Supplementary information for

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

Anonymous author(s)

Affiliation
Adress
email

Table of Contents

A: Comparing computation time of ML-MotEx with a 'brute-force' approach

	Brute-force time (s)	XGBoost model time (s) / rmse (%)
$(CH_3)_4N)_4SiW_{12}O_{40}$[1]	1215	0.14 / 0.010
$[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine)[2]	3376	0.11 / 0.026
$(((CH_3)_2NH_2)_6$ $(Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$[3]	2300	0.12 / 0.032
$((CH_3)_2NH_2)_3(PW_{12}O_{40})$[4]	3983	0.14 / 0.023

Table S1 | The XGBoost model is significantly faster than a 'brute-force' approach while providing high accuracy of the predicted goodness-of-fit values. Here a MacBook Pro with 2.3 GHz Dual-Core Intel Core i5 processor is used to do 10,000 fits on motifs extracted from four different structures. The goodness-of-fit values, R_{wp}, and structure catalogue generated are used to train a XGBoost model which is significantly faster, ~10.000-times, while obtaining high accuracy, 0.023 % root-mean-squared-error (rmse).

Calculating how long time it would take to fit all the possible structure motifs using the brute-force approach:

In order to calculate the overall time, t, that it takes to fit all the structure motifs that can be generated from a starting model containing N atoms, we use the formula:

$$t = \#structure\ motifs \cdot fitting\ time = 2^{N-1} \cdot fitting\ time$$

$(CH_3)_4N)_4SiW_{12}O_{40}$[1] – 24 atoms – 1215 s / 10.000 fits.

$$t = (2^{24} - 1) \cdot \frac{1215}{10000} = 2038432\ s = 23.6\ days$$

$[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine)[2] – 48 atoms – 3376 s / 10.000 fits.

$$t = (2^{48} - 1) \cdot \frac{3376}{10000} = 28147497671\ s = 3011191\ years$$

$(((CH_3)_2NH_2)_6\ (Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$[3] – 48 atoms – 2300 s / 10.000 fits.

$$t = (2^{48} - 1) \cdot \frac{2300}{10000} = 64739244643451\ s = 2051463\ years$$

$((CH_3)_2NH_2)_3(PW_{12}O_{40})$[4] – 72 atoms – 3983 s / 10.000 fits.

$$t = (2^{72} - 1) \cdot \frac{3983}{10000} = 1.88091857012698e21\ s = 59602712821222\ years$$

B: PDF calculation parameters for all the examples provided in the paper

	Example 1: A single C_{60} buckyball	Example 1: A supercell made of 1 unit cells of crystalline C_{60} polymer[5]	Example 2: The $Mo_{36}O_{128}$ starting model	Example 3: A $W_{12}O_{40}$ unit (data from Juelsholt et al.)[6]	Example 3: A $W_{12}O_{40}$ unit (data collected at DanMax)	Example 4: A supercell made of 2x2x3 unit cells of β-Bi_2O_3
r-range (Å)	0 – 30	0 – 30	3 – 12	1.6 – 10	1.6 – 10	2 – 15
r-step	0.1	0.1	0.01	0.01	0.01	0.01
Q_{min} (Å^{-1})	0.5	0.5	0.5	1.2	0.7	0.7

Q_{max} (Å^{-1})	24	24	24	17	20	15.5
Q_{damp} (Å^{-1})	0.02	0.02	0.04	0.04	0.05	0.04
ADP (Å^2)	0.5	0.5	0.3	0.3	0.3	1.35

Table S2 | PDF calculation parameters for all the examples provided in the paper. Note: the r-step was set to 0.01 Å for the data in the plots, but 0.1 Å for some of the simulated data used in the data analysis.

C: Visualisation of the ML-MotEx analysis on the simulated PDF data of the C_{60} buckyball using a continuous colorbar
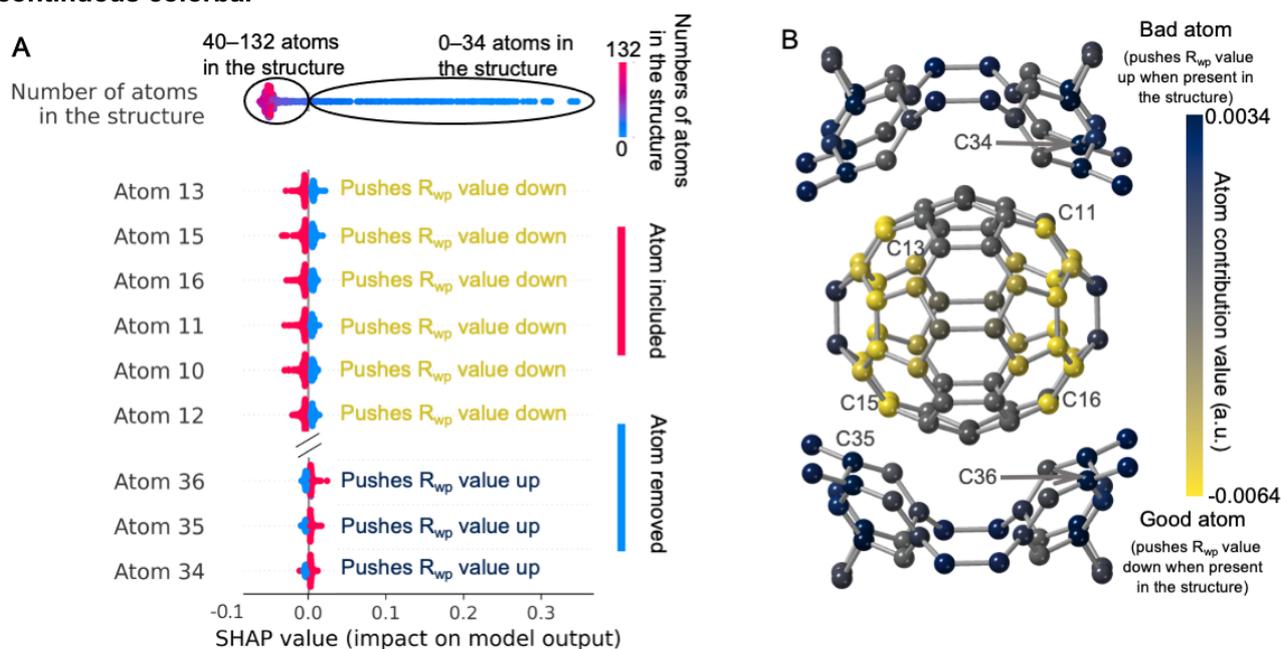
Figure S1 | Summary of the ML-MotEx analysis on simulated C_{60} PDF data. A) Plot of the SHAP values obtained in the C60 analysis, showing if atoms in the starting model are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. The SHAP values are plotted as a violin plot. B) Structural visualisation where the colorbar indicates the atom contribution value of the atom.

D: SHAP value plots from ML-MotEx utilized on experimental PDF data obtained at DanMAX from 0.05 M ammonium metatungstate hydrate in water

Total X-ray scattering data were measured using the RA-PDF[7] geometry at DanMAX (MAX IV, Lund) with X-ray wavelength 0.3542 Å providing a Q_{max} of 20 Å^{-1}. Both the 0.05 M ammonium metatungstate hydrate solution and the water background were transferred to Kapton tubes with an inner diameter of 1.00 mm and data was measured for 15 min. The data were integrated using Fit2D[8] and PDFs were obtained using xPDFSuite.[9]
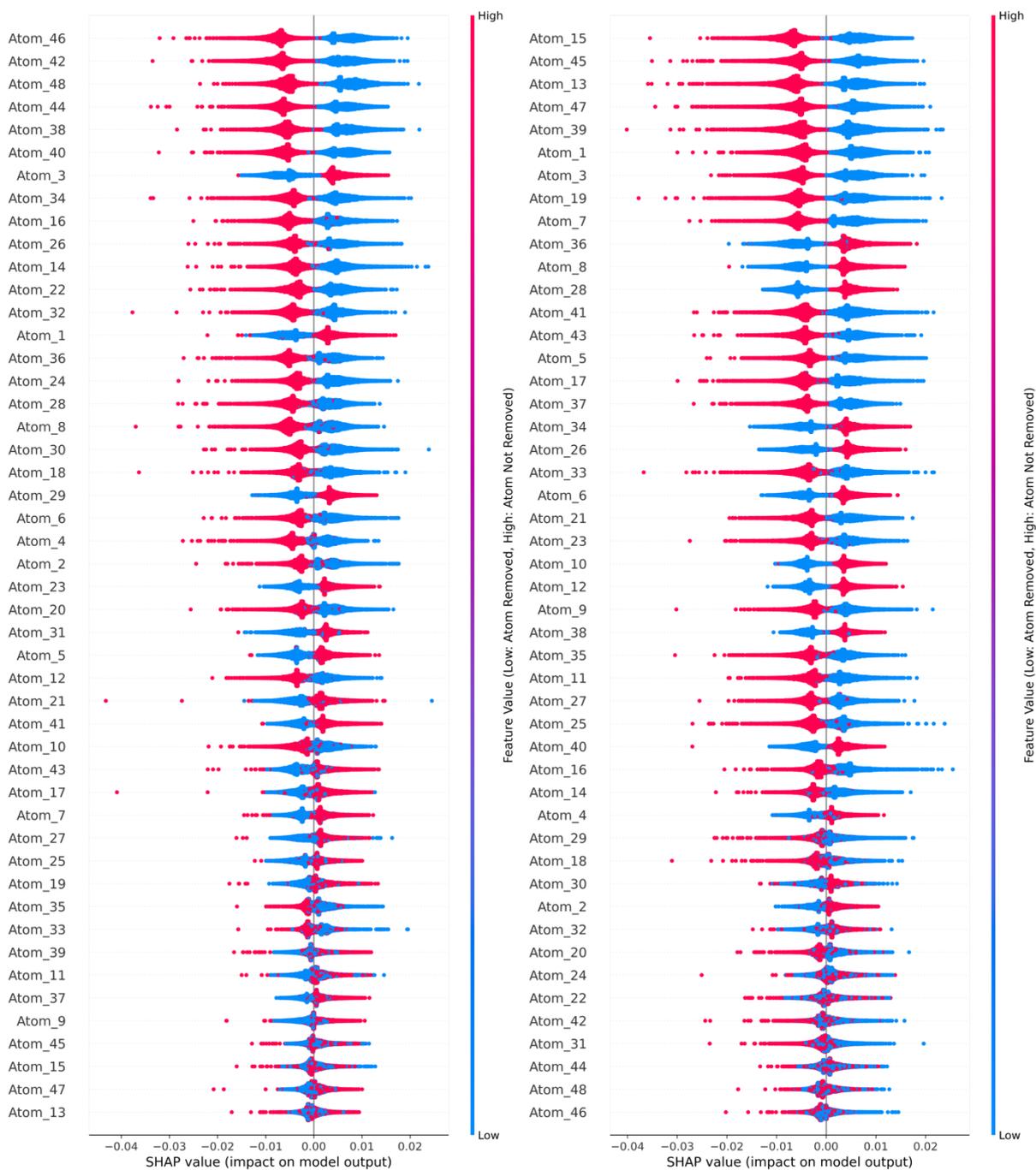
Figure S2 | Plot of the SHAP values obtained in the analysis of the PDF measured at MAX IV on ammonium metatungstate hydrate. The plot is showing if atoms in the starting model Left) [Hpy]$_4$H$_2$[H$_2$W$_{12}$O$_{40}$] (py = pyridine),[2] Right) (((CH$_3$)$_2$NH$_2$)$_6$ (Cu(HCON(CH$_3$)$_2$)$_4$)(GeW$_{12}$O$_{40}$)$_2$)(HCON(CH$_3$)$_2$)$_2$,[3] are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. The SHAP values are plotted as a violin plot.
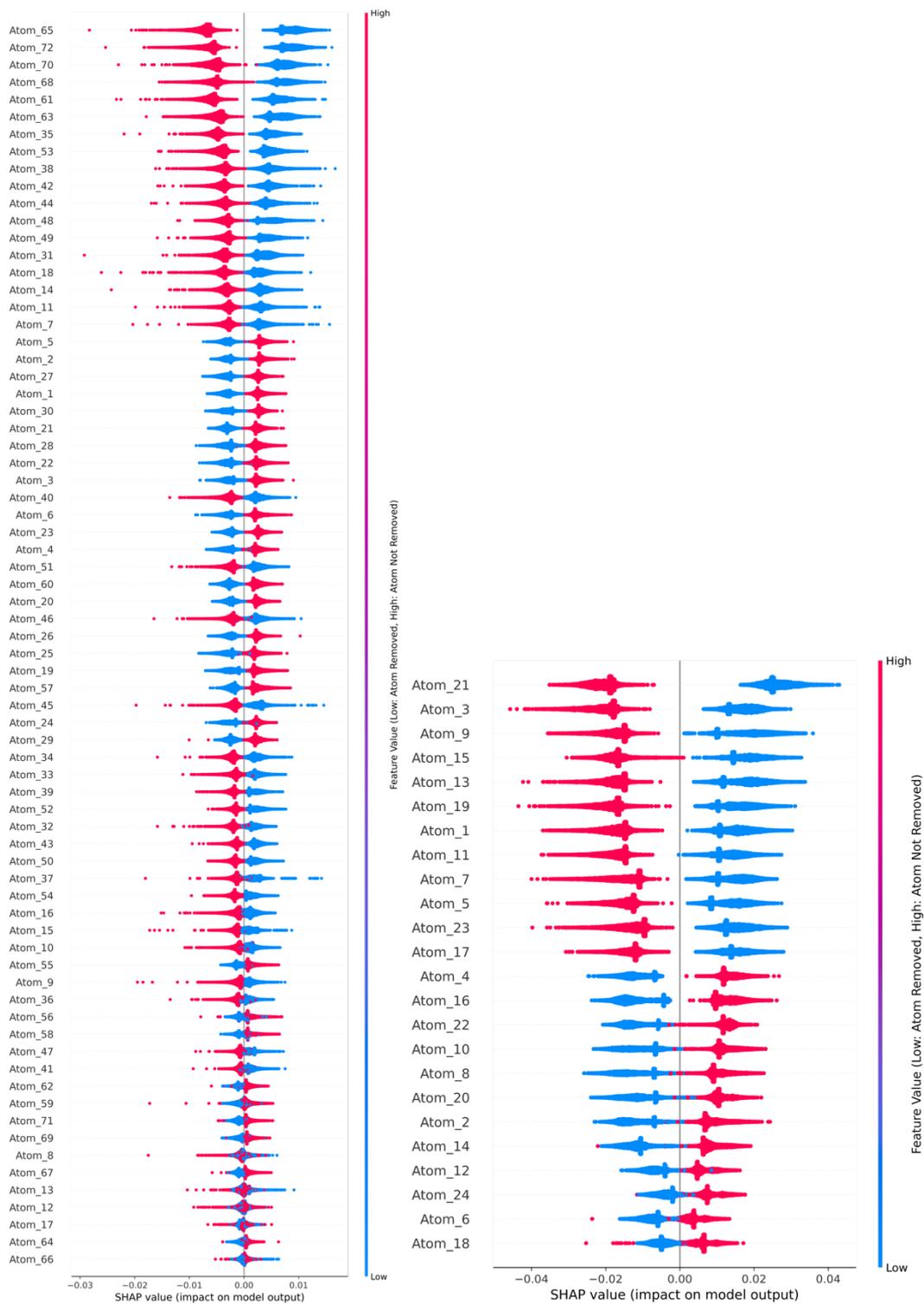
Figure S3 | Plot of the SHAP values obtained in the analysis of the PDF measured at MAX IV on ammonium metatungstate hydrate. The plot is showing if atoms in the starting model Left) $((CH_3)_2NH_2)_3(PW_{12}O_{40})$.,[4] Right) $(CH_3)_4N)_4SiW_{12}O_{40}$,[1] are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. The SHAP values are plotted as a violin plot.

