Electronic Supplementary Material (ESI) for Digital Discovery.

Supplementary Information for:

MLstructureMining: A machine learning tool for structure identification from X-ray pair distribution function data

Emil T. S. Kjær[1], Andy S. Anker[1], Andrea Kirsch[1], Joakim Lajer[1], Olivia Aalling-Frederiksen[1], Simon J. L. Billinge[*2], Kirsten M. Ø. Jensen*[*1]*

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

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

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

Table of Contents

A: The Pair Distribution Function (PDF)

A PDF can be interpreted as a weighted histogram of atom – atom distances, and is the Fourier transform of X-ray, neutron or electron total scattering data. To obtain the PDF, the measured scattering intensity, I(Q), undergoes corrections for fluorescence, incoherent scattering and normalization.[11,17,18] The PDF or G(r) is then obtained by calculating the Fourier transformation over a truncated Q-interval, Q_{min} to Q_{max} as $G(r) = \frac{2}{\pi}\int_{Q_{min}}^{Q_{max}} Q[S(Q)-1]\sin(Q\cdot r)dQ$. Here, the scattering vector Q is dependent on the wavelength of the incoming beam, λ, and the scattering angle, θ, and is defined as $Q = \frac{4\cdot\pi\cdot\sin(\theta)}{\lambda}$.

B: Training and evaluating the MLstructureMining model

For the training, validation, and testing process of MLstructureMining, 100 Pair Distribution Functions (PDFs) of each of the entries in the structure catalogue were simulated. The simulations were done using Latin hypercube sampling[1] with the parameters shown in Table S1. The objective function of MLstructureMining was a multiclass log loss function.[2]

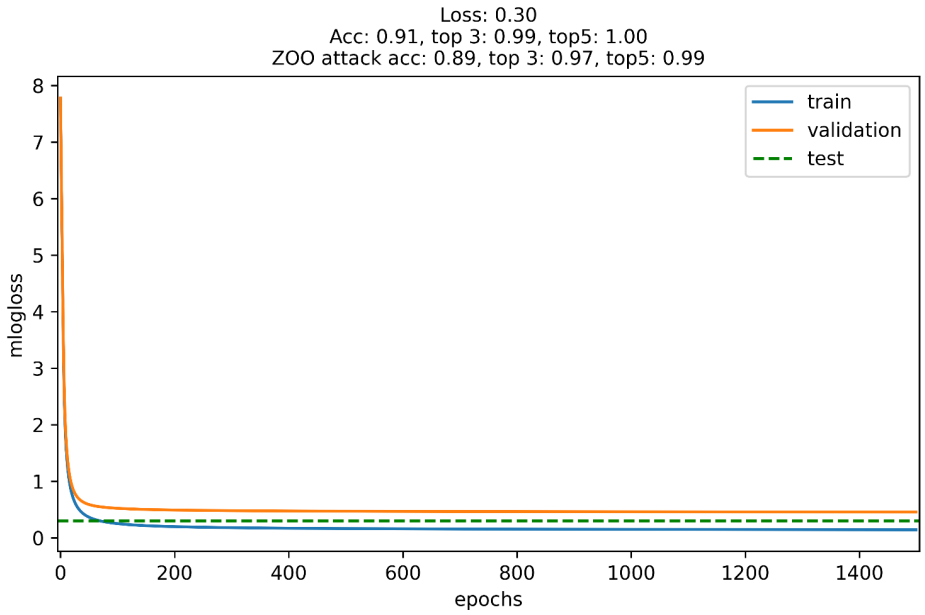

Figure S1 | Training and validation loss curve of MLstructureMining. The training (blue) and validation (orange) loss curves shown as a function of epochs. MLstructureMining obtains an accuracy of 91%, a top-3 accuracy of 99% and a top-5 accuracy of 100% on the test seen which were not seen during training. Additional, MLstructureMining's robustness was tested using Zeroth-Order Optimization (ZOO), which gave an accuracy of 89%, a top-3 accuracy of 97% and a top-5 accuracy of 99%.

B.1.: Pair Distribution Function (PDF) simulation parameters

The PDF simulation parameters are shown in Table S1. All PDFs were simulated using the DiffPy-CMI[3] Python library and Latin hypercube sampling[1] was used to sample the PDF simulation parameter space.

Table S1 | PDF simulation parameter range used for training, validation and testing set. All parameters only containing one value was not changes for any of the simulations. The unit cell parameter is only varying the dimensions of the unit cell (*a*, *b* and *c* sides) and only if they are allowed to change without breaking the symmetry.

Parameter	Value
R_{min} [Å]	0.00
R_{max} [Å]	30.00
R_{step} [Å]	0.10
Q_{min} [Å^{-1}]	0.70
Q_{max} [Å^{-1}]	20.00
Q_{damp} [Å^{-1}]	0.04
δ_2 [Å^2]	2.00
P_{size} [Å]	No size dampening applied
U_{iso} [Å^2]	0.005 – 0.025
Unit cell* [%]	± 4.00
Number of simulated PDF	100

B.2.: XGBoost hyperparameters

To determine the hyperparameters of MLstructureMining, Bayesian optimization was performed using the bayesian-optimization Python library.[4] The Bayesian optimization was run using 10 'init_points' and 20 'n_iter' training a total of 30 models. The allowed range during Bayesian optimization for all hyperparameters are shown in Table S2.

Table S2 | The hyperparameter range used during Bayesian optimization.

Parameter	Range
learning_rate	(0.05, 1.0)
min_child_weight	(0.1, 10)
max_depth	(3, 6)
max_delta_step	(0, 20)
subsample	(0.01, 1.0)
colsample_bytree	(0.01, 1.0)
colsample_bylevel	(0.01, 1.0)
reg_lambda	(0, 10.0)
reg_alpha	(0, 10.0)
gamma	(0, 10.0)

The hyperparameters used for the best XGBoost[5] classifier model are listed below. Parameters not listed in Table S3 are using default values.

Table S3 | XGBoost hyperparameter values.

Parameter	Value
random_state	0
num_class	6062
tree_method	hist

objective	multi:softprob
early_stopping_rounds	25
eval_metric	mlogloss
learning_rate	0.3
min_child_weight	1
max_depth	6
max_delta_step	0
subsample	1
colsample_bytree	1
colsample_bylevel	1
reg_lambda	1
reg_alpha	0
gamma	0

B.3.: Testing robustness of the model using Adversarial Robustness Toolbox (ART)

To further evaluate MLstructureMining the Adversarial Robustness Toolbox (ART)[6] was utilized to test its robustness. Here, Zeroth-Order Optimization (ZOO) was deployed to perform adversarial attacks on the model. The ZOO attack was performed using the test data.

ZOO is an adversarial attack technique that can find weaknesses in a model without having direct access to the model's intricate details. Instead of requiring direct access to the model's internal components, ZOO approximates the model's behavior using only its outputs. By probing the model with various inputs and observing the outputs, ZOO can craft adversarial examples that deceive the model, all without ever peeking inside it. In essence, ZOO is making small changes to the data to test the model's robustness. The parameters used for ZOO are shown in Table S4.

In our study, we utilized the ART library to deploy the ZOO attack on our model, emphasizing the importance of robust machine learning practices.

Table S4 | Input values for ZOO attack.

Parameter	**Value**
confidence	0.0
targeted	False
learning_rate	1e-1
max_iter	10
binary_search_steps	3
initial_const	0.001
abort_early	True
use_resize	False
use_importance	False
nb_parallel	1
batch_size	1
variable_h	0.2

C: The Pearson Correlation Coefficient (PCC)

We use the PCC to obtain a discrete measure of comparison between two continuous functions, PDFs. By providing two equally r-sampled PDFs the PCC will return a value between -1 and 1. A value of 1

corresponds to a perfect linear correlation between the two PDF's. -1 indicates the exact opposite relationship between the two functions. The PCC is defined as shown below:[7]

$$r = \frac{1}{1-n}\sum_{i=0}^{n}\left(\frac{X_i - \bar{X}}{\sigma_x}\right)\left(\frac{Y_i - \bar{Y}}{\sigma_y}\right)$$

The two compared datasets are denoted X and Y. We sum over all points in the two datasets. $\bar{X}$ and $\bar{Y}$ are the mean values of the functions while σ_x and σ_y are their standard deviation. The PCC is scale invariant which makes it ideal for comparing the peak positions of various functions. This results in the PCC being highly sensitive to shifts in peak positions (lattice parameters) and not sensitive to peak intensities (different atomic species). Using the PCC as a measure of similarity between PDF data has been done in multiple occasions.[8-10]

D: Synthesis and data collection:

D.1. Example 1: $CoFe_2O_4$

Cobalt iron oxide was synthesized using a hydrothermal synthesis approach. 0.5 mmol $CoCl_2 \cdot 6H_2O$ (ACS reagent, 98%) and 1.0 mmol $Fe(NO_3)_3 \cdot 9H_2O$ (ACS reagent, ≥98 %) were added to water. 1.0 mmol KOH (ACS reagent, $\geq 85\%$) was added and the solution was sonicated for 30 min, transferred to a Teflon lined autoclave and heated for 2 hours at 60 °C followed by 1 hour at 160 °C. The autoclave was cooled down and the formed powder washed first in pure hexane followed by a mixture of hexane and ethanol (1:3) and dried overnight in ethanol.

