
(III)	$ZnAl_2O_4$	Holgersson (1927)
(IV)	$ZnAl_2O_4$	Vegard & Borlaug (1943)
(V)	$ZnAl_2O_4$	Saalfeld (1964)

Table 4
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 structure (Son & Kwon, 2004).

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)

Threadripper 3990X with 64 cores at 2.9/4.3 GHz or 23 h (82 100 s) on a standard laptop (Intel Core i7-8665U CPU at 1.9/2.11 GHz). The top five structures, shown in Fig. 6, are all spinel structures.

We now proceed to apply *ClusterFinder* to a simulated PDF calculated from the ε-Keggin cluster cut out from an ε-Keggin crystal structure {here $[Al_{13}O_4(OH)_{24}(H_2O)_{12}]_2(V_2W_4O_{19})_3(OH)_2(H_2O)_{27}$; Son & Kwon, 2004} instead of a cut out from the spinel crystal structure. The ε-Keggin obtained in this way is more disordered than that cut out from the spinel crystal structure. The disorder can be seen in both the structures and their PDFs (Figs. 6 and 7), where the PDF simulated from the spinel-derived ε-Keggin (Fig. 6) exhibits sharper 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 structure (Son & Kwon, 2004) (Fig. 7). Again, we use *ClusterFinder* on all ICSD structures containing W, Fe, Mo or Al atoms one by one. Afterwards, it ranks the structures based on their average ΔR^i_{wp} value obtained during the *ClusterFinder* process. Fig. 7 and Table 4 show that the top five structures mainly contain ε-Keggin clusters or are variants of the spinel structure [structures (III) and (V)]. While α-Keggin and ε-Keggin clusters are very similar and only distinct in the different rotational orientations of their four M_3O_{13} units, *ClusterFinder* is able to differentiate between them in starting template structures (I) and (II) where the α-Keggin motif is removed (blue) and the ε-Keggin motifs are kept (yellow).

ClusterFinder can, moreover, discern between the more ordered spinel-obtained motifs (Fig. 6 and Table 3) and the more distorted Keggin crystal structure (Fig. 7 and Table 4), which demonstrates that it is sensitive to minor changes in the PDF. This highlights the level of detailed description attained in this modelling approach.

In Sections F and G in the supporting information, 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.* (2017). For the ceria nanoparticles, the highest ranked structures correspond to bixbyite-type structures, which are related to the fluorite-type structure that CeO_2 would be expected to take. This demonstrates that, while *ClusterFinder* often provides results closely related to the true chemical solution, validation and considerations of

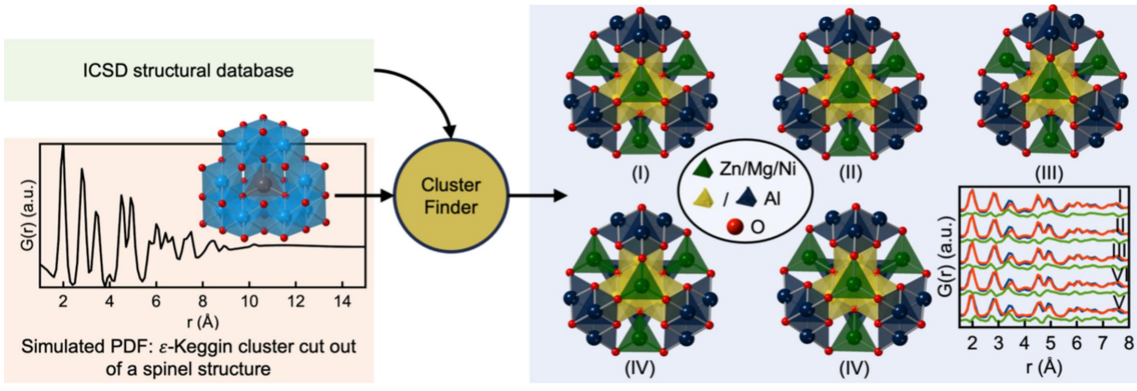


Figure 6
An illustration of how *ClusterFinder* is used to screen the ICSD for the correct starting model for a simulated PDF obtained from an ε-Keggin cluster cut out of a spinel crystal structure (coloured light blue in the left of the figure with Mg in the centre). For each structure in the ICSD, the *ClusterFinder* procedure is performed and the atoms are colour coded based on their impact on the fit quality. Afterwards, the ICSD structures are sorted according to their average ΔR^i_{wp} values during the *ClusterFinder* process. The five candidates with the lowest R_{wp} values are highlighted. More extensive views of the PDF fits, including the calculated R_{wp} values, can be seen in Section D in the supporting information. Atoms different from W, Fe, Mo, Al or O have been omitted for clarity.