6
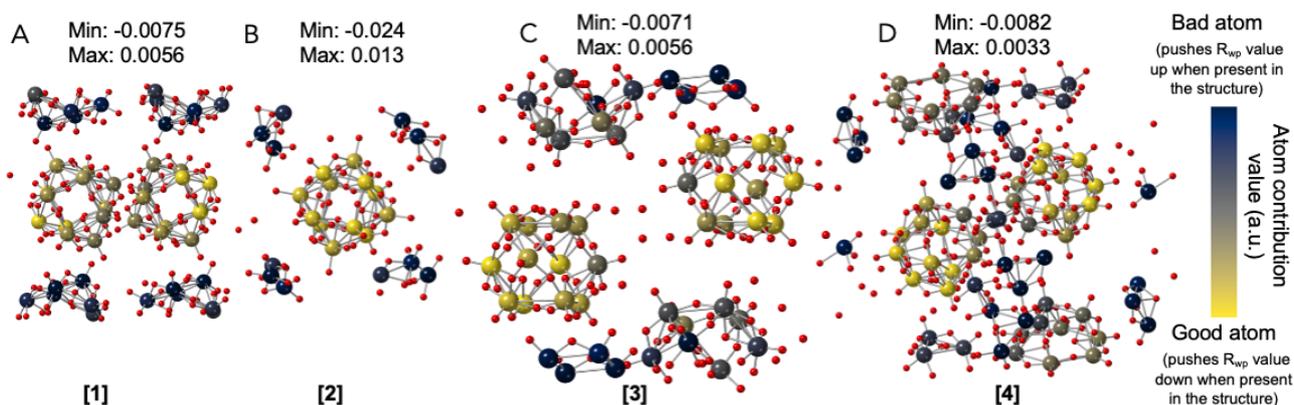
Figure S4 | Results from the ML-MotEx method on a PDF from a solution of ammonium metatungstate hydrate, using four different starting models: A) $[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine),[2] B) $(CH_3)_4N)_4SiW_{12}O_{40}$,[1] C) $(((CH_3)_2NH_2)_6$ $(Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$,[3] D) $((CH_3)_2NH_2)_3(PW_{12}O_{40})$.[4] The colorbar indicates the atom contribution value of the atom. The minimum and maximum values indicated is the minimum and maximum value used for the colorbar.

E: Using ML-MotEx on experimental PDF data from Juelsholt et al.[6] of 0.05 M ammonium metatungstate hydrate in water
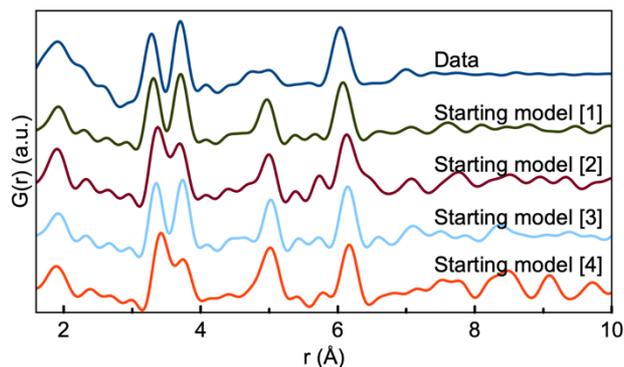
Figure S5 | Comparison of the data and the simulated data of the four different starting models [1]–[4]. The simulation parameters mimic typical values of a PDF dataset and can be seen in Table S1.
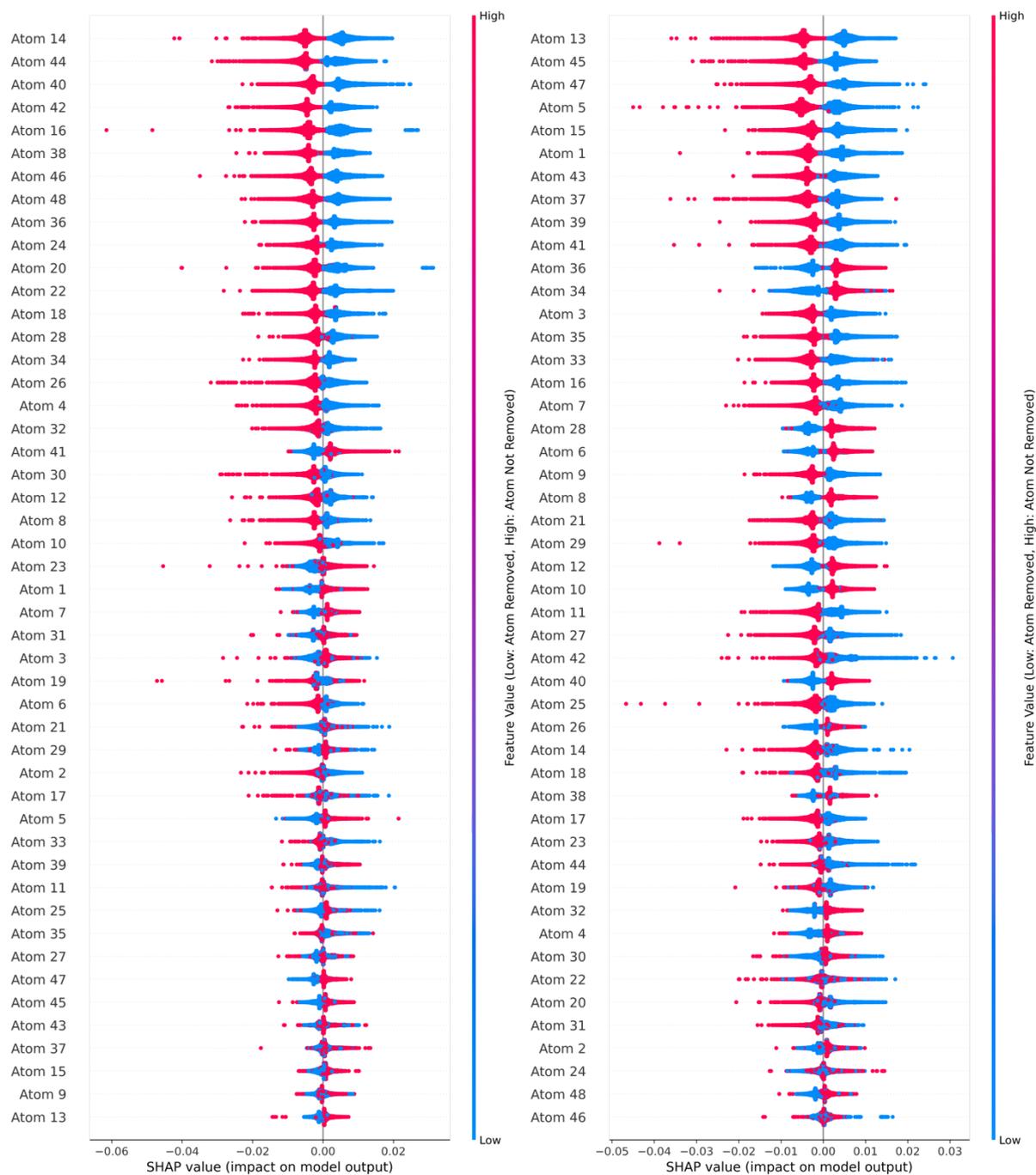
Figure S6 | Plot of the SHAP values obtained in the analysis of the PDF measured at DESY on ammonium metatungstate hydrate. The plot is showing if atoms in the starting model Left) [Hpy]$_4$H$_2$[H$_2$W$_{12}$O$_{40}$] (py = pyridine),[2] Right) (((CH$_3$)$_2$NH$_2$)$_6$ (Cu(HCON(CH$_3$)$_2$)$_4$)(GeW$_{12}$O$_{40}$)$_2$)(HCON(CH$_3$)$_2$)$_2$,[3] are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. The SHAP values are plotted as a violin plot.
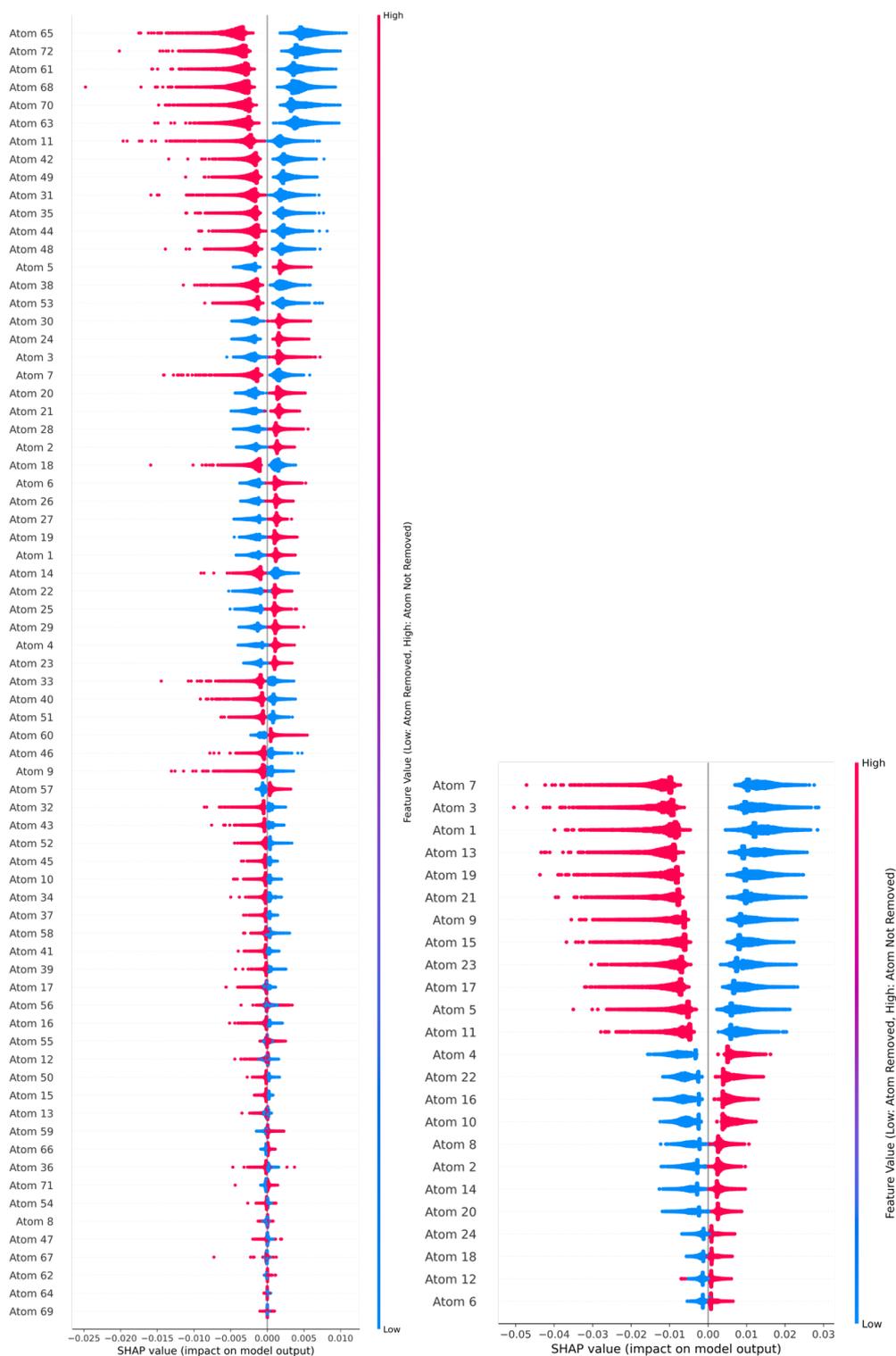
Figure S7 | Plot of the SHAP values obtained in the analysis of the PDF measured at DESY on ammonium metatungstate hydrate. The plot is showing if atoms in the starting model Left) $((CH_3)_2NH_2)_3(PW_{12}O_{40})$.,[4] Right) $(CH_3)_4N)_4SiW_{12}O_{40}$[1] are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. The SHAP values are plotted as a violin plot.