X-ray total scattering data were collected using a Panalytical Empyrean Series 2 diffractometer equipped with an Ag-source with X-ray wavelength 0.56 Å and a GaliPIX detector. The PDF was generated using PDFgetX3[11] with Q_{min} = 1.6 Å^{-1} and Q_{max} = 17.5 Å^{-1}, Q_{damp} = 0.04 Å^{-1} and r_{poly} = 0.9 Å.

D.2. Example 2: CeO_2

$[Ce_6(\mu_3\text{-}O)_4(\mu_3\text{-}OH)_4(NH_3CH_2COO)_8\text{-}(NO_3)_4(H_2O)_6]Cl_8 \cdot 8H_2O$ crystals were dissolved in DMSO at 80 °C until fully dissolved. 0.05 M NaOH were added while stirring vigorously for 3 minutes. The powder was lastly annealed at 60 °C for 3 hours. The powder was transferred to a Kapton tube with an inner diameter of 1.05 mm and X-ray total scattering data were collected using the RA-PDF geometry with x-ray wavelength 0.2072 Å at beamline P02.1 at PETRAIII, DESY, Hamburg. The data were integrated using Fit2D,[12] and PDFs were obtained using PDFgetx3 using Q_{min} = 0.7 Å^{-1} and Q_{max} = 24 Å^{-1}, Q_{damp} = 0.04 Å^{-1} and r_{poly} = 0.9 Å.[11]

D.3. Example 3: W_5O_{14}

The data presented in the article is the last frame of an *in situ* PDF series on the formation of W_5O_{14}. To synthesize the tungsten oxide nanoparticles the precursor consisted of WCl_6 (≥99.99 %, Sigma–Aldrich) was dissolved in 15 mL of isopropanol to reach a 0.3 M tungsten concentration. The precursor solution was continuously mixed until all of the WCl_6 powder was dissolved, and the solution turned dark blue.

The synthesis was done in a custom-made reaction cell specifically design for performing *in situ* X-ray total scattering experiments. The setup is similar to the one described by Becker et al.[12,13] The precursor suspension was injected into a fused silica tube with 0.7 mm inner diameter and 0.09 mm wall thickness. Throughout the synthesis the pressure was kept constant at 100 bar using an HPLC pump. To apply heat during the reaction a heat gone was placed centered underneath the silica tube. Heat was applied after approximately 30 seconds and kept stable once the reaction reached 310 °C.

The scattering experiment was performed at the P02.1, PETRA III beamline at DESY in Germany, using a wavelength of 0.2072 Å and a sample to detector distance of 210 mm. The detector used is A Perkin Elmer XRD1621 area detector with a pixel size of 0.2×0.2 mm. The sample and detector setup used was the Rapid Acquisition Pair Distribution Function (RA-PDF) setup.[14]

The calibration was done in Fit2D[15] while the azimuthal integration was done with PyFAI[16] and the PDF was calculated using PDFgetX3 with Q_{min} = 0.7 Å^{-1} and Q_{max} = 15 Å^{-1}, Q_{damp} = 0.04 Å^{-1} and r_{poly} = 0.9 Å.[11]

D.4. Example 4: $Bi_2Fe_4O_9$

The precursor sample used for the *in situ* heating experiments was synthesized using the sol-gel method. First, 5 mmol $Bi(NO_3)_3 \cdot 5\,H_2O$ (reagent grade, 98% Sigma-Aldrich) and 10 mmol $Fe(NO_3)_3 \cdot 9\,H_2O$ (reagent grade, 98% Sigma-Aldrich), 11.25 mmol complexing agent (meso-Erythritol, ≥99% Sigma-Aldrich) are dissolved in 27.5 mL deionized water. Heating the solution under stirring to 100 °C leads to a gel, which is subsequently heated at 250 °C for 18 h in an oven to remove residual organics. The resulting solid powdered precursor is then ground and used for the *in situ* total scattering experiments. These were carried out at beamline P21.1 at PETRA III/DESY in Hamburg, Germany using a wavelength of 0.12203 Å, a Perkin-Elmer XRD1621 area detector with a sample detector distance of ca. 500 mm. The powder sample was filled into a quartz capillary with a diameter of 2 mm. The capillary was positioned horizontally, left open at one end, and only half filled to allow for gas exchange. The powder was hold in place using quartz wool and heated with a hot air blower from room temperature to 700 °C with a heating rate of 20 K/min and kept for ca. 2 h. Data were integrated using pyFAI[16] and the PDFs obtained using xPDFsuite[12] with a Q_{min} = 0.1 Å^{-1}, Q_{max} = 21.5 Å^{-1}, Q_{damp} = 0.04 Å^{-1} and r_{poly} = 0.9. The scattering pattern of an empty quartz capillary was used for background subtraction.

E: Baseline refinements of experimental PDF's

The baseline fits for the four different experimental PDFs are shown in Fig. S2 and each fits refined parameters are shown in Table S5. The real-space Rietveld refinement of $CoFe_2O_4$, CeO_2, W_5O_{14} $Bi_2Fe_4O_9$ are shown in Fig. S2.

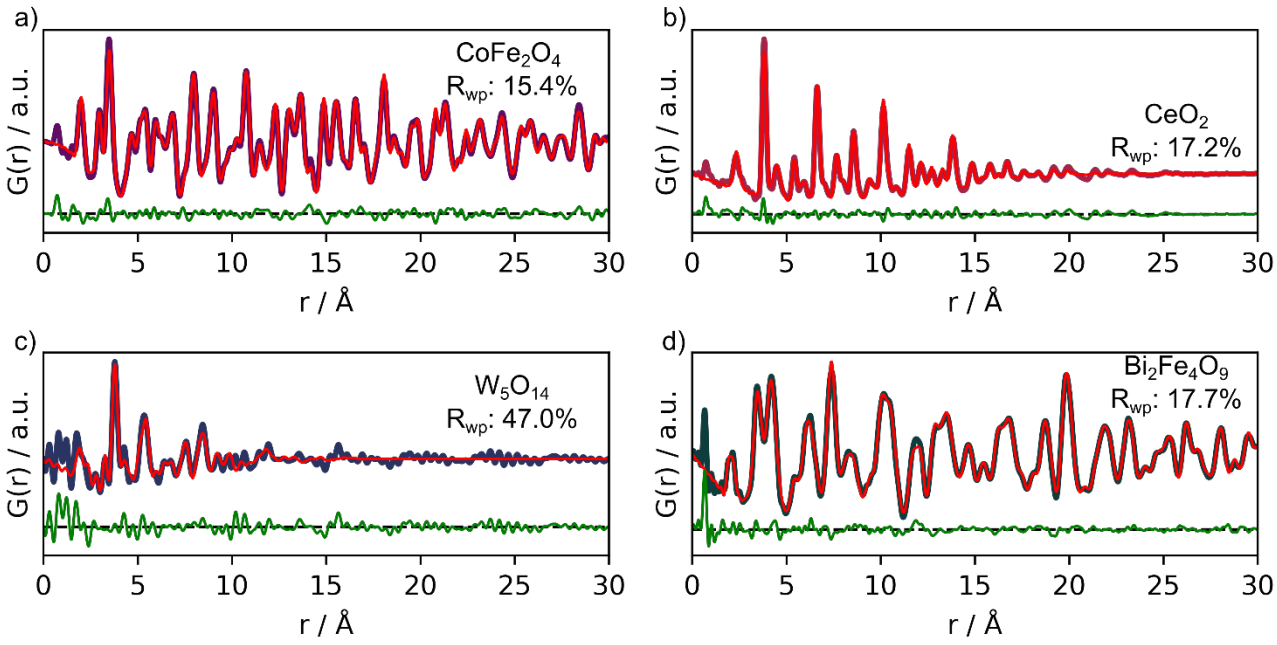

Figure S2 | Plot of the baseline PDF refinements for the four experimental PDFs. a) baseline fit of experimental PDF 1: spinel $CoFe_2O_4$ with a R_{wp} of 15.4%, b) baseline fit of experimental PDF 2: fluorite CeO_2 with a R_{wp} of 17.2%, c) baseline fit of experimental PDF 3: W_5O_{14} with a R_{wp} of 47.0% and d) baseline fit of experimental PDF 4: mullite $Bi_2Fe_4O_9$ with a R_{wp} of 15.4%.

Table S5 | Fitted baseline parameters of the PDF refinements for the four experimental PDF. Real-space Rietveld refinements are shown in Fig. S2.

Structure	R_{wp} [%]	Space group	Scale	*a* [Å]	*b* [Å]	*c* [Å]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
$CoFe_2O_4$	15.4	$Fd\bar{3}m$	0.59	8.4	-	-	93	2.9	0.012	0.027
CeO_2	17.2	$Fd\bar{3}m$	0.64	5.4	-	-	23	3.4	0.008	0.046
W_5O_{14}	48.2	$P\bar{4}21m$	0.72	23.4	-	3.8	19	2.7	0.005	0.005
$Bi_2Fe_4O_9$	17.7	$Pbam$	0.19	8.0	8.5	6.0	319	3.5	0.022	0.066

E.1. Size estimation of $CoFe_2O_4$ and $Bi_2Fe_4O_9$

To estimate size of the experimental syntheses of $CeFe_2O_4$ and $Bi_2Fe_4O_9$ real-space Rietveld refinement from 0 – 60 Å was performed. The fits are shown in Fig. S3 and the fitted parameters in Table S6.

