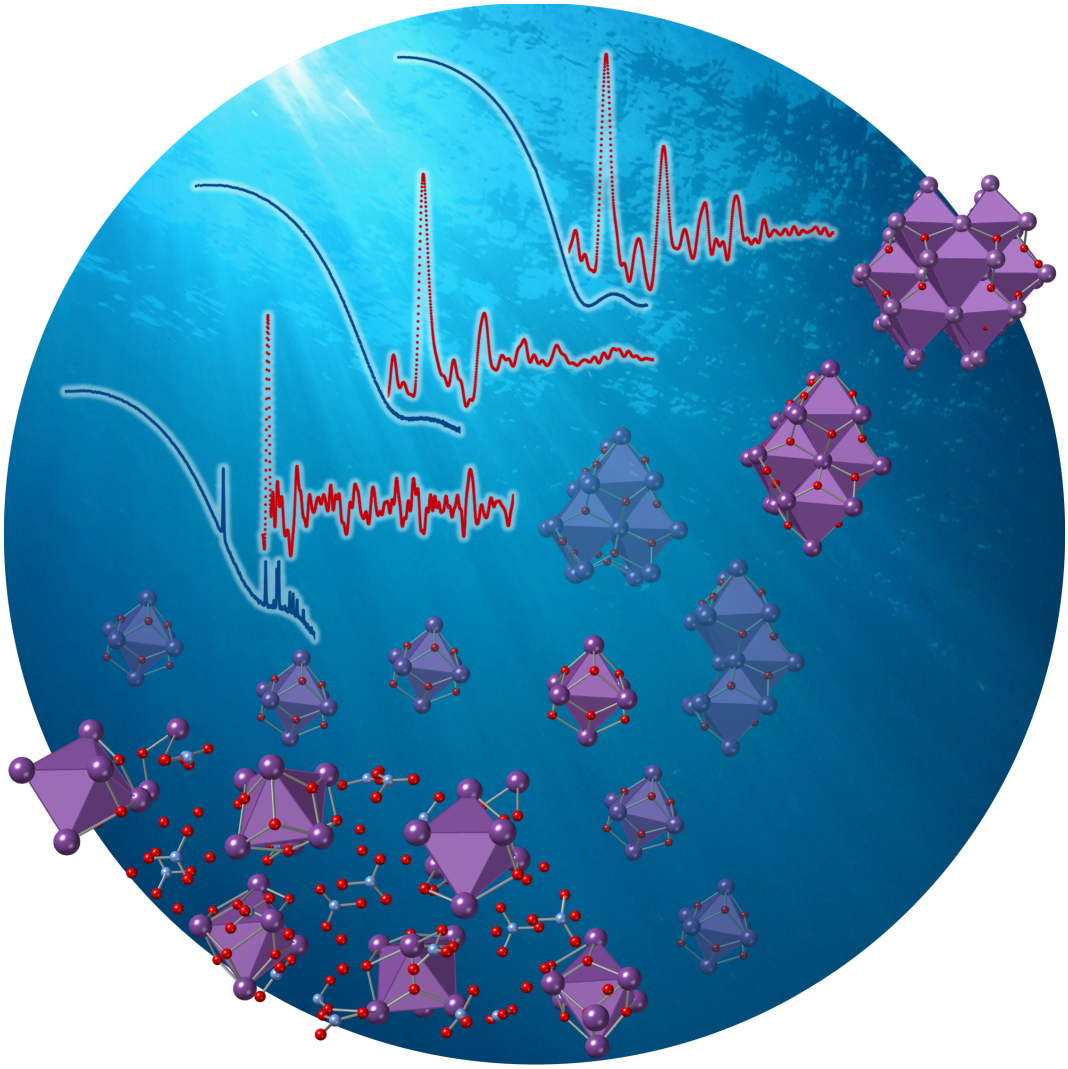
UNIVERSITY OF COPENHAGEN
NANO-SCIENCE CENTER & DEPARTMENT OF CHEMISTRY

# Pushing the boundaries of Nanocluster Modelling from Pair Distribution Function and Small-Angle X-ray Scattering Analysis

Master Thesis submitted by Andy Sode Anker



Supervisor: Kirsten Marie Ørnsbjerg Jensen

Submitted on: 3rd of August 2021

# Table of contents

# Abstract

From the Stone Age to the Iron Age, the materials we use have always defined our civilization. Today, we call our civilization the Silicon Age due to our large consumption of silicon in electrical components. Materials Science is therefore an incredibly important research field with a major role in the development of our civilization.

For millenaries, the way we have designed new materials has been through the *trial-and-error* approach. However, with the high demands of modern society, we need to develop new and functional materials faster than ever. Therefore, we must understand the properties governed by materials and how to synthesize materials with specific properties. We call this *synthesis by design*.

For the last century, we have been able to characterize the structure of crystalline materials using crystallography but to do *synthesis by design*, we must also understand how materials form. Often, the formation of materials goes through a chemical phase where nanoclusters are present in solution. The structure of nanoclusters in solution cannot be characterized with conventional crystallographic methods and hence, this part of the material formation is poorly understood. In the last decades, it has been shown that total scattering with Pair Distribution Function (PDF) can be used to describe the atomic arrangements of nanoclusters in solution and Small-Angle X-ray Scattering (SAXS) can be used to describe the shape and size distribution of the nanoclusters. However, modelling of PDF and SAXS have been restricted to expert users and rarely a combination of PDF and SAXS has been used. In this Master Thesis, we push the boundaries of PDF and SAXS analysis. It is demonstrated how we can combine the information from PDF and SAXS to get a better understanding of nanoclusters in solution. Specifically, we focus on the formation of the $[Bi_{38}O_{45}]$ cluster by dissolving crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ in dimethylsulfoxid. Furthermore, we show that automated modelling of PDF and SAXS data can provide new insight compared to conventional modelling techniques, and potentially automated modelling can make PDF and SAXS analysis accessible for non-expert users.

# 1 Introduction

Atomically precise nanoclusters have in recent years been massively investigated since the ability to synthetically control the nanocluster sizes[1, 2, 3, 4-7] brings along the ability to reach a range of applications with tuned properties.[8-12, 13, 14, 15, 16, 4, 16, 3] Thereby atomically precise nanoclusters serve as an excellent model system for growth mechanisms and many of the nanoclusters can undergo nucleation to form materials.[3, 17] Developing new materials have in the past millenaries been done using the *trial-and-error* approach. However, if we want to follow the high demand of superior materials by modern society, we must be able to construct materials with specific properties by design. Developing new and functional materials by design is dependent on a thorough understanding of the growth mechanism or nucleation reaction. Despite that, the challenge of understanding growth mechanisms and nucleation reactions at an atomic scale is still an open problem in inorganic- and materials chemistry.[18] Total Scattering (TS) with Pair Distribution Function (PDF) analysis is now widely used as a method for characterizing materials without long-range order, such as nanomaterials,[19, 20] disordered materials[21, 22] or amorphous materials,[23] where conventional crystallographic approaches fail.[24-26] For small, non-periodic objects, e.g., nanoclusters, structure modelling in either reciprocal (Q) or real (r) space can be done by employing the Debye-equation for calculation of scattering intensities, where the position of each atom in the structure is taken into account:

**Equation 1.1**

$$I(Q) = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i(Q) \cdot f_j(Q) \cdot \frac{\sin(Q \cdot r_{ij})}{Q \cdot r_{ij}}$$

Here, $r_{ij}$ is the distance between atom *i* and *j*, and *f* is the atomic scattering factor. The Debye Equation can be applied for both small-angle and wide-angle scattering data (SAXS/WAXS), and a Debye analysis can be done either in reciprocal-space, or real-space by PDF modelling after Fourier transformation of calculated scattering intensities.[27, 28] While the use of the Debye scattering equation is computationally expensive, it can be used to calculate the scattering pattern of any structure. In recent years, the Debye scattering equation has been applied in PDF analysis of e.g. local structural motifs in amorphous materials,[22, 29] metallic nanoclusters,[28, 30, 31] understanding nucleation processes from ionic metal oxido nanoclusters in solution[32] and growth mechanisms of ionic metal oxido nanoclusters in solution.[33, 34] In these studies, a scattering pattern or PDF from a model structure is calculated using the Debye equation, and variables such as atomic positions and atomic displacement parameters (ADPs) may be refined, until a good fit with the data is obtained. We will demonstrate

how this approach can be used to follow the cluster growth from $[Bi_6O_8]$ to $[Bi_{38}O_{45}]$ using *in situ* PDF and SAXS analysis of data measured on crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ dissolved in dimethylsulfoxid. This is a cumbersome process that often is restricted to experts in the field.

Fortunately, automated methods such as structure-mining and cluster-mining have recently appeared to overcome this challenge.[30, 35, 36] In a recent study of amorphous molybdenum oxide materials, Christiansen et al. introduced a new approach, where they automatically generated a large number of $MoO_x$ clusters to fit experimental data to identify dominating structural motifs in the sample.[22] I will refer to this approach as the *automated motif extractor*. They hypothesized that the structural motifs present in amorphous molybdenum oxides can also be found in well-known crystalline structures, and they therefore used crystal structures of molybdenum oxides as starting models, from which they could cut out thousands of different cluster structure models of different sizes, which were all tested against their data. However, both the structure-mining, cluster-mining and the automated motif extractor approach require extensive computational power since they are based on fits of the data, using minimization methods as least-squares algorithm, to a large number of structural models. Ideally, we want automated structure characterization in seconds such that the results can be used directly to effectively adjust the experiments while performing them. Lately, Machine Learning (ML) has both been used to model PDF data by predicting the space group of a structure from the PDF[37] and to predict the atomic coordinates of a mono-metallic cluster.[36] ML is significantly faster than minimization techniques and often the prediction is instant.

Here, we introduce two new approaches of automated modelling, which are both based on ML. The first approach uses simulated data of polyoxometalate (POM) clusters to train a ML model to classify which POM structure a PDF is from. The other approach is a further development of the automated motif extractor. We use machine learning (ML) to evaluate the results from automated motif extractor to identify important structural motifs. **M**achine **L**earning based **Mot**if **Ex**tractor (ML-MotEx) extracts hundreds or thousands of structural models from chosen starting structure, and then fit these individual models to a dataset. The structures and the goodness-of-fit parameter, $R_{wp}$ value, from each fit are handed to an ML algorithm which learns how to predict $R_{wp}$ values based on a structure (A description of the $R_{wp}$ value is given in Appendix III). The ML algorithm is subsequently analysed with the use of SHAP (SHapley Additive exPlanations) values,[38] which provide a measure of how important each structural feature is to the prediction of the goodness-of-fit value. The ML-MotEx algorithm outputs quantified values of how important each atom in a model is for the structure to yield a low $R_{wp}$ value in the given fitting algorithm. Both ML procedures presented in this Master Thesis are done completely automated, in semi-real experimental time and without human bias. It is thereby a step towards high-throughput real-time automated analysis of PDF analysis.

# 2 Using Scattering to Characterise the Nanocluster Structure: A 'Complex' Problem

## 2.1 Introduction

As stated in Chapter 1, it is important to resolve the structure of nanoclusters at an atomic level to understand cluster growth and nucleation processes with the overall goal of doing synthesis of new and functional materials by design. The Debye equation has recently been used with PDF and SAXS analysis in both reciprocal-space and real-space to elucidate the structure of nanoclusters in solution.[28, 30, 32] This Chapter introduces the scattering theory behind the Debye Equation. The strengths and weaknesses are discussed of presenting Debye scattering data in either reciprocal- or real-space. It is illustrated that some structures are too complex to be characterized by a single technique alone but by applying data from multiple techniques, the structure can be characterized.

## 2.2 Scattering from an Electron to a Material

### Scattering from an Electron

Scattering is a well-known phenomenon that is happening all around us. For example, light scattering is accountable for the color of the blue sky due to the scattering of sunlight from the water droplets and it is accountable for the white color of milk due to scattering of light from colloidals in suspension. For materials chemists, scattering has been an exceptionally important tool to characterise the structure of materials. However, materials chemists do not only use light scattering to characterize the structure of materials. Material chemists utilize a range of scattering phenomena as light, electron, neutron and X-ray scattering. In this section, we will focus on X-ray scattering, however, the principles of X-ray scattering are very similar to those of electron and neutron scattering.

From X-ray scattering with in-house equipment to synchrotrons, all build on the interaction between X-ray and the electron cloud. We will start by considering the scattering of X-rays from a single electron as illustrated in Figure 2.1. The oscillating electromagnetic wave, X-ray, induces an oscillation of the electron, hence an acceleration. From fundamental electromagnetic theory, it is known that accelerated charged particles emit electromagnetic waves. Therefore, the electron will emit X-rays. The outgoing X-ray wave of the electron will have a spherical propagation with a phase shift of $\pi$ to the incoming X-ray wave.[39-41]

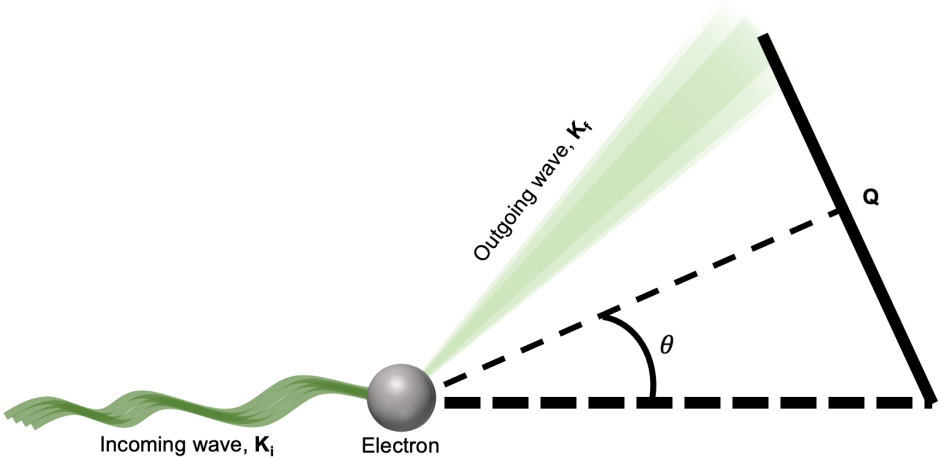



**Figure 2.1: The incoming wave, $k_i$, is scattered from the electron as outgoing wave $k_f$. The change in wavevectors, $k_f - k_i$, is referred to as the scattering vector, Q.**

As seen in Figure 2.1 the scattering process can be described by the scattering vector, **Q**, which is the difference of the incident X-ray beam $\mathbf{k_i}$ and the scattered X-ray beam, $\mathbf{k_f}$. During a scattering event, two types of scattering occur: elastic and inelastic scattering. The illustrated process, where the size of $\mathbf{k_i}$ and $\mathbf{k_f}$ is the same, is only valid for elastic coherent scattering i.e., where the energy of the X-ray beam is unchanged, and the phase shift is constant during the scattering event (π for free electrons). This process is referred to as Thomson scattering. Compton scattering is another common process that happens during a scattering event. This is incoherent and inelastic.[39] In this work, we will focus on the Thomson scattering, which contains the structural information of the materials, while the Compton scattering can be seen as a diffuse background in the data. For Thomson scattering, the magnitude of the wavevectors, $\mathbf{k_i}$ and $\mathbf{k_f}$, is given by $\frac{2\pi}{\lambda}$. From geometrical considerations (Figure 2.1), the scattering vector, **Q**, can be derived.

**Equation 2.1**

$$Q = 2ksin(\theta) = \frac{4\pi}{\lambda}\sin(\theta), k = \frac{2\pi}{\lambda}$$

### Scattering from an Atom

So far, we have only considered the scattering event from a single electron, which is considered as scattering from a point. However, most atoms consist of numerous electrons. When we consider scattering from multiple electrons in the same atom, we must include the effect of constructive or destructive interference. The interference effects will be dependent on the morphology of the electron

cloud and the wavelength. Consequently, the scattering amplitude will be **Q**-dependent and is defined as the atomic form factor, f(**Q**). Effectively the atomic form factor is a Fourier transform of all electron positions, **r**, in the electron cloud, $\rho$.[42, 43]

**Equation 2.2**

$$f(\boldsymbol{Q}) = \int \rho(\boldsymbol{r}) \cdot e^{i\boldsymbol{Q}\cdot\boldsymbol{r}} dr$$

From this equation, we can derive that when the scattering vector goes towards 0 $\text{Å}^{-1}$, $\boldsymbol{Q} \to 0\ \text{Å}^{-1}$, then the atomic form factor goes toward the number of electrons in that atom, $f(\boldsymbol{Q} = 0\ \text{Å}^{-1}) \to Z$. For X-rays, the atomic form factor rapidly decreases with increasing scattering vector, **Q**, due to the spatial extension of the electron cloud. This is also the case for electron scattering since electrons also scatter from the electron cloud. However, this is not the case for neutrons that scatter from the nuclei since the nuclei can practically be seen as a point in space without extension. Consequently, the neutron form factor is a Fourier transform of a point scatterer (the nuclei) which results in a constant.

### Scattering from a Material

We will now extend the theory of scattering from a single atom to real materials which consist of multiple atoms. In order to do so, we sum the scattering contribution of each atom, j, in the material with position, **r**. The scattering amplitude, $\psi$, is thereby the sum of the Fourier transform of all atomic form factors in the material.

**Equation 2.3**

$$\psi(\boldsymbol{Q}) = \sum_{j=1}^{N} f_j(\boldsymbol{Q}) \cdot e^{i(\boldsymbol{Q}\cdot\boldsymbol{r}_j)}$$

By measuring the scattering amplitude, it is possible to extract the atomic position of each atom in the material. In practice, it is not the amplitude that is measured during a scattering experiment but the scattering intensity, I. However, the scattering intensity is proportional to the magnitude of the squared scattering amplitude.[43]

**Equation 2.4**

$$I(\boldsymbol{Q}) \propto |\psi(\boldsymbol{Q})|^2$$

In 1915, Debye derived that for materials where the scattering electrons are isotopically distributed in the material, as electrons normally are in a material, the intensity distribution can be written as:[27, 44]

**Equation 2.5**

$$I(\boldsymbol{Q}) = \sum_{i,j}^{N} f_i(Q) \cdot f_j(Q) \cdot \frac{\sin(Q \cdot \boldsymbol{r}_{ij})}{Q \cdot \boldsymbol{r}_{ij}}$$

By examining the intensity of the scattered X-rays from a material dependent on **Q**, we can thereby retrieve information about the structure of the given material.

## 2.3 The two-headed scattering data representation: reciprocal- and real-space

Traditionally, the mathematical expressions of scattering intensity of a material are expressed in Q-space as seen in the sections above. A simplified explanation for this is that the interaction between the material and the scattered wave happens in the *scattering vector space* i.e. Q-space. This is why data from a scattering measurement is measured in Q-space (or $2\theta$). Conventional diffraction experiments as Single Crystal Diffraction (SCD) or Powder Diffraction (PD) are both measured and modelled in Q-space. However, it is also possible to present scattering data in r-space using a Fourier transformation. Fourier transformations are usually used to convert the variable of a dataset between time and frequency or as seen in the previous section it can be used to convert the electron cloud distribution into an atomic form factor. A Fourier transform is thereby a mathematical tool to 'change the space' of a dataset. Figure 2.2 illustrates simulated scattering data of a [$Bi_{38}O_{45}$] cluster in Q-space and the ideal Fourier transformation of the data from $Q = 0$ Å$^{-1}$ to $Q = \infty$ Å$^{-1}$. Simulation details are given in Appendix I.

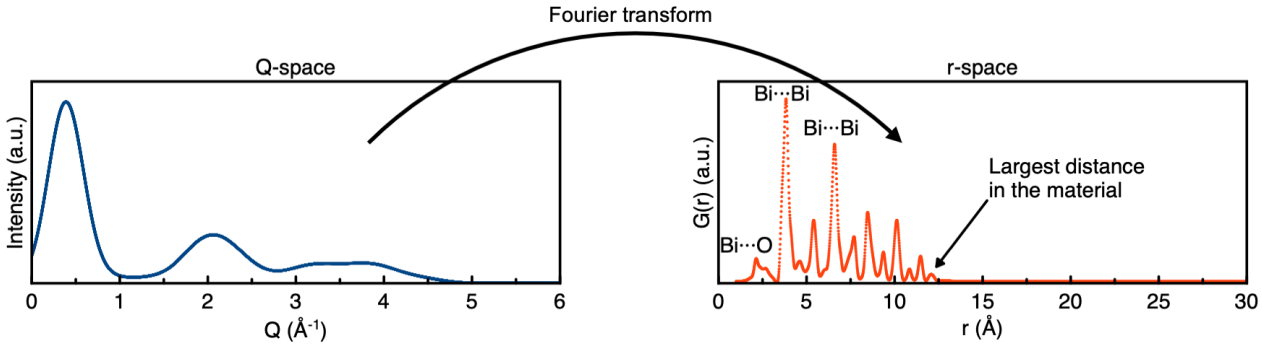



**Figure 2.2: Simulated TS data of the [$Bi_{38}O_{45}$] cluster in Q-space is Fourier transformed to real-space data, the PDF.**

While the data in the two figures contain the same information, the representation of the information has been changed. Diffraction peaks in the scattering pattern in Q-space, also referred to as Bragg peaks, originate from a periodic order of the atoms in the material. Since the scattering pattern in Figure 2.2 does not contain Bragg peaks but only broad features, we can conclude that there is no

periodic order of the atoms in the [$Bi_{38}O_{45}$] cluster. The Q-space and r-space are inverse, which means that intensity in the low Q-regime corresponds to larger distances in real space. The large scattering intensity in the small-angle regime thereby tells us that the material consists of large units. This could for example be from the scattering of the nanoclusters.

In r-space it is easier to extract information about the local order of the material. For example, one can easily identify peaks at r = 3.8 Å, 6.6 Å, 8.7 Å and 10.2 Å in Figure 2.2, which corresponds to frequent distances in the material. Based on the intensity, G(r), of these peaks, one can also index the scattering atom because bismuth scatters significantly more than oxygen. The scattering peaks with high intensity is from frequent Bi···Bi distances in the structure, while the smaller peaks are from Bi···O distances in the material. Furthermore, one can determine the size of the material by simply identifying the peak with the highest r-value, as seen about r = 12.2 Å in Figure 2.2. Q-space and r-space are thereby complementary data representations that highlight different information about the structure of a material.

## 2.4 Using Total Scattering with Pair Distribution Function Analysis to reveal the information underneath the Bragg peaks

As demonstrated in the previous section, it is a powerful tool to analyse both the scattering data in Q-space and r-space since they highlight complementary information. However, in reality, obtaining useful r-space data from a scattering pattern in Q-space is a challenging process. Figure 2.3 shows an example of Q-space scattering data of the [$Bi_{38}O_{45}$] cluster and the Fourier transformed r-space data. Details about the experiment and how it has been transformed to r-space are given in Appendix I.