9
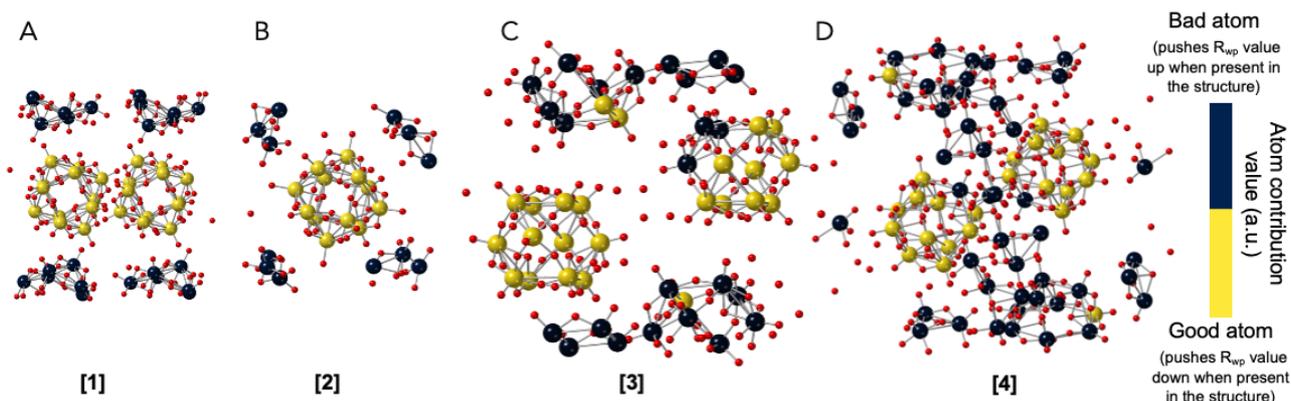
Figure S8 | Structural visualisation of kept and removed atoms using the colorbar to indicate the atom contribution value of each atom in the structure. ML-MotEx has been used on 4 different starting models but with the same fitting-algorithm. A) The $[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine)[2], B) $(CH_3)_4N)_4SiW_{12}O_{40}$[1], C) $(((CH_3)_2NH_2)_6 (Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$[3], D) $((CH_3)_2NH_2)_3(PW_{12}O_{40})$[4]. *The colorbar is discrete colouring the A) 24, B) 12, C) 24 and D) 24 atoms with the lowest atom contribution value yellow and the rest black.* Results from the analysis of data published by Juelsholt et al.[6]

	MacBook Pro with 2.3 GHz Dual-Core Intel Core i5 processor (s)	2.9/4.3 GHz 64-core AMD Ryzen Threadripper 3990 X (s)
$(CH_3)_4N)_4SiW_{12}O_{40}$[1]	3376	106
$[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine)[2]	1215	47
$(((CH_3)_2NH_2)_6 (Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$[3]	2300	87
$((CH_3)_2NH_2)_3(PW_{12}O_{40})$[4]	3983	166

Table S3 | The code is highly scalable, allowing significantly faster computer time on computers with many cores. Here a normal laptop is compared with a workstation by doing 10,000 fits on motifs extracted from 4 different structures. Results from the analysis of data published by Juelsholt et al.[6]

F: How can we determine if the starting model is good?
A limitation of ML-MotEx is that it needs a starting model that contains the right structural motif to work. How can we know that the starting model that is provided does contain the right structural motif?
To investigate this question, we here do an experiment where we, on purpose, use a range of starting models without the right structural motif and compare it to results using good starting models.
We use the PDF data from example 3, 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.[6] In the main text, we show that 4 different starting models, all containing the α-Keggin cluster, allows to extract the α-Keggin cluster using ML-MotEx. We here show what happens if we use other starting models that do not contain an α-Keggin cluster. To do so, we use starting models created from; 1) the $Na_5(H_7W_{12}O_{42})(H_2O)_{20}$ crystal (containing a Dodecatungstate icosahydrate cluster),[10] 2) the $(Ba(H_2O)_2(H(N(CH_3)_2)CO)_3)_2(W_{10}O_{32})(H(N(CH_3)_2)CO)_2$ crystal,[11] (containing a decatungstate polyoxometalate) 3) the $(NH_4)_{10}(H_2W_{12}O_{42})(H_2O)_4$ crystal,[12] (containing a paratungstate cluster) and 4) $Cs_2K_{0.5}P_2W_{18}O_{62}C_3O_6$ (containing a Dawson type cluster).[13]
As a positive control (starting models including the α-Keggin cluster) we use the 5) $(CH_3)_4N)_4SiW_{12}O_{40,}$[1] crystal,[3] 6) $[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine) crystal,[2] 7) C) $(((CH_3)_2NH_2)_6 (Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$ crystal[3] and the 8) $((CH_3)_2NH_2)_3(PW_{12}O_{40})$ crystal.[4]

Figure S9 shows the SHAP analysis on the 8 starting models using the same experimental PDF.
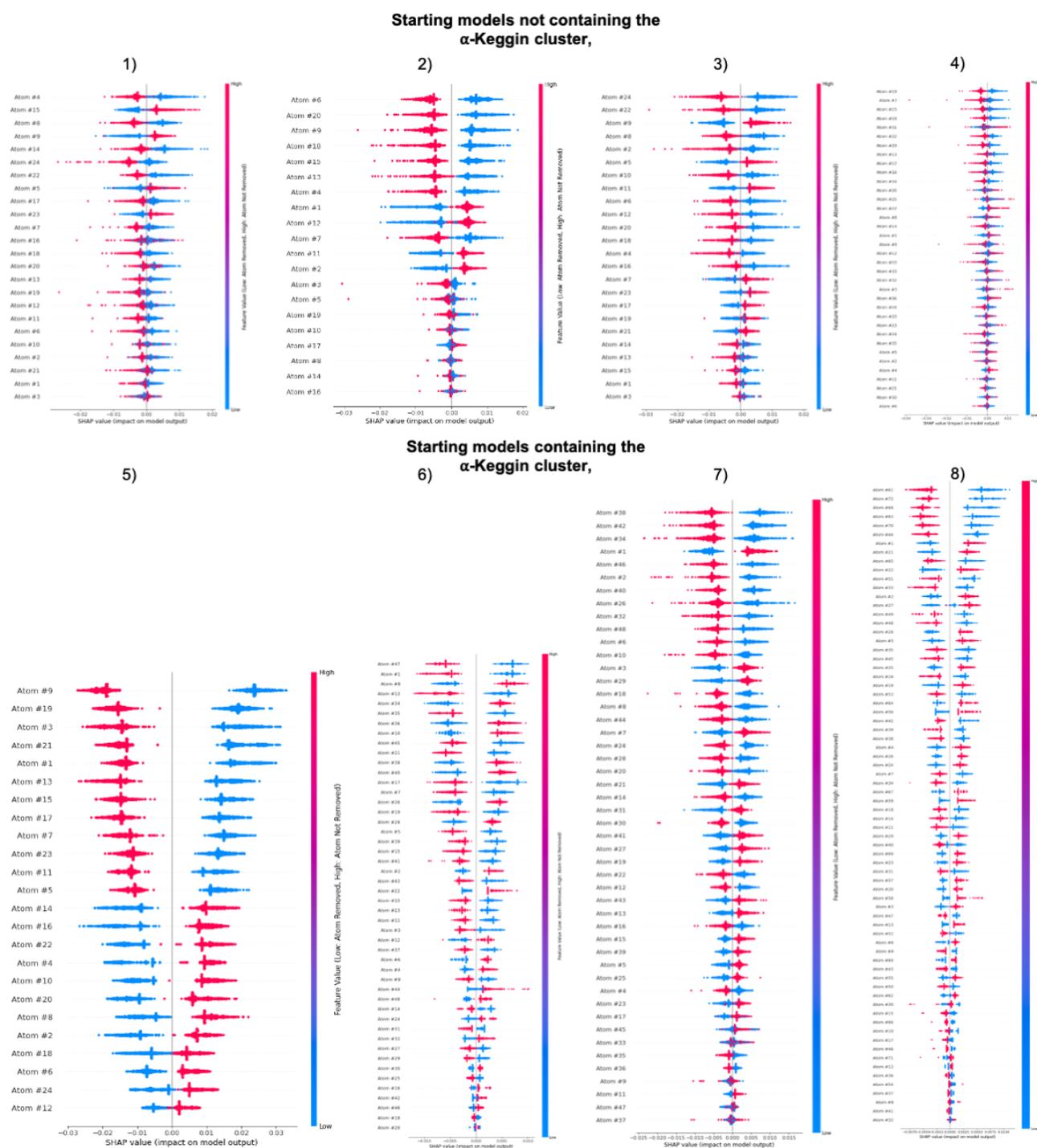
Figure S9 | SHAP summary plot of ML-MotEx used on an experimental PDF measured on a α-Keggin cluster using different starting models created from; 1) the $Na_5(H_7W_{12}O_{42})(H_2O)_{20}$ crystal,[10] 2) the $(Ba(H_2O)_2(H(N(CH_3)_2)CO)_3)_2(W_{10}O_{32})(H(N(CH_3)_2)CO)_2$ crystal,[11] 3) the $(NH_4)_{10}(H_2W_{12}O_{42})(H_2O)_4$ crystal,[12] 4) $Cs_2K_{0.5}P_2W_{18}O_{62}C_3O_6$[13] and as a positive control the 5) $(CH_3)_4N)_4SiW_{12}O_{40},$[1] crystal,[3] 6) $[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine) crystal,[2] 7) C) $(((CH_3)_2NH_2)_6 (Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$ crystal[3] and the 8) $((CH_3)_2NH_2)_3(PW_{12}O_{40})$ crystal.[4]

We observe that the SHAP values are more scattered across positive and negative values when the starting model does not contain the α-Keggin cluster, i.e. the correct structural motif. This is seen when comparing the SHAP values obtained for models 1-4 (not containing Keggin structures) with those from models 5-8. We can quantify this effect by calculating an uncertainty (root mean square / RMS) on the atom contribution value:

$$\text{atom contribution value RMS} = (\text{SHAP}^{RMS}_{\text{average-kept}}{}^2 - \text{SHAP}^{RMS}_{\text{average-removed}}{}^2)^{0.5}$$

We can now calculate the ratio of the atom contribution value and the uncertainty on the atom contribution value, which we refer to as the confidence factor as it tells how confident we can be about including/excluding that atom in a structural motif:

$$\text{Confidence factor} = \text{atom contribution value} / \text{atom contribution value RMS}$$

We can also calculate an average confidence factor for the starting model and use that as a proxy for how good the starting model is together with the R_{wp} values. Figure S10 shows the 8 starting models presented with their average confidence factors and the R_{wp} value of the best fit in the structure catalogue. We see that the 4 starting models containing the α-Keggin cluster all have larger confidence factors than the 4 starting models that do not. We also observe that the R_{wp} value of the best fitting structure is in general lower if the α-Keggin cluster is found in the starting model.

Bad atom
(pushes R_{wp} value up when present in the structure)

Atom contribution value (a.u.)

Good atom
(pushes R_{wp} value down when present in the structure)

1)
Confidence factor = 1.02
R_{wp} = 70 %

2)
Confidence factor = 1.61
R_{wp} = 73 %

3)
Confidence factor = 1.49
R_{wp} = 74 %

4)
Confidence factor = 0.54
R_{wp} = 59 %

5)
Confidence factor = 4.46
R_{wp} = 53 %

6)
Confidence factor = 2.06
R_{wp} = 46 %

7)
Confidence factor = 3.14
R_{wp} = 51 %

8)
Confidence factor = 2.98
R_{wp} = 67 %

Figure S10 | Results of using ML-MotEx on an experimental PDF measured on a α-Keggin cluster and different starting models created from; 1) the $Na_5(H_7W_{12}O_{42})(H_2O)_{20}$ crystal,[10] 2) the $(Ba(H_2O)_2(H(N(CH_3)_2)CO)_3)_2(W_{10}O_{32})(H(N(CH_3)_2)CO)_2$ crystal,[11] 3) the $(NH_4)_{10}(H_2W_{12}O_{42})(H_2O)_4$ crystal,[12] 4) $Cs_2K_{0.5}P_2W_{18}O_{62}C_3O_6$[13] and as a positive control the 5) $(CH_3)_4N)_4SiW_{12}O_{40}$,[1] crystal,[3] 6) $[Hpy]_4H_2[H_2W_{12}O_{40}]$ (py = pyridine) crystal,[2] 7) C) $(((CH_3)_2NH_2)_6 (Cu(HCON(CH_3)_2)_4)(GeW_{12}O_{40})_2)(HCON(CH_3)_2)_2$ crystal[3] and the 8) $((CH_3)_2NH_2)_3(PW_{12}O_{40})$ crystal.[4]

G: Example 3: *Identification of structural motifs in disordered molybdenum oxides*

Christiansen et al. have recently used the brute-force automated modelling method to identify structural motifs in disordered molybdenum oxides from PDF analysis.[14] Here we show that by reassessing the data with ML-MotEx, we can reproduce the results from Christiansen et al.[14] but in an automated way that allows analysis of the resulting structure model using SHAP values. Figure S11A shows the difference-PDF (d-PDF) obtained from amorphous molybdenum oxide supported on γ-Al_2O_3 nanoparticles (15 w% Mo), where the signal from the γ-Al_2O_3 nanoparticles has been subtracted. The d-PDF thus only reflects the structure of the supported material. The aim is to develop a structural model for the amorphous MoO_x. In the previous paper, different starting models were tested, which were all based on structures of molybdenum-based polyoxometalates (POMs) built from [MoO_4] tetrahedra and [MoO_6] octahedra. The analysis showed that the best fitting models did not contain tetrahedral motifs. Instead, the brute-force automated modelling approach hinted to a unit of three edge-sharing [MoO_6] octahedra, or a 'triad', to be present in the structure. However, the use of the computationally expensive brute-force method limited the number of atoms that could be included in the starting model. This meant that a range of different smaller starting models were used to test different structure hypotheses. With ML-MotEx, we can instead test much larger systems and thereby include several different structural motifs at the same time in one starting model, as well as a quantitatively analyse the results using SHAP values. We therefore use a larger POM as starting model namely the entire $Mo_{36}O_{128}$ cluster cut out of the $K_8(Mo_{36}O_{112}(H_2O)_{16})\cdot(H_2O)_{37}$ crystal structure,[15] which contains a range of different chemical topologies. Figure S11A shows the simulated data from the $Mo_{36}O_{128}$ cluster which has some similarities to the experimental PDF and Figure S11B shows the structure of the $Mo_{36}O_{128}$ cluster.
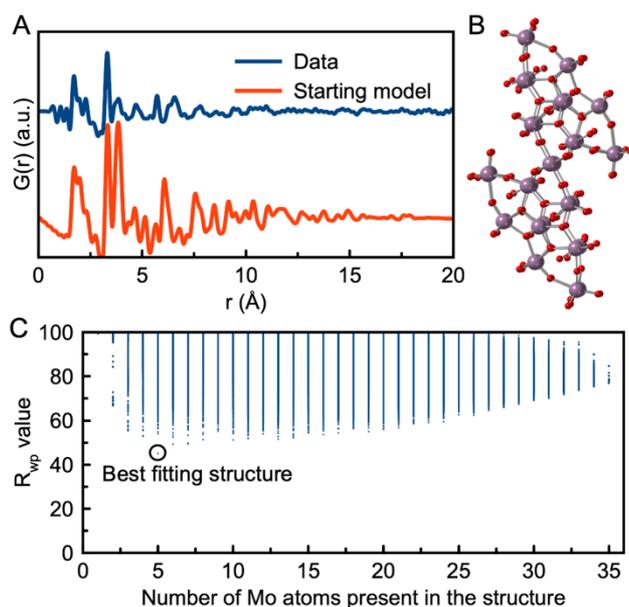
Figure S11: Analysis of experimental PDF from disordered molybdenum oxide. A) Comparison of experimental PDF from a disordered molybenum oxide,[14] and simulated data from $Mo_{36}O_{128}$ cluster, used as starting model. The simulation parameters mimic typical values of a PDF dataset and can be seen in section B in the Supplementary Information. B) Structure of the $Mo_{36}O_{128}$ cluster. C) R_{wp} values obtained in the fits using the $Mo_{36}O_{128}$ structure catalogue, plotted as a function of number of atoms in the structure motifs.