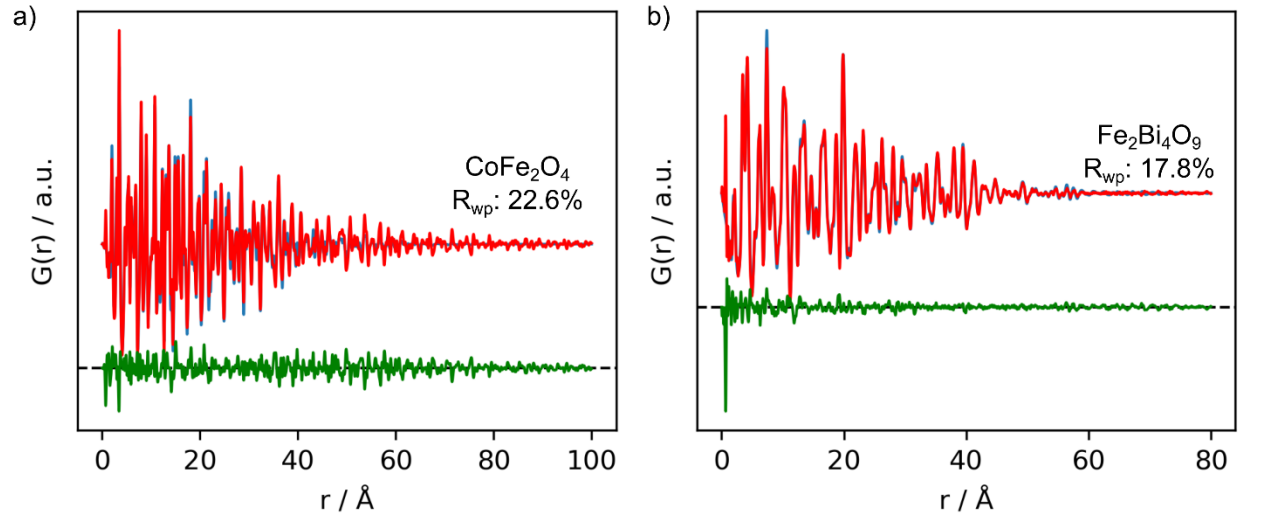

Figure S3 | Long range real-space Rietveld refinement of $Fe_2Bi_4O_9$ and $CoFe_2O_4$. a) baseline fit of example 1: spinel $CoFe_2O_4$ with a R_{wp} of 22.6% and b) baseline fit of example 4: mullite $Bi_2Fe_4O_9$ with a R_{wp} of 17.8%.

Table S6 | Fitted baseline parameters of the PDF refinements for the four examples. Real-space Rietveld refinements are shown in Fig. S2.

Structure	R_{wp} [%]	Scale	*a* [Å]	*b* [Å]	*c* [Å]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
$CoFe_2O_4$	22.6	0.55	8.4	-	-	174	3.3	0.013	0.029
$Bi_2Fe_4O_9$	17.8	0.20	8.0	8.5	6.0	230	3.2	0.022	0.062

F: Real-space Rietveld refinements of predicted structures

Real-space Rietveld refinement of MLstructureMining's top-5 predictions for each of the four examples.

F.1.: $CoFe_2O_4$ fit parameters

R_{wp} values and fitted parameters for $CoFe_2O_4$ are shown in the Table S7 below.

Table S7 | Fitted structure parameters of top-5 structure predictions on $CoFe_2O_4$.

Structure [COD ID]	R_{wp} [%]	Scale	*a* [Å]	*c* [Å]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1536758	16.9	0.58	5.9	8.4	98	2.0	0.011	0.019
1537073	17.7	0.59	8.4	-	87	1.8	0.012	0.017
1541403	17.3	0.57	8.4	-	89	1.9	0.012	0.015
5910031	49.7	0.28	8.4	-	250	5.0	0.009	0.069
1539596	27.5	0.49	8.4	-	106	2.1	0.013	0.010

F.2.: CeO_2 fit parameters

R_{wp} values and fitted parameters for CeO_2 are shown in the Table S8 below.

Table S8 | Fitted structure parameters of top-5 structure predictions on CeO_2.

Structure [COD ID]	R_{wp} [%]	Scale	*a* [Å]	*c* [Å]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1006067	16.5	0.63	3.8	19.0	24	3.5	0.007	0.044
1527617	17.3	0.76	5.5	10.7	24	3.5	0.007	0.029
1527729	96.2	0.34	9.3	-	21	2.2	0.007	3.590
2102840	17.9	0.66	3.8	5.5	24	3.8	0.007	0.055
1537009	16.1	0.61	5.4	-	24	3.8	0.008	0.015

F.3.: W_5O_{14} fit parameters

R_{wp} values and fitted parameters for W_5O_{14} are shown in the Table S9 below.

Table S9 | Fitted structure parameters of top-5 structure predictions on W_5O_{14}.

Structure [COD ID]	R_{wp} [%]	Scale	*a* [Å]	*b* [Å]	*c* [Å]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
9014025	66.9	0.73	5.4	-	10.6	8	3.0	0.005	16.470
7230340	67.2	0.56	24.1	5.3	5.3	12	3.0	0.002	0.000
1536855	61.6	1.23	5.4	10.4	13.5	11	4.6	0.005	17.999
2311027	75.1	0.45	8.7	-	-	7	4.7	0.000	0.028
9007595	56.2	0.57	5.55	11.7	5.2	12	3.3	0.000	0.005

F.4.: $Bi_2Fe_4O_9$ fit parameters

R_{wp} value and fitted parameters for $Bi_2Fe_4O_9$ are shown in the Table S10 below.

Table S10 | Fitted structure parameters of top-5 structure predictions on $Bi_2Fe_4O_9$.

Structure [COD ID]	R_{wp} [%]	Scale	*a* [Å]	*b* [Å]	*c* [Å]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
4342599	23.6	0.19	8.0	8.5	6.0	187	3.6	0.020	0.076
2104768	30.4	0.20	8.0	8.5	6.0	172	3.3	0.020	0.052
2002314	22.8	0.22	8.0	8.5	6.0	186	2.4	0.021	0.084
2002219	64.9	0.25	8.3	-	6.0	53	3.6	0.048	0.283
2106245	78.2	0.24	8.5	5.3	10.7	64	1.0	0.021	0.031

F.5.: $Bi_2Fe_4O_9$ fit parameters for NMF component 1

R_{wp} value and fitted parameters for $Bi_2Fe_4O_9$ are shown in the Table S11 below.

Table S11 | Fitted structure parameters of top-5 structure predictions on NMF component 1.

Structure [COD ID]	R_{wp} [%]	Scale	*a* [Å]	*b* [Å]	*c* [Å]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
4342599	26.6	0.16	8.0	8.5	6.0	187	3.7	0.019	0.117
2002219	64.2	0.20	8.3	-	6.0	54	3.6	0.044	0.345
2104768	34.6	0.16	8.0	8.5	6.0	161	2.8	0.019	0.058
9000738	55.8	0.22	8.3	-	18.2	65	1.0	0.026	0.086
9008148	22.8	0.17	8.0	8.5	6.0	172	2.7	0.020	0.095

G: PDF in the cloud (PDFitc) benchmark tests

This section shows the benchmark results produced by PDF in the cloud (PDFitc). Each of the four experimental PDF have been tested and fitted.

G.1.: $CoFe_2O_4$

Table S12 | PDFitc's settings for the benchmark test on the experimental PDF of $CoFe_2O_4$. This query resulted in a total of 151 structures and 70 structures had a R_{wp} below 50%.

Scattering type	Composition	Optional parameter	Type of PDF
X-ray	Fe-O	rmin=0 rmax=30 qmin=1.6 qmax=17.5 spd=93	Experimental

Table S13 | PDFitc's top-5 structure predictions for $CoFe_2O_4$.

Rank	Composition	Space group	R_{wp}	COD ID
1.	Fe_3O_4	$P2/c$	17.1%	1532800
2.	Fe_3O_4	$Fd\bar{3}m$	17.4%	1513301
3.	Fe_3O_4	$R\bar{3}m$	17.4%	1526955
4.	Fe_3O_4	$Fd\bar{3}m$	17.7%	9006199
5.	Fe_3O_4	$Fd\bar{3}m$	17.8%	9016805

Table S14 | Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of $CoFe_2O_4$.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*β* [°]	p_{size} [Å]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1532800	5.9	6.0	16.7	90.5	108	0.009	0.021
1513301	8.4	-	-	-	80	0.011	0.020
1526955	5.9	-	14.6	-	93	0.010	0.022
9006199	8.4	-	-	-	80	0.011	0.022
9016805	8.4	8.3	5.9	-	80	0.011	0.023

G.2.: CeO_2

Table S15 | PDFitc's settings for the benchmark test on the experimental PDF of CeO_2. This query resulted in a total of 10 structures and 6 structures had a R_{wp} below 50%.