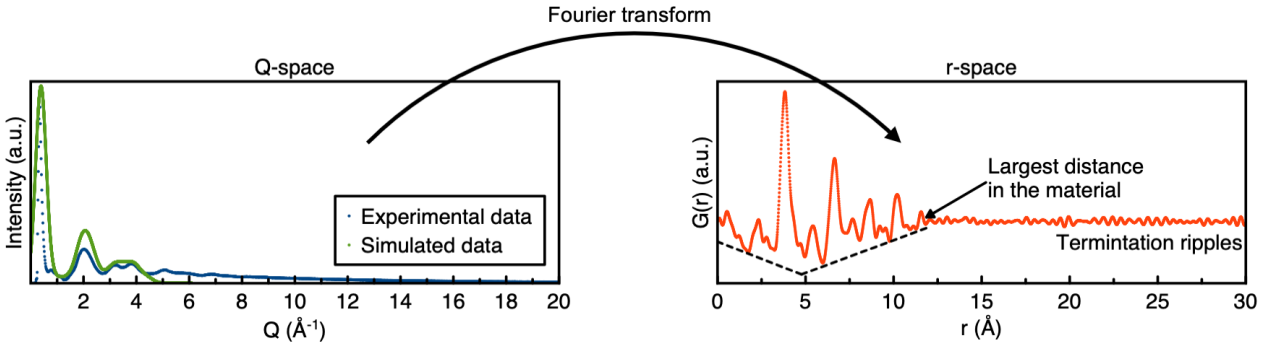



**Figure 2.3: Experimental TS data of the [$Bi_{38}O_{45}$] cluster in Q-space is Fourier transformed to real-space data, the PDF.**

The experimental Q-space data of the [$Bi_{38}O_{45}$] cluster is very similar to the simulated data, but it has a background signal originating from incoherent scattering. However, the structural information is in the coherent scattering. We introduce the *Total Scattering Structure Function*, S(**Q**), which only uses the coherent scattering:[42]

**Equation 2.6**

$$S(\boldsymbol{Q}) = \frac{I_{coh}(\boldsymbol{Q}) + \langle f(\boldsymbol{Q})\rangle^2 - \langle f(\boldsymbol{Q})^2\rangle}{N\langle f(\boldsymbol{Q})\rangle^2}$$

S(**Q**) normalizes the coherent elastic scattering intensity, $I_{coh}$, with the average scattering power, $\langle f(\boldsymbol{Q})\rangle$. For X-ray scattering data this results in an amplified signal in high Q which contains the local structural information. The scattering intensity in the high Q regime is low due to the Q-dependency of the atomic scattering factors. The $\langle f(\boldsymbol{Q})\rangle^2 - \langle f(\boldsymbol{Q})^2\rangle$ term arises from an imperfect cancellation in destructive and positive interference between scattering by different elements.[42]

The next mathematical treatment of the data is by calculating the *Reduced Total Scattering Function*, F(Q), where the high Q regime is further enhanced:[42]

**Equation 2.7**

$$F(\boldsymbol{Q}) = Q(S(\boldsymbol{Q}) - 1)$$

F(**Q**) is the function we Fourier transform in order to obtain the *Reduced Atomic Pair Distribution Function*, G(**r**), or also referred to as the PDF:[42]

**Equation 2.8**

$$G(\boldsymbol{r}) = \frac{2}{\pi}\int_0^{\infty} F(\boldsymbol{Q})\sin(\boldsymbol{Q}\cdot\boldsymbol{r})\,dQ$$

However, here we encounter another issue as a real experiment does not contain data from $Q = 0$ Å$^{-1}$ to $Q = \infty$ Å$^{-1}$. In reality, we therefore use a PDF including artefacts from a limited $Q_{range}$ of our experiments:[42]

**Equation 2.9**

$$G(\boldsymbol{r}) = \frac{2}{\pi}\int_{Q_{min}}^{Q_{max}} F(\boldsymbol{Q})\sin(\boldsymbol{Q}\cdot\boldsymbol{r})\,d\boldsymbol{Q}$$

The artefacts seen from the limit of $Q_{min}$ are illustrated with a dotted line (baseline) in Figure 2.3. The data in the low Q-regime contains information about the size and shape of the particles, which therefore affects the baseline of the PDF. The limiting $Q_{range}$ ($Q_{min} > 0$ Å$^{-1}$ and $Q_{max} < \infty$ Å$^{-1}$) results in 'termination ripples' also illustrated in Figure 2.3. In order to decrease the effect of these artefacts, we perform TS experiments where the measured $Q_{range}$ is as large as possible. In order to achieve a large $Q_{range}$, we often use the rapid-acquisition PDF (RA–PDF)[45] setup where a 2D detector is moved as close to the sample as possible. Using RA–PDF together with a low wavelength, Equation 2.1, makes it possible to obtain high Q-values often up to 20 – 30 Å$^{-1}$. This could in principle also have been done with the use of a point detector, however, an advantage of using a 2D detector is the achievement of higher intensity which makes it possible to do measurements with second resolution.

We have now shown that it is a challenging process to obtain useful data in r-space. However, with TS experiments, we can overcome this challenge with minor artefacts in the data. The data in r-space is a great tool to highlight complementary features to the Q-space data, especially the structural information of the local order.

## 2.5 When the Problem is too 'Complex' to a Single Scattering Technique

While TS data is an extremely powerful technique to characterize the structure of most materials, the information in a TS experiment is not always sufficient to characterize the structure. Sometimes, we simply need data from multiple experiments to determine the atomic structure. Figure 2.4 demonstrates such an example. The figure compares total X-ray scattering plotted in Q-space, X-ray PDF and SAXS data of the $[Bi_{18}O_{36}]$ and the $[Bi_{22}O_{38}]$ cluster, Figure 2.4D–E. The simulation details are given in Appendix I. The two clusters have no long-range order, and they have a very similar local structure built up by $[Bi_6O_8]$ octahedra, why it is difficult to distinguish them in both the total X-ray scattering data plotted in Q-space and the X-ray PDF data, Figure 2.4A–B. The most prominent difference of the scattering from the two clusters in Figure 2.4A–B is the scattering about Q = 0.45 $Å^{-1}$, which is also the part of the signal which gives rise to the SAXS pattern illustrated in Figure 2.4C. However, due to instrumental limitations, the SAXS pattern is rarely measured during a TS

experiment and in practice these two measurements are done separately. The two clusters have distinct shapes, which makes them easily distinguishable in the SAXS data, Figure 2.4C.

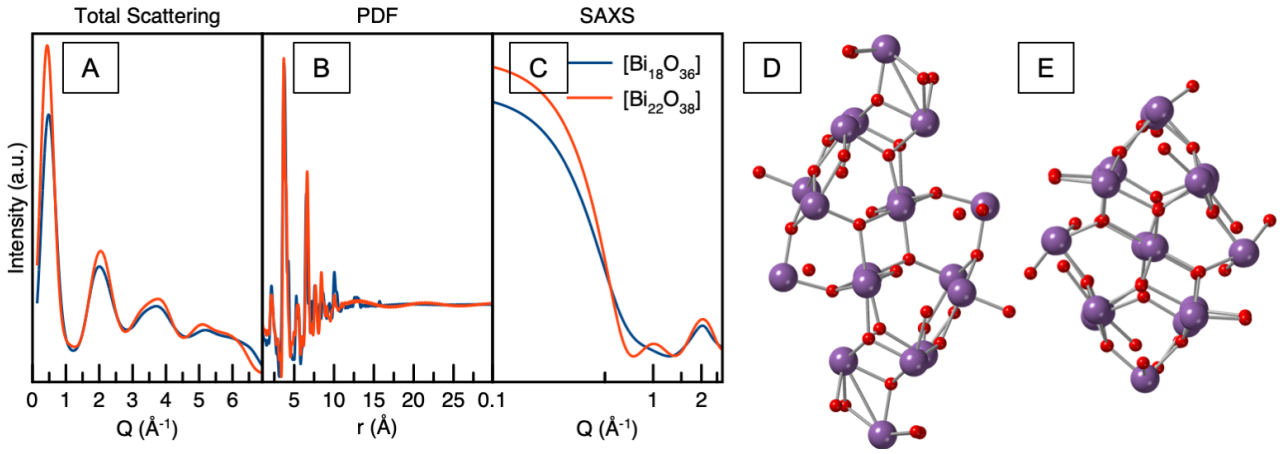



**Figure 2.4: A) Comparison of the simulated total X-ray scattering, X-ray PDF and SAXS data of B) the [$Bi_{18}O_{36}$] cluster and C) the [$Bi_{22}O_{38}$] cluster.**[5, 46]

Nonetheless, the SAXS data does not describe the local order of the structure, which can be modelled with PDF data. A combination of PDF and SAXS data is therefore ideal to fully describe both the atomic structure and the morphology of the clusters. Even when combining the X-ray PDF and SAXS, it can be difficult to describe the atomic positions and vibrations of the oxygen atoms since oxygen $({f_O}^{X-ray}(Q = 0\ \text{Å}^{-1}) = 8)$ scatters significantly less than bismuth $({f_{Bi}}^{X-ray}(Q = 0\ \text{Å}^{-1}) = 83)$. Therefore, one needs to do neutron scattering experiments to get a better description of the oxygen atoms where the bound coherent scattering length is much more similar for oxygen and bismuth $({f_O}^{Neutron}(Q = 0\ \text{Å}^{-1}) = 5.803\ \text{fm})$ and $({f_{Bi}}^{Neutron}(Q = 0\ \text{Å}^{-1}) = 8.532\ \text{fm})$.[41] For a comprehensive description of the atomic structure of these clusters, one would thereby need both TS data with PDF and SAXS from both X-rays and neutrons.

# 3 Conventional Modelling of Scattering Data

We have now demonstrated that Total Scattering (TS) with Pair Distribution Function (PDF) analysis and Small-Angle X-ray Scattering (SAXS) can be powerful tools to describe the structure of a material. However, finding the right atomic structure is a cumbersome process, which often requires expert knowledge. Ideally, we want automated structure characterization of PDF and SAXS data, while doing the measurements, to effectively screen the nanocluster structure during cluster growth or nucleation reactions.

The conventional process to find the right model is to first search the literature of what the expected outcome of the synthesis is. Afterwards, the structural model must be found in a database as the Inorganic Crystal Structure Database (ICSD) or the Crystallography Open Database (COD) and in the end one must fit the model to the dataset. Sometimes this process is straightforward and can be done in hours. Often, this is not the case. As a scientist, we often pursue to synthesize a material, which cannot be found in the database. In that case, one must create the model. Another issue can be that the structure during a synthesis is different than expected. In that case, one must fit several structures from the database to the dataset to find the right model and maybe also create some structural models to supplement the modelling with. Another issue discussed in section 2.5 is that several structures can give reasonable fits to the dataset. A final common problem is the fitting process of the model to the dataset. This is normally done with dedicated software such as PDFGui[47], Diffpy-CMI[48], SASVIEW[49, 50], RMCProfile[51] or TOPAS[52]. The fitting process itself can take a long time. Furthermore, the scientist often has datasets from multiple techniques, which all have to match with the same model. Overall, the conventional process of finding the right model to describe the data is a cumbersome process that often takes months and is constrained to expert users. Therefore, we need new approaches to analyse the data which is faster and more efficient than the conventional minimization approaches such as automated analysis that can be done while doing the measurements.

# 4 Using Machine Learning to Automate Modelling of Scattering Data

When we have a chemical structure, we can analytically calculate its scattering pattern. However, going from a scattering pattern to a chemical structure is a more difficult process that needs modelling. This problem is called the *inverse problem*. While the inverse problem is often a difficult problem to solve for human people, it is an easy task for a ML algorithm. It has previously been shown that ML can automate the process from scattering data to structural model.[36, 53, 54] ML serves as an excellent framework, which can tackle data from all sorts of techniques and data presentations for example both PDF and SAXS data in both reciprocal-space and real-space. The ML algorithm can be trained on numerous examples to learn trends in the dataset, which it can use to classify structures given a scattering pattern. This type of machine learning, where it is trained on examples with a known answer, is called supervised learning. We will demonstrate how supervised learning can be used to model both PDF and PDF data individually but also combined. Furthermore, we will discuss how various techniques can be used to extract reasons behind the predictions of a ML algorithm with a focus on Shapley additive explanations (SHAP).

## 4.1 Gradient Boosting Decision Trees[1]

Gradient Boosting Decision Trees (GBDT) is a type of supervised learning algorithm, which means it can predict a variable $\hat{\boldsymbol{y}}$ based on the input feature list, **x**. This is done with the use of decision trees, which predict variable, $\hat{\boldsymbol{y}}$, using yes/no questions[55] as shown in Figure 4.1. Here it is demonstrated how a tree-based method can be used to distinguish four different cluster structures based on their PDF. If a PDF is given as input feature list, **x,** to a decision tree, it can for example ask "Does the PDF have a Bi – O peak?" (Figure 4.1 top question) and thereby distinguish between the clusters containing oxygen atoms from the clusters not containing oxygen atoms.

[1] This Chapter is based on the theory presented in https://xgboost.readthedocs.io/en/latest/tutorials/model.html

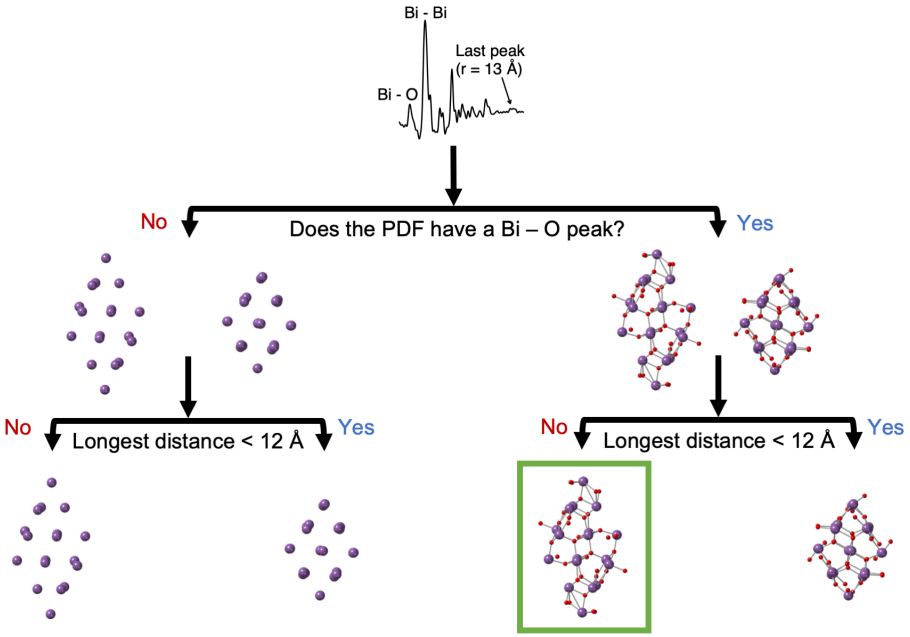



**Figure 4.1: Example of the use of a GBDT algorithm to predict the cluster structure, $\hat{y}$, based on its PDF as the input feature list, x. The green box indicates the structure that would be predicted by this decision tree from this PDF.**

The trees are made by minimizing an error function on already labelled data, or ground truths, **y**. This process is referred to as training the algorithm.

**Equation 4.1**

$$obj = Error_{Rec} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right)$$

Where obj or error function is a function summing all the i[th] trees with their prediction values, $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(\boldsymbol{t})}$, at step t. $\mathbf{y_i}$ is the true prediction value. The error function used to calculate the error value between $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(\boldsymbol{t})}$ and $\mathbf{y_i}$ is annotated **l**. The user-defined input of the maximum of trees is given as **n**.

To understand the training process, let us consider an example where we train the algorithm to predict a 3[rd] degree polynomial (green curve) including noise as shown in Figure 4.2. In this case, we define the variable M as the number of variables that are used in the polynomial to fit the data. However, in the GBDT algorithm, this transfers to how complex the algorithm is. It is thereby a measure of how complex the GBDT algorithm is i.e. number of trees, depth of trees, leaf's per tree etc. When no training has been done, the algorithm will always predict an average value as visualized in Figure 4.2A. However, after some training (M = 1, Figure 4.2B) the algorithm still predicts a line

but now with the right slope. This period, where the algorithm has not finished training, is referred to as underfitting. After some training (M = 3), the algorithm catches the trend of a 3rd degree polynomial. This is where the algorithm is trained to an optimal point. But if the algorithm is trained further, it will keep minimizing the error function and thereby describe all the points in the data which it has been trained on (training data), Figure 4.2D. While this behavior describes the training data very well (or perfectly), it would not generalize well to other points in a 3rd degree polynomial (the green curve). This behavior where the algorithm does predict perfectly on the training set but does not generalize well is referred to as overfitting. There are multiple tools to avoid overfitting. The most common tool is to split the data into training and validation set. Typically, this is a split of 80 % and 20 % of the data. For every iteration the algorithm has been trained, it can be validated on the validation set in order to determine if it is in the underfit or overfit regime. Thereby, the optimal training iterations can be determined, in this case M = 3. Another way of avoiding overfitting is to increase the size of the training set on the cost of longer training time.

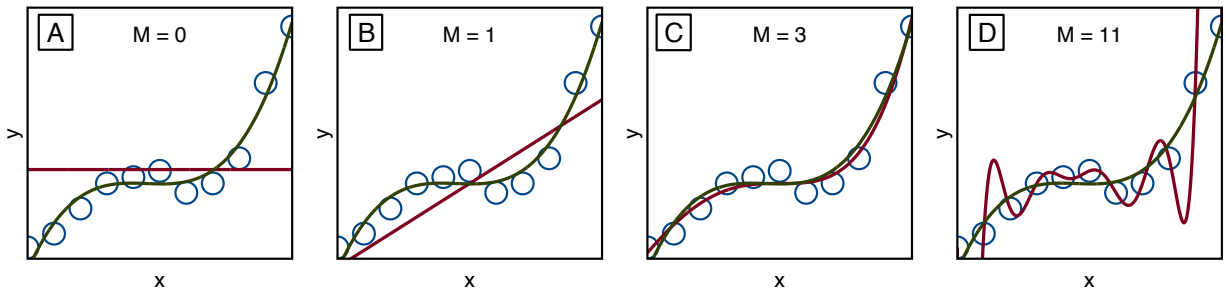



**Figure 4.2: Modelling of a 3rd degree polynomial with noise using A) A constant, B) A line, C) A 3rd-degree polynomial and D) A 11th-degree polynomial.**

## 4.2 Feature Ranking using SHapley Additive exPlanations analysis

To analyse the nature of the GBDT algorithm or any other ML algorithms, it is very useful to know which features are important for the algorithm in its predictions. This research field is called interpretable machine learning and the method we will highlight is called feature ranking.

The most intuitive form of feature ranking is permutation feature importance, which measures the increase in prediction error when values of a feature are permuted.[56, 57] However, permutation feature importance is highly inaccurate on correlated data.[56]

Another feature ranking method is Shapley values which come from coalitional game theory.[58] A Shapley value quantifies how much a feature contributes to a prediction compared to the overall

prediction. The Shapley value is calculated by calculating weighted and summed feature contributions over all the possible feature value combinations.

**Equation 4.2**

$$\phi_j(v) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(v\left(S \cup \{x_j\}\right) - v(S)\right)$$

Where **x** is the feature values, **p** is the total number of features, **S** is the subset of features and **v(S)** is the prediction for feature values in the feature subset, **S**. The Shapley value does not need any form of user-defined inputs, however, they are computationally expensive to calculate.[56] An alternative to Shapley values are SHAP values, which is a fast and accurate approximation of Shapley values.[38] The SHAP value is thereby a quantification of how much the input feature in a ML model pushes the prediction up or down from the average prediction. All the advantages of Shapley values do also account for SHAP values, however, it is an approximation that imposes smaller issues with highly correlated features. Another issue with SHAP values is that the feature value under investigation is randomized to calculate the Shapley values, which not necessarily is a good assumption. Both issues are under research but, to my knowledge, no python modules that are easy to access have been developed yet.[59, 60] In general, interpretable machine learning is not an exact science, why we must interpret the models with great care.[61]

# 5 Following Structural Changes during the Growth of Atomically Precise Metal Oxido Clusters from combined PDF and SAXS analysis[2]

Atomically precise bismuth oxido nanoclusters have recently been extensively investigated due to the well-controlled synthesis of bismuth oxido clusters in various sizes,[1, 2 3 4-7] and bismuth have been used in applications as radiopaque materials[4, 16] and photocatalysis.[8-12] The ability to control the size of bismuth oxido clusters also makes them an excellent model system to understand growth mechanisms and nucleation reactions. The bismuth oxido clusters are reported to be built up by $[Bi_xO_y]^{z+}$ units with varying charge and nuclearity.[2] The clusters are stabilized by a negative charged ligand shell. We will in this Thesis neglect the charge and ligand shell and focus on the structural motifs of the bismuth oxido clusters. The clusters investigated here are built by edge-sharing octahedral $[Bi_6O_8]$ units. Figure 5.1B illustrates a single octahedral unit, while Figure 5.1C–D illustrates the $[Bi_{22}O_{26}]$ cluster which consists of 6 octahedral units and the $[Bi_{38}O_{45}]$ which is built by 13 octahedral units. Generally, the nanoclusters can be considered as cut-outs of the fluorite-type structures as the δ- and β-$Bi_2O_3$.[2, 3, 62] Metal oxido clusters with different elements than bismuth are also built of the octahedral $[M_6O_x]$ units as $[Ce_{22}O_{24}]$,[63] $[Ce_{38}O_{54}]$,[13] $[U_{38}O_{56}]$,[64] and $[Pu_{38}O_{56}]$.[65, 66] The $[M_6O_x]$ building block is therefore anticipated to play a central part in the cluster growth.

While the cluster structures are well-known from single crystal diffraction, it is more challenging to study the cluster growth in solution. Sattler et al. has investigated the cluster growth from $[Bi_6O_8]$ to $[Bi_{38}O_{45}]$ clusters using Electrospray Ionization Mass Spectrometry (ESI-MS),[67] and identified the $[Bi_6O_8]$, the $[Bi_{22}O_{26}]$ and the $[Bi_{38}O_{45}]$ as particularly stable cluster species. They also identified other fragments of bismuth oxido species in minor concentrations. However, while ESI-MS is very sensitive to small concentrations of charged fragments, it cannot be used to follow structural changes during a cluster growth reaction. Extended X-ray Absorption Fine Structure (EXAFS) and Nuclear Magnetic Resonance (NMR) can be used to characterize the atomic structure of clusters in solution[68-71] but both techniques are limited to the characterization of the atomic structure in the short range.

[2] This chapter is based on some of the results from *Anker et. al., Structural Changes during the Growth of Atomically Precise Metal Oxido Nanoclusters from Combined Pair Distribution Function and Small-Angle X-ray Scattering Analysis, Angew. Chem. Int. Ed.* **2021,** *60*, 2-12. Which is also included as Appendix V.

Here, we use *in situ* TS with PDF and SAXS analysis to follow the formation of the $[Bi_{38}O_{45}]$ cluster, by dissolving crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$[72] (Figure 5.1A) in dimethylsulfoxid (DMSO). PDF is an excellent technique to characterize nanoclusters in solution because it is sensitive to the atomic structure of the clusters.[65, 70] However, PDF is not sensitive to cluster shape and size distribution. Therefore, similar experiments have been done using SAXS which is highly sensitive to shape and size distribution.[19, 48, 73, 74] To analyse the data, we have developed a new tool to identify intermediate species and map the reaction pathway. This automated intermediate extractor finds the best fitting structural motif that is contained in the $[Bi_{38}O_{45}]$ cluster for every timestep of the reaction. This approach does not identify the $[Bi_6O_8]$ cluster motif as a stable intermediate, but it finds intermediate species with a stable core of $[Bi_{22}O_{26}]$. Furthermore, combined modelling of PDF and SAXS data is introduced with the use of the Debye equation. The combined modelling confirms the $[Bi_{38}O_{45}]$ cluster as the reaction product and the $[Bi_{22}O_{26}]$ cluster as the intermediate and it is used to yield a robust phase fraction of the two phases. Overall, automated modelling of combined PDF and SAXS data seems like a promising approach to understand cluster growth mechanisms in solution.

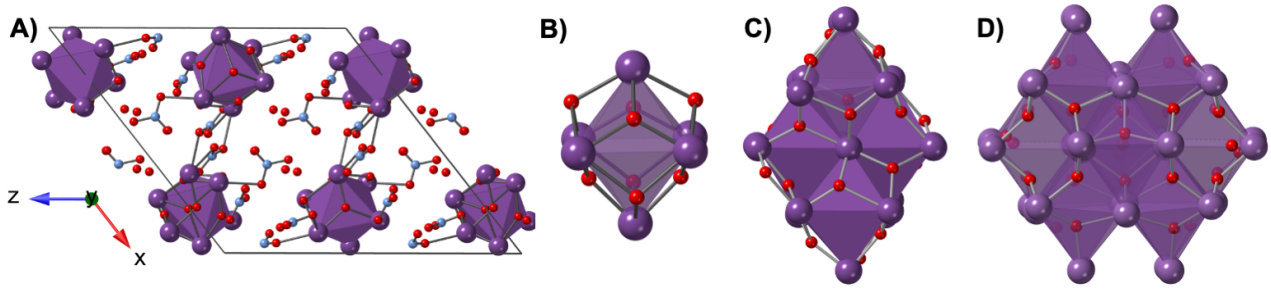



**Figure 5.1: The structure of the A) $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ crystal.[72] B) single octahedral $[Bi_6O_8]$ unit. C) $[Bi_{22}O_{26}]$[5] cluster. D) $[Bi_{38}O_{45}]$[15] cluster. Bismuth is shown in purple, oxygen in red, and nitrogen in blue for all structures. Hydrogen atoms have been omitted for clarity. Adapted from Anker et al.[33]**

## 5.1 Using conventional refinement approaches to analyse the growth of bismuth oxido nanoclusters

Figure 5.2A–B shows *in situ* TS data in Q-space and r-space obtained during the cluster growth reaction of $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ in DMSO at $T$ = 80 °C. Experimental details are given in Appendix II. The two data representations highlight complimentary information. Figure 5.2A shows Bragg peaks in the first 10 min of the reaction, which indicates a crystalline phase. It would be

intuitive that the crystalline phase at the beginning of the reaction is $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$, why a pattern of $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ were simulated and compared with the data in Q-space at $t$ = 0 min, Figure 5.2D. The experimental dataset possesses very similar peaks as the simulated dataset of crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$, however, the background of the experimental dataset makes it difficult to model with for example a Rietveld refinement. Therefore, the data were investigated in r-space instead. Background subtraction of the data in r-space is shown in Appendix II. The crystalline phase is observed in the r-space data, Figure 5.2B, as peaks up to large r-values. Figure 5.2E demonstrates a fit of the crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ phase to the first dataset of the *in situ* dataset in r-space. After $t \sim 7$ min the crystalline precursor begins to dissolve as seen from the disappearing of the Bragg peaks in Figure 5.2A and by the disappearing of the peaks at high r-values in Figure 5.2B. Modelling in Q-space was not attempted due to the very broad scattering signal as seen in Figure 5.2A and further highlighted in Figure 5.2D. The broad features in the Q-space data are most likely due to diffuse scattering. Instead, the last frame of the data in r-space was modelled with a $[Bi_{38}O_{45}]$ cluster, confirming that this is the product of the reaction, Figure 5.2F.