We apply ML-MotEx to the molybdenum oxide system in the same manner as we did to the C_{60} buckyball. First, we used the starting model to make a catalogue of candidate structure motifs, as described in detail in the Methods section. These are all fit to the experimental PDF, and the results are used to train the GBDT

model. The fits are made with the same fitting algorithm as used in the paper from Christiansen et al.[14] Figure S11C illustrates the R_{wp} values of the fits, plotted as a function of the number of molybdenum atoms present in the structural motif. The best fitting models contain 5–7 molybdenum atoms. The model that fits the data with the lowest R_{wp} value (45 %) can be identified as a Mo_5O_{24} structure as shown in section H in the Supporting Information. However, it is difficult to justify that this structural model is unique representing the structure in the sample, purely based on the R_{wp} value.

We therefore use step 3 and 4 of ML-MotEx to analyse the results of the ensemble of fits. The resulting SHAP values are shown in Figure S12A. The plot should be interpreted in the same way as for the C_{60} example: Each atom is assigned a SHAP value in each of the fits in the training set. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. When considering the amplitudes of the SHAP values, we see that the atoms labelled with 14, 15, 19, and 20 are marked as very important by ML-MotEx. When these atoms are present in the structure (red), they all have large negative SHAP values, indicating that their presence in the model pushes the R_{wp} down. When they are not present in the structure (blue), they all have large positive SHAP values, also indicating that they should be present in the structure to obtain a good fit. Atom 22 and 23 are examples of atoms that ML-MotEx do not suggest keeping in the structure. As seen from the SHAP values, its presence pushes up the R_{wp} value.

Based on the SHAP analysis, atom contribution values were calculated. The results are visually illustrated in Figure S12B, where the molybdenum atoms in the structure are coloured yellow if the atom contributed to a better fit quality, otherwise it is coloured black. Figure S12B clearly shows a specific motif that ML-MotEx wants to keep in the model. The yellow molybdenum atoms are all part of a 'triad' structure, where three $[MoO_6]$ octahedra share edges, and all oxygen atoms that bond to 3 or 4 Mo atoms are connected to yellow molybdenum atoms. This is further illustrated in section H in the Supplementary Information. Specifically, the resulting structural unit that ML-MotEx wants to keep is similar to heptamolybdate $[Mo_7O_{24}]^{6-}$, which can be described as several triads connected through edge-sharing. These results indicate that a motif of connected edge-sharing triads as shown in Figure S12C are important in order to describe the data of the disordered molybdenum oxides, which was also found by Christiansen et al.[14] We note here that when fitting this model to the PDF itself, we cannot describe the medium-range order present in the PDF. The ML-MotEx rather allows identifying the main local motifs in the data.
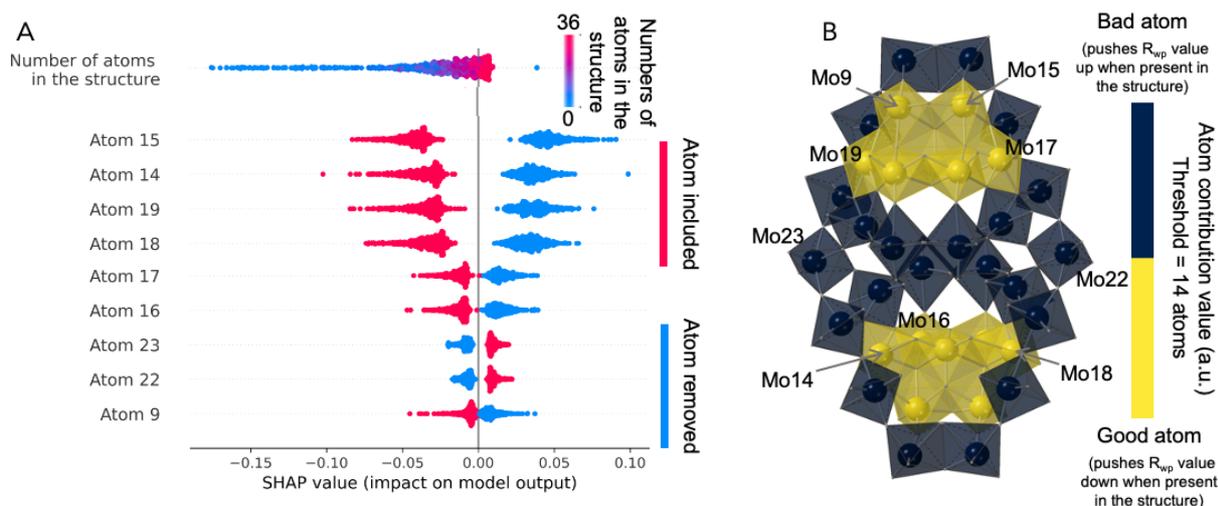
Figure S12: Summary of the ML-MotEx analysis of experimental PDF from disordered molybdenum oxide. A) Plot of the SHAP values obtained in the molybdenum oxide analysis, showing if atoms in the starting model $Mo_{36}O_{128}$ are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. The SHAP values are plotted as a violin plot. B) Structural visualisation of kept (yellow) and removed (black) atoms.

Section H in the Supplementary Information shows a similar representation but where the atom contribution values are directly shown from a continuous colour bar.

H: Best fitting model of the $Mo_{36}O_{128}$ system obtained with brute-force modelling
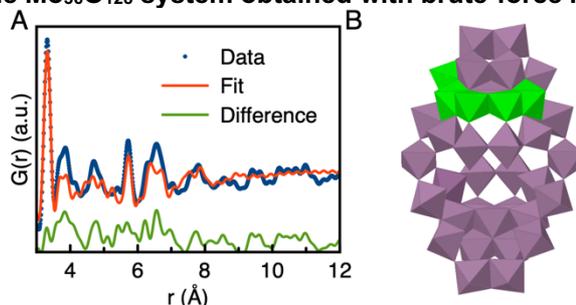
Figure S13 | Using a brute-force approach to extract the structure motifs on an experimental PDF from disordered molybdenum oxide which has the best fit quality. A) Fit of the best fitting model, B) the Mo_5O_{24} cluster (green atoms), extracted with brute-force modelling of the data obtained from amorphous molybdenum oxide supported on γ-Al2O3 nanoparticles (15 w% Mo). The signal from the γ-Al2O3 nanoparticles has been subtracted. R_{wp} = 45.1 %

Example system	$Mo_{36}O_{128}$
r-range (Å)	3–12
Scale	0.28
ADP_{Bi} (Å^2)	0.34
ADP_O (Å^2)	0.37
Isotropic expansion	1.00
Delta2 (Å^2)	6.00

Table S4 | Refinement details to Figure S14. Refinement details from the fit of the Mo_5O_{24} cluster, extracted with brute-force modelling of the data obtained from amorphous molybdenum oxide supported on γ-Al2O3 nanoparticles (15 w% Mo), where the signal from the γ-Al2O3 nanoparticles has been subtracted.
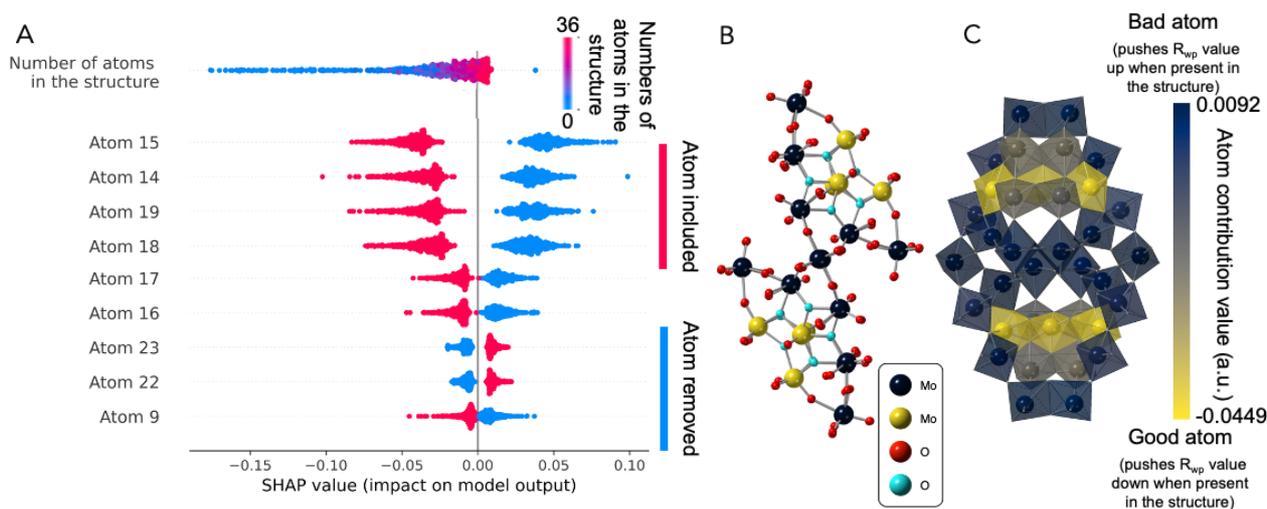
Figure S14 | Summary of the ML-MotEx analysis on an experimental PDF from disordered molybdenum oxide. A) Plot of the SHAP values obtained in the molybdenum oxide analysis, showing if atoms in the starting model $Mo_{36}O_{128}$ are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. The SHAP values are plotted as a violin plot. B) Structural visualisation of kept and removed atoms. The colorbar is discrete colouring the 14 atoms with the lowest atom contribution value yellow and the rest black. The turquoise oxygen atoms are either bonded to 3 or 4 Mo atoms while red oxygen atoms are bonded to less than 3 Mo atoms.C) Structural visualisation where the colorbar indicates the atom contribution value of the atom.

I: Example 4: Extracting structural motifs from ionic bismuth oxido clusters in solution using a 'cookie-cutter' strategy

In example 3, we showed that ML-MotEx can be applied for ionic clusters in solution, and it was possible to identify the Keggin-motifs from different POM starting clusters. In this final example, we will show another example of an ionic cluster in solution, namely bismuth oxido clusters in solution. Specifically, we study the $[Bi_{38}O_{45}]$ (Figure S15A), which is well known to be a stable cluster in solution.[16, 17] If considering the structure of $[Bi_{38}O_{45}]$, it can be seen that this can be described as a cut-out of the β-Bi_2O_3 structure, Figure S15B.[16, 17] The cut-out of the $[Bi_{38}O_{45}]$ cluster in the 2x2x3 supercell of the β-Bi_2O_3 structure is visualized in yellow in Figure S15B. We can therefore apply the ML-MotEx method for analysis of a PDF from this cluster, using a starting model reflecting the β-Bi_2O_3 structure. The starting model is thus a discrete model consisting of 2x2x3 β-Bi_2O_3 unit cells.

Figure S15C shows the experimental PDF obtained from a solution of $[Bi_{38}O_{45}]$ clusters compared to the simulated PDF data of the discrete starting model built from 2x2x3 unit cells of the β-Bi_2O_3 structure. As expected, the local range of the PDFs from the two structures are highly similar, but peaks to higher r-values are present for the β-Bi_2O_3 starting model. This confirms that the local structure is the same in the $[Bi_{38}O_{45}]$ sample as the starting model, but that the cluster size is smaller.
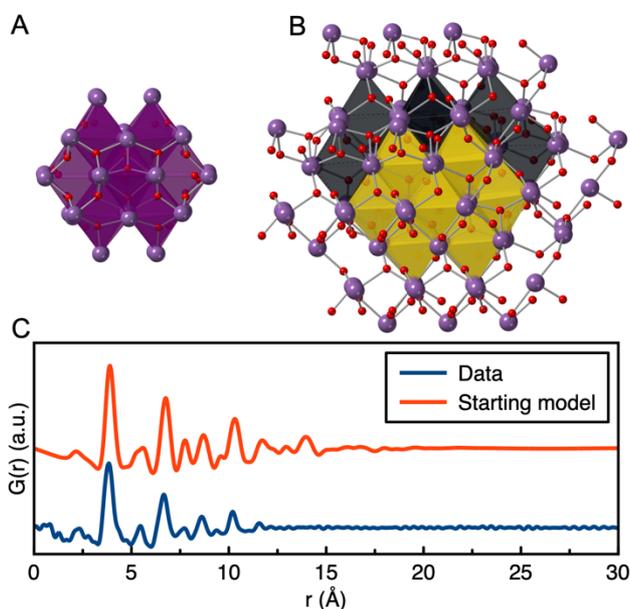
Figure S15 | The experimental PDF from bismuth oxido clusters in solution. A) The $[Bi_{38}O_{45}]$ cluster. B) The β-Bi_2O_3 structure starting model, where the yellow corresponds to the $[Bi_{38}O_{45}]$ cluster. C) Comparison of the experimental PDF of the $[Bi_{38}O_{45}]$ cluster and simulated data of a model consisting of 2x2x3 β-Bi_2O_3 unit cells, treated as a discrete model The simulation parameters mimic typical values of a PDF dataset and can be seen in Table S2.