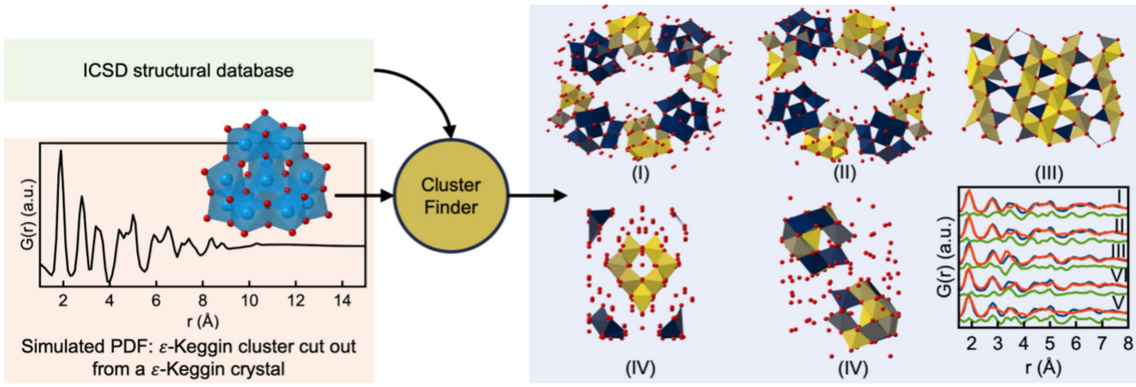


Figure 7
An illustration of how *ClusterFinder* is used to screen the ICSD for the correct starting model for a simulated PDF obtained from an ε-Keggin cluster cut out of an ε-Keggin crystal structure (coloured light blue in the left of the figure). For each structure in the ICSD, the *ClusterFinder* procedure is performed and the atoms are colour coded based on their impact on the fit quality. Afterwards, the ICSD structures are sorted according to their average ΔR^{i}_{wp} values 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 supporting information. Oxygen atoms are coloured red. Other atoms than W, Fe, Mo, Al or O have been omitted for clarity.

structure relations are still required in the data analysis process.

4. Conclusions

We have introduced a new automated structure selection approach called *ClusterFinder* for identifying suitable starting models for analysis and refinement of PDFs from nanoclusters. The premise of *ClusterFinder* is that the structure of a nanocluster can probably be described as a fragment of an already published crystal structure, and it thus screens crystal structures and identifies fragments for further analysis. The structure found by *ClusterFinder* is not necessarily a unique solution to the PDF, but *ClusterFinder*'s automated process ensures a systematic and extensive screening of a range of possible structures.

ClusterFinder is inspired by our previously developed algorithms, *LIGA* and *ML-MotEx*, but is significantly faster, facilitating screening of large databases for cluster identification in minutes. Our study demonstrates *ClusterFinder*'s efficacy as a robust tool for extracting appropriate starting models from extensive structural databases like the ICSD. By applying *ClusterFinder* to PDFs from various nanoclusters, such as α-Keggin clusters, ε-Keggin clusters, ionic $[Bi_{38}O_{45}]$ clusters and ceria nanoparticles, we have showcased its abilities in effectively ranking and selecting the most relevant structure models based on fit quality.

All the data supporting this study are available either within the paper, as supporting information or on the associated GitHub to the paper, https://github.com/AndySAnker/ClusterFinder. The code supporting this study is also available on the same associated GitHub.

5. Related literature

For further literature related to the supporting information, see Anker *et al.* (2021), Artini *et al.* (2014), Chakraborty *et al.* (2006), Coduri *et al.* (2013), Estes *et al.* (2016), Juhás *et al.* (2013), Labidi *et al.* (2008), Rademacher *et al.* (2001), Radosavljević-Evans *et al.* (2002), Sasaki *et al.* (2004) and Yang *et al.* (2014).