However, as shown in Figure 2.4, the model that fits the TS data is not necessarily unique. Therefore, we conducted a similar SAXS experiment to confirm the results. TS and SAXS datasets obtained from similar experiments performed at lower temperatures (T = 30–60 °C) can be found in Appendix II. Figure 5.2C shows the *in situ* SAXS experiment obtained from crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ in DMSO at $T$ = 80 °C. Again, Bragg peaks are observed at the beginning of the reaction, which broadens after approximately 7 minutes. The Bragg peaks at the beginning of the reaction have the same Q-values as for the TS experiments, indicating that the crystal is $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$. Again, the last frame of the experiment was modelled with the $[Bi_{38}O_{45}]$ cluster, Figure 5.2G, confirming the results from the PDF modelling. The PDF and SAXS modelling details are described in Appendix II.

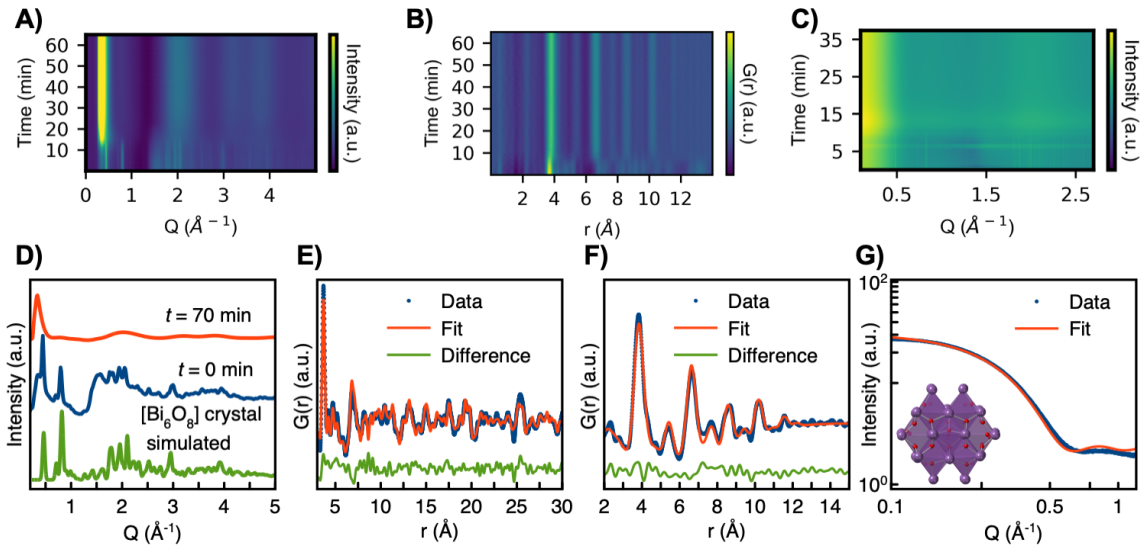



**Figure 5.2: The background subtracted *in situ* X-ray TS data obtained during the experiment performed at *T* = 80 °C plotted as a function of time. While Bragg peaks are seen at the beginning of the reaction, only diffuse features are seen after dissolution. A) Data presented in Q-space B) Data presented in r-space. C) The logarithmic background subtracted SAXS data obtained during the experiment performed at *T* = 80 °C plotted as a function of time. While Bragg peaks are seen at the beginning of the reaction, only diffuse features are seen after dissolution. D) The simulated X-ray TS data of the crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$, compared to background subtracted data obtained after *t* = 0 min and *t* = 70 min of the *T* = 80 °C *in situ* experiment. E) The fit ($R_{wp}$ = 29.0 %) of the $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ crystal structure to the PDF obtained after *t* = 0 in the *in situ* experiment at *T* = 80 °C. F) The fit of the $[Bi_{38}O_{45}]$ cluster structure to the PDF obtained at the end of the experiment at *t* = 70 min. $R_{wp}$ = 16.6 %. G) The fit ($R_{wp}$ = 3.1 %) of the $[Bi_{38}O_{45}]$ cluster structure to the SAXS data obtained after *t* = 37 min at the *in situ* experiment at *T* = 80 °C. The SAXS data are plotted in a log-log plot. Adapted from Anker et al.[33]**

We have now established how crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ dissolves into a cluster in DMSO, which grows to the $[Bi_{38}O_{45}]$ cluster. However, we have yet not investigated the intermediate cluster species during the growth mechanism. Figure 5.3A–B shows chosen PDF and SAXS data through the *in situ* reaction. The data represents the crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ in the beginning of the reaction and the $[Bi_{38}O_{45}]$ cluster in the end of the reaction. The intermediate dataset has similar motifs as the last dataset but the PDF damps at a lower r-value indicating a smaller species. Furthermore, the intermediate species are more disordered than the $[Bi_{38}O_{45}]$ cluster seen from the asymmetric peak in Figure 5.3D. Therefore, we suggest that the intermediate must be structurally correlated with the $[Bi_{38}O_{45}]$ cluster but present as a smaller and more disordered species.

[$Bi_6O_8$] cluster has previously been theoretically predicted to form in DMSO with molecular dynamic simulations[1] and experimentally it has been described as particularly stable.[62, 67] Therefore, the PDF and SAXS scattering patterns were simulated from the [$Bi_6O_8$] cluster and compared to data through the reaction, Figure 5.3A–B. The simulated data of the [$Bi_6O_8$] cluster does not match any of the datasets and it is too small based on the lack of long-range peaks in the PDF. Oppositely, the intermediate PDF or SAXS data is not well described by the [$Bi_{38}O_{45}$] cluster either, Figure 5.3C and Figure 5.3E.

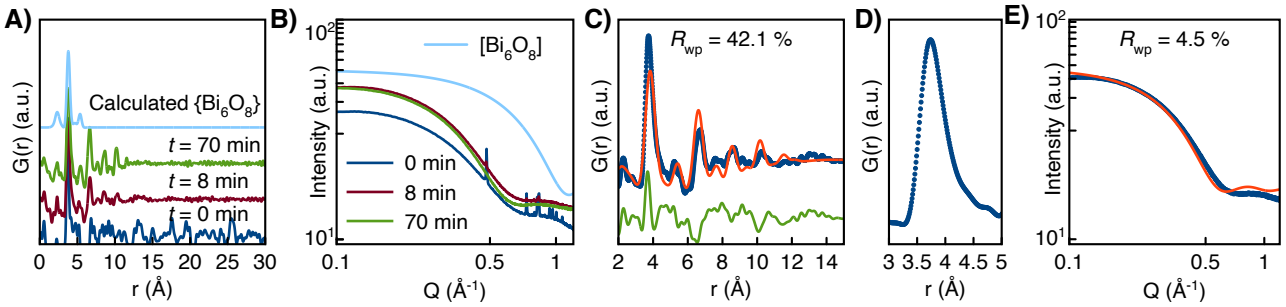



**Figure 5.3: Comparisons between calculated patterns from an isolated [$Bi_6O_8$] unit and experimental (A) PDFs and (B) SAXS data (plotted on a log-log scale). C) The fit of the [$Bi_{38}O_{45}$] cluster structure to the PDF obtained just after the dissolution of the crystalline starting material, $t$ = 8 minutes, for the *in situ* experiment obtained at $T$ = 80 °C. D) The asymmetric peak at $r$ = 3.8 Å in the PDF. E) The fit of the [$Bi_{38}O_{45}$] cluster structure to the SAXS data (plotted on a log-log scale) obtained just after the dissolution of the crystalline precursor $t$ = 16 minutes, for the *in situ* experiment obtained at $T$ = 80 °C. Adapted from Anker et al.[33]**

To understand the size evolution through the reaction, we applied SAXS form factor analysis of the *in situ* SAXS data using a spherical model with a lognormal size distribution. Figure 5.4 shows the change of diameter and dispersity of the clusters through the reaction. Modelling details and two examples of fits are given in Appendix II. During reaction, the clusters grow and get more monodisperse. We therefore attempt to fit other bismuth oxido clusters with a similar local structure as the [$Bi_{38}O_{45}$] but with a smaller size. Figure 5.5 illustrates fits of various cluster sizes from [$Bi_6O_8$] to the [$Bi_{38}O_{45}$] to the intermediate PDF dataset. It is observed how the [$Bi_{18}O_{36}$] and [$Bi_{22}O_{38}$] models make the best fits, but a unique cluster structure is still not determined.

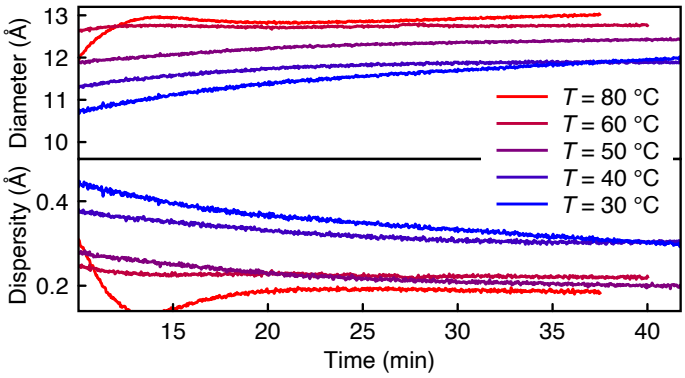



**Figure 5.4: The diameter and dispersity of the clusters, plotted as a function of reaction time at 5 different reaction temperatures. The values are obtained from a SAXS form factor analysis using a spherical model with a lognormal size distribution. The data analyzed here were obtained at SWING at Soleil, but similar results from the analysis of data from ID02, ESRF are shown in Appendix II. Adapted from Anker et al.[33]**

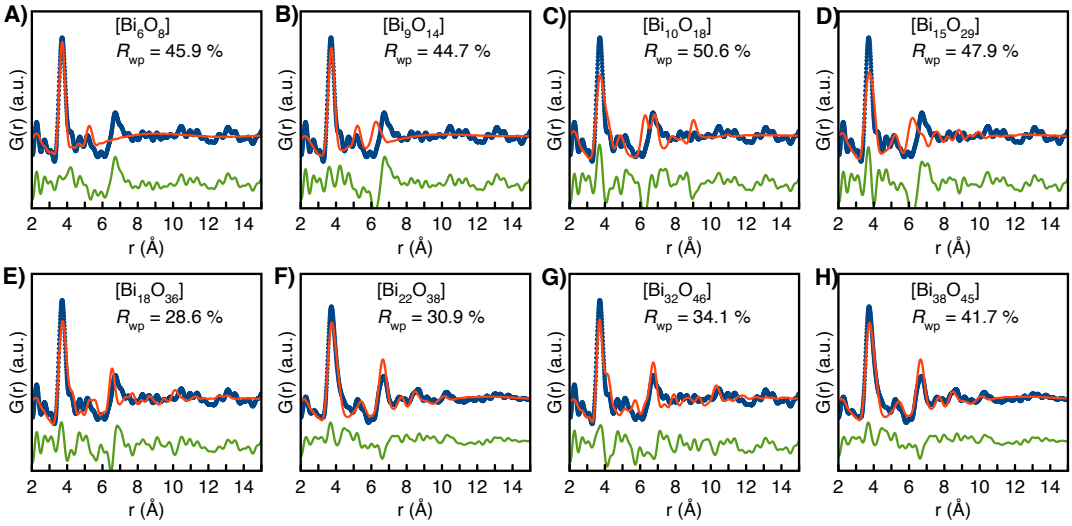



**Figure 5.5: The PDF obtained after *t* = 8 min for the *in situ* experiments at *T* = 80 °C is fitted to different bismuth oxido clusters, which were cut out of their corresponding crystal structures.[5, 7, 15, 46, 72, 75, 76] Adapted from Anker et al.[33]**

## 5.2 Developing 'automated intermediate extractor' to determine the intermediate cluster

We have now seen that the intermediate species has the same local motif as the $[Bi_{38}O_{45}]$ built up by edge-sharing $[Bi_6O_8]$ units and that it has a size approximately about the size of the $[Bi_{18}O_{36}]$ and $[Bi_{22}O_{36}]$ cluster. However, to get an overview of the intermediate species during the reaction, we created an algorithm that automatically extracts the intermediate structure which I refer to as *automated intermediate extractor*. Automated modelling has previously been proposed by Yang et al.[35] and Banerjee et al.[30] who automated characterization of crystal structures and metal nanoparticles. In another recent study, Christiansen et al. extract structural motifs of disordered molybdenum particles by automatically fitting fragments of known polyoxometalate structures to the PDF.[22] I will refer to this approach as *automated motif extractor*.

The method uses the methodology in the latter paper. We have adapted automated motif extractor to extract intermediate species of *in situ* PDF data. The principle of the algorithm is illustrated in Figure 5.6. First, the last frame of the dataset is modelled, which we know corresponds to the $[Bi_{38}O_{45}]$ cluster. The goal is to validate that there is no smaller structural motif with the same local order that fits the data better than the $[Bi_{38}O_{45}]$ cluster. This is checked by removing atoms systematically from the structure and fit them to the data. In order to test all combinations of atoms, which is done in automated motif extractor, $38^2 - 1$ fits have to be made. Since this is not computationally feasible, we restrict the algorithm to remove up to 8 atoms from the outer shell of the $[Bi_{38}O_{45}]$ cluster. Finally, all oxygen atoms which were not bonded to at least a single Bi atom were removed. The longest allowed bond length was set to 2.5 Å, which is the longest Bi···O distance observed in the experimental PDFs. The best-fitting model is chosen based on the $R_{wp}$ value. When this procedure is done, the process is proceeded for the next frame in the dataset (starting from the last and going towards the first frame). The best-fitted structure from the previous frame is now added 3 atoms in the outer shell and the procedure is continuing with removing up to 8 atoms.

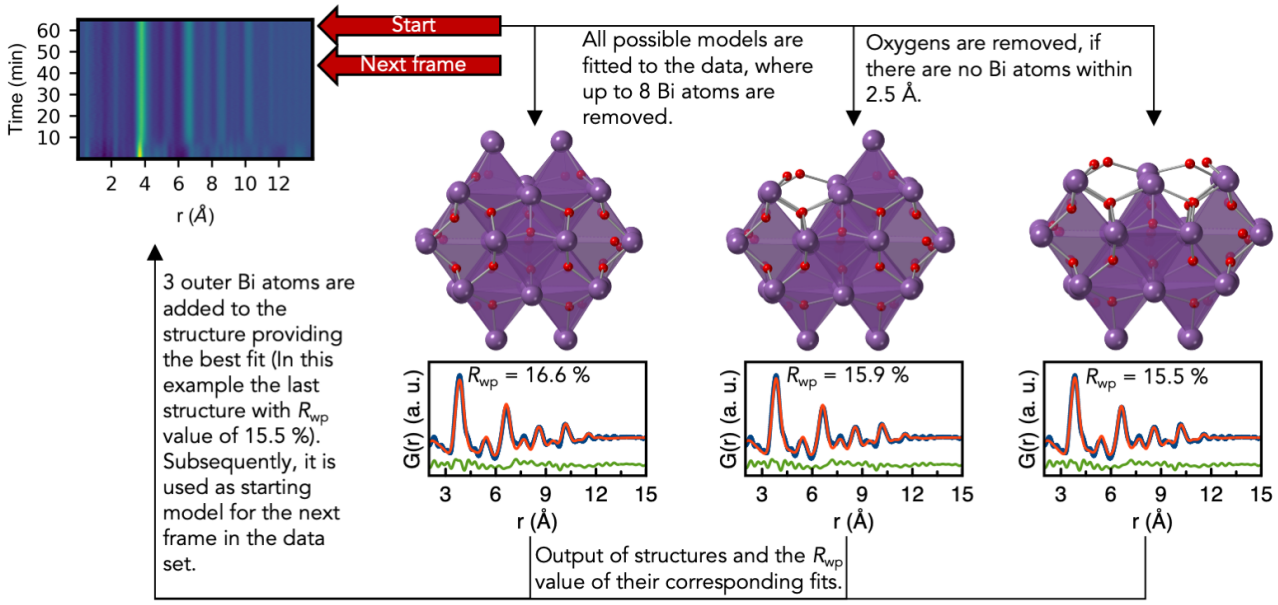



**Figure 5.6: The pseudo-code of the automated modelling process shows that: firstly, the last frame of the *in situ* dataset is modelled with the [$Bi_{38}O_{45}$] cluster structure, and all structures where up to 8 Bi atoms from the outer shell is removed together with non-bonding oxygens. This is $2^8$ = 256 candidate structures in total. The best-fitting structure is added three Bi atoms and used as the starting-point for the fitting of the second-last frame. The addition of the Bi-atoms is done in order to prevent false minima in the algorithm. The process is repeated for all frames in the reaction. Adapted from Anker et al.[33]**

This modelling approach gives an overview of the best fitting intermediate species through the reaction that has the same structural motif as the [$Bi_{38}O_{45}$] cluster. Figure 5.7A–C shows the number of Bi atoms of the best fitting structures during the reaction for the *in situ* PDF dataset obtained with $T$ = 80 °C, 60 °C and 30 °C. For the high temperature dataset, $T$ = 80 °C, large structures with about 32 atoms are preferred, while plateaus of structures with 26 atoms and 22 atoms are stable for the dataset obtained at $T$ = 60 °C and 30 °C. Results for the datasets obtained at $T$ = 40 °C and 50 °C are shown in Appendix II. A histogram of the occurrence of cluster sizes through the reactions is illustrated in Figure 5.7D. A large portion of clusters in the range 20–25 atoms are found. Figure 5.7E–H shows fits of chosen dominant cluster species from the automated intermediate extractor. While the structures fit the data well, they are unphysical with 'dangling' Bi atoms but they all contain the same $Bi_{22}$ motif in the backbone of the structure. This is especially clear for structure III and IV, which is compared to a $[Bi_{22}O_{26}]^5$ structure in Figure 5.8.

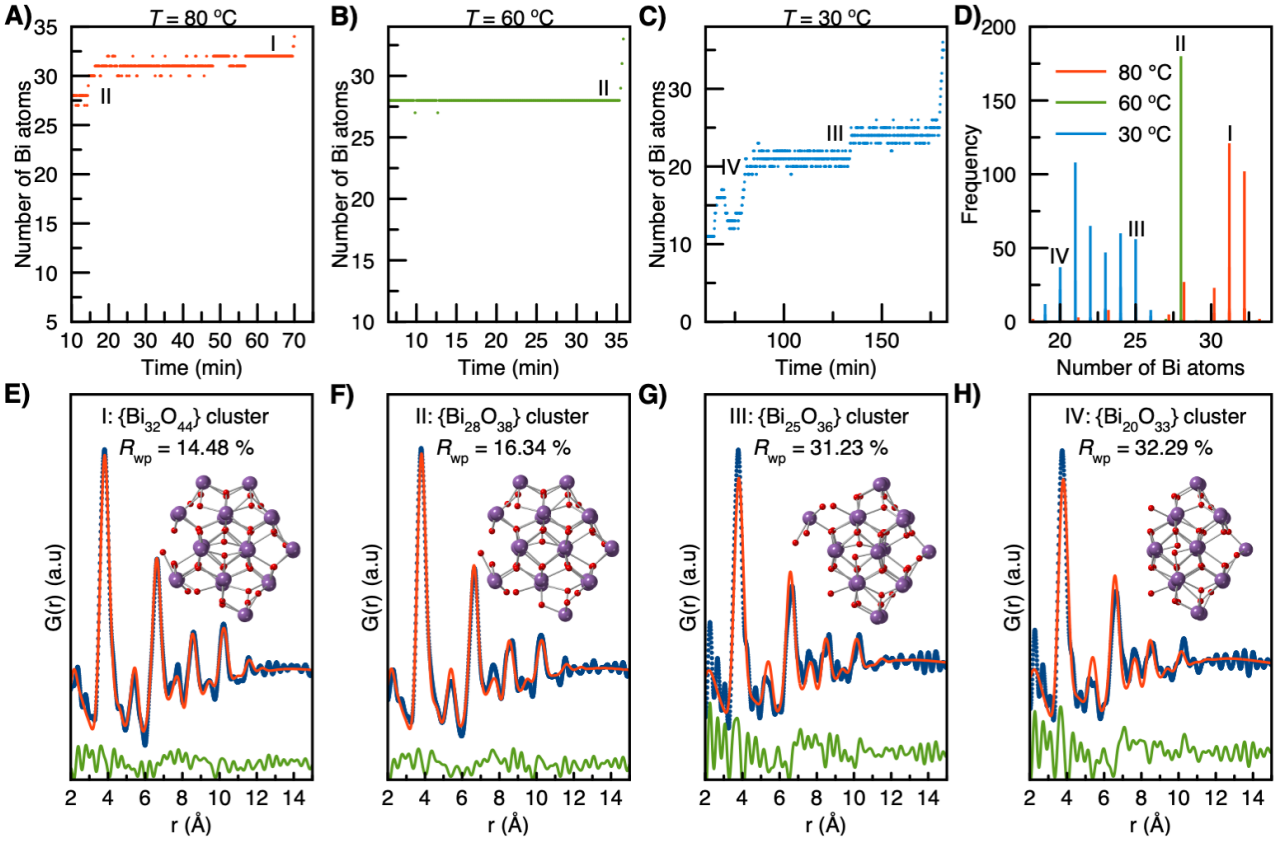



**Figure 5.7: Number of bismuth atoms in the best fitting structure for each frame is plotted versus the reaction time for the experiments done at A) *T* = 80 °C, B) 60 °C and C) 30 °C. D) A histogram of each unique best fitting structure found with the automated modelling process through the reaction. Frequency is the number of times the specific cluster size (number of Bi atoms) is present during the reaction. Some sizes are present multiple times. E–H) Representative clusters and fits that have been obtained in the automated modelling process. Adapted from Anker et al.[33]**

When modelling the intermediate datasets with the physical $[Bi_{22}O_{26}]$ structure, cut out of a $[Bi_{22}O_{26}(OSiMe_2tBu)_{14}]$ crystal,[5] a comparable fit to the unphysical structures are obtained, Figure 5.8B–C. The automated intermediate extractor thereby indicates that the intermediate is a $[Bi_{22}O_{26}]$ cluster that grows through the reaction. Other cluster species such as the $Bi_{28}$ and $Bi_{32}$ also appear through the analysis, however, these do not appear physical as closed-shell clusters. The results are supported by the SAXS form factor analysis, which also showed a growing size through the reaction and a decreasing size dispersity of the particles indicating that there can be multiple cluster species present simultaneously in the beginning of the reaction.

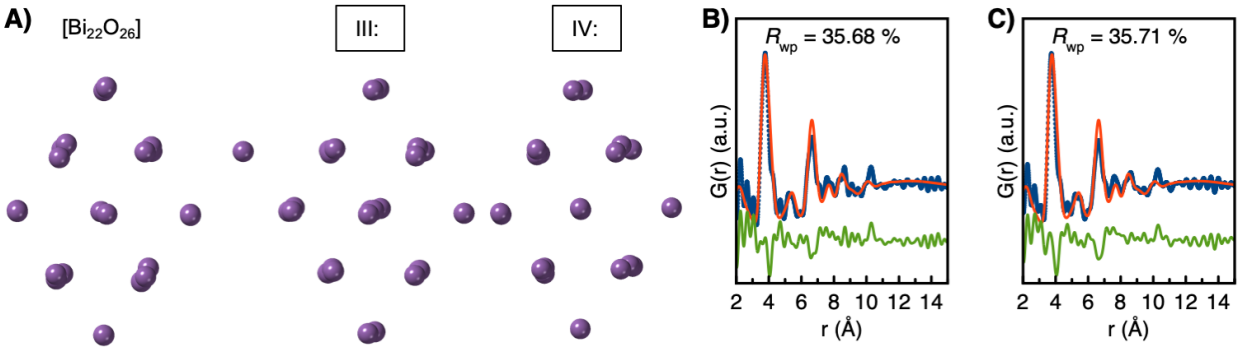



**Figure 5.8: A) The [$Bi_{22}O_{26}$] cluster from Mansfeld et al.[5] has a similar structure to structure III and IV, found from the automated modelling method. The fit to the *in situ* PDF ($T$ = 30 °C) obtained at B) $t$ = 152 min and C) $t$ = 103 min using the [$Bi_{22}O_{26}$] cluster as the model. Adapted from Anker et al.[33]**

It has now been established that the intermediate cluster species of the reaction is a [$Bi_{22}O_{26}$] cluster and the product of the reaction is the [$Bi_{38}O_{45}$] structure. Therefore, it was investigated if the data is better described with a 2-phase system of the [$Bi_{22}O_{26}$] and [$Bi_{38}O_{45}$] cluster compared to the best fitting average structure obtained from the automated intermediate extractor.

In Figure 5.9 data from four steps during the reaction is investigated. The best-fitting average cluster model from the automated motif extractor is compared with a 2-phase refinement of the [$Bi_{22}O_{26}$] and [$Bi_{38}O_{45}$] cluster using PDF and SAXS Debye modelling and SAXS form factor analysis. Late in the reaction, the one-structure model is very similar to the 2-phase refinement. However, earlier in the reaction, $T$ = 30 °C, the 2-phase refinement is describing the SAXS data significantly better. Thereby, it can be concluded that the 2-phase refinement using the [$Bi_{22}O_{26}$] and [$Bi_{38}O_{45}$] cluster yields the best description of the data throughout the entire reaction. Modelling details of the refinements are shown in Appendix II. Appendix II also contain further argumentation for the 2-phase fit describes the data over the entire reaction of the $T$ = 80 °C experiment better than the individual phases.