In order to only produce physically sensible clusters from the β-Bi_2O_3 structure, the strategy for generating the structural motifs is different than for the other examples. Rather than using permutations over different atoms in the structure, we first cut out geometrically reasonable shapes using a 'cookie-cutter' strategy. Instead of deleting a number of random atoms, a range of reasonable shapes and sizes are made from the discrete β-Bi_2O_3 structure. Four shapes are used; spherical, cubic, ellipsoidal and disk. All the atoms in the starting model were kept or removed if they were inside the geometric object or not. Finally, oxygen atoms were removed if they had no bismuth atoms within the threshold distance of 2.5 Å which is the shortest Bi···O distance in the experimental PDF. Further details on the 'cookie-cutter' algorithm can be found in section N in the Supporting Information.

We first cut out 5950 unique clusters and fit them to the PDF. As in the previous examples, we can first analyse the R_{wp} values without using ML. Figure 16A shows all the R_{wp} values obtained from the PDF fits, where the fit quality is again plotted versus the number of bismuth atoms. The refinement parameters are given in Table S1. It is evident that the R_{wp} values are lowest when a structure of ca. 38 Bi atoms is used to fit the data, however, the curve is not very shallow. In Figure S16B, we plot the R_{wp} value plotted for the 4 different shape categories used in the 'cookie-cutter' strategy, where the colour of the point indicates the sphericity. Sphericity is a measure of how spherical an object is. A sphere has a sphericity of 1, while a cube has a sphericity of 0.806 and the sphericity of an ellipsoid depends on its radii. The equation and a detailed description of sphericity are given in section O in the Supporting Information.

The plot shows that cube-shaped structures generally result in a higher R_{wp} value than the other shapes, ruling out this shape for describing the cluster structure. The spheres generally give good fits, but the structures providing the lowest R_{wp} values are not spherical. The structures resulting in the lowest R_{wp} values are ellipsoids or disks, however, common for those two shapes is that the best fitting structures have high sphericity, meaning that the structure is very close to being spherical. In Figure S16C, we further analyse the results of fitting spherical models to the data. The figure shows the R_{wp} value plotted as a function of radius, showing a minimum at ca. 7.2 Å. Similar plots of the volume and ellipsoidal radius are shown in section P in the Supporting Information. The best structure found with the brute-force approach is shown in Figure S17A where the yellow polyhedra represent parts of the [$Bi_{38}O_{45}$] structure.
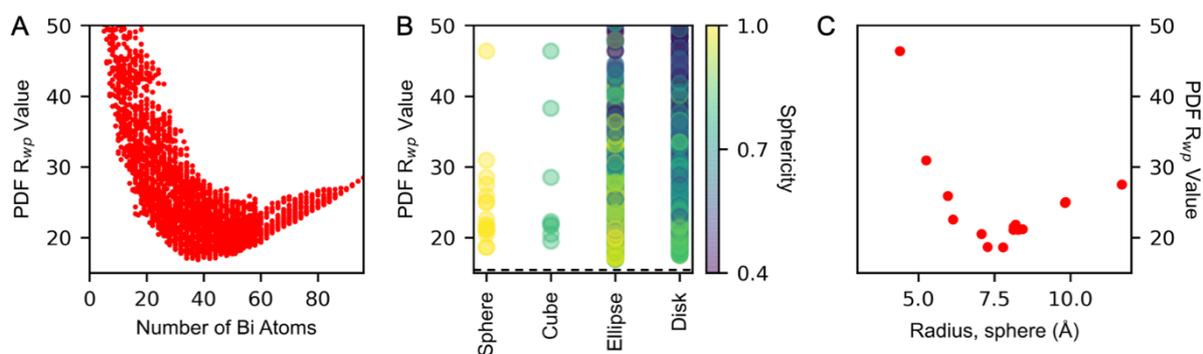
Figure S16 | Using a brute-force approach to extract structure motifs on an experimental PDF from bismuth oxido clusters in solution. Overview of the R_{wp} values of the PDF fits versus important structural parameters. A) The R_{wp} values plotted versus the number of bismuth atoms present in the structure, B) The R_{wp} value plotted versus the shape and sphericity of the structures and C) The R_{wp} value plotted versus the radius of the structures made with a spherical shape.

We now again move to ML-analysis for further analysis. Again, this allows taking all 5950 fits into account when finding an appropriate structure motif that fits the data, rather than only considering a few of the fits in the minimum in the R_{wp} plot. Section Q in the Supporting Information shows the results of the SHAP analysis of the motifs made from the discrete starting model built from 2x2x3 unit cells of the β-Bi_2O_3 structure and their

17

corresponding R_{wp} values. Afterward, the atom contribution values are again calculated and since the best fitting structure from the brute-force approach had 38 Bi atoms, we choose to keep the 38 atoms with the lowest atom contribution value. Figure S17B illustrates the structure when all other than the 38 Bi atoms with the lowest SHAP value are removed. This structure fits the data with an R_{wp} value of 17.4 %, i.e a slightly higher value than the motif identified from the brute-force calculations, however, it contains more of the original [$Bi_{38}O_{45}$] structure than the brute-force model does, making it more intuitive for an experienced chemist to 'finish' the model by constructing closed-shell structures. Fitting parameters are given in section K in the Supporting Information. The fits of the [$Bi_{38}O_{45}$] are shown in comparison in section S in the Supporting Information. From this example, it is clear that the automated brute-force routine itself reveals valuable insight into the structure and shape of the cluster in question. However, it is a cumbersome and biased process to extract the information from the many fits to build a structure that fits the data, and only the best fit is usually considered. By applying ML-MotEx we obtain a reasonable structural estimate, which can be used with chemical knowledge to construct the right model that fits the data.
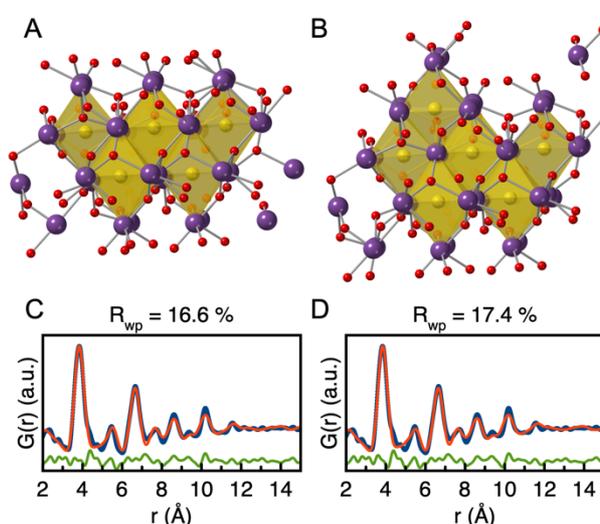
Figure S17 | Summary of the brute-force and ML-MotEx analysis on an experimental PDF from bismuth oxido clusters in solution. Structural visualisation of the kept atoms in the model built from 2x2x3 unit cells of β-Bi_2O_3 by A) brute-force approach and B) ML-MotEx. The polyhedral, which are part of the real [$Bi_{38}O_{45}$] cluster are shown in yellow.

J: How large a catalogue of candidate structure motifs does ML-MotEx need as training data to output reasonable results

When the number of structures in the structure catalogue is increased, the confidence factor goes up. The confidence factor can thereby be used to tell how many structures are needed in the structure catalogue.
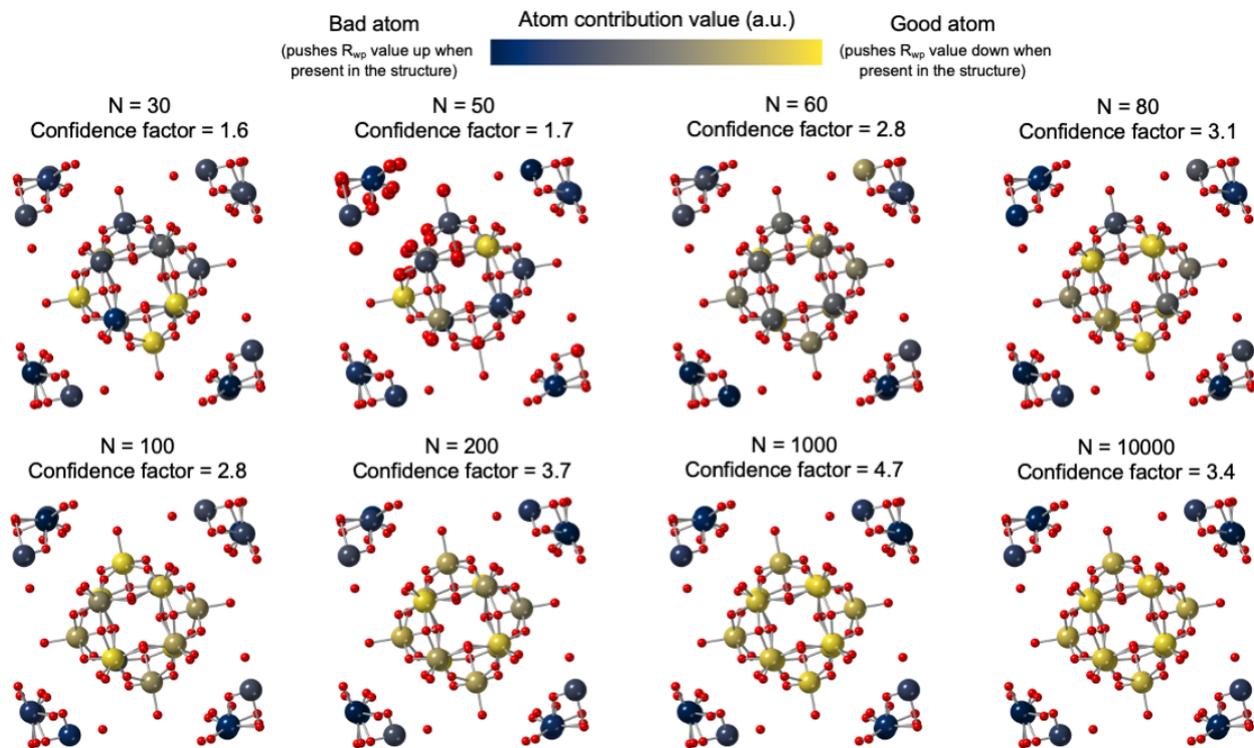
Figure S18 | Results of using ML-MotEx on an experimental PDF measured on a α-Keggin cluster and various amount of structures in the structure catalogue generated from the same starting model.

K: Fitting variables for all the examples provided in the paper

	C_{60} buckyball	$Mo_{36}O_{128}$	α-Keggin	β-Bi_2O_3
PDF				
r-range (Å)	0 – 30	3 – 12	1.6 – 10	2 – 15
r-step	0.01	0.01	0.01	0.01
Q_{min} (Å^{-1})	0.5	0.5	1.2	0.7
Q_{max} (Å^{-1})	24	24	17	15.5
Q_{damp} (Å^{-1})	0.02	0.04	0.04	0.04
Q_{broad} (Å^{-1})	-	-	0.01	-
Oxygen threshold (Å)	-	2.5	2.5	2.5
PDF Scaling factor	0 – ∞ (1)	0 – ∞ (1)	0 – ∞ (1)	0 – 1.5 (0.9)
Delta2 (Å^2)	0 – 2 (1)	0.5 – 6 (5)	0 – ∞ (0)	-
ADP$_{C/Mo/w/Bi}$ (Å^2)	0 – 2 (0.5)	0.3 – 3 (0.3)	0 – ∞ (0.4)	0.01 – 3 (1.35)
ADP$_O$ (Å^2)	-	0.3 – 4 (0.5)	0 – ∞ (0.4)	2
Isotropic expansion	0.98 – 1.02 (1.00)	0.98 – 1.02 (1.00)	0.98 – 1.01 (1.00)	0.98 – 1.02 (1.00)

Table S5 | Data specific parameters and fitting parameters for the 4 examples provided in the paper. The initial values of the parameters are written inside the parentheses, while the fitting range is written outside the parentheses.