Acknowledgements

We acknowledge the 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 on beamline P02.1 at PETRA III, and we thank Martin Etter and Jozef Bednarcik for assistance in using the beamline. Author contributions are as follows: ASA contributed to all aspects of the paper; ASA, UFJ and FLJ wrote the code; KMØJ and SJLB procured funding; SJLB and KMØJ supervised the project; all authors contributed to the writing of the manuscript. The authors declare no competing interests.

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). Work in the Billinge group was supported by the US National Science Foundation (grant No. DMREF-1922234). We are grateful to the Villum Foundation for financial support through a Villum Young Investigator grant (No. VKR00015416). Funding from the Danish Ministry of Higher Education and Science through the SMART Lighthouse is gratefully acknowledged.

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). In *Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG)*, 24 August 2020, San Diego, California, USA (virtual). New York: Association for Computing Machinery. https://www.mlgworkshop.org/2020/.

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.

Anker, A. S., Kjær, E. T. S., Juelsholt, M. & Jensen, K. M. Ø. (2024). *J. Appl. Cryst.* **57**, 34–43.

Artini, C., Pani, M., Lausi, A., Masini, R. & Costa, G. A. (2014). *Inorg. Chem.* **53**, 10140–10149.

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

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.

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

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

Castillo-Blas, C., Moreno, J. M., Romero-Muñiz, I. & Platero-Prats, A. E. (2020). *Nanoscale*, **12**, 15577–15587.

Chakraborty, K. R., Krishna, P. S. R., Chavan, S. V. & Tyagi, A. K. (2006). *Powder Diffr.* **21**, 36–39.

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.

Coduri, M., Scavini, M., Allieta, M., Brunelli, M. & Ferrero, C. (2013). *Chem. Mater.* **25**, 4278–4289.

Du, P., Kokhan, O., Chapman, K. W., Chupas, P. J. & Tiede, D. M. (2012). *J. Am. Chem. Soc.* **134**, 11096–11099.

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

Estes, S. L., Antonio, M. R. & Soderholm, L. (2016). *J. Phys. Chem. C*, **120**, 5810–5818.

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., 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., Davis, T., Farrow, C. L. & Billinge, S. J. L. (2013). *J. Appl. Cryst.* **46**, 560–566.

Juhás, P., Farrow, C., Yang, X., Knox, K. & Billinge, S. (2015). *Acta Cryst.* A**71**, 562–568.

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

Juhás, 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). *Digit. Discov.* **2**, 69–80.

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

Labidi, O., Drache, M., Roussel, P. & Wignacourt, J.-P. (2008). *Solid State Sci.* **10**, 1074–1082.

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). *Z. Anorg. Allg. Chem.* **619**, 961–968.

Magnard, N. P. L., Anker, A. S., Aalling-Frederiksen, O., Kirsch, A. & Jensen, K. M. Ø. (2022). *Dalton Trans.* **51**, 17150–17161.

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). *Appl. Mater. Interfaces*, **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.

Proffen, Th. & Neder, R. B. (1997). *J. Appl. Cryst.* **30**, 171–175.

Proffen, Th. & Neder, R. B. (1999). *J. Appl. Cryst.* **32**, 838–839.

Rademacher, O., Göbel, H., Ruck, M. & Oppermann, H. (2001). *Z. Kristallogr. New Cryst. Struct.* **216**, 29–30.

Radosavljevic Evans, I., Tao, S., Irvine, J. T. S. & Howard, J. A. K. (2002). *Chem. Mater.* **14**, 3700–3704.

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

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

Saalfeld, H. (1964). *Z. Kristallogr. Cryst. Mater.* **120**, 476–478.

Sasaki, T., Ukyo, Y., Kuroda, K., Arai, S., Muto, S. & Saka, H. (2004). *J. Ceram. Soc. Jpn*, **112**, 440–444.

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.

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

Vegard, L. & Borlaug, A. (1943). *Avhandlinger/Norske Videnskaps-Akademi, Matematisk-Naturvidenskapelig Klasse*. Oslo: Dybwad [in Komm.].

Weber, M., Schlesinger, M., Walther, M., Zahn, D., Schalley, C. A. & Mehring, M. (2017). *Z. Kristallogr. Cryst. Mater.* **232**, 185–207.

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

Yang, X., Juhas, P., Farrow, C. L. & Billinge, S. J. (2014). arXiv:1402.3163.

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.