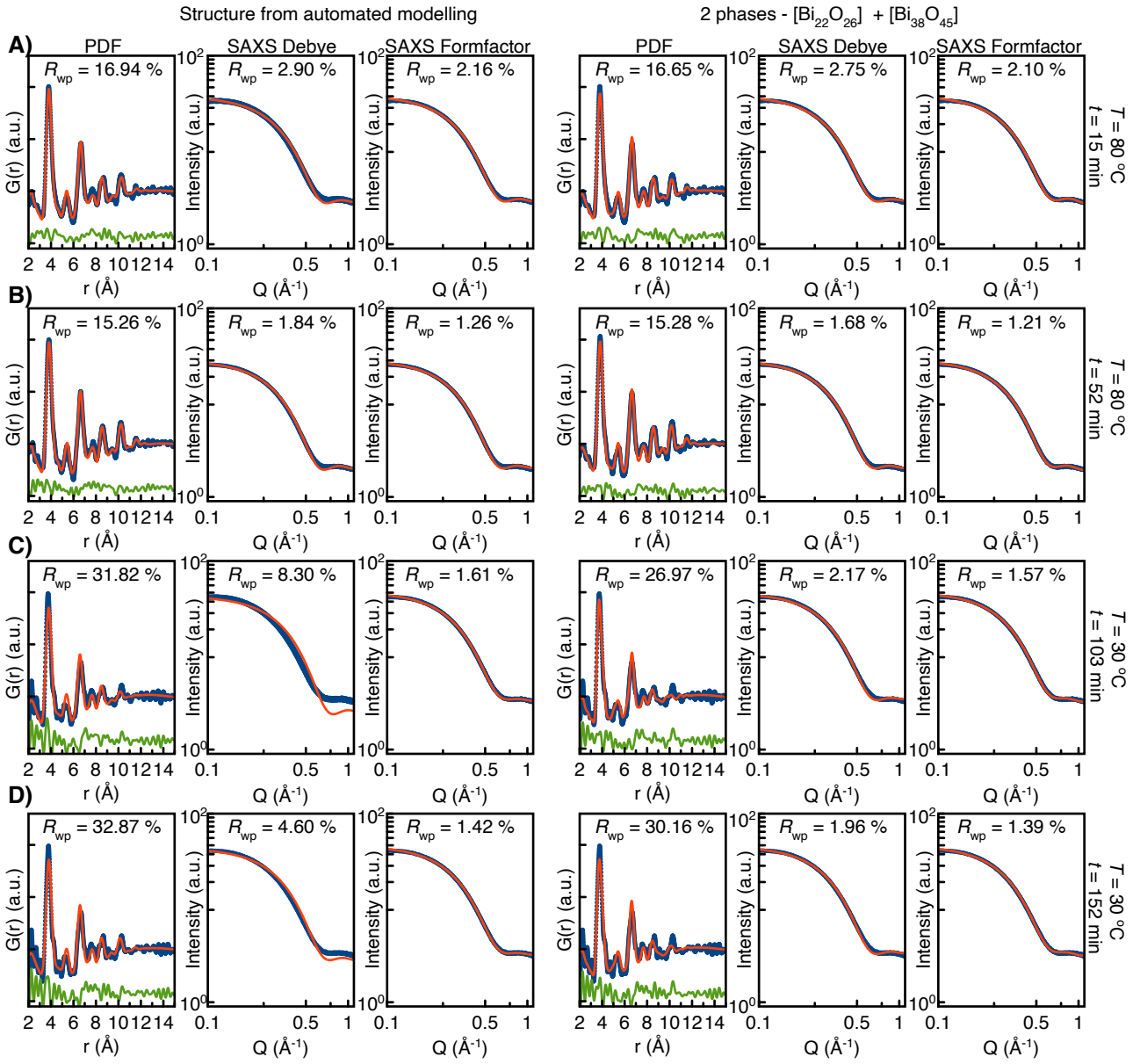



**Figure 5.9: Comparison of fits using the best fitting cluster structure identified with the automated modelling process or using a 2-phase model with the [$Bi_{22}O_{26}$[ and [$Bi_{38}O_{45}$] cluster are used to selected PDFs and SAXS patterns obtained during the *in situ* experiments at A – B) $T = 80$ ° and C–D) $T = 30$ °C. The fits to the PDF and SAXS data using the structure models obtained from the automated modelling process are shown to the left (modelled using the Debye function), together with a simple SAXS form factor analysis of the SAXS data. The fits to the PDF and SAXS data using a two-structure model of [$Bi_{38}O_{45}$] and [$Bi_{22}O_{26}$] are shown to the right (modelled using the Debye function), with a SAXS form factor analysis, here including two particle shapes corresponding to the [$Bi_{38}O_{45}$] (spherical) and [$Bi_{22}O_{26}$] (ellipsoid) clusters. Adapted from Anker et al.[33]**

## 5.3 Using combined PDF and SAXS modelling to achieve robust and accurate results

Now that it has been established that the best description of the reaction is a 2-phase refinement with the $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ cluster, the phase fractions during reaction can be determined. Figure 5.10 shows the phase fraction of the two clusters during the reaction at temperatures $T$ = 80 °C, 60 °C and 30 °C. Results for $T$ = 40 °C and 50 °C are shown in Appendix II. The red curves indicate the results from PDF analysis, the blue curves indicate results from SAXS analysis, and the green is from combined analysis of the PDF and SAXS data. After the crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ is dissolved in DMSO, the $[Bi_{22}O_{26}]$ phase is dominant. However, during the reaction, the $[Bi_{22}O_{45}]$ cluster species grow to the $[Bi_{38}O_{45}]$ cluster which ends as the dominant phase at high temperature, $T$ = 80 °C, or after a long time at room temperature, Appendix II, which also shows that the results are similar if the $[Bi_6O_4(OH)_4(NO_3)_6(H_2O)_2]\cdot(H_2O)$ crystal is used as the crystalline precursor. At all three temperatures, the results from the PDF and SAXS analysis follow the same trend, indicating that the kinetics of the experiments are close to equivalent. Therefore, it is intuitive to try combined modelling of the PDF and SAXS data, which results are indicated with green curves.

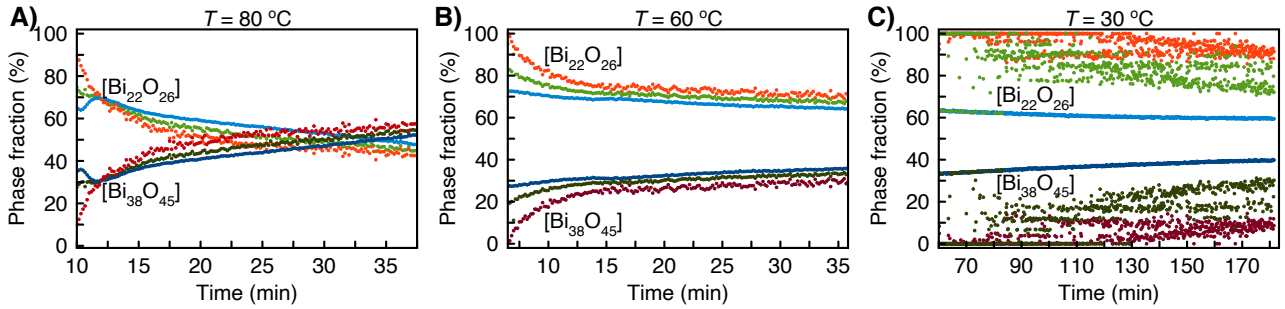



**Figure 5.10: The phase fractions of the $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ clusters, plotted as a function of time for experiments at A) $T$ = 80 °C, B) 60 °C and C) 30 °C. The results from PDF analysis are plotted in red, from SAXS analysis in blue, and from complex modelling in green. Adapted from Anker et al.[33]**

Combined modelling of data from multiple techniques is in the literature referred to as 'complex modelling'.[48, 51, 73, 74, 77] Complex modelling of PDF and SAXS data have previously been shown to yield a more robust description of shape and size than PDF modelling alone.[73] A similar approach has been used in this project, using the Diffpy-CMI framework, which was originally written to enable complex modelling of data from multiple techniques.[48] However, Farrow et al. modelled the data using a crystalline model applying a 'shape' function to mimic the finite size of the nanocrystal,[73] whereas we model the data using the Debye equation.

To my knowledge a yet unsolved problem when doing complex modelling has been the weighting of the datasets. In this project, the optimal weighting of the datasets has been determined by making a 'calibration' of the weights. Figure 5.11 shows the effect of using various weights for PDF and SAXS data in the complex modelling approach when modelling the data obtained after $t$ = 37 min from the *in situ* experiment at $T$ = 80 °C. When the weights are in a ratio 1:1 (PDF:SAXS), the PDF fit is identical to fitting the PDF data individually, however, the fit of the SAXS data is not optimal. When using a weight ratio of 1:1000, the SAXS data is well-described, but the PDF data is not. In order to achieve satisfying fits for both techniques a weight between 1 and 1000 has to be used.

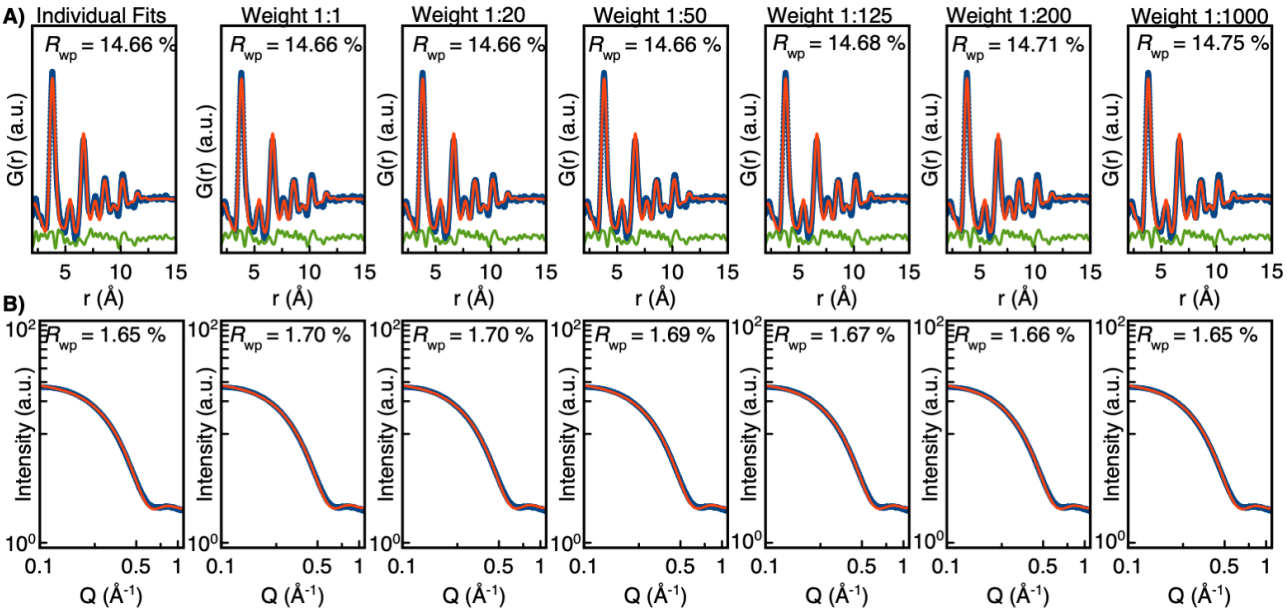



**Figure 5.11: The fits of the PDF & SAXS data obtained during *in situ* experiments after $t$ = 70 min (PDF) and $t$ = 37 min (SAXS) at $T$ = 80 °C using complex modelling with different weights. Here the weight of 1:20 means that the residuals from the SAXS fit are multiplied by 20 times in the fitting process. Adapted from Anker et al.[33]**

To find the optimal spot where both the PDF and SAXS data are satisfied in the fitting process, a calibration curve is made as shown in Figure 5.12. The $R_{wp}$ value of the PDF and SAXS fit is plotted versus the weight ratio used. When lower weight ratios are used, the PDF data is fitted well, while the SAXS data is fitted well for higher weight ratios. However, when using a weight ratio of 1:125, both datasets achieve low $R_{wp}$ values.

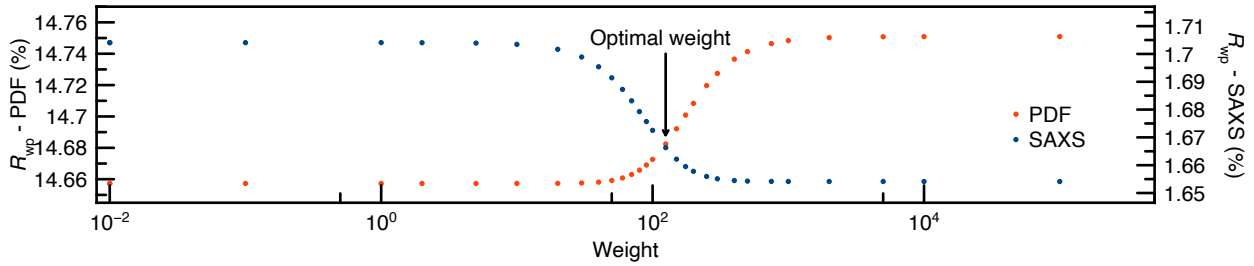



**Figure 5.12: The $R_{wp}$ values of the PDF and SAXS fit on the $T$ = 80 °C, $t$ = 70 min (PDF) and $t$ = 37 min (SAXS), *in situ* data from complex modelling with different weights shows that there is an optimal weight, which favours both techniques. Adapted from Anker et al.[33]**

This weight was therefore used in the complex modelling procedure shown in Figure 5.10. Similar analysis of phase fraction versus reaction time were made using other weights from weight ratio 1:1 to 1:1000 shown in Figure 5.13. The same trend as Figure 5.11 and Figure 5.12 is seen. When the weight is not calibrated, the results of the complex modelling approach mimic the results of one of the datasets individually. However, when an optimal weight ratio of 1:125 is used, the complex modelling approach does optimize the fitting quality of both datasets. It is thereby concluded that a weight ratio of 1:125 must be used to get the best overall description of the PDF and the SAXS data.

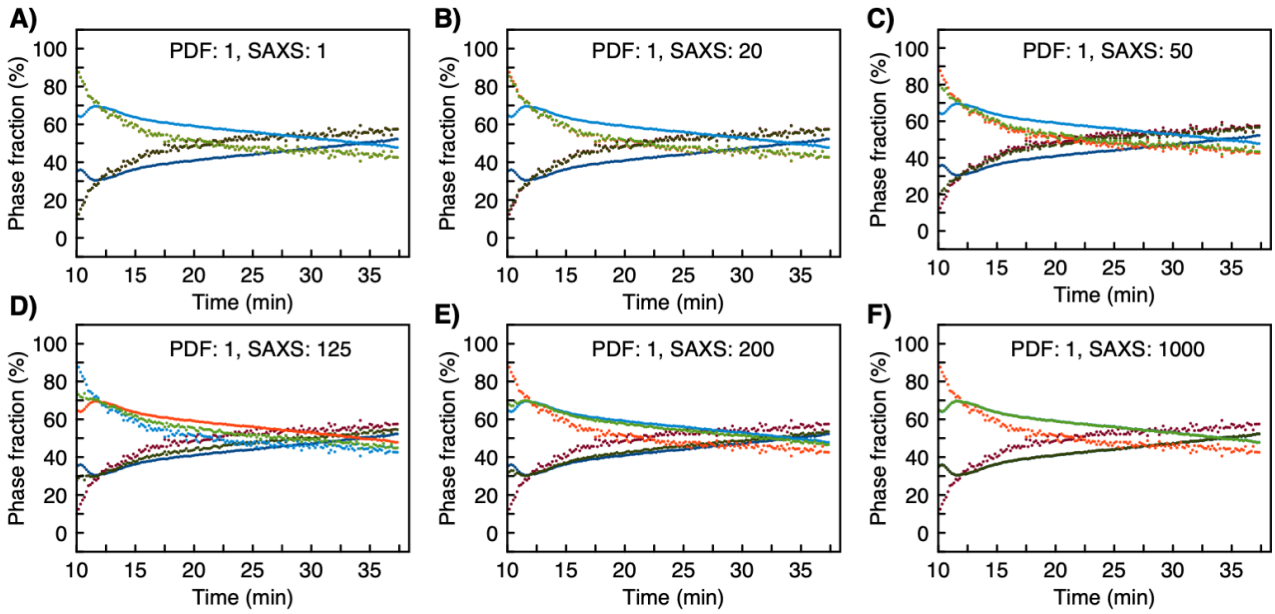



**Figure 5.13: Varying the weights of the PDF or SAXS data in the complex modelling affects the results significantly. Adapted from Anker et al.[33]**

Now, it has been established that the crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ dissolves in DMSO and forms the intermediate cluster species $[Bi_{22}O_{26}]$ faster than the time resolution for our experiments. However, it cannot be excluded that any intermediate species smaller than $[Bi_{22}O_{26}]$ exists. Afterwards, a slow growth from $[Bi_{22}O_{26}]$ to $[Bi_{38}O_{45}]$ is happening, which's kinetics is temperature dependent. The reaction mechanism is illustrated in Figure 5.14.

The chemical insight gained in this study can potentially be transferred to other metal oxido cluster systems or especially the ones of the fluorite type. The approach of using both PDF and SAXS analysis have shown great potential to study nanoclusters in solution and the modelling approaches presented here can both be used to automatically extract intermediates and to combine modelling of PDF and SAXS data using the Debye equation.

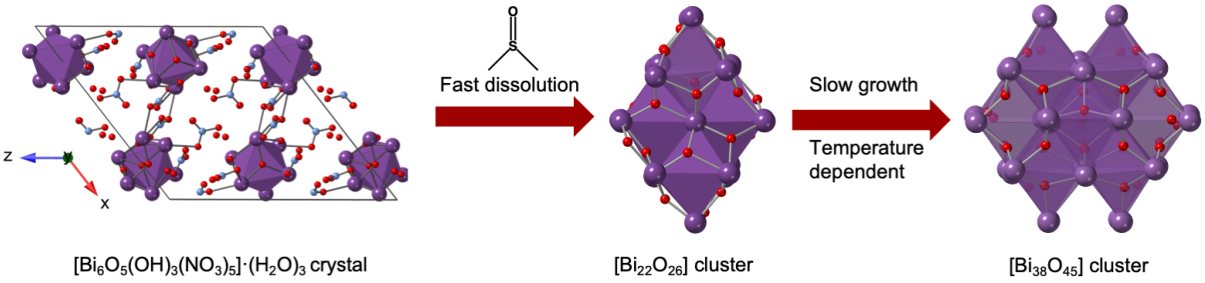



**Figure 5.14: The proposed reaction mechanism. The crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ dissolves in DMSO and goes through a disordered $[Bi_{22}O_{26}]$ cluster intermediate before forming atomically precise $[Bi_{38}O_{45}]$ clusters. It cannot be excluded that other intermediate clusters exist during the reaction; however, they could not be identified in our experiments. Adapted from Anker et al.[33]**

# 6 Using a Gradient Boosting Decision Tree Algorithm together with SHapley Additive exPlanations to Extract Structural Motifs from PDF and SAXS data

We have seen in Chapter 5 that automated modelling can be used to extract the intermediate species during a reaction by fitting a large number of candidate structures built from the $[Bi_{38}O_{45}]$ cluster. The dataset is fitted sequently in reverse order. Firstly, the $[Bi_{38}O_{45}]$ cluster is tested with all candidate structures where up to 8 Bi atoms from the outer shell are removed. The best-fitting candidate is then used as a starting point for the next frame. Here, 3 atoms are added to the outer shell whereafter the procedure is redone. This process is iteratively repeated for the entire dataset. However, the presented method is computationally expensive, and the candidate structures are limited to all structures with up to 8 Bi atoms removed limiting, whereas we in principle would like to test all candidate structures where up to 38 Bi atoms are removed. With this drastic constrain, the algorithm still takes a few days to run on a MacBook Pro 2017. For each frame, the algorithm produces $2^8$ fits, which each yields an $R_{wp}$ value (A description of the $R_{wp}$ value is given in Appendix III). For the automated motif extractor, it is only the model with the lowest $R_{wp}$ value, which is used for further analysis, thereby discarding the information of all other fits.

Here, we introduce a new approach using machine learning (ML) to evaluate results from automated modelling. **M**achine **L**earning based **Mot**if **Ex**tractor (ML-MotEx), which is illustrated in Figure 6.1, first builds a catalogue of hundreds or thousands of candidate structure motifs which are all 'cutouts' from a chosen starting structure, and then fit these individual models to a dataset. The structures and $R_{wp}$ values from each fit are handed to an ML algorithm applying gradient boosting decision trees (GBDTs), which learns to predict $R_{wp}$ values based on an atomic structure model. This allows expanding the structural space that can be evaluated compared to the brute-force methods. Most importantly, the ML-MotEx algorithm then outputs quantified values of how important each atom or feature in the starting structure is for the fit to yield a low $R_{wp}$ value with the given fitting algorithm. This is done by using SHAP (SHapley Additive exPlanations) values, as discussed in detail below.[38]

ML-MotEx is significantly faster and less biased than the brute-force intermediate extractor due to the ability to use the information from all the fits made. Furthermore, we extend the method to use a 'cookie-cutter' strategy to generate a catalogue of candidate structure motifs. We demonstrate that it is possible to obtain knowledge of dominating structural motifs from PDF data in an automated

manner using ML. We show a simple example on a single dataset where the [$Bi_{38}O_{45}$] cluster is cut out of the β-$Bi_2O_3$ structure with the use of ML-MotEx. However, in principle, the ML-MotEx approach can be applied to find intermediate species in an *in situ* dataset, allowing more robust results than the brute-force intermediate extractor presented in Chapter 5.2.

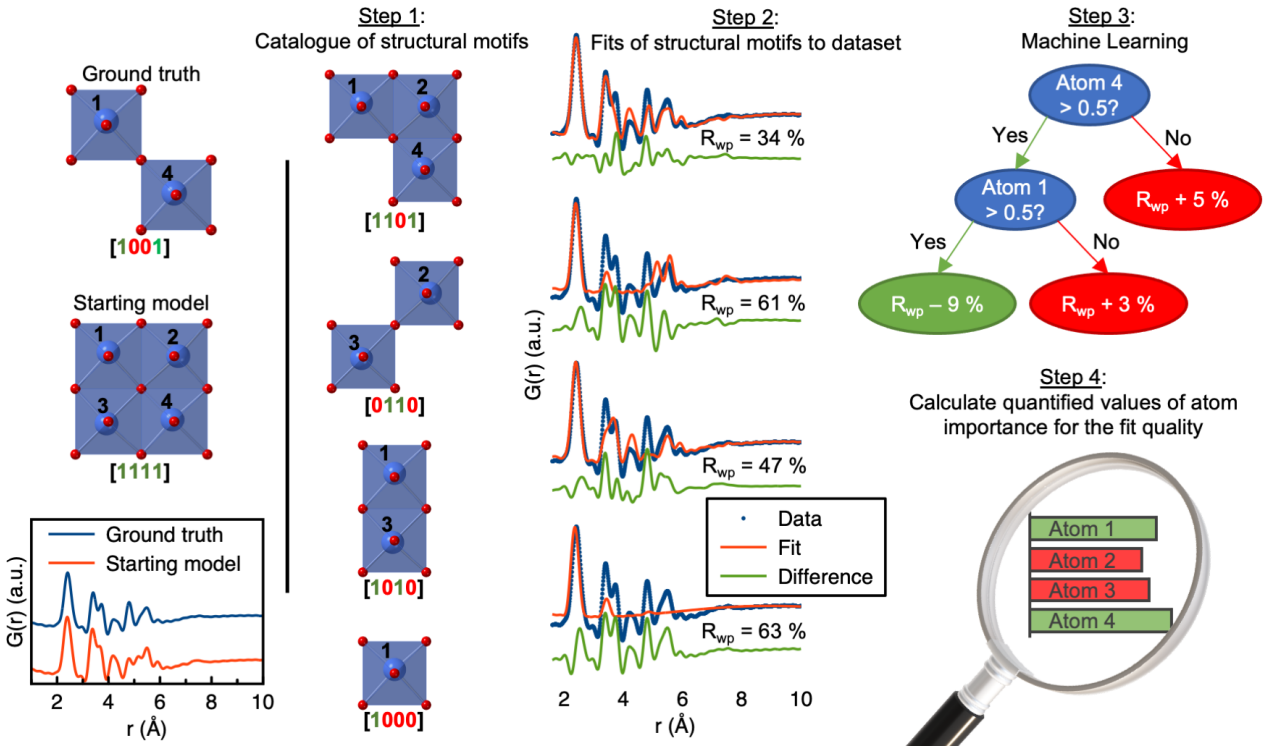



**Figure 6.1: Illustration of how ML-MotEx extracts the structural motif that fits the data from a starting model by generating a catalogue of candidate structure motifs, fitting them to the data, training a machine learning algorithm to predict the $R_{wp}$ value and finally calculating quantified values of atom importance for the fit quality.**

## 6.1 From input to output: ML-MotEx in 4 steps

*Overview of the algorithm*

ML-MotEx consists of four parts, which are all fully automatic. These four parts are shown in the simplified pseudo-code of the algorithm in Figure 6.2. In the first part, a starting model is used to generate a catalogue of candidate structure motifs. In the second part, a fitting script is used to fit the generated candidate structure motifs to the dataset. The fitting parameters can be seen in Appendix III. In the third part, the ML algorithm is optimized and trained, while SHAP values of the structural features in the models are calculated in the fourth part. The output of the algorithm is thus the starting model along with SHAP values indicating the importance of each individual atom in the structure, or

in other words; how much each individual atom affects the $R_{wp}$ value either positively or negatively. A further description of the individual steps of the algorithm is given in the next sections.