The output of the fit is a R_{wp} value reflecting the quality of the fit:

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

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

L: Example of how structure motifs can be extracted from a starting model with 4 metal atoms coordinated to oxygen and used as input to the GBDT model

Index	Number of atoms	Atom #1	Atom #2	Atom #3	Atom #4	R_{wp}
1	3	1	1	0	1	34 %
2	2	0	1	1	0	61 %
3	2	1	0	1	0	47 %
4	1	1	0	0	0	63 %
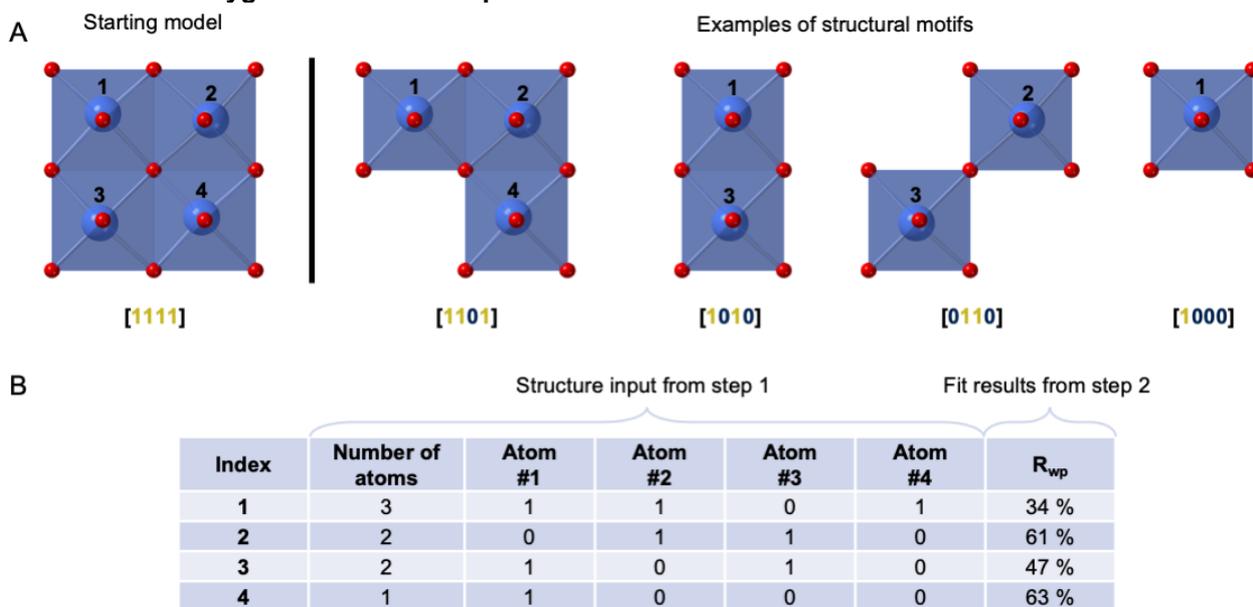
Figure S19: Example of how structure motifs can be extracted from a starting model with 4 metal atoms coordinated to oxygen and used as input to the GBDT model. A) The metal atoms are permuted randomly by creating an array of zeros and ones, where 0 refers to a deleted atom and 1 refers to an atom that is kept in the structure. Oxygen atoms are removed if they do not bond to any metal atoms within a distance threshold that is set by the user. Note that the metal atoms (blue) are slightly distorted from the centre of the octahedra. B) Example of how the four structures from Figure 8A and Figure 1 are given as input to the GBDT model which predicts the R_{wp} value.

M: Effect of changing the hyperparameters of the Gradient Boosting Decision Tree (GBDT) algorithm or the seed

Figure S20 | Effect of changing the hyperparameters of the Gradient Boosting Decision Tree (GBDT) algorithm. Results of using ML-MotEx on four different starting models on the same experimental **PDF** using the same 10.000 fits but various hyperparameters for the GBDT algorithm before calculating the **SHAP-** and the atom contribution value**s**.

Figure S21 | Effect of changing the seed of ML-MotEx.

N: Cookie-Cutter algorithm
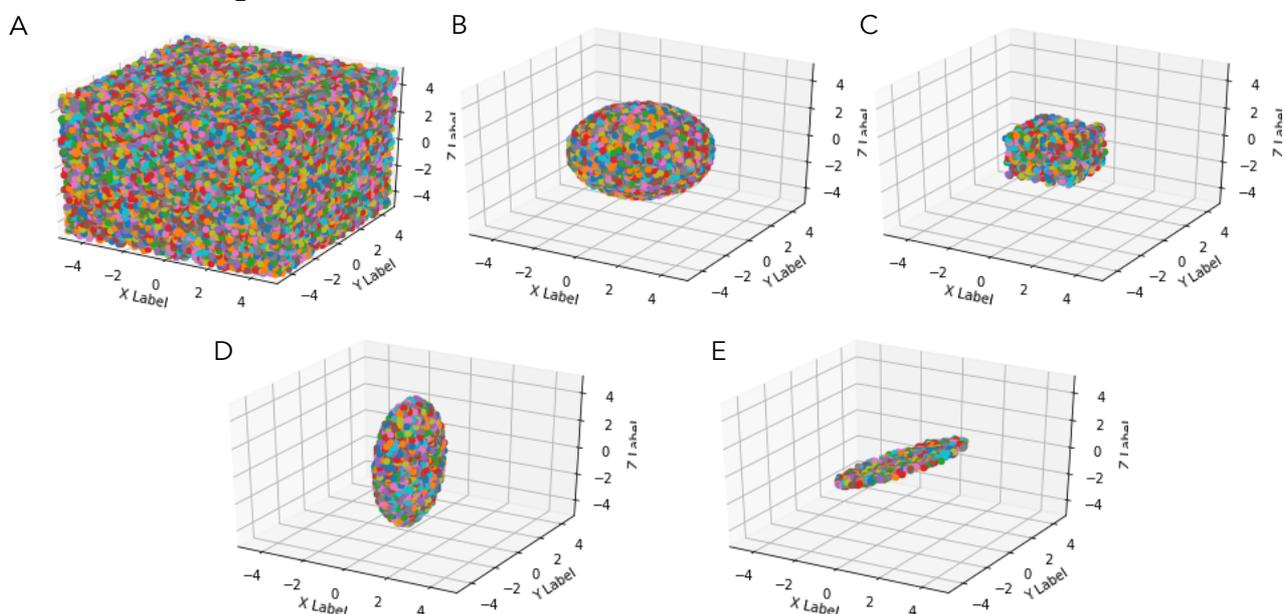
Figure S22 | The 'cookie-cutter strategy. A) The initial box with atoms 0.1 Å apart. B–E shows 4 examples of shapes that can be cut out with the cookie-cutter tool B) the sphere, C) the cube, D) the ellipsoid, E) the disk. In all examples, all atoms from outside this object in A) have been removed.

Sphere	
Radius (Å)	2 – 12
Cube	
Height and length (Å)	3 – 16
Ellipsoid	
Radius1, Radius2, Radius3 (Å)	0 – 20
Disk	
Radius (Å)	0 – 20
Height (Å)	0 – 20
Angle (°)	0 – 360

Table S6 | Parameter thresholds for sampling geometries with spherical-, cubic-, ellipsoidal- or a disk shape. The parameters are sampled from a uniform distribution in this range.

O: Sphericity

The sphericity, Ψ, of an object was defined in 1935 by Wadell and measures how spherical a geometry is. It can be calculated with the following formula[18]:

(2)
$$\Psi = \frac{\pi^{\frac{1}{3}}(6V_p)^{\frac{2}{3}}}{A_p}$$

Where V_p is the volume of the object and A_p is the surface area of the object.

P: Analysis of the PDF fits of data from the [Bi38O45] cluster
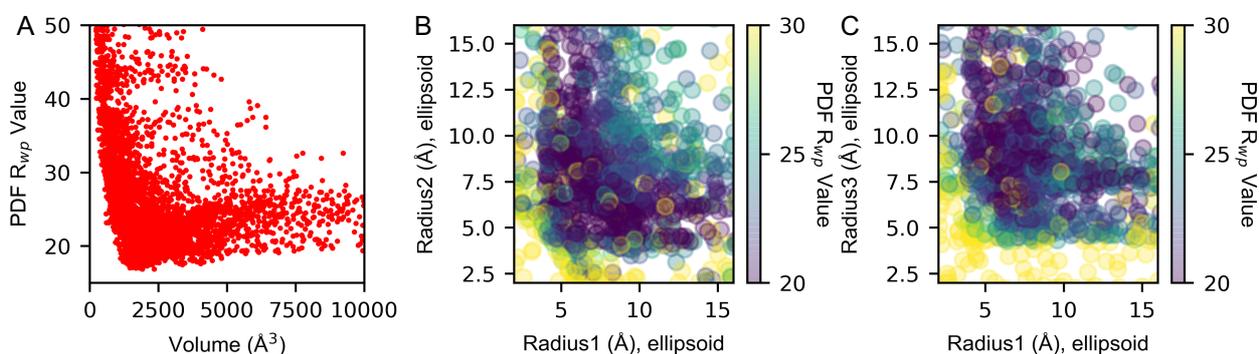
Figure S23 | Using a brute-force approach to extract structure motifs on an experimental PDF from bismuth oxido clusters in solution. Overview of the R_{wp} values of the PDF fits versus important structural parameters. A) The R_{wp} values versus the volume of the structure, B) The R_{wp} value versus radius1 and radius2 in the ellipsoidal shaped structures and C) The R_{wp} value versus radius1 and radius3 in the ellipsoidal shaped structures.

Q: SHAP value summary plot from ML-MotEx applied on a PDF from the [Bi$_{38}$O$_{45}$] cluster
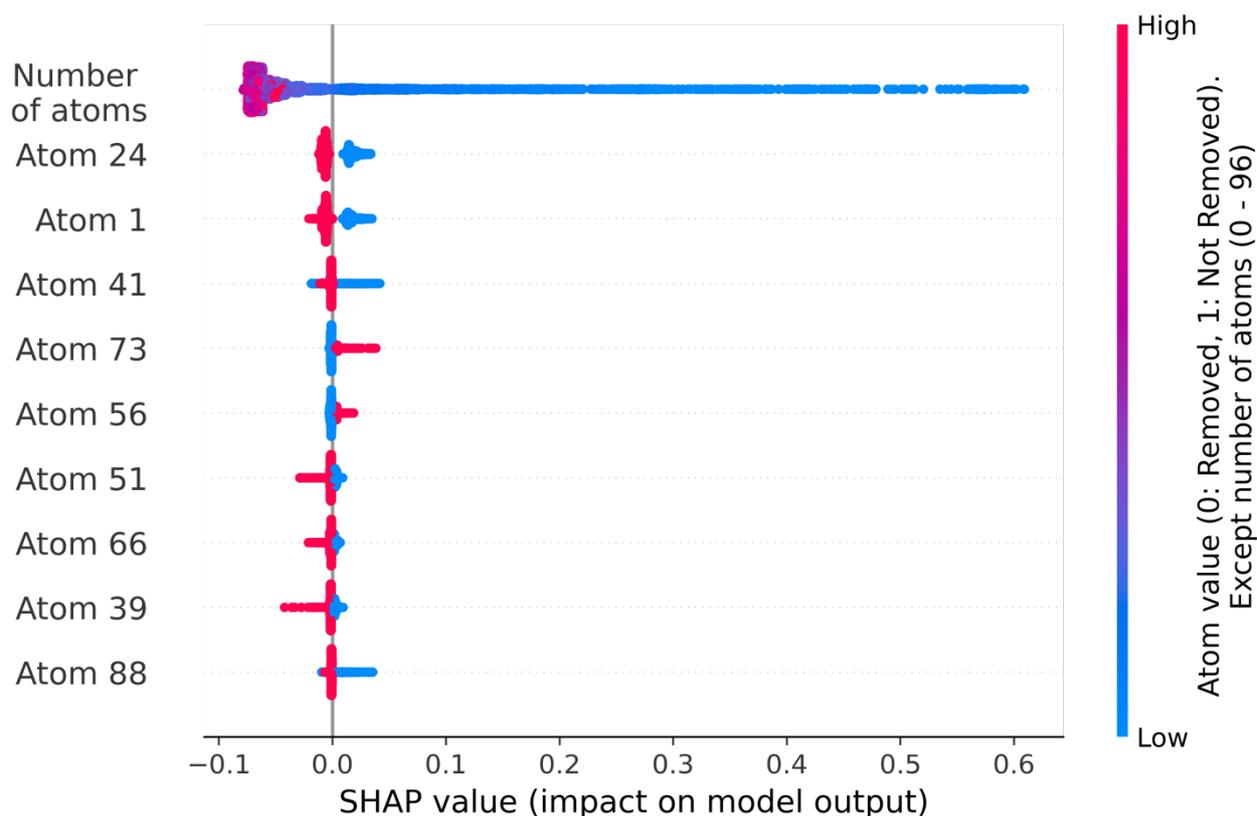
Figure S24 | Plot of the SHAP values obtained in the analysis of the bismuth oxido PDF. The plot is showing if atoms in the starting model 2x2x3 supercell of the β-Bi$_2$O$_3$ are favourable for the fit quality. For the models where the atom is not present in the model, the SHAP value is shown in blue, while it is shown in red for the atoms where it is present in the model. The SHAP values are plotted as a violin plot.

R: Fitting details of the fit using the best model found with the brute-force approach and the best model found with the ML-MotEx approach

Example system	Brute-Force	ML-MotEx
r-range (Å)	2 – 15	2 – 15
Scale	0.89	0.86
ADP$_{Bi}$ (Å^2)	1.71	1.71
ADP$_O$ (Å^2)	Fixed to 1	Fixed to 1
Isotropic expansion	0.98	0.98

Table S7 | Refinement details the best fitting cluster extracted with brute-force modelling and ML-MotEx to the background subtracted PDF measured on the [Bi$_{38}$O$_{45}$] cluster.

S: Fits of the PDF data using the [Bi$_{38}$O$_{45}$] cluster as model
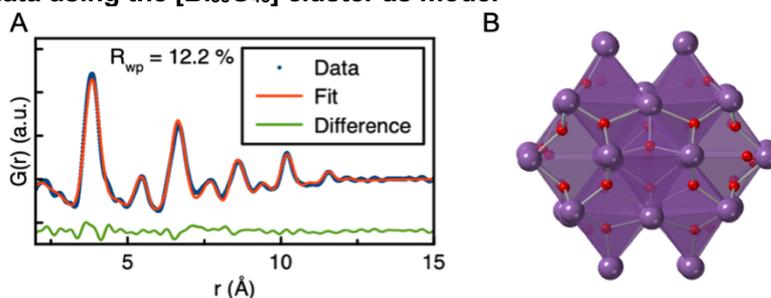
Figure S25 | [Bi$_{38}$O$_{45}$] fit. Fit of the A) PDF data using the B) [Bi$_{38}$O$_{45}$] cluster as model.

Example system	PDF
r-range (Å)	2 – 15
Scale	0 – 10 (0.91)
ADP$_{Bi}$ (Å^2)	0 – 4 (1.3)
ADP$_O$ (Å^2)	Fixed to 1
Isotropic expansion	0.95 – 1.05 (1.01)

Table S8 | Refinement details the fit of the [Bi$_{38}$O$_{45}$] cluster to the background subtracted PDF data from the [Bi$_{38}$O$_{45}$] cluster. The initial guesses of the parameters are written inside the parentheses, while the fitting range is written outside the outside the parentheses.