Scattering type	Composition	Optional parameter	Type of PDF
X-ray	Ce-O	rmin=0 rmax=30 qmin=0.7 qmax=24 spd=23	Experimental

Table S16 | PDFitc's top-5 structure predictions for CeO_2.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$Ce_4O_{6.64}$	$Fd\bar{3}m$	18.1%	1521459
2.	CeO_2	$Fd\bar{3}m$	19.0%	4343161
3.	CeO_2	$Fd\bar{3}m$	19.0%	9009008
4.	CeO_2	$Fd\bar{3}m$	19.0%	1562989
5.	$Ce_{11}O_{20}$	$P\bar{1}$	32.8%	1521460

Table S17 | Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of CeO_2.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*α* [°]	*β* [°]	*γ* [°]	p_{size} [Å]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1521459	5.4	-	-	-	-	-	23	0. 007	0. 032
4343161	5.4	-	-	-	-	-	22	0.007	0.032
9009008	5.4	-	-	-	-	-	22	0.007	0.032
1562989	5.4	-	-	-	-	-	22	0.007	0.032
1521460	6.7	10.3	6.6	90.6	98.7	97.7	26	0.002	0.048

G.3.: W_5O_{14}

Table S18 | PDFitc's settings for the benchmark test on the experimental PDF of W_5O_{14}. This query resulted in a total of 25 structures and 5 structures had a R_{wp} below 50%.

Scattering type	Composition	Optional parameter	Type of PDF
X-ray	W-O	rmin=0 rmax=30 qmin=0.7 qmax=15 spd=19	Experimental

Table S19 | PDFitc's top-5 structure predictions for W_5O_{14}.

Rank	Composition	Space group	R_{wp}	COD ID
1.	W_5O_{14}	$P\bar{4}21m$	44.9%	1527783
2.	$W_{18}O_{49}$	$P2/m$	45.4%	1001678
3.	$W_{18}O_{49}$	$P2/m$	45.4%	1528166
4.	$W_{18}O_{49}$	$P2/m$	46.2%	1538315
5.	$W_{10}O_{29}$	$P2/m$	48.3%	1538317

Table S20 | Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of W_5O_{14}.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*β* [°]	p_{size} [Å]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1527783	23.5	-	3.8	-	17	0. 006	0. 000
1001678	18.4	3.8	14.3	118.1	20	0.004	0.030
1528166	18.4	3.8	14.3	118.1	20	0.004	0.026
1538315	18.4	3.8	14.3	118.3	19	0.004	0.026
1538317	12.2	3.8	22.6	94.2	16	0.003	0.036

G.4.: $Bi_2Fe_4O_9$

For the experimental PDF of $Bi_2Fe_4O_9$ three benchmarks test cases we performed to screen a suitable chemical space.

Table S21 | PDFitc's settings for the three benchmark tests on the experimental PDF of $Bi_2Fe_4O_9$.

Case #	Scattering type	Composition	Optional parameter	Type of PDF	Total # structures	R_{wp} below 50%
1	X-ray	Fe-*-O9	rmin=0 rmax=30 qmin=0.7 qmax=15 spd=19	Experimental	9	2
2	X-ray	Fe4-*-O		Experimental	22	2
3	X-ray	*-Bi2-O		Experimental	115	2

Table S22 | Case 1: PDFitc's top-5 structure predictions for $Bi_2Fe_4O_9$.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$Fe_4Bi_2O_9$	$Pbam$	21.6%	1530918
2.	$Fe_4Bi_2O_9$	$Pbam$	21.9%	9008148
3.	$Fe_4P_2O_9$	$P21/c$	98.2%	1534301
4.	$Fe_2(SeO_3)_3$	$P63/m$	98.7%	1542288
5.	$Fe(PO_3)_3$	Cc	98.9%	1520966

Table S23 | Case 1: Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of $Bi_2Fe_4O_9$.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*β* [°]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1530918	8.0	8.5	6.0	-	163	0.018	0.021	0.042
9008148	8.0	8.5	6.0	-	153	0.019	0.020	0.053
1534301	6.7	11.2	9.5	105.6	-	-	-	-
1542288	7.3	-	7.4	-	-	-	-	-
1520966	13.4	19.0	9.6	19.0	-	-	-	-

Table S24 | Case 2: PDFitc's top-5 structure predictions for $Bi_2Fe_4O_9$.

Rank	Composition	Space group	R_{wp}	COD ID

1.	$Fe_4Bi_2O_9$	$Pbam$	21.6%	1530918
2.	$Fe_4Bi_2O_9$	$Pbam$	21.9%	9008148
3.	$Fe_4As_2O_{11}$	$P\bar{1}$	90.4%	9009251
4.	$Fe_8Cl_{1.35}O_{16}$	$C2/m$	95.5%	9001319
5.	$Fe_4As_5O_{13}$	$P\bar{1}$	95.6%	9004184

Table S25 | Case 2: Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of $Bi_2Fe_4O_9$.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*α* [°]	*β* [°]	*γ* [°]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1530918	8.0	8.5	6.0	-	-	-	163	0.018	0.021	0.042
9008148	8.0	8.5	6.0	-	-	-	153	0.019	0.020	0.053
9009251	6.5	6.6	5.0	105.7	98.5	110.2	-	-	-	-
9001319	10.6	3.0	10.4	-	91.0	-	-	-	-	-
9004184	9.0	10.2	9.1	60.5	111.6	80-5	-	-	-	-

Table S26 | Case 3: PDFitc's top-5 structure predictions for $Bi_2Fe_4O_9$.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$Fe_4Bi_2O_9$	$Pbam$	21.6%	1530918
2.	$Fe_4Bi_2O_9$	$Pbam$	21.9%	9008148
3.	$Ga_4Bi_2O_9$	$Pbam$	81.1%	1530919
4.	$Ga_4Bi_2O_9$	$Pbam$	84.3%	2002314
5.	$Ga_4Bi_2O_9$	$Pbam$	84.5%	4342602

Table S27 | Case 3: Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of $Bi_2Fe_4O_9$.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1530918	8.0	8.5	6.0	163	0.018	0.021	0.042
9008148	8.0	8.5	6.0	153	0.019	0.020	0.053
1530919	8.0	8.3	6.0	-	-	-	-
2002314	8.0	8.3	5.9	-	-	-	-
4342602	8.0	8.3	5.9	-	-	-	-

G.5.: $Bi_2Fe_4O_9$ NMF component

For the experimental PDF of $Bi_2Fe_4O_9$ three benchmarks test cases we performed to screen a suitable chemical space.

Table S28 | PDFitc's settings for the three benchmark tests on the PDF NMF component of $Bi_2Fe_4O_9$.

Case #	Scattering type	Composition	Optional parameter	Type of PDF	Total # structures	R_{wp} below 50%
1	X-ray	Fe-*-O9	rmin=0 rmax=30 qmin=0.7	Experimental	9	2
2	X-ray	Fe4-*-O		Experimental	22	2

3	X-ray	*-Bi2-O	qmax=15 spd=19	Experimental	115	2

Table S29 | Case 1: PDFitc's top-5 structure predictions for NMF component of $Bi_2Fe_4O_9$.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$Fe_4Bi_2O_9$	$Pbam$	22.5%	1530918
2.	$Fe_4Bi_2O_9$	$Pbam$	23.6%	9008148
3.	$Fe_2(CO)_9$	$P63/m$	98.7%	1010480
4.	$Fe_2(SeO_3)_3$	$P63/m$	98.9%	1542288
5.	$Fe_4P_2O_9$	$P21/c$	99.2%	1534301

Table S30 | Case 1: Fitted parameters of top-5 structures obtained from the benchmark test for the PDF NMF component of $Bi_2Fe_4O_9$.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*β* [°]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1530918	8.0	8.5	6.0	-	163	0.019	0.020	0.044
9008148	8.0	8.5	6.0	-	153	0.020	0.019	0.062
1010480	6.5	-	16.03	-	-	-	-	-
1542288	7.5	-	7.6	-	-	-	-	-
1534301	6.8	11.1	9.6	106.7	-	-	-	-

Table S31 | Case 2: PDFitc's top-5 structure predictions for NMF component of $Bi_2Fe_4O_9$.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$Fe_4Bi_2O_9$	$Pbam$	22.5%	1530918
2.	$Fe_4Bi_2O_9$	$Pbam$	23.6%	9008148
3.	$Fe_4As_2O_{11}$	$P\bar{1}$	89.9%	9009251
4.	$Fe_8Cl_{1.35}O_{16}$	$C2/m$	96.0%	9001319
5.	$Fe_4As_5O_{13}$	$P\bar{1}$	97.2%	9004184

Table S32 | Case 2: Fitted parameters of top-5 structures obtained from the benchmark test for the PDF NMF component of $Bi_2Fe_4O_9$.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*α* [°]	*β* [°]	*γ* [°]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1530918	8.0	8.5	6.0	-	-	-	163	0.019	0.020	0.044
9008148	8.0	8.5	6.0	-	-	-	153	0.020	0.019	0.062
9009251	6.5	6.6	5.0	105.8	98.4	110.2	-	-	-	-
9001319	10.6	3.0	10.4	-	91.0	-	-	-	-	-
9004184	9.1	10.2	9.1	60.4	111.4	80.3	-	-	-	-

Table S33 | Case 3: PDFitc's top-5 structure predictions for NMF component of $Bi_2Fe_4O_9$.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$Fe_4Bi_2O_9$	$Pbam$	22.5%	1530918
2.	$Fe_4Bi_2O_9$	$Pbam$	23.6%	9008148
3.	$Ga_4Bi_2O_9$	$Pbam$	81.7%	1530919
4.	$Ga_4Bi_2O_9$	$Pbam$	85.0%	2002314
5.	$Ga_4Bi_2O_9$	$Pbam$	85.6%	4342602

Table S34 | Case 3: Fitted parameters of top-5 structures obtained from the benchmark test for the PDF NMF component of $Bi_2Fe_4O_9$.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
1530918	8.0	8.5	6.0	163	0.019	0.020	0.044
9008148	8.0	8.5	6.0	153	0.020	0.019	0.062
1530919	8.0	8.3	6.0	-	-	-	-
2002314	8.0	8.3	5.9	-	-	-	-
4342602	8.0	8.3	5.9	-	-	-	-

H: Structure predictions for CeO_2, W_5O_{14}, $Bi_2Fe_4O_9$ and NMF component 1

Top-5 of MLstructureMining's structure predictions for the experimental PDFs of CeO_2, W_5O_{14}, $Bi_2Fe_4O_9$ and NMF component 1.