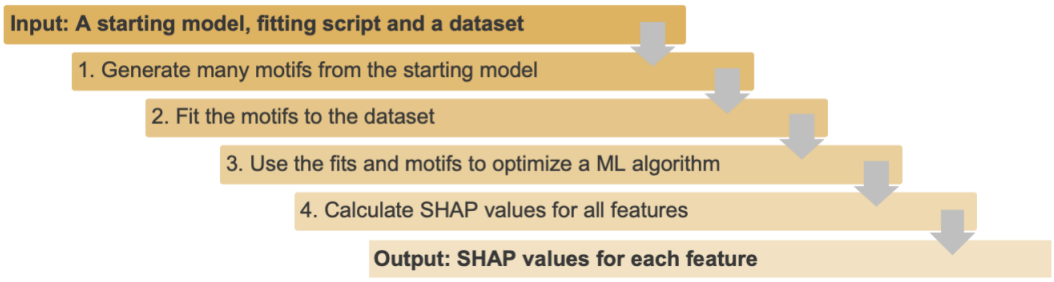



**Figure 6.2: Pseudo-code that describes the four steps of the ML-MotEx algorithm.**

*Part 1: Creation of a catalogue of candidate structure motifs*

The first step in ML-MotEx is to use a starting structure model to generate a catalogue of candidate structure motifs, which are all fitted to the data. 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 brute-force automated intermediate extractor. Rather than using permutations over different atoms in the structure, we 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 β-$Bi_2O_3$ structure. Four shapes are used; spherical, cubic, ellipsoidal and disk in appropriate sizes concerning the starting model. All the atoms in the starting model were kept or removed according to if they were inside the geometric object or not. Further details about the cookie-cutter algorithm can be found in Appendix III. One of the advantages of the cookie-cutter strategy is that we get structural parameters like shape, size, volume and sphericity, from the candidate structure motifs, which we fit the data. 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 an ellipsoids sphericity depends on its radii. The equation and a detailed description of sphericity are given in Appendix III. The cookie-cutter procedure is used to generate the catalogue of candidate structure motifs, while the corresponding fits are done with a fitting algorithm to yield a $R_{wp}$ value.

*Part 2: Fitting the catalogue of candidate structure motifs to the data*

In the next step, we fit each of the structures in the catalogue of candidate structure motifs to the experimental PDF. We here use the Python-based program Diffpy-CMI[48] for PDF fitting, but this could equally be done with other software packages such as DISCUS,[78, 79] TOPAS,[52] or any other

program that can be used to model X-ray or neutron scattering data. The output of the fit is a $R_{wp}$ value reflecting the quality of the fit.

*Part 3: Predicting $R_{wp}$ values using Gradient Boosting Decision Trees*

We use GBDTs to evaluate the results from the fits of the catalogue of the candidate structure motifs to the experimental PDF. The aim of this step is for the algorithm to learn how to predict $R_{wp}$ values, generated in part 2, based on the input which is the catalogue of candidate structure motifs generated in part 1.

GBDT is a tool that can do classification or regression using decision trees. In this work, we are using the GBDT tool to do the regression task of predicting the $R_{wp}$ value based on the structural input given as zeros or ones. The optimization is done by making trees of 'yes' and 'no' questions on whether to keep an atom in the structure or not (based on the resulting $R_{wp}$ value). An example of a simple tree can be seen in Figure 6.1. When atom 4 is present in the structure, the GBDT algorithm will predict a $R_{wp}$ value that is 5 % lower than if atom 4 is not present in the structure. In the same way, it will predict an $R_{wp}$ value that is 12 % lower if atom 1 is present in the structure. In reality, the GBDT algorithm will make 100 trees and do a weighting of them.

In this project, XGBoost[55] has been used as the GBDT algorithm with the use of default parameters except for learning rate and max depth, which were optimized with the use of Bayesian optimization using 50 iterations and cross-validation split on 3.[80] A lower iteration number and fewer cross-validation splits can be used to achieve faster optimization time, on the cost of performance. The process of training the GBDT algorithm is done on 80 % of the data. The last 20 % of the data is used to validate the performance of the algorithm. The GBDT algorithm performance is better with a large amount of training data, which in this tool is provided by creating a larger catalogue of candidate structure motifs and fitting them to the data.

*Part 4: Quantifying the effect of structural features with SHAP values*

SHAP values are used to analyze the $R_{wp}$ values resulting from the process described above. The amplitude of a SHAP value contains information about *how* important a structural feature is, while the sign of the SHAP value reflects whether the feature affects the $R_{wp}$ value of the fit towards 1 (poor fit) or 0 (perfect fit), in other words *why* it is important. In this work, the results can be visually inspected as the atoms in the starting model are colored according to their SHAP values using green for low SHAP values (tendency to keep atom, pushing $R_{wp}$ down) and red for high SHAP values (tendency to remove atom, pushing $R_{wp}$ up). Any 3D visualization tools for structural models, which can visualize XYZ files, can be used such as VESTA[81] or CrystalMaker.[82]

## 6.2 Extracting Structural Motifs from PDF data of the [$Bi_{38}O_{45}$] cluster

It is well known that [$Bi_{38}O_{45}$] (Figure 6.3A) is a stable cluster structure.[62, 83] If considering the cluster structure, it can be seen that this can be described as a cut-out of the β-$Bi_2O_3$ structure, Figure 6.3B.[62, 83] We can therefore apply the ML-MotEx method for analysis of a PDF from this cluster, using a β-$Bi_2O_3$ starting model. The starting model is thus a discrete model consisting of 2x2x3 unit cells of the β-$Bi_2O_3$ structure.

Figure 6.3C 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 illustrates 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.

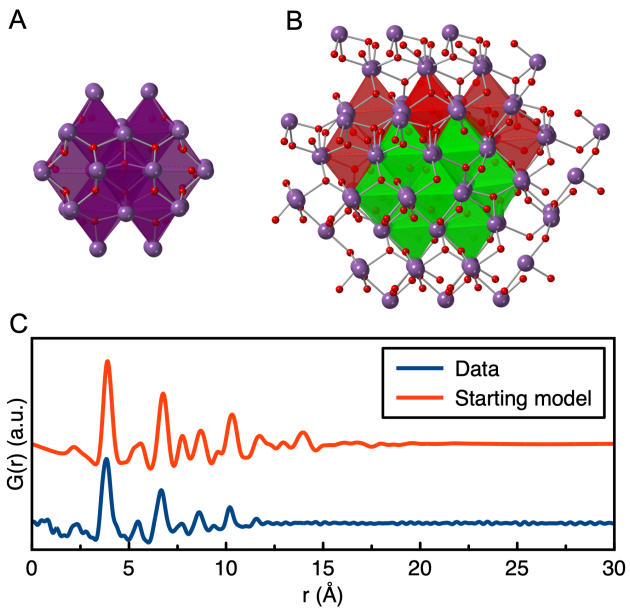



**Figure 6.3: A) The [$Bi_{38}O_{45}$] cluster B) The β-$Bi_2O_3$ structure with highlighted polyhedra in red and green, where the green corresponds to the [$Bi_{38}O_{45}$] cluster. Some polyhedra are drawn as balls and stick instead, to enhance visibility. C) Comparison of the experimental PDF of the [$Bi_{38}O_{45}$] cluster and simulated data of a 2x2x3 unit cell of β-$Bi_2O_3$. The simulation parameters mimic typical values of a PDF dataset and can be seen in Appendix III.**

After constructing and fitting these models to the PDF, we can analyze the results from the brute-force approach before moving to ML. Figure 6.4A shows all the $R_{wp}$ values obtained from the PDF fits, where the fit quality is plotted versus the number of bismuth atoms. The refinement parameters

are given in Appendix III. It is evident that the $R_{wp}$ values are lowest when a structure of 38 Bi atoms is used to fit the data, however, the curve is not very shallow. In Figure 6.4B, the $R_{wp}$ values are plotted for the 4 different shape categories used in the 'cookie-cutter' strategy. The color of the point indicates the sphericity. The plot shows that cube-shaped structures generally result in 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 6.4C, the results of fitting spherical models to the data is further analyzed. 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 Appendix III.

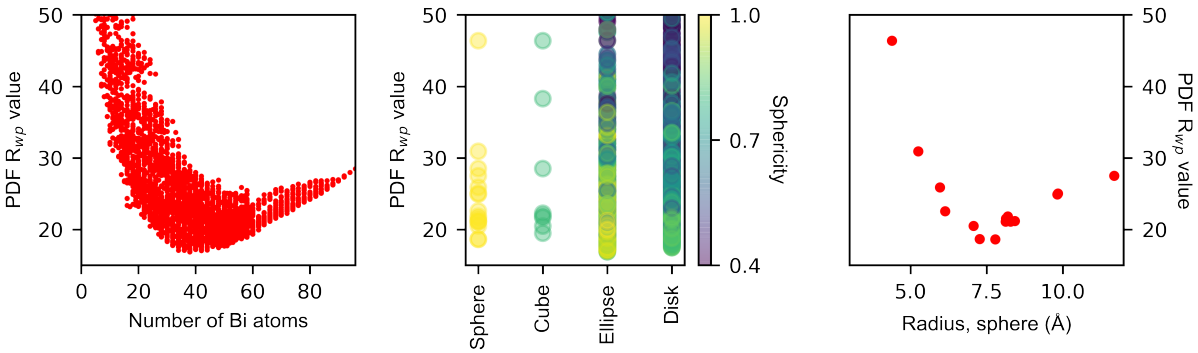



**Figure 6.4: Overview of the $R_{wp}$ values of the PDF fits versus important structural parameters. A) The $R_{wp}$ values versus the number of bismuth atoms present in the structure, B) The $R_{wp}$ value versus the shape and sphericity of the structures and C) The $R_{wp}$ value versus the spherical radius of the structures made with a spherical shape.**

We now again move to ML-analysis for further atomistic understanding. Again, this allows taking all 5950 fits into account when finding an appropriate structural motif that fits the data. Figure 6.5 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 corresponding $R_{wp}$ values.

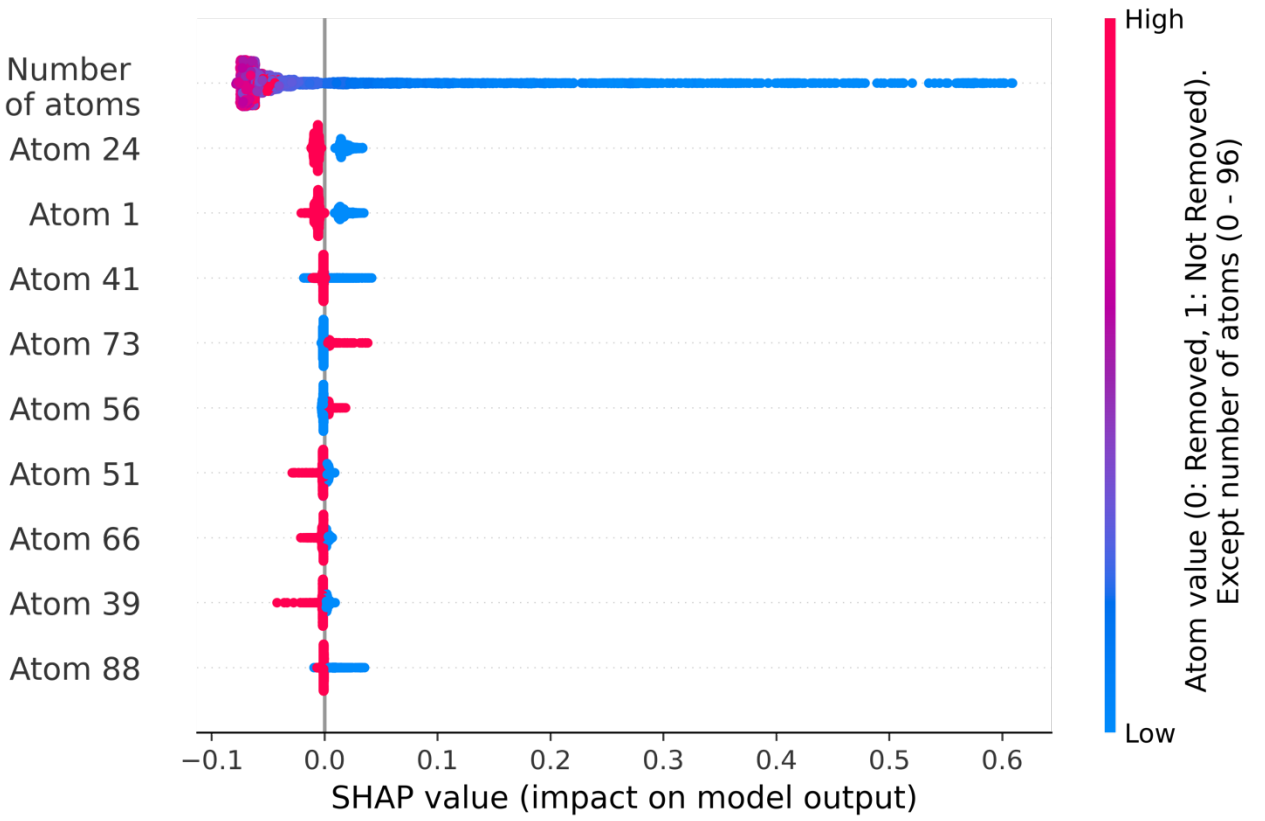



**Figure 6.5: SHAP summary plot from ML-MotEx utilized on PDF data from the [$Bi_{38}O_{45}$] cluster.**

The SHAP value plot shows that the model does not prefer a low number (blue color) atoms in the model, however, when more than approx. 20 atoms is reached, it does not make a large difference in how many atoms are included in the model. The plot also tells the effect of each atom's contribution to the prediction of the $R_{wp}$ value of the model. For example, atom 24 has negative SHAP values when it is included in the candidate structure motif. We interpret that atom 24 contributes to a low $R_{wp}$ value. We can thereby construct the structure of the 38 Bi atoms with the lowest SHAP value when they are included in the model. This model is illustrated in Figure 6.6B.

In Figure 6.6, the result from the brute-force approach (Figure 6.6A) is compared to the result of the ML-MotEx approach (Figure 6.6B). The best structure found with the brute-force approach is shown in Figure 6.6A where the green octahedra represent parts of the [$Bi_{38}O_{45}$] structure. This structure can be fitted to the PDF data with an $R_{wp}$ value of 16.6 % as seen in Figure 6.6C. Figure 6.6B 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 % (Figure 6.6D), however it contains

more of the original [$Bi_{38}O_{45}$] structure than the brute-force model does. The fitting parameters and fits of the [$Bi_{38}O_{45}$] are shown in Appendix III.

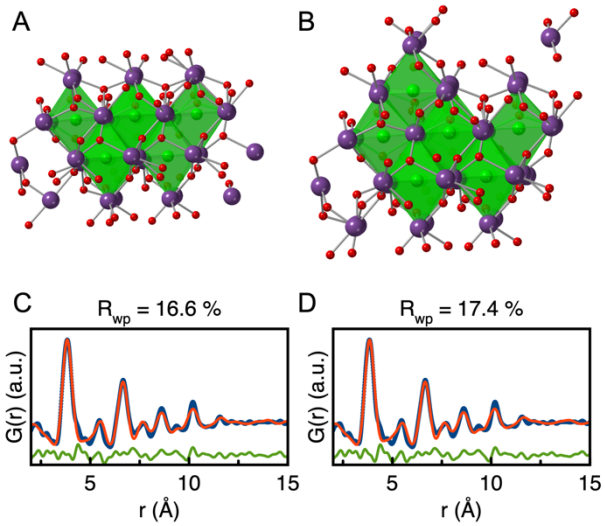



**Figure 6.6: Structural visualization of the kept atoms in the 2x2x3 unit cell of β-$Bi_2O_3$ with highlighted polyhedra from the [$Bi_{38}O_{45}$] cluster using A) The Brute-force approach. B) ML-MotEx approach. The fits to the dataset of the [$Bi_{38}O_{45}$] cluster using C) The cluster illustrated in A. D) The cluster illustrated in B.**

In conclusion, the brute-force routine itself of making a catalogue of candidate structure motifs and investigating the $R_{wp}$ values reveals valuable insight into the structure. However, it is still a cumbersome process to use all this information to build a structure that fits the data. One can simply use the best fit, however, it is a biased approach, and it does not necessarily represent the majority of the structures with the best fits. By applying ML-MotEx a reasonable structural estimate is obtained, which can be used with chemical knowledge to construct the right model that fits the data.

We also tested the ML-MotEx approach on SAXS data. However, as shown in Appendix III, the missing atomistic information means that the SAXS results are less meaningful, and PDF is thus better suited for identifying which atoms should be included in the model using the ML-MotEx approach than SAXS data is.

# 7 Using a Gradient Boosting Decision Tree Algorithm to screen large Databases of matching Polyoxometalate Clusters to a PDF

Characterizing the atomic model in PDF data via conventional data analysis methods was in Chapter 5 demonstrated to be a highly complicated and time-consuming process.

Here, We present POMFinder, an algorithm that can automatically identify polyoxometalate (POM) clusters from PDF data. The algorithm has been trained on simulated PDF data of 443 POM cluster fragments, which are cut out of metal oxide crystals from the Crystallography Open Database (COD) and the Inorganic Crystal Structure Database (ICSD). The algorithm predicts the right structure 94.0 % of the time on simulated data, and it provides us with promising results on experimental PDF data. We furthermore show that the method can be extended to jointly use data from multiple scattering techniques instead of modelling the data separately: the 'complex modelling' approach. We use PDF and SAXS data. We demonstrate that the algorithm can instantly predict the structure of free $[H_2W_{12}O_{40}]^{6-}$ ions with the $\alpha$-Keggin structure.[32] POMFinder enables automated PDF and SAXS modelling, which can be used directly while measuring the data without the need for any prior knowledge.

## 7.1 From input to output: POMFinder in 3 steps

*Overview of how POMFinder was created*

The goal of this project is to create a ML algorithm that can match an atomic structure to an experimental dataset. We have chosen to focus on PDF and SAXS data as the structural characterization techniques, however, in principle the dataset can be any information that can be used to characterize an atomic structure. We used a self-created POM database but this could be any database of any structure type.

A pseudo-code of how to create POMFinder can be seen in Table 7.1. The algorithm consists of 3 parts, which are building a POM database, simulating data from the database and training a GBDT algorithm on the simulated data. The structural database of POM clusters is built from CIF files from both COD and ICSD using chemical restraints appropriate for POM clusters. Afterwards, a number of datasets of each POM cluster are simulated using Latin hypercube sampling[84] and a GBDT

algorithm[55] is trained to classify which POM structure the dataset matches. In the following sections, we will elaborate on how POMFinder is created.

| **How to create POMFinder** |
|---|
| **Input** : Structural Database (COD + ICSD) & chemical restraints (if any) |
| 1. Build a database of polyoxometalate clusters |
| a. Remove crystals which does not fulfill the set chemical restraints |
| b. Cut out clusters from crystal structures in the database |
| c. Remove clusters which does not fulfill the set chemical restraints |
| d. Remove all equivalent clusters based on a similarity score |
| 2. Simulate N datasets per structure (N is normally set to 100) (dataset can be any structural information as PDF or SAXS etc.) |
| 3. Train a GBDT algorithm to classify which structure a dataset match to |
| **Output** : A tool that can output a cluster-structure based on a dataset |

**Table 7.1: Pseudo-code that describes how to create the POMFinder algorithm.**

*Part 1: Building a database of polyoxometalate clusters*

The COD and ICSD databases contain a total of 412,738 and 166,831 crystallographic information files (CIFs). First, the database is screened for CIFs that fulfill the desired chemical restraints. We chose to restrain the structures to have the same number of metal and oxygen atoms as the reported structures in a review about POM clusters in solution by N. I. Gumerova and A. Rompel.[85] This restrained the database to 56 different metal – oxygen compositions, yielding 1281 CIFs in total. A full list of metal – oxygen compositions used for this project is shown in Appendix IV. Clusters were cut out of the remaining CIFs by creating a 2x2x2 unit cell of the crystal and extracting all clusters of atoms that were not bonded to other atoms in the structure. A 2x2x2 unit cell is needed to capture all clusters, which are split upon several unit cells. All clusters that did not fulfill the chemical restraints (the 56 different metal – oxygen compositions) were removed. Figure 7.1 illustrates an example of a cluster that was cut out of a porous framework based on Keggin polyoxoanions, $K_2NaH_2[BW_{12}O_{40}]\cdot 12H_2O$.[86]

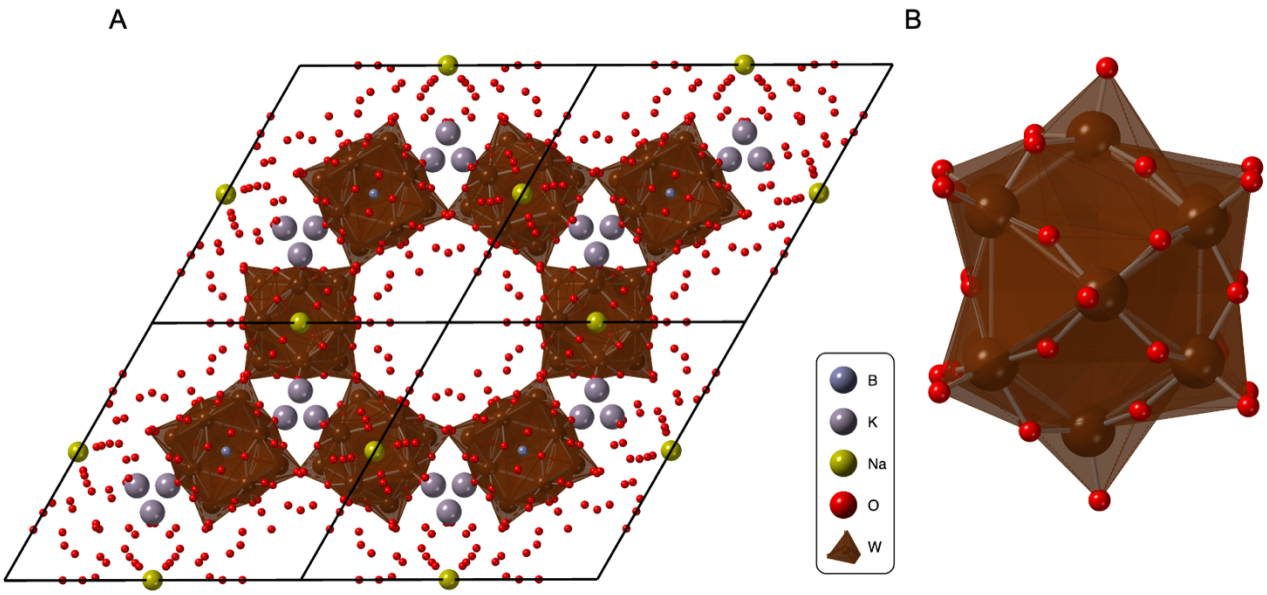



**Figure 7.1: An example of a cluster fragment from a crystal structure. A) The crystal structure of a porous framework based on Keggin polyoxoanions, $K_2NaH_2[BW_{12}O_{40}]\cdot 12H_2O$,[86] and the corresponding B) cluster.**

This procedure yielded 969 structures, that potentially are polyoxometalate clusters. Be aware that it is not guaranteed that the clusters are perfectly cut out of the structure with this routine, which makes it important to inspect the results of POMFinder from a chemical point of view. In order to remove similar structures, all structures were compared using the Pearson value correlation of simulated PDFs of the structures. This process is done iteratively by simulating a PDF of the 1st and 2nd cluster with the parameters given in Appendix IV and comparing their absolute Pearson value correlation. The Pearson value correlation is a measure from -1 to 1 of how correlated two datasets are, where -1 represents inverse datasets and 1 represents identical datasets. If the absolute Pearson value correlation is higher than 0.99, the 2nd cluster is not included in the database. The 3rd cluster is compared to the 1st and 2nd cluster by the same procedure, etc. The value of 0.99 was defined by manual inspecting structures, their corresponding PDFs and the Pearson value correlation. Examples of 3 structures, their corresponding simulated PDFs and the PDFs Pearson value correlation can be seen in Appendix IV. This process was done with all the 926 structures yielding 443 unique structures.