T: Enlarged SHAP plots
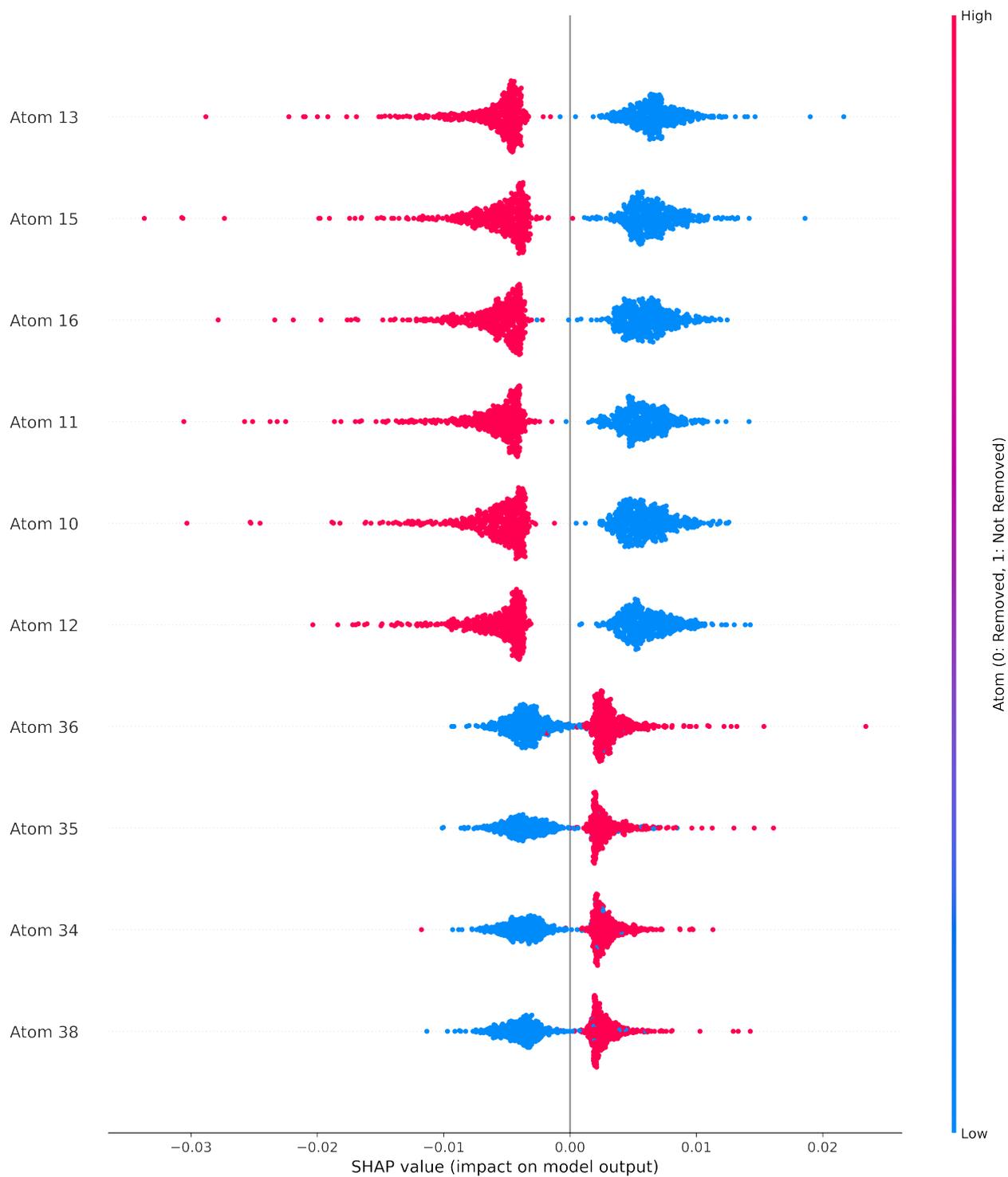
Figure S26 | Enlarged summary SHAP plot of Figure 5A removing the number of atoms to highlight the effect of each atom.
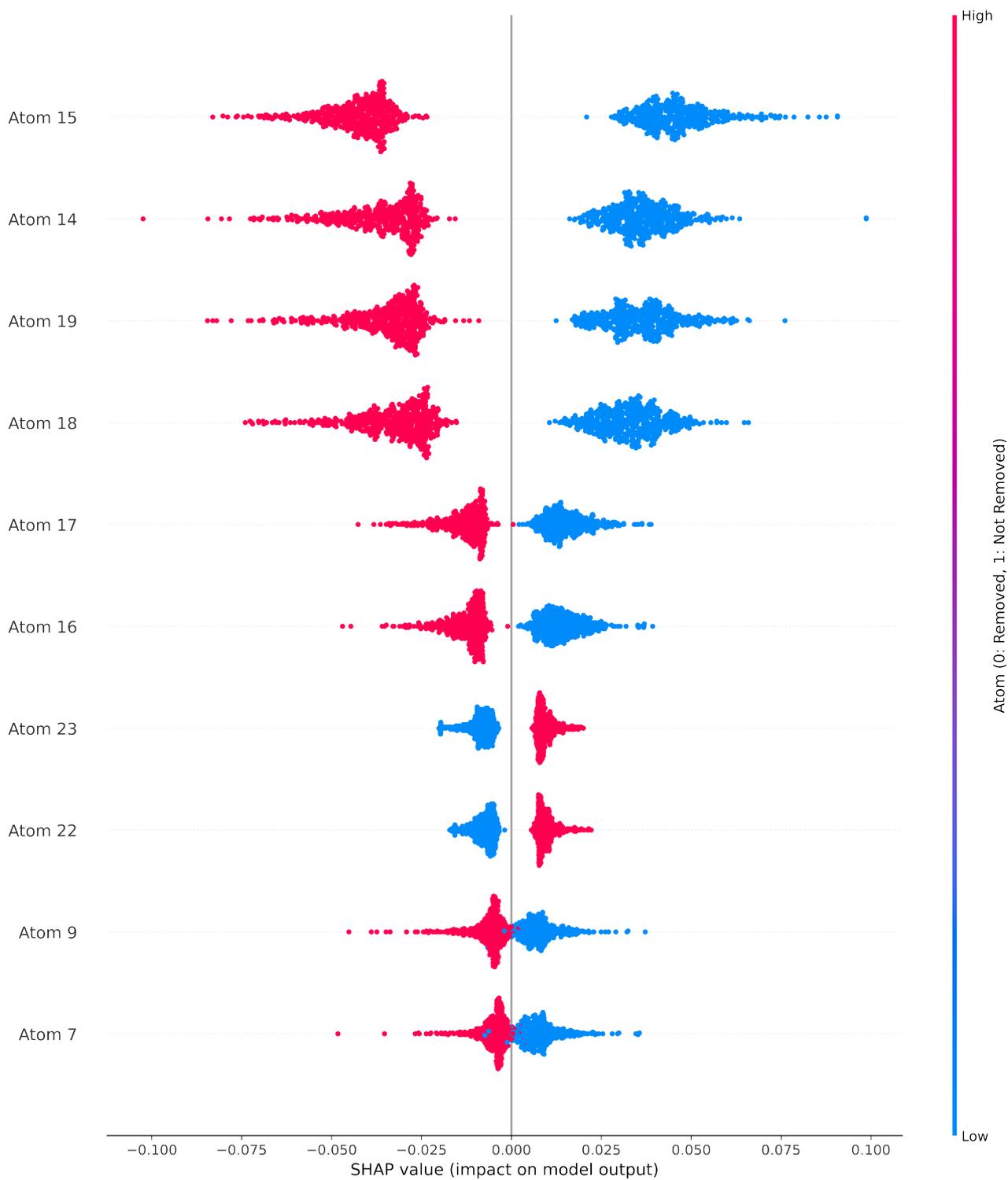
Figure S27 | Enlarged summary SHAP plot of Figure S13A removing the number of atoms to highlight the effect of each atom.
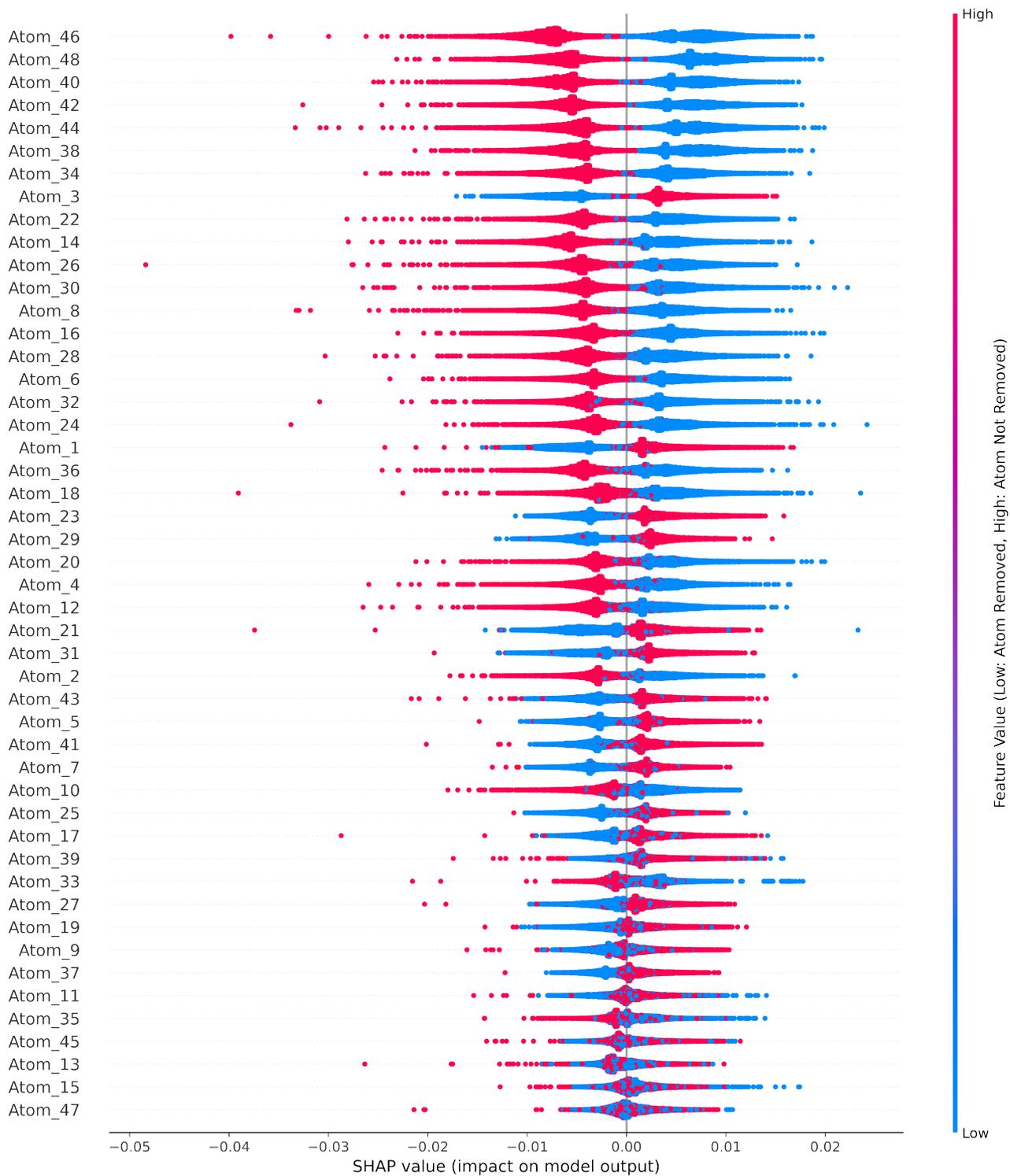
Figure S28 | Enlarged summary SHAP plot of Figure S2, left removing the number of atoms to highlight the effect of each atom.
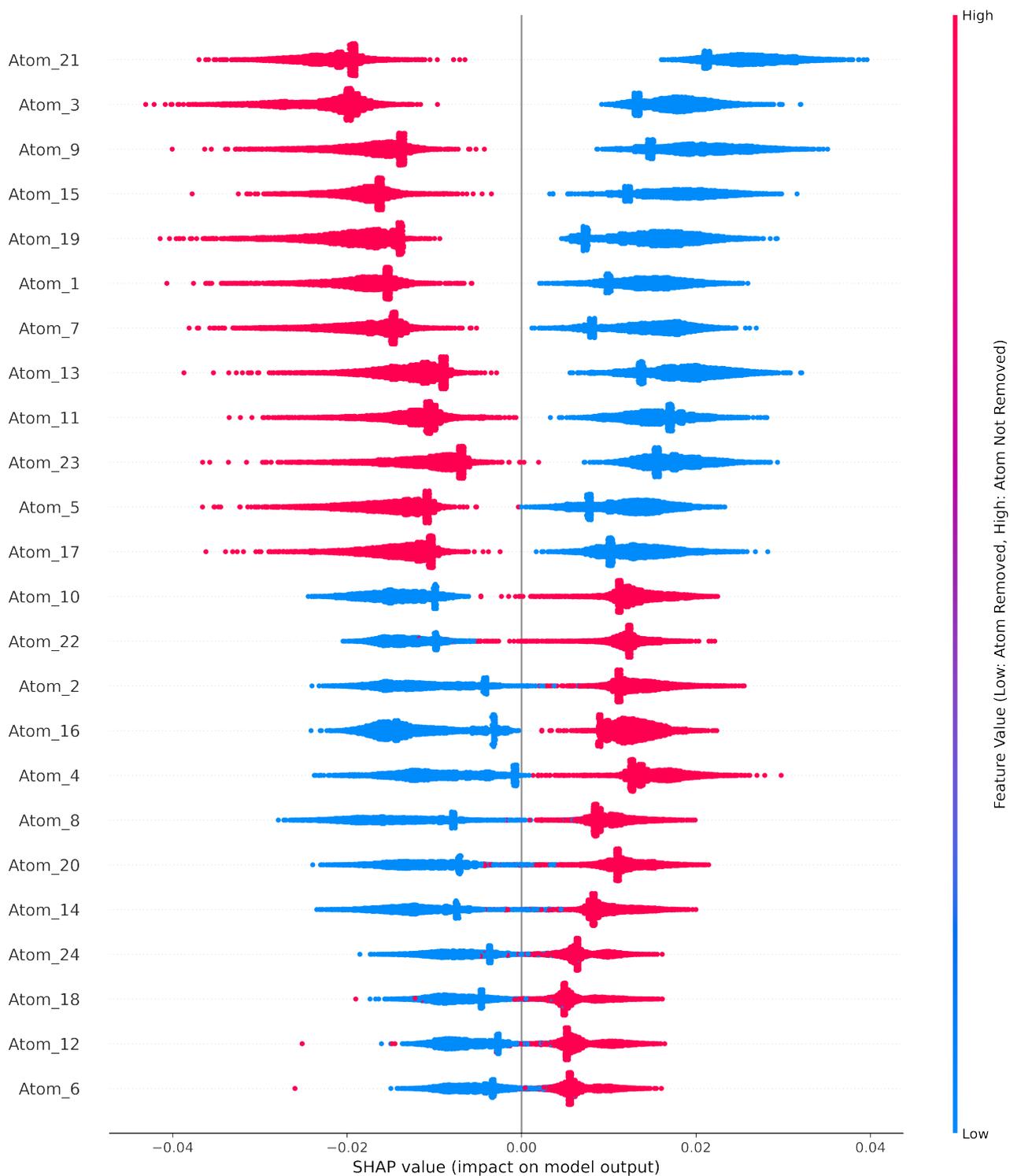
Figure S29 | Enlarged summary SHAP plot of Figure S3,right removing the number of atoms to highlight the effect of each atom.
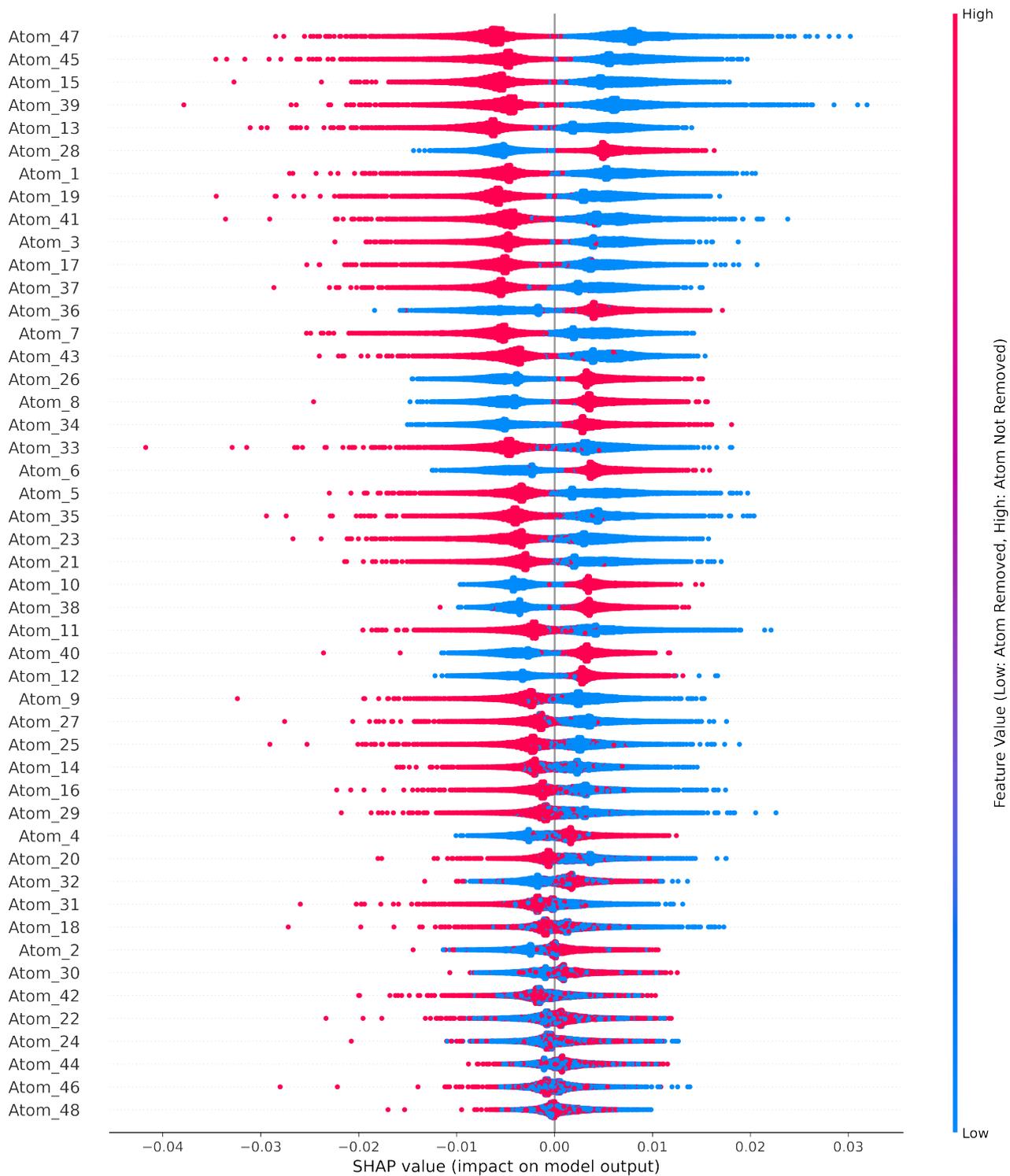
Figure S30 | Enlarged summary SHAP plot of Figure S2,right removing the number of atoms to highlight the effect of each atom.
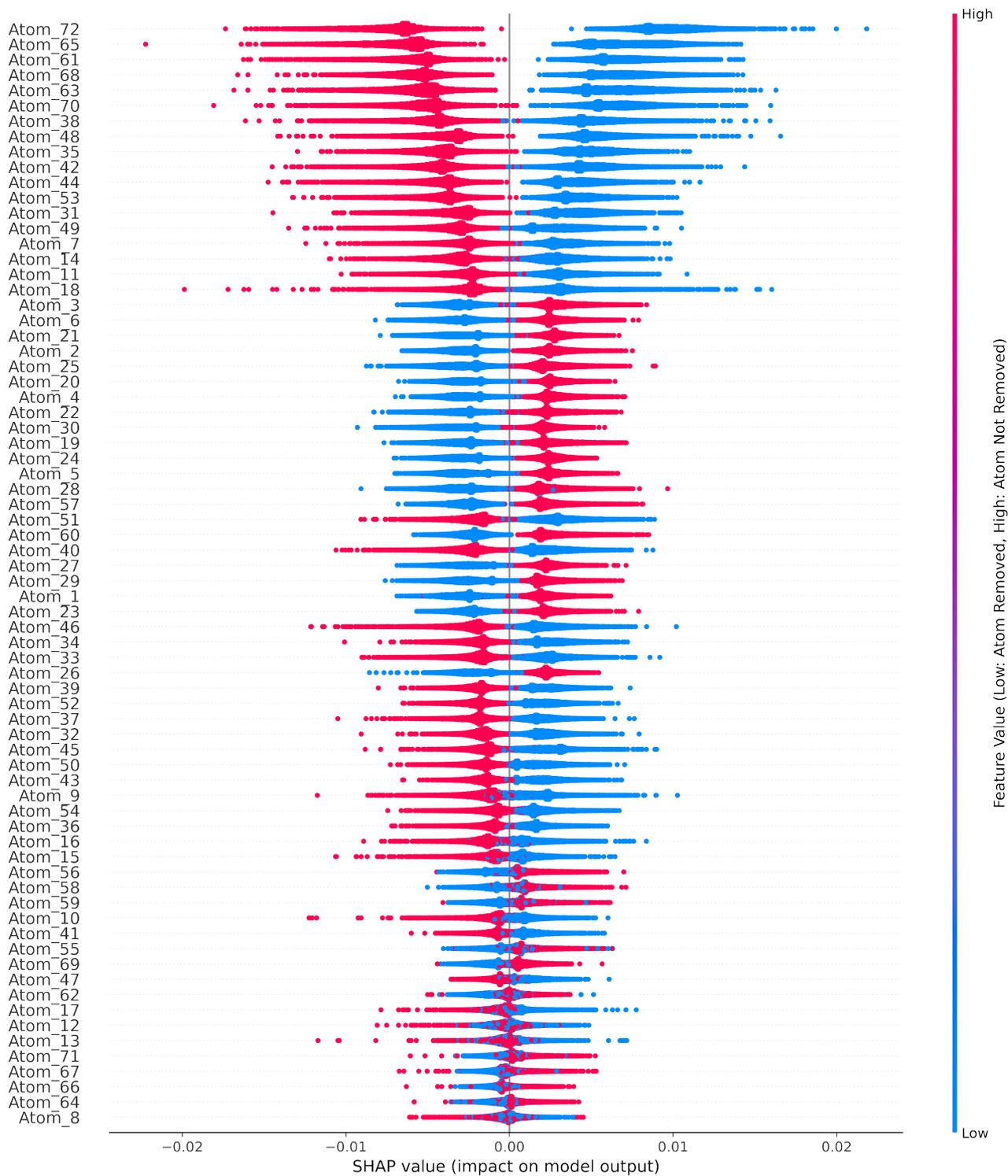
Figure S31 | Enlarged summary SHAP plot of Figure S3,left removing the number of atoms to highlight the effect of each atom.