Table S35 | Top-5 structure predictions for CeO_2.

Rank	Composition	Space group	Probability	R_{wp}	COD ID
1.	$La_{1.2}U_{0.8}O_4$	$R\bar{3}m$	41.7	16.5%	1006067
2.	$Cu_{3.75}Hg_{1.75}S_8Sn_2$	$I\bar{4}2m$	7.2	17.3%	1527617
3.	CdH_6O_6Pb	$Pn\bar{3}$	3.0	96.2%	1527729
4.	$Ce_{0.5}O_2Zr_{0.5}$	$P42/nmc$	0.9	17.9%	2102840
5.	Bi_2O_3	$Pn\bar{3}m$	0.7	16.1%	1537009

Table S36 | Top-5 structure predictions for W_5O_{14}.

Rank	Composition	Space group	Probability	R_{wp}	COD ID
1.	O_4PbW	$I41/a$	1.2	66.9%	9014025
2.	$Bi_2O_9W_2$	$Pbcn$	1.1	67.2%	7230340
3.	$Fe_3S_4Tl_2$	$Ibam$	1.1	61.6%	1536855
4.	$H_{15}Th_4$	$I\bar{4}3d$	1.1	75.1%	2311027
5.	O_4SnW	$Pnna$	0.8	56.2%	9007595

Table S37 | Top-5 structure predictions for $Bi_2Fe_4O_9$.

Rank	Composition	Space group	Probability	R_{wp}	COD ID
1.	$AlBi_2Ga_3O_9$	$Pbam$	32.9	23.6%	4342599
2.	$Bi_2Ga_4O_9$	$Pbam$	12.2	30.4%	2104768
3.	$Bi_2Ga_4O_9$	$Pbam$	5.9	22.8%	2002314
4.	Bi_2O_4Pd	$I4cm$	2.6	64.9%	2002219
5.	O_8S_2Zr	$Pbam$	2.3	78.2%	2106245

Table S38 | Top-5 structure predictions for NMF component 1.

Rank	Composition	Space group	Probability	R_{wp}	COD ID
1.	$AlBi_2Ga_3O_9$	$Pbam$	16.0	26.6%	4342599
2.	Bi_2O_4Pd	$I4cm$	15.3	64.2%	2002219
3.	$Bi_2Ga_4O_9$	$Pbam$	9.3	34.6%	2104768
4.	$As_{0.4}Fe_{4.56}O_{12}S_{0.84}Sb_{3.84}Zn_{0.2}$	$P42/mbc$	7.2	55.8%	9000738
5.	$Bi_2Fe_4O_9$	$Pbam$	5.2	22.8%	9008148

I: Comparing the PDF of CeO_2 with $La_{1.2}U_{0.8}O_4$

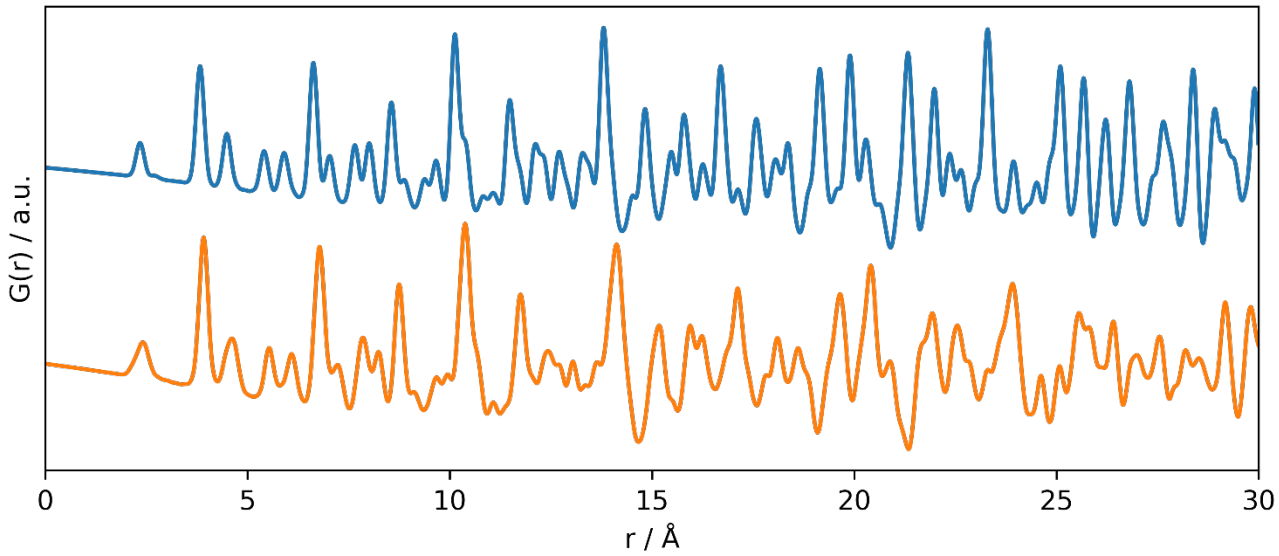

Figure S4 | Simulated PDFs of CeO_2 and $La_{1.2}U_{0.8}O_4$. The PDF of CeO_2 (blue) and $La_{1.2}U_{0.8}O_4$ (orange) show similar structural peaks. However, the similarity decreases as a function of r. Due to the difference in electron density, that ration between the metal – metal peaks and the oxygen – oxygen peaks vary.

J: Real-space Rietveld refinement of Ce constrained structure search for CeO_2

Table S39 | Fitted structure parameters of top-5 structure predictions on $CoFe_2O_4$.

Structure [COD ID]	Composition	Space group	R_{wp} [%]	Scale	*a* [Å]	*c* [Å]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
2102840	$Ce_{0.5}Zr_{0.5}O_2$	$P42/nmc$	17.9	0.66	3.8	5.5	24	3.8	0.007	0.055
2102840	CeO_2	$P42/nmc$	16.1	0.61	3.9	5.4	24	3.8	0.007	0.034

K: Top-3 structure prediction for every frame in the $Bi_2Fe_4O_9$ *in situ* experiment