*Part 2: Simulating data from the POM structures*

For each structure, a number of PDF or SAXS data are simulated, N, with a broad range of instrumental parameters using Latin hypercube sampling[84] to determine the parameters. Diffpy-CMI was used to simulate the datasets.[48] The range of simulation parameters of the PDF and SAXS data is given in Appendix IV. The PDFs are normalized to have the largest peak at G(r) = 1, and G(r ≤ 1 Å) = 0 since it is very unlikely to have atomic distances that contribute to the signal in this range of the PDF. The SAXS data are normalized to have the largest intensity at I(Q) = 1. Figure 7.2 illustrates the *ex situ* data of 0.05 M ammonium metatungstate hydrate, $(NH_4)_6[H_2W_{12}O_{40}]\cdot H_2O$, solution before and after normalization.

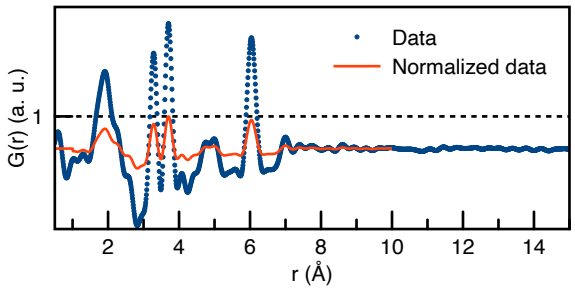



**Figure 7.2: Comparison of the experimental dataset of a 0.05 M ammonium metatungstate hydrate, $(NH_4)_6[H_2W_{12}O_{40}]\cdot H_2O$, solution before and after normalization.**

*Part 3: Training process of POMFinder*

The simulated datasets and their corresponding instrumental parameters are input in a GBDT algorithm used to predict the structure. The GBDT algorithm used is XGBoost with default parameters except for the learning rate which was set to 0.3 and the early stop criterion of 5 rounds without improvement.[55] The problem is thereby a 443 class classification problem with 443×N simulated datasets. For each structure, 2 random of the N simulated PDFs are set aside during training of the model and later used as validation- and test set. The validation set is used to validate when the GBDT algorithm has converged.

The loss curve (multiclass log loss), for the case where N = 100, is plotted in Appendix IV, which shows that the algorithm can predict the training data with 100 % accuracy, while the validation data is predicted with a small error. The concluding accuracy of the algorithm can be determined on the test set, which is data that the model has not been trained or validated on i.e. comparable to how POMFinder can be used for experimental data. When POMFinder is trained on 100 PDFs for each structure, the accuracy on the test set is 94.0 %. However, this is based on simulated PDF data, whereas the interest is to predict POMs from experimental data, which is more challenging since

experimental data can contain noise, background scattering and scattering from other chemical species.

## 7.2 Description of how to use POMFinder

When POMFinder has been created, it is a very simple tool to use where everything is fully automated with no requirement of any knowledge about PDF or ML. As seen in Table 7.2, the scientist simply provides a dataset as input to POMFinder and it will return a structure as output. The dataset can currently be a X-ray PDF, neutron PDF, electron PDF or a SAXS dataset. The output will be given in XYZ format providing the elements and coordinates of all the atoms in the structure.

| **How to use POMFinder** |
|---|
| **Input** : A dataset |
| 1. Give the dataset as input to the algorithm |
| **Output** : A structure |

**Table 7.2: Pseudo-code that describes how to use the POMFinder algorithm.**

## 7.3 Characterizing the POM structure from experimental X-ray PDF data by screening the COD & ICSD database in a second

We demonstrate the power of POMFinder on an experimental *ex situ* dataset of a 0.05 M ammonium metatungstate hydrate, $(NH_4)_6[H_2W_{12}O_{40}]\cdot H_2O$, solution which is known to yield free $[H_2W_{12}O_{40}]^{6-}$ ions with the $\alpha$-Keggin structure.[32] Keggin structures, an example is seen in Figure 7.1B, have the chemical composition $[XM_{12}O_{40}]^{n-}$, where X is a tetrahedrally coordinated cationic central atom in the middle of the cluster, M is the metal atom of the cluster and n is the negative charge of the cluster. Keggin clusters are divided into five rotational isomers, with increasing degree of edge-sharing: $\alpha, \beta, \gamma, \delta$ and $\varepsilon$.

The input to POMFinder is the normalized PDF. The output is how probable it is that the normalized PDF originates from each of the 443 POM structures in our POM database, with a total sum of 100 %. We decided to investigate the five structures with the highest probability assigned by POMFinder, which are most likely to be the right POM structure based on the POMFinder predictions. Figure 7.3 shows the five best predictions, together with a comparison of the normalized PDF of the

experimental dataset and the simulated PDF of the POM structure. The calculated PDF of the first three structures is very comparable to the normalized PDF, whereas the last two structures have scattering signals up to high r distances, which is not seen in the data. The last two structures are therefore concluded to be too large. The three best candidate structures are all fragments from CIF files containing the Keggin structure. Some of them are not perfectly cut out though. The first two structures represent the $\alpha$-Keggin structure, while the third represents the $\beta$-Keggin structure. The best candidate structure, Figure 7.3A, is a lacunary Keggin with three out of four triads of the $\alpha$-Keggin structure. This could indicate that the last triad in the structure is dynamic. This is supported by the 5$^{th}$ best prediction, which is a fragment built of two lacunary Keggins. The best candidate also has a Si atom in the center, which impacts the peak intensity at r = 3.3 Å and r = 3.7 Å. From this point, one needs to use chemical intuition to build a proper model to describe the data. In this synthesis, hydrogen is the only cation in the reaction, which could be stabilized in the center of the cluster. Additionally, the best candidate structures have an overweight of corner-sharing octahedra. Therefore, we conclude that the right model for the experimental dataset is $[H_2W_{12}O_{40}]^{6-}$ with the $\alpha$-Keggin structure.

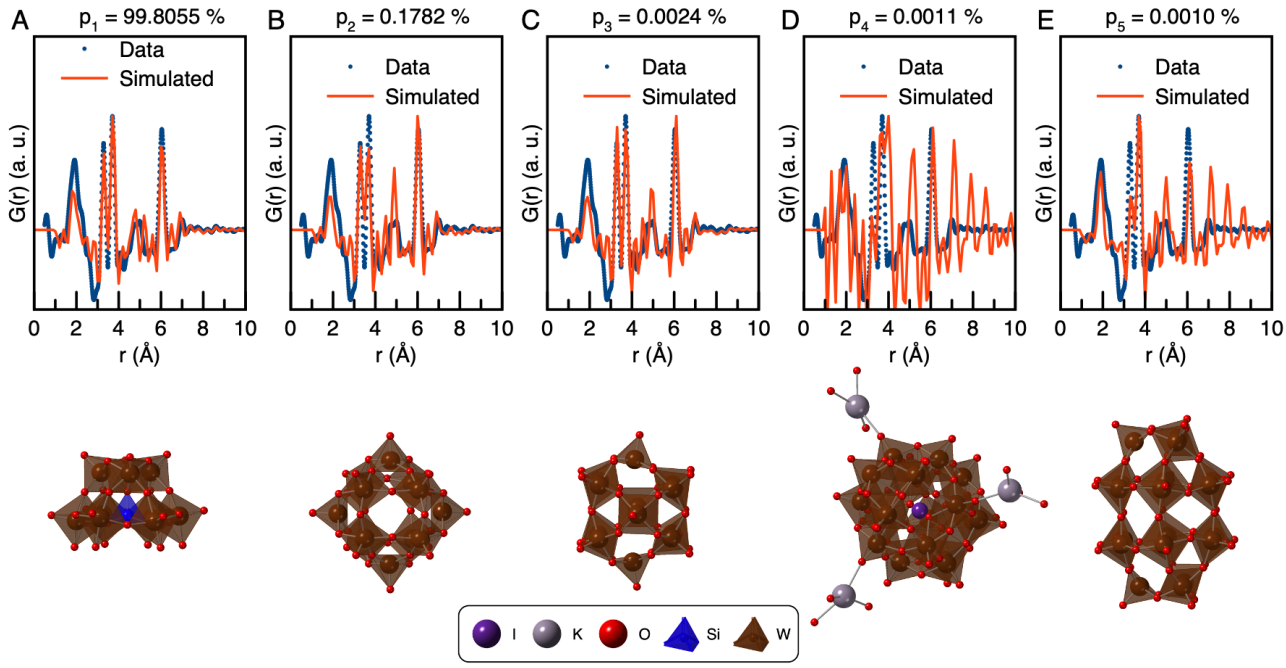



**Figure 7.3: Comparison of the normalized PDF of the 0.05 M ammonium metatungstate solution and the calculated PDF of A) a $W_9SiO_{34}$ Keggin-based fragment from the $Na_2[C(NH_2)_3]_2[((CH_3)_2Sn(H_2O))_3(A\text{-}\alpha\text{-}SiW_9O_{34})]\cdot 10H_2O$ crystal[87], B) a $W_{12}O_{36}$ fragment from the $K_5(CoW_{12}O_{40})\cdot 20H_2O$ crystal[88], C) a $W_{12}O_{36}$ fragment from a crystal structure[86] of a porous framework based on Keggin polyoxoanions, $K_2NaH_2[BW_{12}O_{40}]\cdot 12H_2O$, D) a $W_{18}I_3K_3O_{75}$ fragment from the first crystallographically characterized[89] tungstatoperiodate cluster of the form $\beta^*\text{-}[H_3W_{18}O_{56}(IO_6)]^{6-}$, and E) a $W_{18}P_2O_{89}$ fragment from the Dawson type POM[90] $Li_6[\alpha\text{-}P_2W_{18}O_{62}]\cdot 28H_2O$.**

POMFinder thereby provides a tool that can predict the metal oxido cluster from a PDF while measuring the data, making it an ideal approach for structure characterization during a cluster growth reaction or in the early stage of a nucleation process.

## 7.4 We investigate how POMFinder works by using SHAP values to interpret the features

We have now established that POMFinder is a tool that can predict the POM structure of experimental PDFs of simple solutions. Yet, we do not fully understand how POMFinder makes predictions. In order to get this understanding, we use SHAP analysis. SHAP is a measure of feature importance, which yields information about how the ML algorithm exploits the individual features ($Q_{min}$, $Q_{max}$,

$Q_{damp}$ and $r_{range}$ = 0 – 10 Å) to make its prediction. The amplitude of the SHAP value can be interpreted as *how* important the feature is or *how* large a difference it makes to the predictions. The sign of the SHAP value tells if the feature is confirming or disqualifying the specific structure as a match to the dataset, or *why* the feature is important. In this project we only analyse the amplitude of the SHAP value. Figure 7.4A shows a SHAP analysis of the simulated PDF of the $\alpha$-Keggin clusters. The two most important features are the G(r) value at r = 6.0 Å and r = 3.6 Å. Figure 7.4B shows a comparison of the SHAP values with the simulated PDF. The highest SHAP values, thereby the most important features, are at G(r) values corresponding to the 1$^{st}$ and 2$^{nd}$ metal – metal shell distances. Therefore, we conclude that the ML algorithm predominantly bases its predictions on the intensities of the PDF peaks describing the 1$^{st}$ and 2$^{nd}$ metal – metal shell which is the dominant peaks in the PDF because W scatters significantly more than O. The model thereby bases its predictions according to the scattering theory outlined in Chapter 2.

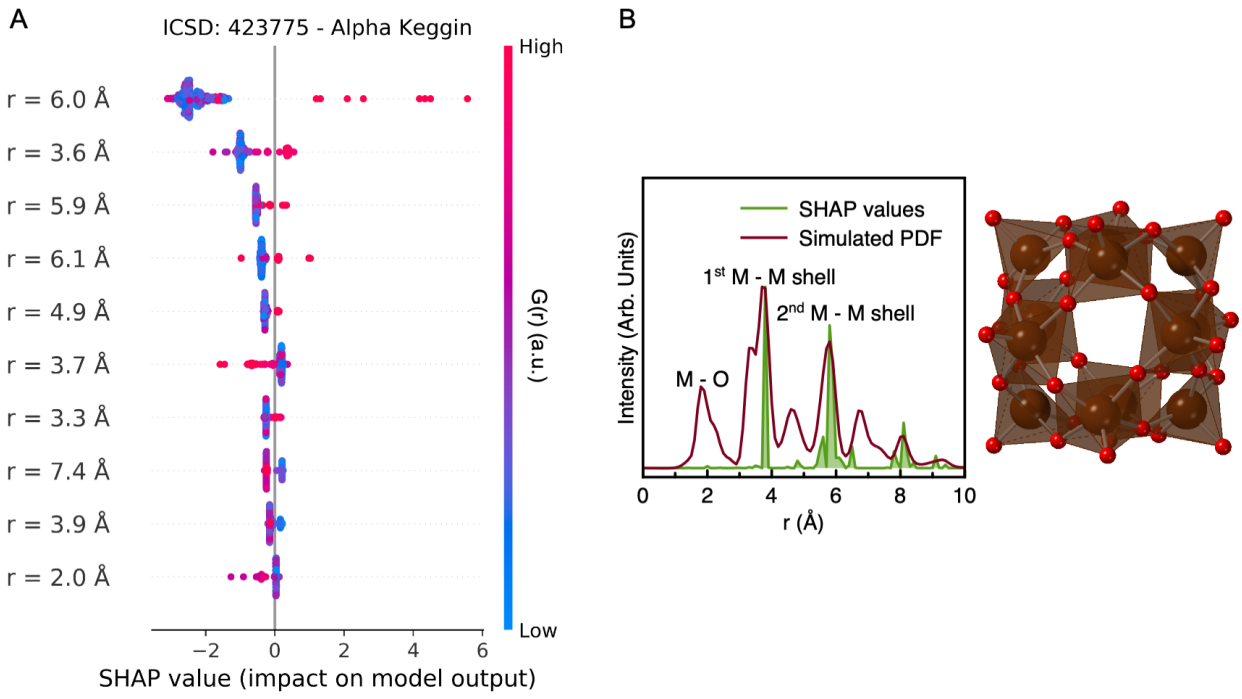



**Figure 7.4: A) SHAP summary plot of the A) $\alpha$-Keggin cluster. B) Comparison of the SHAP values and the simulated PDF for the $\alpha$-Keggin.**

## 7.5 Demonstrating that the POMFinder framework can easily be upscaled to include data from other techniques

We have now established that POMFinder does base its prediction on scattering theory learned from 98 X-ray PDFs on each POM structure in our database, all with different $Q_{min}$, $Q_{max}$ and $Q_{damp}$ parameters. It is interesting to investigate how many PDFs are needed for it to learn the scattering theory and make accurate predictions. Figure 7.5 shows POMFinder's performance on the test set when POMFinder is trained on 2, 3, 5, 8 and 98 PDFs per POM structure. The trend is clear: The performance of POMFinder increases when it is trained on more data.

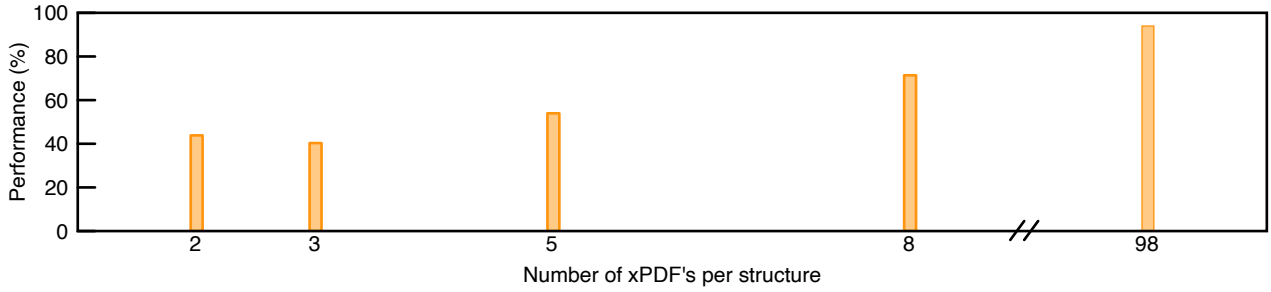



**Figure 7.5: The performance of the model trained with various amounts of X-ray PDFs per structure. Appendix IV shows the mean and standard deviation based on five iterations where the model has been trained on different simulated PDFs and predictions have been done on the same test set.**

It has previously been shown that combined modelling of data from multiple scattering techniques can provide more robust results than the modelling of the data from individual scattering techniques.[33, 48, 51, 73, 74, 77] However, it is a cumbersome process to do combined modelling of data from multiple scattering techniques and it can be difficult to weigh the contribution from each dataset.[33, 48, 77] ML provides a unique framework, which weights the datasets automatically. Furthermore, it is extremely easy to transform an ML algorithm from handling one dataset to another. We demonstrate the flexibility of POMFinder by replacing the X-ray PDF (xPDF) data with neutron PDF (nPDF), electron PDF (ePDF), X-ray SAXS (SAXS), a combination of xPDF/SAXS and a combination of xPDF/SAXS/nPDF. The simulation parameters for the SAXS data are given in Appendix IV. The simulation parameters for the nPDF and ePDF data are the same as for the xPDF data given in Appendix IV.

Figure 7.6 shows the performance of POMFinder when using various amounts of data as well as changing the scattering technique the data is from. Generally, POMFinder has a comparable

performance when trained on xPDF, nPDF and ePDF data. The performance of POMFinder based on PDF or SAXS data cannot be compared, since it will be related to the range of noise included in the simulated datasets. With the given parameters in this study, Appendix IV, POMFinder is comparable in performance when it is trained on PDF and SAXS data. However, we can conclude that when combining information from both PDF and SAXS experiments, the performance increases. This example shows the flexibility of the POMFinder and demonstrates the power of combined modelling.

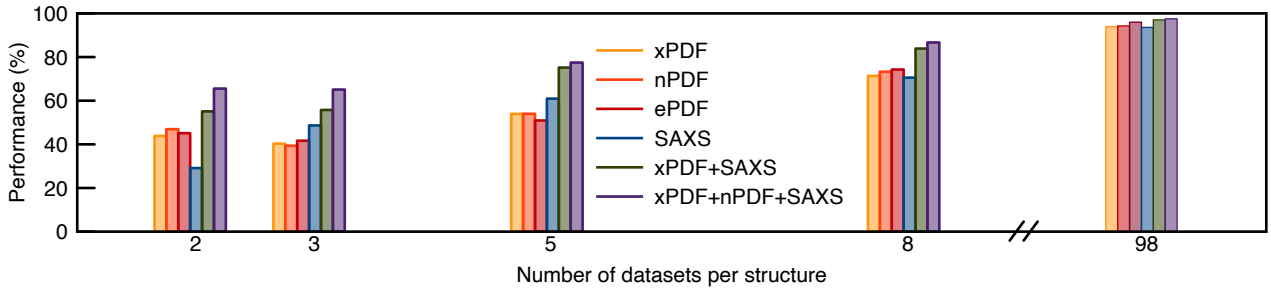



**Figure 7.6: The performance of the model trained with various simulated datasets and different amounts of datasets per structure. Appendix IV shows the mean and standard deviation based on five iterations where the model has been trained on different simulated PDFs and predictions have been done on the same test set.**

# 8 Concluding remarks

I have in this Master Thesis described how TS with PDF analysis and SAXS data are powerful techniques to characterise the atomic structure of nanoclusters. The measurements can be done in seconds allowing studies of cluster growth and nucleation mechanisms, which are important to understand in order to make new and functional materials. We have demonstrated that by using conventional PDF and SAXS analysis using minimization approaches as least-square refinements, it is possible to follow the cluster growth from $[Bi_6O_8]$ to $[Bi_{38}O_{45}]$ when $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ crystals are dissolved in DMSO.

However, this modelling approach is restricted to expert users, and it is a time-consuming way to model the data. We have demonstrated that automated modelling using least-squares minimization approaches can extend the information that is normally gained from conventional modelling. In this Thesis, it was used to identify the intermediate species as the $[Bi_{22}O_{26}]$ cluster. Furthermore, I have demonstrated that combined PDF and SAXS modelling provides more robust results than analysis of the individual techniques. However, these methods are highly dependent on expert users given the right parameters to the algorithms and combined modelling of PDF and SAXS data must be weighted appropriately.

We have also shown that both problems can be solved by replacing the least-squares minimization approach with ML. ML provides a unique framework, which can predict structural models instantly and it weights datasets from multiple techniques automatically. We have shown this by training an algorithm to predict the POM structure from a PDF in 94 % of the cases in the test set, while it can predict the cluster structure in 97 % of the cases if it is trained on both PDF and SAXS data. We have furthermore presented **M**achine **L**earning based **Mot**if **Ex**tractor (ML-MotEx) that can identify the important structural motif in a PDF pattern. These methods are some of the first to automate the analysis of scattering data using ML. The ML methods used here are tree-based algorithms, which are intuitive and easy to use. However, I imagine that more complicated neural networks as generative models will outperform the tree-based methods in the future. Generative models can both predict the chemical structure that matches a dataset but they are also able to map a space of chemical structures including properties and synthesis. ML using generative models can consequently be the tool that combines synthesis, structure and properties in a single 2D/3D map, which can freely be drawn from to do synthesis by design.

# Acknowledgements

I would first like to thank my supervisor Assoc. Prof. Dr. Kirsten M. Ø. Jensen, who has motivated me, guided me, and corrected every sentence I have written and every word I have said throughout my master period and two years afore. I believe that the care of a supervisor can be measured by red marks in a manuscript. I deeply appreciate all the time you have put aside to shape me with corrections and a large number of scientific discussions. I also appreciate the large scientific freedom you have given me both regarding choosing my research topics but also regarding summer schools and conferences you have taken me to. I could not wish for a better supervisor!

I also consider myself very lucky for the great collaborators and mentors I had through my Master Thesis. I would especially like to thank Dr. Martin Schmiele and Dr. Erik Brok for assisting me and teaching me in everything related to Small-Angle X-ray Scattering in this Thesis and Asst. Prof. Dr. Raghavendra Selvan for guiding me in the world of big data science. I would like to thank Prof. Dr. Michael Mehring from Technische Universität Chemnitz and his two colleges Dr. Marcus Weber and Thomas Rico, who have synthesized the Bi crystals used to conduct some of the experiments presented in this Thesis, and who have participated in fruitful discussions about Bi chemistry.

This project could not have been done without the financial support from Danscatt, the European Research Council and the Siemens Foundation. Thank you.

I would like to thank beamline P02.1 and P07 at Petra III, the beamline SWING at Soleil, the beamline ID02 at European Synchrotron Radiation Facility and the Advanced Photon Source for conducting experience at their instruments.

I would like to thank Dr. Jette, Mikkel, Emil and Olivia for proofreading this Thesis and for being the prime reason for loving my work together with the rest of the Nanostructure Group. You are all fantastic colleagues and great friends! I also consider myself very lucky that I had Dr. Troels Lindahl Christiansen as a co-supervisor. You have your own style of supervision, which fits well with me, and I had a great time with you in the group both as a colleague and a friend.

I would like to thank my family, friends and my outstanding girlfriend, Klara, for always being supportive and proud of me. A special thanks should be given to the *Concert group* for being a great study group in courses during the past years of my studies. Based on the number of concerts we have seen and the amount of cake we ate, we should in hindsight have been called the *cake group*. My Master Thesis would not be complete without thanking Ehlers bakery, who has supported me with endless studenterbrød or ‘usuals’.

In the end, I would like to thank my censor, Assoc. Prof. Dr. Ivano E. Castelli, for taking his time to read and grade my Thesis.