U: References

1.   Joachim, F.; Axel, T.; Rosemarie, P., Strukturen und Schwingungsspektren des Tetramethylammonium-α-dodekawolframatosilikats und des Tetrabutylammonium-β-dodekawolframatosilikats: Structures and Vibrational Spectra of Tetramethylammonium α-

Dodecatungstosilicate and Tetrabutylammonium β-Dodecatungstosilicate. *Z. Naturforsch.* **1981,** *36* (2), 161-171.

2.   Niu, J.; Zhao, J.; Wang, J.; Bo, Y., Syntheses, spectroscopic characterization, thermal behavior, electrochemistry and crystal structures of two novel pyridine metatungstates. *J. Coord. Chem.* **2004,** *57* (11), 935-946.

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

4.   Busbongthong, S.; Ozeki, T., Structural Relationships among Methyl-, Dimethyl-, and Trimethylammonium Phosphododecatungstates. *Bull. Chem. Soc. Jpn.* **2009,** *82* (11), 1393-1397.

5.   Chen, X.; Yamanaka, S., Single-crystal X-ray structural refinement of the `tetragonal' C_{60} polymer. *Chem. Phys. Lett.* **2002,** *360* (5), 501-508.

6.   Juelsholt, M.; Lindahl Christiansen, T.; Jensen, K. M. Ø., Mechanisms for Tungsten Oxide Nanoparticle Formation in Solvothermal Synthesis: From Polyoxometalates to Crystalline Materials. *J. Phys. Chem. C* **2019,** *123* (8), 5110-5119.

7.   Chupas, P. J.; Qiu, X.; Hanson, J. C.; Lee, P. L.; Grey, C. P.; Billinge, S. J. L., Rapid-acquisition pair distribution function (RA-PDF) analysis. *J. Appl. Cryst.* **2003,** *36* (6), 1342-1347.

8.   Hammersley, A. P., FIT2D : a multi-purpose data reduction, analysis and visualization program. *J. Appl. Cryst.* **2016,** *49* (2), 646-652.

9.   Yang, X.; Juhas, P.; Farrow, C. L.; Billinge, S. J. L., xPDFsuite: an end-to-end software solution for high throughput pair distribution function transformation, visualization and analysis. *eprint arXiv:1402.3163* **2014**, arXiv:1402.3163.

10.   Redrup, K. V.; Weller, M. T., Hydrothermal routes to new sodium hydrogen polytungstates. *Dalton Trans.* **2009,** (23), 4468-4472.

11.   Poimanova, O. Y.; Radio, S. V.; Bilousova, K. Y.; Baumer, V. N.; Rozantsev, G. M., Equilibria in the acidified aqueous dimethylformamide solutions of tungstate anion. Synthesis, crystal structure and characterization of novel decatungstate [Ba(H2O)2(C3H7NO)3]2[W10O32]·(C3H7NO)2. *J. Coord. Chem.* **2015,** *68* (1), 1-17.

12.   Evans, H. T., Crystal structure studies of isopoly and heteropoly oxometalates. Structure of the paradodecatungstate molecule and its environment in crystals. In *Polyoxometalates: From Platonic Solids to Anti-Retroviral Activity*, Springer: 1994; pp 71-86.

13.   Ogiwara, N.; Tomoda, M.; Miyazaki, S.; Weng, Z.; Takatsu, H.; Kageyama, H.; Misawa, T.; Ito, T.; Uchida, S., Integrating molecular design and crystal engineering approaches in non-humidified intermediate-temperature proton conductors based on a Dawson-type polyoxometalate and poly (ethylene glycol) derivatives. *Nanoscale* **2021,** *13* (17), 8049-8057.

14.   Christiansen, T. L.; Kjær, E. T. S.; Kovyakh, A.; Röderen, M. L.; Høj, M.; Vosch, T.; Jensen, K. M. Ø., Structure analysis of supported disordered molybdenum oxides using pair distribution function analysis and automated cluster modelling. *J. Appl. Cryst.* **2020,** *53* (1), 148-158.

15.   Krebs, B.; Paulat-Böschen, I., The structure of the potassium isopolymolybdate $K_8[Mo_{36}O_{12}(H_2O)_{16}]\cdot nH_2O$ (n = 36···40). *Acta Crystallogr. B* **1982,** *38* (6), 1710-1718.

16.   Miersch, L.; Schlesinger, M.; Troff, R. W.; Schalley, C. A.; Rüffer, T.; Lang, H.; Zahn, D.; Mehring, M., Hydrolysis of a Basic Bismuth Nitrate—Formation and Stability of Novel Bismuth Oxido Clusters. *Chem. Eur. J.* **2011,** *17* (25), 6985-6990.

17.   Mehring, M., Metal Oxido Clusters of Group 13–15 Elements. In *Clusters – Contemporary Insight in Structure and Bonding*, Dehnen, S., Ed. Springer International Publishing: Cham, 2017; pp 201-268.

18.   Wadell, H., Volume, Shape, and Roundness of Quartz Particles. *The Journal of Geology* **1935,** *43* (3), 250-280.