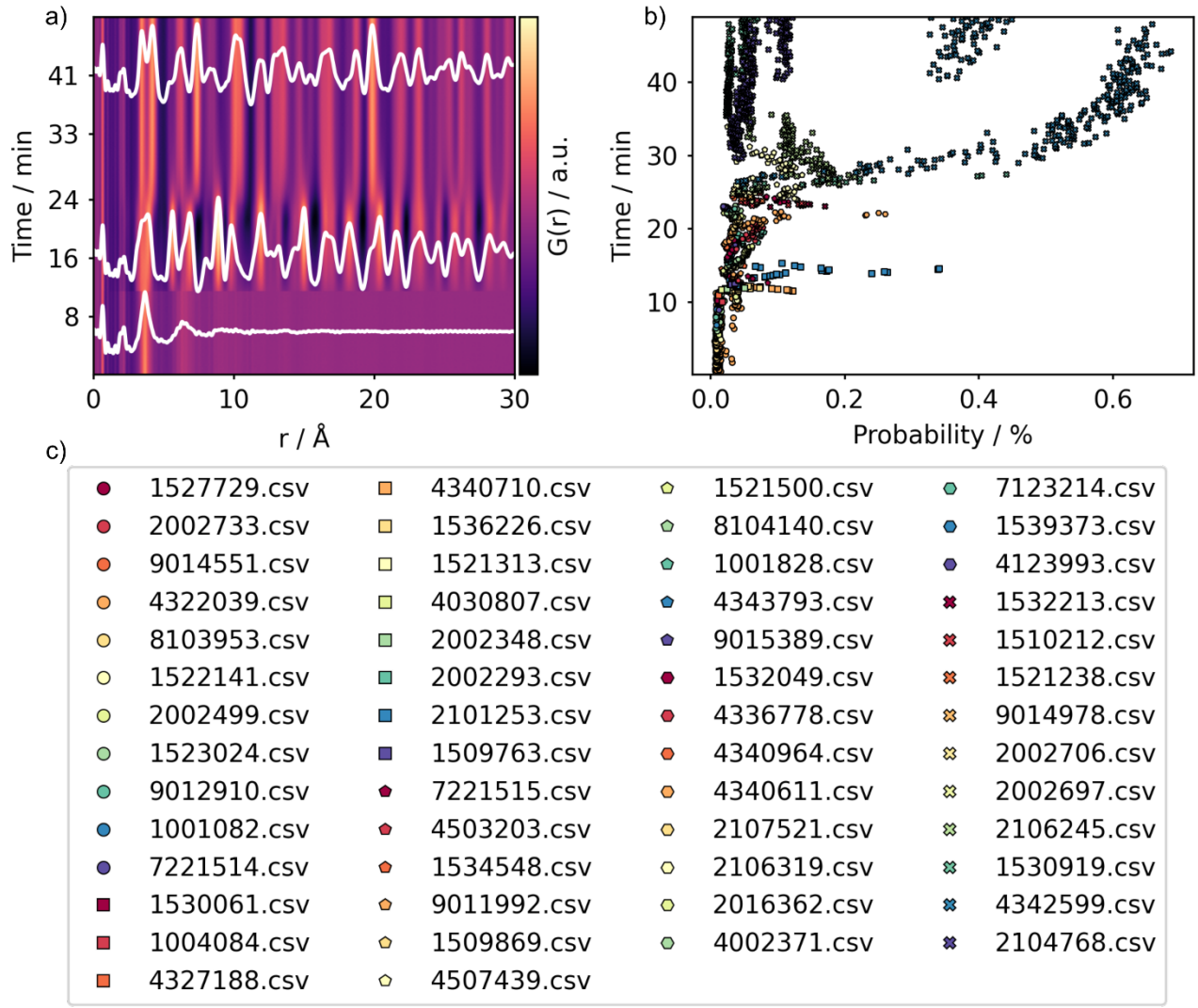

Figure S5 | *In situ* experiment of example 4, $Bi_2Fe_4O_9$, with frame wise top-3 prediction from MLstructureMining**.** a) *in situ* experiment from the formation of $Bi_2Fe_4O_9$. The three distinct phases (precursor, intermediate and product) are highlighted in white. b) The time plotted as function of the probability output for the top-3 predictions. c) all legends for plot b) with the COD identification codes.

L: Structure analysis of the $BiFeO_3$ intermediate and NMF component 2

This section shows the results of the structure analysis of the intermediate phase ($BiFeO_3$) and the reconstructed NMF component obtained from the *in situ* series of $Bi_2Fe_4O_9$.

L.1.: Baseline refinement of $BiFeO_3$ and NMF component 2

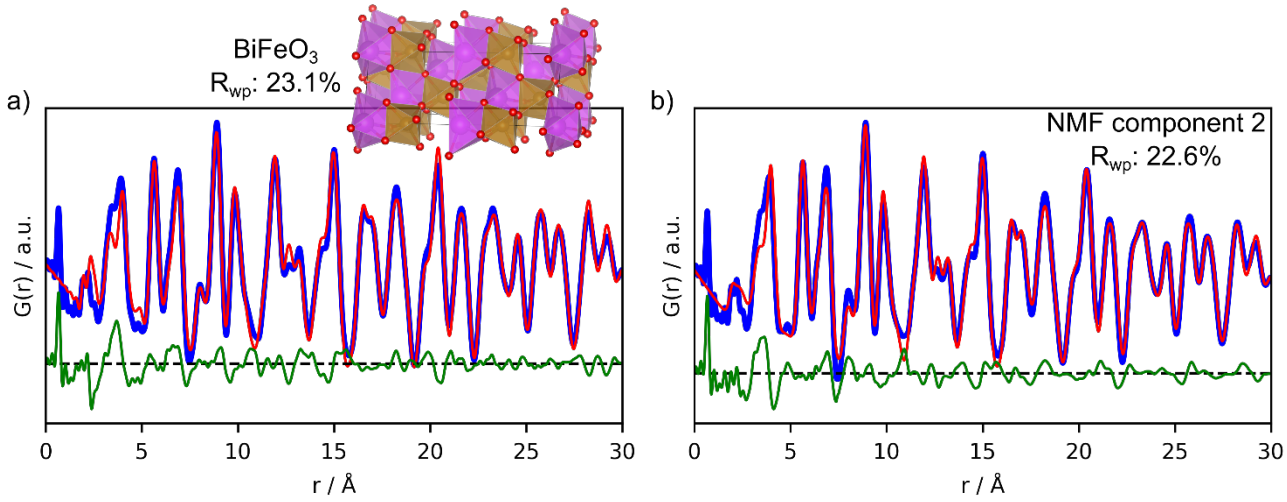

Figure S6 | Plots of the baseline PDF refinements for the BiFeO3 and NMF component 2. a) baseline fit of intermediate phase if the *in situ* series of $Bi_2Fe_4O_9$: $BiFeO_3$ with a R_{wp} of 23.1%, and b) baseline fit NMF components 2: $BiFeO_3$ with a R_{wp} of 22.6%. The structures used for both fits are shown above a).

L.2.: Structure predictions for $BiFeO_3$ and NMF component 2

Table S40 | Top-5 structure predictions for $BiFeO_3$.

Rank	Composition	Space group	Probability	R_{wp}	COD ID
1.	$BiFeO_3$	$R3c$	7.9	18.3%	4336778
2.	$BiMnO_3$	$C12/c1$	3.4	34.5%	4340611
3.	$Ca_{2.667}Nb_{1.333}O_6$	$P121/c1$	3.2	38.8%	1521500
4.	O_3PbTi	$P4mm$	3.2	28.0%	2107521
5.	$O_3PbTi_{0.1}Zr_{0.9}$	$R3c$	1.6	30.0%	2106319

Table S41 | Top-5 structure predictions for NMF component 2.

Rank	Composition	Space group	Probability	R_{wp}	COD ID
1.	$Co_{0.5}NdO_3Pt_{0.5}$	$P121/n1$	8.9	41.3%	9015792
2.	$BiMnO_3$	$C12/c1$	3.9	34.2%	4340611
3.	$BiFeO_3$	$R3c$	3.2	21.1%	4336778
4.	$Ca_{2.667}Nb_{1.333}O_6$	$P121/c1$	2.6	41.5%	1521500
5.	O_3PbTi	$P4mm$	2.3	29.9%	2107521

L.3.: Real-space Rietveld refinements of predicted structures for $BiFeO_3$ and NMF component 2

Table S42 | Fitted structure parameters of top-5 structure predictions on $BiFeO_3$.

Structure [COD ID]	R_{wp} [%]	Scale	*a* [Å]	*b* [Å]	*c* [Å]	*β* [°]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
4336778	18.3	0.15	5.6	-	14.0	-	144	3.5	0.026	0.032
4340611	34.5	0.14	9.8	5.6	9.9	108.9	300	4.0	0.023	0.024
1521500	38.8	0.20	5.6	5.7	9.7	126.1	300	1.0	0.022	0.022
2107521	28.0	0.15	4.0	-	4.0	-	167	3.9	0.31	0.428

2106319	30.0	0.14	5.6	-	13.9	-	236	3.7	0.027	0.337

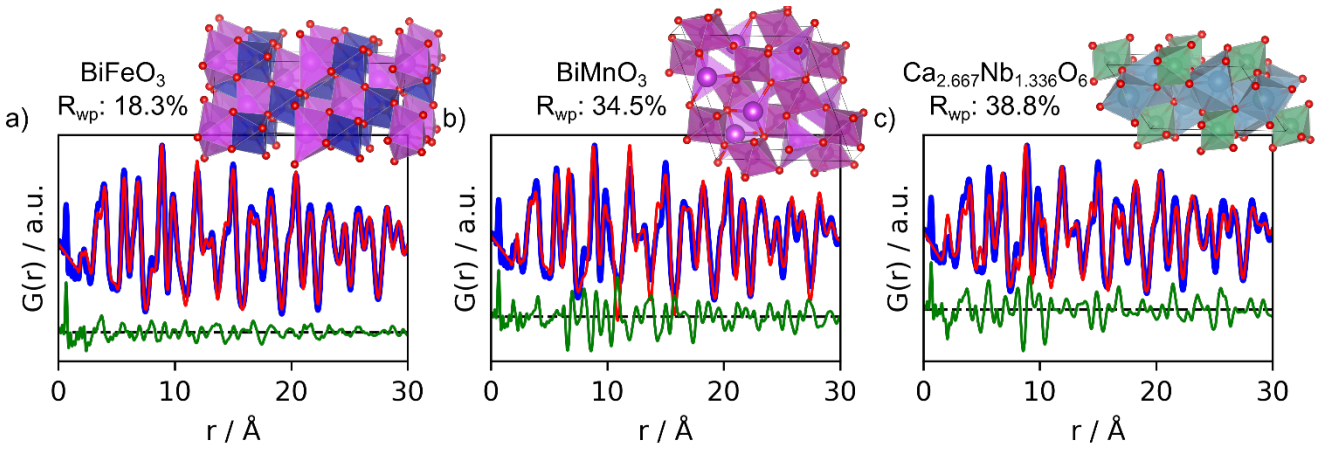

Figure S7 | Top-3 real-space Rietveld refinement of the intermediate, BiFeO3, phase of the *in situ* series of $Bi_2Fe_4O_9$. a) $BiFeO_3$ with an R_{wp} of 18.3%, b) $BiMnO_3$ with an R_{wp} of 34.5% and c) $Ca_{2.667}Nb_{1.336}O_6$ with an R_{wp} of 38.8%, fit parameters can be seen in Table S41. The structure used for each fit is shown above the fit.