# References


1. Walther, M.; Zahn, D., Molecular Mechanisms of $[Bi_6O_4(OH)_4](NO_3)_6$ Precursor Activation, Agglomeration, and Ripening towards Bismuth Oxide Nuclei. *Eur. J. Inorg. Chem.* **2015,** *2015* (7), 1178-1181.
2. Mehring, M., From molecules to bismuth oxide-based materials: Potential homo- and heterometallic precursors and model compounds. *Coord. Chem. Rev.* **2007,** *251* (7), 974-1006.
3. Schlesinger, M.; Schulze, S.; Hietschold, M.; Mehring, M., Metastable β-$Bi_2O_3$ nanoparticles with high photocatalytic activity from polynuclear bismuth oxido clusters. *Dalton Trans.* **2013,** *42* (4), 1047-1056.
4. Miersch, L.; Rüffer, T.; Mehring, M., Organic-inorganic hybrid materials starting from the novel nanoscaled bismuth oxido methacrylate cluster $[Bi_{38}O_{45}(OMc)_{24}(DMSO)_9]\cdot 2DMSO\cdot 7H_2O$. *Chem. Commun.* **2011,** *47* (22), 6353-6355.
5. Mansfeld, D.; Mehring, M.; Schürmann, M., From a Monomeric Bismuth Silanolate to a Molecular Bismuth Oxo Cluster: $[Bi_{22}O_{26}(OSiMe_2tBu)_{14}]$. *Angew. Chem. Int. Ed.* **2005,** *44* (2), 245-249.
6. Pathak, A.; Blair, V. L.; Ferrero, R. L.; Mehring, M.; Andrews, P. C., Bismuth(iii) benzohydroxamates: powerful anti-bacterial activity against Helicobacter pylori and hydrolysis to a unique $Bi_{34}$ oxido-cluster $[Bi_{34}O_{22}(BHA)_{22}(H\text{-}BHA)_{14}(DMSO)_6]$. *Chem. Commun.* **2014,** *50* (96), 15232-15234.
7. Kou, X.; Wang, X.; Mendoza-Espinosa, D.; Zakharov, L. N.; Rheingold, A. L.; Watson, W. H.; Brien, K. A.; Jayarathna, L. K.; Hanna, T. A., Bismuth Aryloxides. *Inorg. Chem.* **2009,** *48* (23), 11002-11016.
8. Matthews, P. D.; King, T. C.; Wright, D. S., Structure, photochemistry and applications of metal-doped polyoxotitanium alkoxide cages. *Chem. Commun.* **2014,** *50* (85), 12815-12823.
9. Coppens, P.; Chen, Y.; Trzop, E., Crystallography and Properties of Polyoxotitanate Nanoclusters. *Chem. Rev.* **2014,** *114* (19), 9645-9661.
10. Fang, W.-H.; Zhang, L.; Zhang, J., A 3.6 nm $Ti_{52}$–Oxo Nanocluster with Precise Atomic Structure. *J. Am. Chem. Soc.* **2016,** *138* (24), 7480-7483.
11. Yang, S.; Su, H.-C.; Hou, J.-L.; Luo, W.; Zou, D.-H.; Zhu, Q.-Y.; Dai, J., The effects of transition-metal doping and chromophore anchoring on the photocurrent response of titanium-oxo-clusters. *Dalton Trans.* **2017,** *46* (29), 9639-9645.
12. Matthews, P. D.; Li, N.; Luo, H.-K.; Wright, D. S., How Does Substitutional Doping Affect Visible Light Absorption in a Series of Homodisperse $Ti_{11}$ Polyoxotitanate Nanoparticles—A Comment on the Band Gap Determination of the $Fe^{II}$ Cages. *Chem. Eur. J.* **2016,** *22* (13), 4632-4633.
13. Mitchell, K. J.; Abboud, K. A.; Christou, G., Atomically-precise colloidal nanoparticles of cerium dioxide. *Nat. Commun.* **2017,** *8* (1), 1445.
14. Russell-Webster, B.; Abboud, K. A.; Christou, G., Molecular nanoparticles of cerium dioxide: structure-directing effect of halide ions. *Chem. Commun.* **2020,** *56* (40), 5382-5385.
15. Miersch, L.; Rüffer, T.; Schlesinger, M.; Lang, H.; Mehring, M., Hydrolysis studies on bismuth nitrate: synthesis and crystallization of four novel polynuclear basic bismuth nitrates. *Inorg. Chem.* **2012,** *51* (17), 9376-9384.
16. Du, F. Y.; Lou, J. M.; Jiang, R.; Fang, Z. Z.; Zhao, X. F.; Niu, Y. Y.; Zou, S. Q.; Zhang, M. M.; Gong, A. H.; Wu, C. Y., Hyaluronic acid-functionalized bismuth oxide

nanoparticles for computed tomography imaging-guided radiotherapy of tumor. *Int. J. Nanomedicine.* **2017,** *12*, 5973-5992.
17. Weber, M.; Schlesinger, M.; Mehring, M., Evaluation of Synthetic Methods for Bismuth(III) Oxide Polymorphs: Formation of Binary versus Ternary Oxides. *Cryst. Growth Des.* **2016,** *16* (10), 5678-5688.
18. Bøjesen, E. D.; Iversen, B. B., The chemistry of nucleation. *CrystEngComm* **2016,** *18* (43), 8332-8353.
19. Billinge, S. J. L.; Levin, I., The Problem with Determining Atomic Structure at the Nanoscale. *Science* **2007,** *316* (5824), 561-565.
20. Cooper, S. R.; Candler, R. O.; Cosby, A. G.; Johnson, D. W.; Jensen, K. M. Ø.; Hutchison, J. E., Evolution of Atomic-Level Structure in Sub-10 Nanometer Iron Oxide Nanocrystals: Influence on Cation Occupancy and Growth Rates. *ACS Nano* **2020,** *14* (5), 5480-5490.
21. Yang, X.; Masadeh, A. S.; McBride, J. R.; Božin, E. S.; Rosenthal, S. J.; Billinge, S. J. L., Confirmation of disordered structure of ultrasmall CdSe nanoparticles from X-ray atomic pair distribution function analysis. *Phys. Chem. Chem. Phys.* **2013,** *15* (22), 8480-8486.
22. 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.
23. Christiansen, T. L.; Cooper, S. R.; Jensen, K. M. Ø., There's no place like real-space: elucidating size-dependent atomic structure of nanomaterials using pair distribution function analysis. *Nanoscale Adv.* **2020,** *2* (6), 2234-2254.
24. Billinge, S. J. L.; Kanatzidis, M. G., Beyond crystallography: the study of disorder, nanocrystallinity and crystallographically challenged materials with pair distribution functions. *Chem. Commun.* **2004,** (7), 749-760.
25. Toby, B. H.; Dmowski, W.; Egami, T.; Jorgensen, J. D.; Subramanian, M. A.; Gopalakrishnan, J.; Sleight, A. W.; Parise, J. B., Ordering in $Tl_2CaBa_2Cu_2O_8$ and $Tl_2Ba_2CuO_6$ studied by pair distribution function and rietveld analysis. *Physica C* **1989,** *162-164*, 101-102.
26. Keen, D. A.; Goodwin, A. L., The crystallography of correlated disorder. *Nature* **2015,** *521* (7552), 303-309.
27. Scardi, P.; Billinge, S. J. L.; Neder, R.; Cervellino, A., Celebrating 100 years of the Debye scattering equation. *Acta Cryst.* **2016,** *72* (6), 589-590.
28. Jensen, K. M. Ø.; Juhas, P.; Tofanelli, M. A.; Heinecke, C. L.; Vaughan, G.; Ackerson, C. J.; Billinge, S. J. L., Polymorphism in magic-sized $Au_{144}(SR)_{60}$ clusters. *Nat Commun* **2016,** *7*, 11859.
29. Bennett, T. D.; Goodwin, A. L.; Dove, M. T.; Keen, D. A.; Tucker, M. G.; Barney, E. R.; Soper, A. K.; Bithell, E. G.; Tan, J.-C.; Cheetham, A. K., Structure and Properties of an Amorphous Metal-Organic Framework. *Phys. Rev. Lett.* **2010,** *104* (11), 115503.
30. Banerjee, S.; Liu, C.-H.; Jensen, K. M. Ø.; Juhas, P.; Lee, J. D.; Tofanelli, M.; Ackerson, C. J.; Murray, C. B.; Billinge, S. J. L., Cluster-mining: an approach for determining core structures of metallic nanoparticles from atomic pair distribution function data. *Acta Cryst.* **2020,** *76* (1), 24-31.
31. Bertolotti, F.; Moscheni, D.; Migliori, A.; Zacchini, S.; Cervellino, A.; Guagliardi, A.; Masciocchi, N., A total scattering Debye function analysis study of faulted Pt nanocrystals embedded in a porous matrix. *Acta Cryst.* **2016,** *72* (6), 632-644.
32. 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.

33. 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., Structural Changes during the Growth of Atomically Precise Metal Oxido Nanoclusters from Combined Pair Distribution Function and Small-Angle X-ray Scattering Analysis. *Angew. Chem. Int. Ed.* **2021,** *60*, 2-12.
34. Xu, H.; Sommer, S.; Broge, N. L. N.; Gao, J.; Iversen, B. B., The Chemistry of Nucleation: In Situ Pair Distribution Function Analysis of Secondary Building Units During UiO-66 MOF Formation. *Chem. Eur. J.* **2019,** *25* (8), 2051-2058.
35. Yang, L.; Juhas, P.; Terban, M. W.; Tucker, M. G.; Billinge, S. J. L., Structure-mining: screening structure models by automated fitting to the atomic pair distribution function over large numbers of models. *Acta Cryst.* **2020,** *76* (3), 395-409.
36. Anker, A. S.; Kjær, E. T. S.; Dam, E. B.; Billinge, S. J. L.; Jensen, K. M. Ø.; Selvan, R., *Characterising the Atomic Structure of Mono-Metallic Nanoparticles from X-Ray Scattering Data Using Conditional Generative Models*. Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG): 2020.
37. Liu, C.-H.; Tao, Y.; Hsu, D.; Du, Q.; Billinge, S. J. L., Using a machine learning approach to determine the space group of a structure from the atomic pair distribution function. *Acta Cryst.* **2019**.
38. Lundberg, S. M.; Lee, S.-I., A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems* **2017**, 4765-4774.
39. Woolfson, M. M., *An Introduction to X-ray Crystallography*. 2 ed.; Cambridge University Press: Cambridge, 1997.
40. Dinnebier, R. E.; Fitch, A.; Billinge, S. J. L.; Bail, A. L.; Le Bail, A.; Madsen, I.; Cranswick, L. M. D.; Cockcroft, J. K.; Norby, P.; Zuev, A. D., *Powder Diffraction : Theory and Practice*. Royal Society of Chemistry: Cambridge, UNITED KINGDOM, 2008.
41. edited by, E. P., *International tables for crystallography. Volume C, Mathematical, physical and chemical tables*. Third edition. Dordrecht ; Boston ; London : Published for the International Union of Crystallography by Kluwer Academic Publishers, 2004.: 2004.
42. Egami, T.; Billinge, S. J. L., *Underneath the Bragg Peaks, Pergamon* **2012**.
43. Als-Nielsen, J.; McMorrow, D., *Elements of Modern X-Ray Physics*. John Wiley & Sons, Incorporated: Hoboken, UNITED KINGDOM, 2011.
44. Scardi, P.; Gelisio, L., Vibrational Properties of Nanocrystals from the Debye Scattering Equation. *Scientific Reports* **2016,** *6* (1), 22221.
45. 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.
46. Mehring, M.; Mansfeld, D.; Paalasmaa, S.; Schürmann, M., Polynuclear Bismuth-Oxo Clusters: Insight into the Formation Process of a Metal Oxide. *Chem. Eur. J.* **2006,** *12* (6), 1767-1781.
47. Farrow, C. L.; Juhas, P.; Liu, J. W.; Bryndin, D.; Božin, E. S.; Bloch, J.; Th, P.; Billinge, S. J. L., PDFfit2 and PDFgui: computer programs for studying nanostructure in crystals. *J. Phys.: Condens. Matter* **2007,** *19* (33), 335219.
48. Juhas, P.; Farrow, C. L.; Yang, X.; Knox, K. R.; Billinge, S. 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.* **2015,** *71* (Pt 6), 562-568.
49. This work benefited from the use of the SasView application, originally developed under NSF award DMR-0520547. SasView contains code developed with funding from the

European Union's Horizon 2020 research and innovation programme under the SINE2020 project, grant agreement No 654000.
50. http://www.sasview.org/.
51. Tucker, M. G.; Keen, D. A.; Dove, M. T.; Goodwin, A. L.; Hui, Q., RMCProfile: reverse Monte Carlo for polycrystalline materials. *J. Phys.: Condens. Matter* **2007,** *19* (33), 335218.
52. Coelho, A. A., TOPAS and TOPAS-Academic: an optimization program integrating computer algebra and crystallographic objects written in C++. *J. Appl. Cryst.* **2018,** *51* (1), 210-218.
53. Garcia-Cardona, C.; Kannan, R.; Johnston, T.; Proffen, T.; Page, K.; Seal, S. K. In *Learning to Predict Material Structure from Neutron Scattering Data*, 2019 IEEE International Conference on Big Data (Big Data), 9-12 Dec. 2019; 2019; pp 4490-4497.
54. Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M., Insightful classification of crystal structures using deep learning. *Nat Commun* **2018,** *9* (1), 2775.
55. Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery: San Francisco, California, USA, 2016; pp 785–794.
56. Molnar, C., Interpretable Machine Learning - A Guide for Making Block Box Models Explainable. **2021**.
57. Fisher, A.; Rudin, C.; Dominici, F., All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research* **2019,** *20* (177), 1-81.
58. Shapley, L. S., A value for n-person games. *Contributions to the Theory of Games* **1953,** *2* (28), 307-317.
59. Frye, C.; de Mijolla, D.; Cowton, L.; Stanley, M.; Feige, I., Shapley-based explainability on the data manifold. *arXiv preprint arXiv:2006.01272* **2020**.
60. Begley, T.; Schwedes, T.; Frye, C.; Feige, I., Explainability for fair machine learning. *ArXiv* **2020,** *abs/2010.07389*.
61. Dimanov, B.; Bhatt, U.; Jamnik, M.; Weller, A. In *You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods*, SafeAI@ AAAI, 2020.
62. 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.
63. Malaestean, I. L.; Ellern, A.; Baca, S.; Kogerler, P., Cerium oxide nanoclusters: commensurate with concepts of polyoxometalate chemistry? *Chem. Commun.* **2012,** *48* (10), 1499-1501.
64. Falaise, C.; Volkringer, C.; Vigier, J. F.; Beaurain, A.; Roussel, P.; Rabu, P.; Loiseau, T., Isolation of the Large {Actinide}$_{38}$ Poly-oxo Cluster with Uranium. *J. Am. Chem. Soc.* **2013,** *135* (42), 15678-15681.
65. Soderholm, L.; Almond, P. M.; Skanthakumar, S.; Wilson, R. E.; Burns, P. C., The structure of the plutonium oxide nanocluster $[Pu_{38}O_{56}Cl_{54}(H_2O)_8]^{14-}$. *Angew. Chem. Int. Ed. Engl.* **2008,** *47* (2), 298-302.
66. Wilson, R. E.; Skanthakumar, S.; Soderholm, L., Separation of Plutonium Oxide Nanoparticles and Colloids. *Angew. Chem. Int. Ed.* **2011,** *50* (47), 11234-11237.
67. Sattler, D.; Schlesinger, M.; Mehring, M.; Schalley, C. A., Mass Spectrometry and Gas-Phase Chemistry of Bismuth-Oxido Clusters. *ChemPlusChem* **2013,** *78* (9), 1005-1014.
68. Hennig, C.; Ikeda-Ohno, A.; Kraus, W.; Weiss, S.; Pattison, P.; Emerich, H.; Abdala, P. M.; Scheinost, A. C., Crystal Structure and Solution Species of Ce(III) and Ce(IV)

Formates: From Mononuclear to Hexanuclear Complexes. *Inorg. Chem.* **2013,** *52* (20), 11734-11743.
69. Takao, K.; Takao, S.; Scheinost, A. C.; Bernhard, G.; Hennig, C., Formation of Soluble Hexanuclear Neptunium(IV) Nanoclusters in Aqueous Solution: Growth Termination of Actinide(IV) Hydrous Oxides by Carboxylates. *Inorg. Chem.* **2012,** *51* (3), 1336-1344.
70. Hu, Y.-J.; Knope, K. E.; Skanthakumar, S.; Soderholm, L., Understanding the Ligand-Directed Assembly of a Hexanuclear $Th^{IV}$ Molecular Cluster in Aqueous Solution. *Eur. J. Inorg. Chem.* **2013,** *2013* (24), 4159-4163.
71. Hennig, C.; Takao, S.; Takao, K.; Weiss, S.; Kraus, W.; Emmerling, F.; Scheinost, A. C., Structure and stability range of a hexanuclear Th(IV)–glycine complex. *Dalton Trans.* **2012,** *41* (41), 12818-12823.
72. Lazarini, F., The crystal structure of a bismuth basic nitrate, $[Bi_6O_5(OH)_3](NO_3)_5 \cdot 3H_2O$. *Acta Cryst.* **1978,** *34* (11), 3169-3173.
73. Farrow, C.; Shi, C.; Juhás, P.; Peng, X.; Billinge, S. J. L., Robust structure and morphology parameters for CdS nanoparticles by combining small-angle X-ray scattering and atomic pair distribution function data in a complex modeling framework. *J. Appl. Cryst.* **2014,** *47* (2), 561-565.
74. Farrow, C. L.; Billinge, S. J. L., Relationship between the atomic pair distribution function and small-angle scattering: implications for modeling of nanoparticles. *Acta Cryst.* **2009,** *65* (3), 232-239.
75. Rogow, D. L.; Fei, H.; Brennan, D. P.; Ikehata, M.; Zavalij, P. Y.; Oliver, A. G.; Oliver, S. R. J., Hydrothermal Synthesis of Two Cationic Bismuthate Clusters: An Alkylenedisulfonate Bridged Hexamer, $[Bi_6O_4(OH)_4(H_2O)_2][(CH_2)_2(SO_3)_2]_3$ and a Rare Nonamer Templated by Triflate, $[Bi_9O_8(OH)_6][CF_3SO_3]_5$. *Inorg. Chem.* **2010,** *49* (12), 5619-5624.
76. Mehring, M.; Paalasmaa, S.; Schürmann, M., Structural Relationships in High-Nuclearity Heterobimetallic Bismuth-Oxo Clusters. *Eur. J. Inorg. Chem.* **2005,** *2005* (24), 4891-4901.
77. Krayzman, V.; Levin, I.; Tucker, M. G., Simultaneous reverse Monte Carlo refinements of local structures in perovskite solid solutions using EXAFS and the total scattering pair-distribution function. *J. Appl. Cryst.* **2008**.
78. Proffen, T.; Neder, R. B., DISCUS, a program for diffuse scattering and defect structure simulations – update. *J. Appl. Cryst.* **1999,** *32* (4), 838-839.
79. Proffen, T.; Neder, R. B., DISCUS: a program for diffuse scattering and defect-structure simulation. *J. Appl. Cryst.* **1997,** *30* (2), 171-175.
80. Nogueira, F., Bayesian Optimization: Open source constrained global optimization tool for Python. 2014.
81. Momma, K.; Izumi, F., VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Cryst.* **2011,** *44* (6), 1272-1276.
82. Palmer, D. C., Visualization and analysis of crystal structures using CrystalMaker software. *Z. Kristallogr.* **2015,** *230* (9-10), 559-572.
83. 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.
84. Bouhlel, M. A.; Hwang, J. T.; Bartoli, N.; Lafage, R.; Morlier, J.; Martins, J. R. R. A., A Python surrogate modeling framework with derivatives. *Advances in Engineering Software* **2019,** *135*, 102662.
85. Gumerova, N. I.; Rompel, A., Polyoxometalates in solution: speciation under spotlight. *Chem. Soc. Rev.* **2020**.

86. Xin-Bao, H.; Zhi-Ming, Z.; Zhi-Shu, W.; Huan, Z.; Hui, D.; En-Bo, W., A 3D purely inorganic porous framework based on Keggin polyoxoanions. *Inorg. Chem. Commun.* **2012,** *18*, 47-49.
87. Piedra-Garza, L. F.; Reinoso, S.; Dickman, M. H.; Sanguineti, M. M.; Kortz, U., The first 3-dimensional assemblies of organotin-functionalized polyanions. *Dalton Trans.* **2009,** (31), 6231-6234.
88. Glass, E. N.; Fielden, J.; Kaledin, A. L.; Musaev, D. G.; Lian, T.; Hill, C. L., Extending Metal-to-Polyoxometalate Charge Transfer Lifetimes: The Effect of Heterometal Location. *Chem. Eur. J.* **2014,** *20* (15), 4297-4307.
89. Long, D.-L.; Song, Y.-F.; Wilson, E. F.; Kögerler, P.; Guo, S.-X.; Bond, A. M.; Hargreaves, J. S. J.; Cronin, L., Capture of Periodate in a $\{W_{18}O_{54}\}$ Cluster Cage Yielding a Catalytically Active Polyoxometalate $[H_3W_{18}O_{56}(IO_6)]^{6-}$ Embedded with High-Valent Iodine. *Angew. Chem. Int. Ed.* **2008,** *47* (23), 4384-4387.
90. Kato, C.; Nishihara, S.; Tsunashima, R.; Tatewaki, Y.; Okada, S.; Ren, X. M.; Inoue, K.; Long, D. L.; Cronin, L., Quick and selective synthesis of $Li_6[\alpha\text{-}P_2W_{18}O_{62}]\cdot 28H_2O$ soluble in various organic solvents. *Dalton Trans* **2013,** *42* (32), 11363-6.
91. *Generated using CrystalDiffract®: a powder diffraction program for Mac and Windows. CrystalMaker Software Ltd, Oxford, England (www.crystalmaker.com).*
92. Miersch, L.; Rüffer, T.; Schlesinger, M.; Lang, H.; Mehring, M., Hydrolysis Studies on Bismuth Nitrate: Synthesis and Crystallization of Four Novel Polynuclear Basic Bismuth Nitrates. *Inorg. Chem.* **2012,** *51* (17), 9376-9384.
93. 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.* **2013,** *46* (2), 560-566.
94. Christensen, A. N.; Chevallier, M.-A.; Skibsted, J.; Iversen, B. B., Synthesis and characterization of basic bismuth(III) nitrates. *J. Chem. Soc., Dalton Trans.* **2000,** (3), 265-270.
95. Hammersley, A. P., FIT2D : a multi-purpose data reduction, analysis and visualization program. *J. Appl. Cryst.* **2016,** *49* (2), 646-652.
96. 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.
97. Wadell, H., Volume, Shape, and Roundness of Quartz Particles. *The Journal of Geology* **1935,** *43* (3), 250-280.

# Appendix I: Supplementary Information for Chapter 2

### Simulation details of Q-space and r-space data of the [$Bi_{38}O_{45}$] cluster

The Q-space data has been simulated in CrystalDiffract[91] using a wavelength of 0.20717 Å and a 1.2 nm sized [$(Bi_{38}O_{45}(NO_3)_{24}((CH_3)_2SO)_{26})((CH_3)_2SO)_4)((Bi_{38}O_{45}(NO_3)_{24}((CH_3)_2SO)_{24})((CH_3)_2SO)_4$] crystal, mimicking the size of a single cluster.[92] Unfortunately, CrystalDiffract[91] cannot simulate data to higher $Q_{range}$ than shown in the plot. The r-space data has been simulated in Diffpy-CMI[48] using a [$Bi_{38}O_{45}$] cluster and the following parameters: $Q_{min}$ = 0 $Å^{-1}$, $Q_{max}$ = 9999 $Å^{-1}$, $Q_{damp}$ = 0.01 $Å^{-1}$, ADP = 0.03 $Å^2$. $Q_{damp}$ is a Gaussian dampened function which models the instrumental dampening.[48]

### Details of transformation from Q-space to r-space data of the [$Bi_{38}O_{45}$] cluster

The data was transformed to r-space from Q-space using PDFgetX3[93] using the following parameters: $R_{poly}$ = 0.9 Å, $Q_{max,\ instrument}$ = 17.5 $Å^{-1}$, $Q_{max}$ = 17 $Å^{-1}$, $Q_{min}$ = 0.7 $Å^{-1}$ and a chemical composition of $Bi_{38}O_{45}$. $Q_{max,\ instrument}$ is the highest Q value used in the data to calculate S(Q) and $R_{poly}$ is the threshold for what data is used to estimate the incoherent scattering in the data with a polynomial function.[93]

### Simulation details of total scattering, PDF and SAXS data of the [$Bi_{18}O_{36}$] cluster and the [$Bi_{22}O_{38}$] cluster

The total scattering data has been simulated in CrystalDiffract[91] using a wavelength of 0.20717 Å and a 1.2 nm sized [$Bi_{18}O_{18}(OSiMe_3)_{18}$]·$2C_7H_8$ crystal and a 1.2 nm sized [$Bi_{22}O_{26}(OSiMe_2tBu)_{14}$] crystal, mimicking the size of single clusters. Unfortunately, CrystalDiffract[91] cannot simulate data to higher $Q_{range}$ than shown in the plot. The PDF data was simulated using Diffpy-CMI[48] with the following parameters: $Q_{min}$ = 0.7 $Å^{-1}$, $Q_{max}$ = 20 $Å^{-1}$, $Q_{damp}$ = 0.03 $Å^{-1}$, ADP = 0.3 $Å^2$. The SAXS data were simulated with Diffpy-CMI[48] using a constant background to describe the Compton scattering of 1.5. Both the PDF and SAXS data have been calculated with the Debye equation using simulation parameters mimicking experimental data.

# Appendix II: Supplementary Information for Chapter 5[3]

### Bismuth oxido cluster synthesis

Synthesis of the bismuth oxido cluster $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ was carried out according to a literature procedure under modified conditions.[94] $Bi(NO_3)_3\cdot 5H_2O$ ($m$ = 50 g, $n$ = 0.1 mol) was dissolved in aqueous nitric acid ($c$ = 0.9 M, $V$ = 50 mL) followed by the addition of deionized water ($V$ = 1000 mL) to give a colorless reaction mixture. After stirring for 24 h, the colorless solid was separated by filtration and washed with aqueous nitric acid ($c$ = 0.63 M, $V$ = 100 mL) and deionized water ($V$ = 600 mL). After drying of the washed solid in vacuo ($p = 10^{-3}$ mbar, $T$ = 60 °C, $t$ = 1 h), the bismuth oxido cluster $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ was obtained as a colorless powder ($m$ = 21 g, $n$ = 12 mmol, $\eta$ = 70 % based on bismuth in $Bi(NO_3)_3\cdot 5H_2O$).