Table S43 | Fitted structure parameters of top-5 structure predictions on NMF component 2.

Structure [COD ID]	R_{wp} [%]	Scale	*a* [Å]	*b* [Å]	*c* [Å]	*β* [°]	*γ* [°]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
9015792	41.3	0.10	5.6	5.7	8.0	-	91.7	120	3.1	0.021	0.059
4340611	34.2	0.09	9.8	5.6	9.8	108.9	-	135	4.6	0.025	0.391
4336778	21.1	0.09	5.6	-	14.0	-	-	107	3.5	0.026	0.052
1521500	41.5	0.12	5.6	5.7	9.7	126.1	-	300	1.0	0.021	0.021
2107521	29.9	0.09	4.0	-	4.0	-	-	115	3.9	0.31	0.598

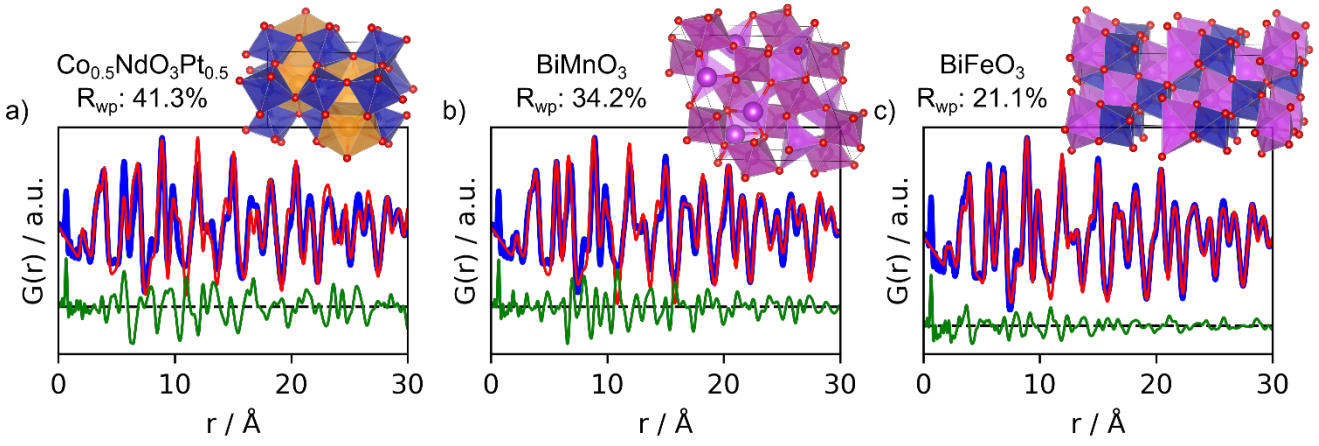

Figure S8 | Top-3 real-space Rietveld refinement of NMF component 2 of the *in situ* series of $Bi_2Fe_4O_9$. a) $Co_{0.5}NdO_3Pt_{0.5}$ with an R_{wp} of 41.3%, b) $BiMnO_3$ with an R_{wp} of 34.2% and c) $BiFeO_3$ with an R_{wp} of 21.1%, fit parameters can be seen in Table S42. The structure used for each fit is shown above the fit.

L.4.: PDFitc benchmark tests for $BiFeO_3$ and NMF component of $BiFeO_3$

For the experimental PDF of $Bi_2Fe_4O_9$ two benchmarks test cases we performed to screen a suitable chemical space.

Table S44 | PDFitc's settings for the three benchmark tests on the experimental PDF of $Bi_2Fe_4O_9$.

Case #	Scattering type	Composition	Optional parameter	Type of PDF	Total # structures	R_{wp} below 50%
1	X-ray	Fe-*-O3	rmin=0 rmax=30 qmin=0.1 qmax=21.5 spd=300	Experimental	202	6
2	X-ray	Bi-*-O3		Experimental	62	6

Table S45 | Case 1: PDFitc's top-5 structure predictions for $BiFeO_3$.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$FeBiO_3$	$P1$	17.7%	4333972
2.	$FeBiO_3$	$R3c$	18.0%	4336776
3.	$FeBiO_3$	$R3c$	18.1%	4336775
4.	$FeBiO_3$	$R3c$	18.1%	7233675
5.	$FeBiO_3$	$R3c$	18.3%	4501315

Table S46 | Case 1: Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of $BiFeO_3$.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*α* [°]	*β* [°]	*γ* [°]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
4333972	5.7	5.7	5.6	60.1	59.5	59.6	154	0.011	0.026	0.032
4336776	5.6	-	13.9	-	-	-	118	0.013	0.027	0.030
4336775	5.6	-	13.9	-	-	-	118	0.027	0.013	0.031
7233675	5.6	-	13.9	-	-	-	115	0.014	0.027	0.031
4501315	5.6	-	13.9	-	-	-	115	0.012	0.027	0.034

Table S47 | Case 2: PDFitc's top-5 structure predictions for $BiFeO_3$.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$FeBiO_3$	$P1$	17.7%	4333972
2.	$FeBiO_3$	$R3c$	18.0%	4336776
3.	$FeBiO_3$	$R3c$	18.1%	4336775
4.	$FeBiO_3$	$R3c$	18.1%	7233675
5.	$FeBiO_3$	$R3c$	18.3%	4501315

Table S48 | Case 2: Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of $BiFeO_3$.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*α* [°]	*β* [°]	*γ* [°]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
4333972	5.7	5.7	5.6	60.1	59.5	59.6	154	0.011	0.026	0.032
4336776	5.6	-	13.9	-	-	-	118	0.013	0.027	0.030

4336775	5.6	-	13.9	-	-	-	118	0.027	0.013	0.031
7233675	5.6	-	13.9	-	-	-	115	0.014	0.027	0.031
4501315	5.6	-	13.9	-	-	-	115	0.012	0.027	0.034

Table S49 | PDFitc's settings for the three benchmark tests on NMF component 2.

Case #	Scattering type	Composition	Optional parameter	Type of PDF	Total # structures	R_{wp} below 50%
1	X-ray	Fe-*-O3	rmin=0 rmax=30 qmin=0.1 qmax=21.5 spd=104	Experimental	202	6
2	X-ray	Bi-*-O3		Experimental	62	6

Table S50 | Case 1: PDFitc's top-5 structure predictions for NMF component 2.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$FeBiO_3$	$P1$	20.1%	4333972
2.	$FeBiO_3$	$R3c$	21.2%	4336776
3.	$FeBiO_3$	$R3c$	21.2%	4336775
4.	$FeBiO_3$	$R3c$	21.2%	7233675
5.	$FeBiO_3$	$R3c$	21.4%	4501315

Table S51 | Case 1: Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of NMF component 2.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*α* [°]	*β* [°]	*γ* [°]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
4333972	5.7	5.7	5.6	60.0	60.0	59.8	105	0.012	0.026	0.040
4336776	5.6	-	13.9	-	-	-	93	0.013	0.026	0.043
4336775	5.6	-	13.9	-	-	-	104	0.014	0.026	0.044
7233675	5.6	-	13.9	-	-	-	94	0.015	0.026	0.045
4501315	5.6	-	13.9	-	-	-	104	0.013	0.026	0.054

Table S52 | Case 2: PDFitc's top-5 structure predictions for NMF component 2.

Rank	Composition	Space group	R_{wp}	COD ID
1.	$FeBiO_3$	$P1$	20.1%	4333972
2.	$FeBiO_3$	$R3c$	21.2%	4336776
3.	$FeBiO_3$	$R3c$	21.2%	4336775
4.	$FeBiO_3$	$R3c$	21.2%	7233675
5.	$FeBiO_3$	$R3c$	21.4%	4501315

Table S53 | Case 2: Fitted parameters of top-5 structures obtained from the benchmark test for the experimental PDF of NMF component 2.

Structure [COD ID]	*a* [Å]	*b* [Å]	*c* [Å]	*α* [°]	*β* [°]	*γ* [°]	p_{size} [Å]	$U_{iso\ Fe}$ [Å^2]	$U_{iso\ Bi}$ [Å^2]	$U_{iso\ O}$ [Å^2]
4333972	5.7	5.7	5.6	60.0	60.0	59.8	105	0.012	0.026	0.040
4336776	5.6	-	13.9	-	-	-	93	0.013	0.026	0.043

4336775	5.6	-	13.9	-	-	-	104	0.014	0.026	0.044
7233675	5.6	-	13.9	-	-	-	94	0.015	0.026	0.045
4501315	5.6	-	13.9	-	-	-	104	0.013	0.026	0.054

M: Principal component analysis and non-negative matrix factorization on *in situ* $Bi_2Fe_4O_9$ with precursor

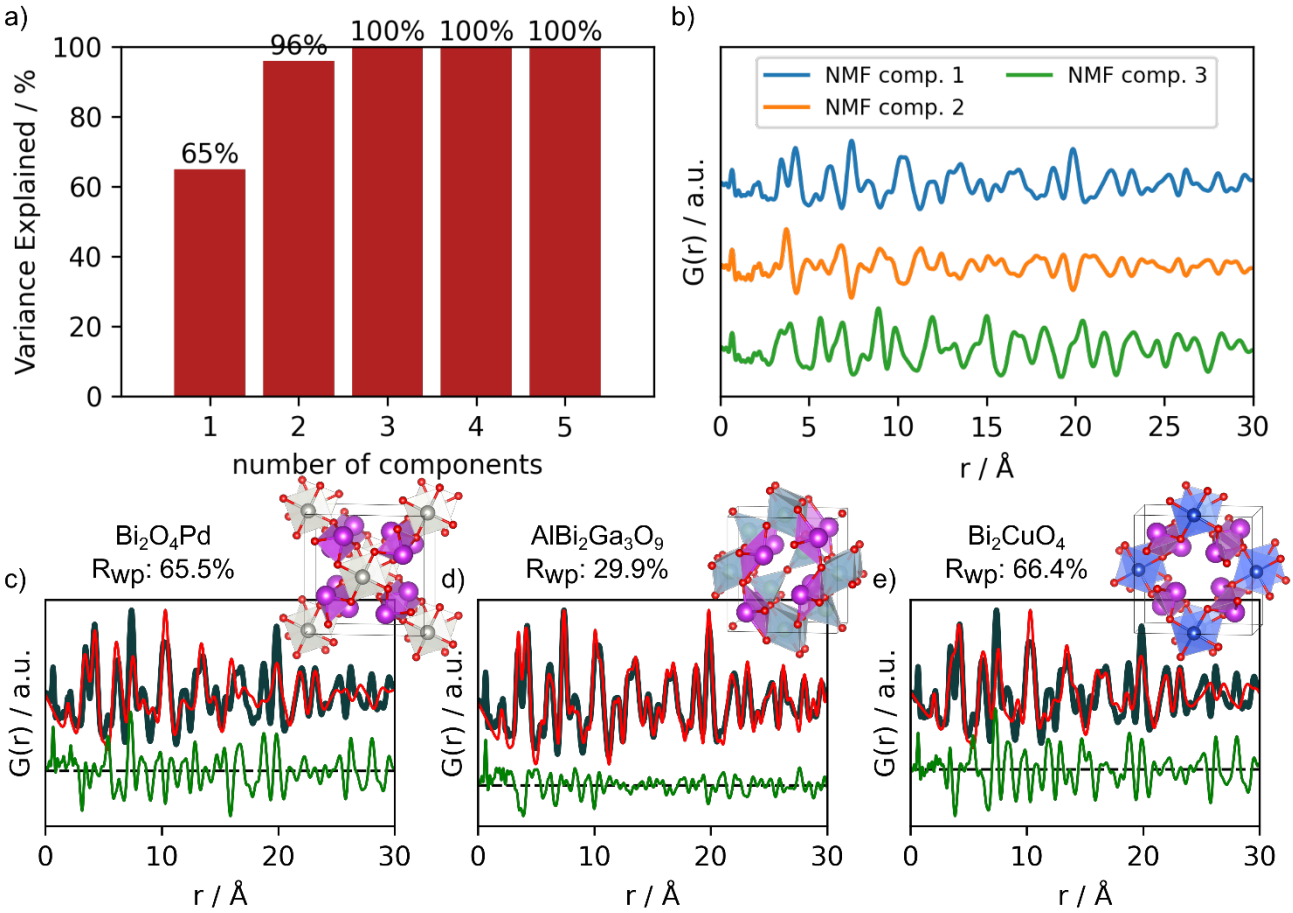

Figure S9 | Three component PCA and NMF structure characterization the *in situ* series of $Bi_2Fe_4O_9$. a) The cumulative variance explained by the PCA components and b) show the reconstructed NMF components. The real-space Rietveld refinements of top-3 predictions are shown in c), d) and e), fit parameters can be seen in Table S54. The fits are performed on NMF component 1.

Table S54 | Top-5 structure predictions for NMF component 1 with precursor.

Rank	Composition	Space group	Probability	R_{wp}	COD ID
1.	Bi_2O_4Pd	$I4cm$	23.7%	65.5%	2002219
2.	$AlBi_2Ga_3O_9$	$Pbam$	11.7%	29.9%	4342599
3.	Bi_2CuO_4	$P4/ncc$	6.9%	66.4%	1004051
4.	$As_{0.4}Fe_{4.56}O_{12}S_{0.84}Sb_{3.84}Zn_{0.2}$	$P42/mbc$	4.6%	57.6%	9000738
5.	$Bi_2Fe_4O_9$	$Pbam$	3.6%	24.2%	1530918

Table S55 | Fitted structure parameters of top-5 structure predictions on NMF component 1 with precursor.

Structure [COD ID]	R_{wp} [%]	Scale	*a* [Å]	*b* [Å]	*c* [Å]	p_{size} [Å]	δ_2 [Å^2]	$U_{iso\ M}$ [Å^2]	$U_{iso\ O}$ [Å^2]
2002219	65.5	0.18	8.3	-	6.0	58	3.6	0.040	0.512
4342599	29.9	0.15	8.0	8.5	6.0	288	3.7	0.019	0.164
1004051	66.4	0.15	8.3	-	6.1	66	4.2	0.036	4.882
9000738	57.6	0.20	8.3	-	18.2	71	1.0	0.024	0.092
1530918	24.2	0.16	8.0	8.5	6.0	243	2.6	0.020	0.062

References

1 Bouhlel, M. A., Hwang, J., Bartoli, N., Lafage, R., Morlier, J. & Martins, J. A Python surrogate modeling framework with derivatives. *Adv. Eng. Softw.* (2019). https://doi.org:10.1016/j.advengsoft.2019.03.005

2 Bishop, C. M. & Nasrabadi, N. M. *Pattern recognition and machine learning*. Vol. 4 (Springer, 2006).

3 Juhás, P., Farrow, Christopher L., Yang, X., Knox, Kevin R. & Billinge, Simon J. L. Complex modeling: a strategy and software program for combining multiple information sources to solve ill posed structure and nanostructure inverse problems. *Acta Cryst. A* **71**, 562-568 (2015). https://doi.org:10.1107/S2053273315014473

4 Nogueira, F. Bayesian Optimization: Open source constrained global optimization tool for Python. *GitHub*, https://github.com/fmfn/BayesianOptimization (2014).

5 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.

6 Nicolae, M.-I. *et al.* Adversarial Robustness Toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069* (2018).

7 Myers, J., Well, A. & Lorch Jr, R. *Research design and statistical analysis Routledge*. Vol. 3 (Routledge, 2010).

8 Sapnik, A. F. *et al.* Multivariate analysis of disorder in metal–organic frameworks. *Nat. Commun.* **13**, 2173 (2022). https://doi.org:10.1038/s41467-022-29849-6

9 Mukaddem, K. T., Chater, P. A., Devereux, L. R., Al Bahri, O. K., Jain, A. & Cole, J. M. Dye-Anchoring Modes at the Dye···TiO2 Interface of N3- and N749-Sensitized Solar Cells Revealed by Glancing-Angle Pair Distribution Function Analysis. *J. Phys. Chem. C* **124**, 11935-11945 (2020). https://doi.org:10.1021/acs.jpcc.0c02314

10 Kjær, E. T. S. *et al.* In Situ Studies of the Formation of Tungsten and Niobium Oxide Nanoparticles: Towards Automated Analysis of Reaction Pathways from PDF Analysis using the Pearson Correlation Coefficient. *Chem. Methods* **2**, e202200034 (2022). https://doi.org:10.1002/cmtd.202200034

11 Juhas, P., Davis, T., Farrow, C. L. & Billinge, S. J. L. PDFgetX3: a rapid and highly automatable program for processing powder diffraction data into total scattering pair distribution functions. *J. Appl. Cryst.* **46**, 560-566 (2013). https://doi.org:10.1107/S0021889813005190

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

13 Becker, J. *et al.* Experimental setup for in situ X-ray SAXS/WAXS/PDF studies of the formation and growth of nanoparticles in near-and supercritical fluids. *J. Appl. Cryst.* **43**, 729-736 (2010). https://doi.org:10.1107/S0021889810014688

14 Chupas, P. J., Qiu, X., Hanson, J. C., Lee, P. L., Grey, C. P. & Billinge, S. J. Rapid-acquisition pair distribution function (RA-PDF) analysis. *J. Appl. Cryst.* **36**, 1342-1347 (2003). https://doi.org:10.1107/S0021889803017564

15 Hammersley, A. FIT2D: an introduction and overview. *European synchrotron radiation facility internal report ESRF97HA02T* **68**, 58 (1997).

16 Ashiotis, G. *et al.* The fast azimuthal integration Python library: pyFAI. *J. Appl. Cryst.* **48**, 510-519 (2015). https://doi.org:10.1107/S1600576715004306

17 Qiu, X., Thompson, J. W. & Billinge, S. J. L. PDFgetX2: a GUI-driven program to obtain the pair distribution function from X-ray powder diffraction data. *J. Appl. Cryst.* **37**, 678-678 (2004). https://doi.org:10.1107/S0021889804011744

18 Dykhne, T., Taylor, R., Florence, A. & Billinge, S. J. L. Data Requirements for the Reliable Use of Atomic Pair Distribution Functions in Amorphous Pharmaceutical Fingerprinting. *Pharm. Res.* **28**, 1041-1048 (2011). https://doi.org:10.1007/s11095-010-0350-0