Synthesis of the bismuth oxido cluster $[Bi_6O_4(OH)_4(NO_3)_6(H_2O)_2]\cdot H_2O$ was carried out according to a literature procedure under modified conditions.[15] $Bi(NO_3)_3\cdot 5H_2O$ ($m$ = 4.85 g, $n$ = 0.01 mol) was dissolved in aqueous nitric acid ($c$ = 1 M, $V$ = 100 mL) followed by the dropwise addition of aqueous NaOH ($c$ = 0.1 M, $V$ = 100 mL). The resulting mixture was stirred under ambient conditions for $t$ = 0.5 h. Slow diffusion of acetone vapor via the gas phase into the mother liquid gave colorless crystals of $[Bi_6O_4(OH)_4(NO_3)_6(H_2O)_2]\cdot H_2O$ ($m$ = 2.20 g, $n$ = 1.21 mmol, $\eta$ = 73 % based on bismuth in $Bi(NO_3)_3\cdot 5H_2O$).

### PDF experiments

All *in situ* PDF experiments were conducted at beamline P02.1 at PETRAIII, DESY, Hamburg. Total scattering data were collected using the RA-PDF geometry with X-ray wavelength 0.2072 Å.

For studies of cluster growth, crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ ($m$ = 65.7 mg) was suspended in DMSO ($V$ = 1 mL). The suspension was then transferred to Kapton tubes with an inner diameter of 1.05 mm, which were mounted at the beamline. Total scattering data were collected with a time resolution of $t$ = 1 s while heating the capillary to $T$ = 30 °C – 80 °C with a heating rate of $\Delta T$ = 5 °C·min$^{-1}$. The time of the measurement was started when the heating was initiated.

---

[3] This chapter is based on some of the results from *Anker et. al., Structural Changes during the Growth of Atomically Precise Metal Oxido Nanoclusters from Combined Pair Distribution Function and Small-Angle X-ray Scattering Analysis, Angew. Chem. Int. Ed.* **2021,** *60*, 2-12. Which is also included as Appendix V.

Additional, *ex situ* X-ray total scattering measurements were done at beamline P07 at PETRAIII, DESY, Hamburg with an X-ray wavelength of 0.1235 Å. $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ ($m$ = 32.85 mg) was dissolved in DMSO ($V$ = 0.5 mL) at room temperature. To identify the effect of the cluster structure, an additional experiment was done where 32.85 mg of crystalline $[Bi_6O_4(OH)_4(NO_3)_6(H_2O)_2]\cdot H_2O$ was dissolved in 0.5 mL DMSO at room temperature. The samples were maintained undisturbed until measurements after $t$ = 2 days, 4 days, and 11 days. Data were collected in Kapton tubes at room temperature. All X-ray total scattering data were integrated using Fit2D[95]. PDFs were obtained using PDFgetx3[93, 96] and modelling was done using DiffPy-CMI.[48]

### SAXS experiments

The SAXS experiments of cluster growth were done at beamline ID02, ESRF ($\lambda$ = 0.7293 Å) and at SWING, SOLEIL ($\lambda$ = 0.7749 Å). The sample preparation and experiments were done as for the corresponding PDF experiments. The samples were loaded in 1.05 mm Kapton tubes (ESRF) and quartz capillaries (SOLEIL). The data were integrated using beamline specific programs. All data were analyzed with Diffpy-CMI,[48] which uses SASVIEW modules.[50]

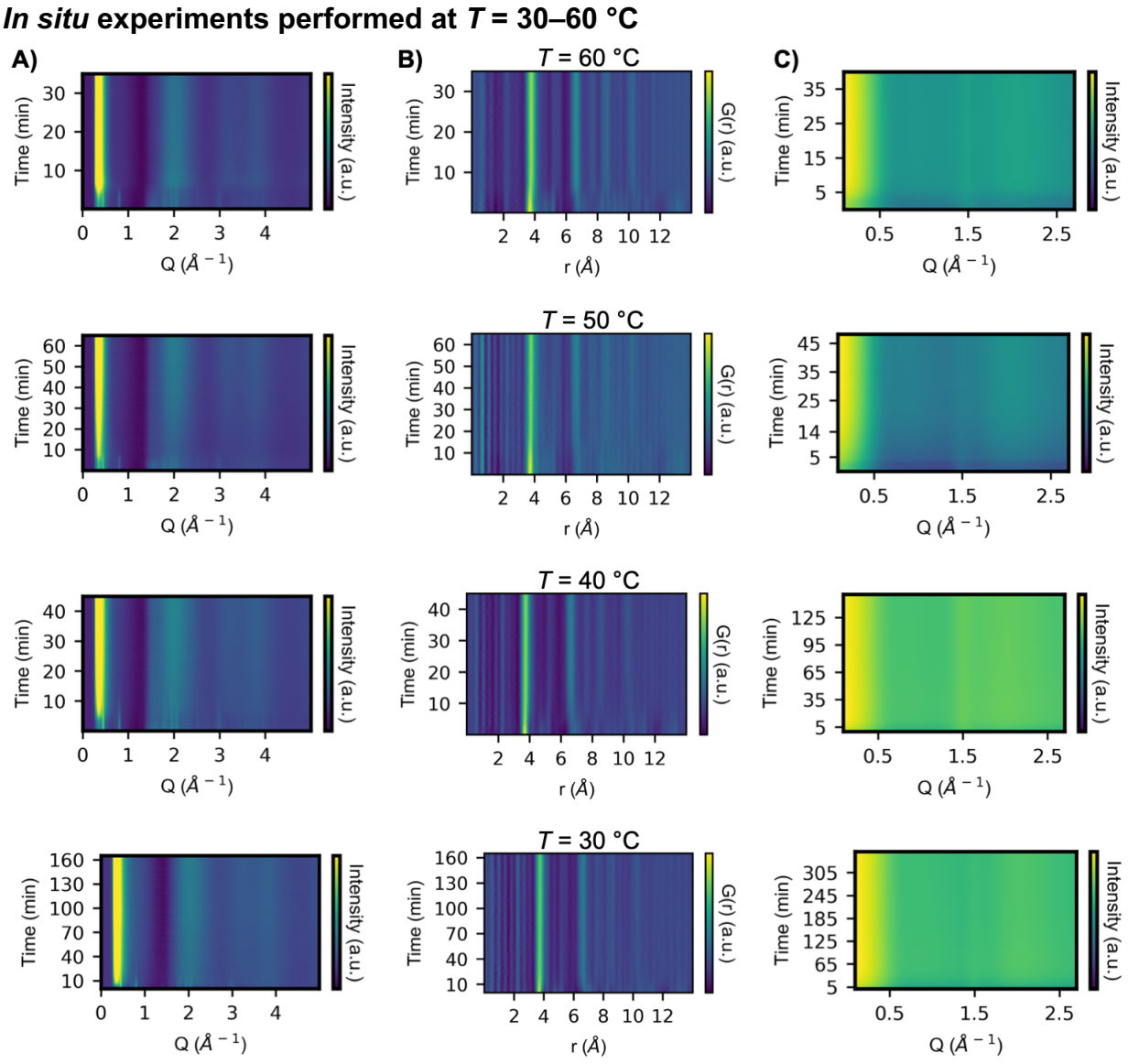



**Figure S1: A) The *in situ* X-ray total scattering data obtained during the experiment performed at T = 30–60 °C plotted as a function of time, which has been background subtracted as seen in Figure S2. Bragg peaks are seen in the beginning of the reaction but only diffuse features are seen after dissolution. Normalization of the colorbar has been done between the lowest intensity and half of the highest intensity to downscale the effect of the peak at Q = 0.34 Å$^{-1}$. B) The corresponding PDFs are plotted as a function of time. C) The background subtracted SAXS data (plotted on a log-normal-log scale) obtained during a similar experiment plotted as a function of time. Adapted from Anker et al.[33]**

## Background subtraction

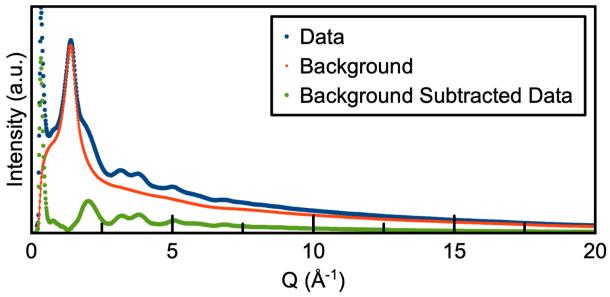



**Figure S2: illustration of how the background signal, a Kapton tube containing DMSO at the relevant temperature, is subtracted from the data in order to isolate the scattering pattern from the clusters. Most of the signal of the background is from DMSO, why a separate Kapton background subtraction is not done. Adapted from Anker et al.[33]**

## PDF and SAXS refinement details, end product *T* = 80 °C

The $[Bi_{38}O_{45}]$ cluster was obtained by cutting an isolated $[Bi_{38}O_{45}]$ cluster out of crystalline $[Bi_{38}O_{45}(NO_3)_{20}(DMSO)_{28}](NO_3)_4 \cdot 4DMSO$.[15]

**Table S1: The refinement details the fit of the $[Bi_{38}O_{45}]$ cluster structure to the background subtracted PDF data obtained at the end of the experiment, *t* = 70 min, at T = 80 °C. For all fits in the paper, the $B_{iso}$ value for oxygen has been fixed to 2 Å$^2$. Adapted from Anker et al.[33]**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Initial scaling factor | 0.5 | Lower boundary | 0.0 | Higher boundary | 1.0 | Refined Value | 0.8 |
| Isotropic ADP – Bi (Å$^2$) | 1.5 | Lower boundary | 0.0 | Higher boundary | 4.0 | Refined Value | 1.5 |
| Isotropic expansion | 1.0 | Lower boundary | $-\infty$ | Higher boundary | $\infty$ | Refined Value | 1.0 |

**Table S2: The refinement details the fit of the [$Bi_{38}O_{45}$] cluster structure to background subtracted SAXS data obtained at the end of the experiment, $t$ = 37 min, at T = 80 °C. Adapted from Anker et al.[33]**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Initial scaling factor | $1 \cdot 10^{-5}$ | Lower boundary | $-\infty$ | Higher boundary | $\infty$ | Refined Value | $2.16 \cdot 10^{-6}$ |
| Constant background | 0.1 | Lower boundary | $-\infty$ | Higher boundary | $\infty$ | Refined Value | 2.1 |

## Sequential SAXS form factor analysis

The SAXS form factor analysis was done in Diffpy-CMI,[48] which uses SASVIEW modules.[50] Spherical particles with a lognormal polydispersity were assumed and fitted with a SAXS form factor in the range $Q$ = 0.1–1.2 Å$^{-1}$. Sequential refinement was done reversely, starting with the last frame. The parameters used are given in Table S3. The uncertainty of the radius and polydispersity were estimated in SASVIEW[50] using a weight on the data points of the absolute value of the intensity. This might lead to uncertainties of the fitted parameters that are too small, but they will be independent of the scaling of the intensity.

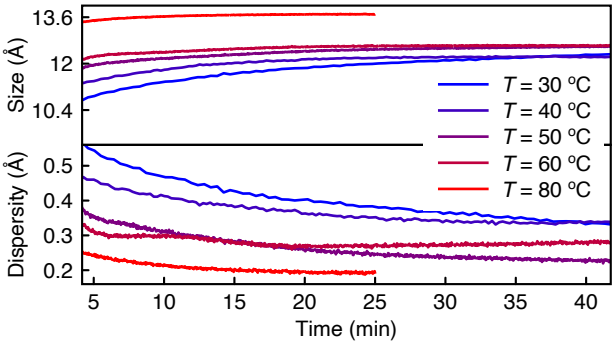



**Figure S3: The size and dispersity of the clusters, plotted as a function of reaction time from each of the *in situ* experiments at $T$ = 30–80 °C. The values are obtained from a SAXS form factor analysis using a spherical model with a lognormal size distribution. The data analyzed here were obtained at beamline ID02, ESRF. Adapted from Anker et al.[33]**

**Table S3: The fitting parameters used for sequential SAXS form factor analysis. Adapted from Anker et al.[33]**

| | | | | | |
|---|---|---|---|---|---|
| Initial scaling factor | $10^4$ | Lower boundary | 0 | Higher boundary | ∞ |
| Initial constant background | 1.0 | Lower boundary | 0.0 | Higher boundary | 10.0 |
| Initial radius | 7 Å | Lower boundary | Radius from last frame -1 Å | Higher boundary | Radius from last frame +1 Å |
| Initial radius dispersity | $1•10^{-2}$ Å | Lower boundary | Radius dispersity from last frame -0.2 Å | Higher boundary | Radius dispersity from last frame +0.2 Å |

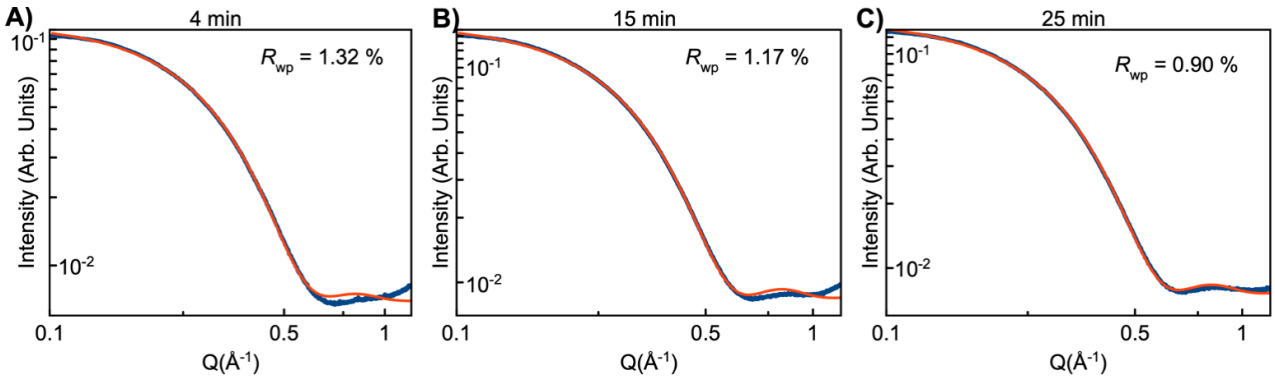



**Figure S4: The SAXS formfactor analysis using a spherical model with a lognormal size distribution at $T$ = 80 °C after A) t = 4 min of reaction. (Radius = 6.6 ± 0.4 Å, polydispersity = 0.10 ± 0.08 Å), B) $t$ = 15 min of reaction. (Radius = 6.7 ± 0.4 Å, polydispersity = 0.09 ± 0.08 Å), C) t = 25 min of reaction (Radius = 6.8 ± 0.4 Å, polydispersity = 0.08 ± 0.09 Å). The modelling were done in SASVIEW.[50] Adapted from Anker et al.[33]**

## Automated intermediate extractor

Test structures were constructed by removing Bi atoms iteratively from the outer shell of the cluster, and all oxygen atoms which were not bonded to at least a single Bi atom were removed. The longest allowed bond length was set to 2.5 Å, which is the longest Bi···O distance observed in the experimental PDFs. The cluster structure giving the best fit was identified and used as the starting-point for the second last frame in the series, where the same procedure was performed. This procedure was used for all frames in *in situ* experiments with $T$ = 30–80 °C. To speed up the calculations, the

maximum number of Bi atoms removed for each time step in the series was limited to 8, and the number of Bi atoms to be added to 3.

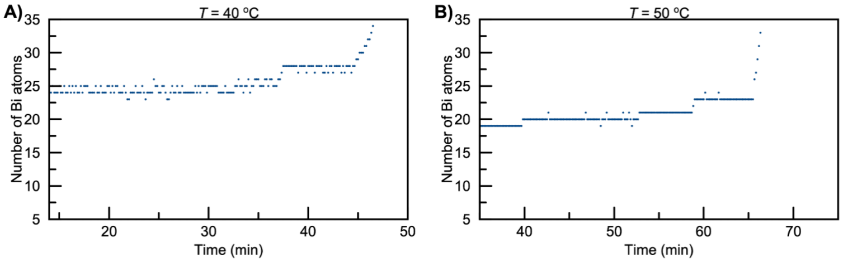



**Figure S5: The number of bismuth atoms in the best possible structure of the frame during the *in situ* experiment is plotted versus reaction time from the $T = 40$ °C and $T = 50$ °C *in situ* PDF & SAXS measurement. Adapted from Anker et al.[33]**

## Comparison of a two-phase model, $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$, with the cluster found from automated intermediate extractor

**Table S4: The fitting parameters used to make the fits in figure 8 of the $[Bi_{32}O_{44}]$ cluster. The $[Bi_{32}O_{44}]$ cluster is fitted to the background subtracted PDF experiment measured at $T = 80$ °C at $t = 52$ min. Adapted from Anker et al.[33]**

| $[Bi_{32}O_{44}]$ cluster | Structure from Permutation | | | 2 phases – $[Bi_{22}O_{26}]$ + $[Bi_{38}O_{45}]$ | | |
|---|---|---|---|---|---|---|
| | Debye PDF | Debye SAXS | Formfactor | Debye PDF | Debye SAXS | Formfactor |
| Scale factor 1 | 0.91 $[Bi_{32}O_{44}]$ | - | - | 0.50 $[Bi_{38}O_{45}]$ | - | - |
| Scale factor 2 | - | - | - | 0.30 $[Bi_{22}O_{26}]$ | - | - |
| ADP (Bi) (Å$^2$) | 1.28 | - | - | 1.22 | - | - |
| ADP (O) (Å$^2$) | 2.00 | - | - | 2.00 | - | - |
| Zoomscale 1 | 1.01 | - | - | 1.00 | - | - |
| Zoomscale 2 | - | - | - | 1.02 | - | - |
| Iqscale 1 | - | $3.15 \cdot 10^{-6}$ $[Bi_{32}O_{44}]$ | $2.25 \cdot 10^{2}$ $[Bi_{32}O_{44}]$ | - | $1.59 \cdot 10^{-6}$ $[Bi_{38}O_{45}]$ | $1.93 \cdot 10^{2}$ $[Bi_{38}O_{45}]$ |
| Iqscale 2 | - | - | - | - | $1.49 \cdot 10^{-6}$ $[Bi_{22}O_{26}]$ | $3.87 \cdot 10^{1}$ $[Bi_{22}O_{26}]$ |

| | | | | | | |
|---|---|---|---|---|---|---|
| Background | - | 1.88 | 1.98 | - | 1.92 | 1. 95 |
| Radius 1 (Å) | - | - | 6.5 ± 0.3 | - | - | 7 ± 1 |
| Radius 2 (Å) | - | - | - | - | - | 5 ± 3 |
| Dispersity 1 (Å) | - | - | $9•10^{-2}$ ± $9•10^{-2}$ | - | - | 0 ± 0 |
| Dispersity 2 (Å) | - | - | - | - | - | 0.2 ± 0.2 |
| $R_{wp}$ (%) | 15.26 | 1.84 | 1.26 | 15.28 | 1.68 | 1.21 |

**Table S5: The fitting parameters used to make the fits in figure 8 of the $[Bi_{28}O_{38}]$ cluster. The $[Bi_{28}O_{38}]$ cluster is fitted to the background subtracted PDF experiment measured at *T* = 80 °C at *t* = 15 min. Adapted from Anker et al.[33]**

| $[Bi_{28}O_{38}]$ cluster | Structure from Permutation | | | 2 phases - $[Bi_{22}O_{26}]$ + $[Bi_{38}O_{45}]$ | | |
|---|---|---|---|---|---|---|
| | Debye PDF | Debye SAXS | Formfactor | Debye PDF | Debye SAXS | Formfactor |
| Scale factor 1 | 0.93 $[Bi_{28}O_{38}]$ | - | - | 0.30 $[Bi_{38}O_{45}]$ | - | - |
| Scale factor 2 | - | - | - | 0.534 $[Bi_{22}O_{26}]$ | - | - |
| ADP (Bi) ($Å^2$) | 1.39 | - | - | 1.19 | - | - |
| ADP (O) ($Å^2$) | 2.00 | - | - | 2.00 | - | - |
| Zoomscale 1 | 1.01 | - | - | 1.01 | - | - |
| Zoomscale 2 | - | - | - | 1.00 | - | - |
| Iqscale 1 | - | $2.64•10^{-6}$ $[Bi_{28}O_{38}]$ | $3.22•10^{2}$ $[Bi_{28}O_{38}]$ | - | $1.70•10^{-6}$ $[Bi_{38}O_{45}]$ | $3.00•10^{2}$ $[Bi_{38}O_{45}]$ |
| Iqscale 2 | - | - | - | - | $3.49•10^{-6}$ $[Bi_{22}O_{26}]$ | $4.21•10^{1}$ $[Bi_{22}O_{26}]$ |
| Background | - | 1.72 | 2.85 | - | 2.70 | 2.77 |
| Radius 1 (Å) | - | - | 6.5 ± 0.3 | - | - | 7 ± 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Radius 2 (Å) | - | - | - | - | - | 4 ± 4 |
| Dispersity 1 (Å) | - | - | $7•10^{-2} \pm 9•10^{-2}$ | - | - | 0 ± 0 |
| Dispersity 2 (Å) | - | - | - | - | - | 0 ± 0 |
| $R_{wp}$ (%) | 16.94 | 2.90 | 2.16 | 16.65 | 2.75 | 2.10 |

**Table S6: The fitting parameters used to make the fits in figure 8 of the $[Bi_{25}O_{36}]$ cluster. The $[Bi_{25}O_{36}]$ cluster is fitted to the background subtracted PDF experiment measured at $T$ = 30 °C at $t$ = 152 min. Adapted from Anker et al.[33]**

| $[Bi_{25}O_{36}]$ cluster | Structure from Permutation | | | 2 phases - $[Bi_{22}O_{26}]$ + $[Bi_{38}O_{45}]$ | | |
|---|---|---|---|---|---|---|
| | Debye PDF | Debye SAXS | Formfactor | Debye PDF | Debye SAXS | Formfactor |
| Scale factor 1 | 0.83 $[Bi_{25}O_{36}]$ | - | - | 0.06 $[Bi_{38}O_{45}]$ | - | - |
| Scale factor 2 | - | - | - | 0.74 $[Bi_{22}O_{26}]$ | - | - |
| ADP (Bi) (Å$^2$) | 1.38 | - | - | 1.01 | - | - |
| ADP (O) (Å$^2$) | 2.00 | - | - | 2.00 | - | - |
| Zoomscale 1 | 1.00 | - | - | 1.01 | - | - |
| Zoomscale 2 | - | - | - | 1.00 | - | - |
| Iqscale 1 | - | $7.97•10^{-6}$ $[Bi_{25}O_{36}]$ | $3.95•10^{2}$ $[Bi_{25}O_{36}]$ | - | $2.24•10^{-6}$ $[Bi_{38}O_{45}]$ | $2.21•10^{2}$ $[Bi_{38}O_{45}]$ |
| Iqscale 2 | - | - | - | - | $3.55 •10^{-6}$ $[Bi_{22}O_{26}]$ | $1.84•10^{2}$ $[Bi_{22}O_{26}]$ |
| Background | - | 2.50 | 3.16 | - | 3.10 | 3.13 |
| Radius 1 (Å) | - | - | 6.1 ± 0.3 | - | - | 7 ± 1 |
| Radius 2 (Å) | - | - | - | - | - | 5 ± 2 |
| Dispersity 1 (Å) | - | - | 1.40 ± 0.07 | - | - | 0 ± 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Dispersity 2 (Å) | - | - | - | - | - | 0.2 ± 0.2 |
| $R_{wp}$ (%) | 32.87 | 4.60 | 1.42 | 30.16 | 1.96 | 1.39 |

**Table S7: The fitting parameters used to make the fits in figure 8 of the [$Bi_{20}O_{33}$] cluster. The [$Bi_{20}O_{33}$] cluster is fitted to the background subtracted PDF experiment measured at *T* = 30 °C at *t* = 103 min. Adapted from Anker et al.[33]**

| [$Bi_{20}O_{33}$] cluster | Structure from Permutation | | | 2 phases - [$Bi_{22}O_{26}$] + [$Bi_{38}O_{45}$] | | |
|---|---|---|---|---|---|---|
| | Debye PDF | Debye SAXS | Formfactor | Debye PDF | Debye SAXS | Formfactor |
| Scale factor 1 | 0.87 [$Bi_{20}O_{33}$] | - | - | 0.08 [$Bi_{38}O_{45}$] | - | - |
| Scale factor 2 | - | - | - | 0.698 [$Bi_{22}O_{26}$] | - | - |
| ADP (Bi) (Å$^2$) | 1.30 | - | - | 0.72 | - | - |
| ADP (O) (Å$^2$) | 2.00 | - | - | 2.00 | - | - |
| Zoomscale 1 | 0.99 | - | - | 0.95 | - | - |
| Zoomscale 2 | - | - | - | 1.00 | - | - |
| Iqscale 1 | - | $1.19 \cdot 10^{-5}$ [$Bi_{20}O_{33}$] | $4.09 \cdot 10^{2}$ [$Bi_{20}O_{33}$] | - | $2.22 \cdot 10^{-6}$ [$Bi_{38}O_{45}$] | $2.46 \cdot 10^{2}$ [$Bi_{38}O_{45}$] |
| Iqscale 2 | - | - | - | - | $3.94 \cdot 10^{-6}$ [$Bi_{22}O_{26}$] | $1.74 \cdot 10^{2}$ [$Bi_{22}O_{26}$] |
| Background | - | 2.03 | 3.27 | - | 3.19 | 3. 24 |
| Radius 1 (Å) | - | - | 6.1 ± 0.3 | - | - | 7 ± 1 |
| Radius 2 (Å) | - | - | - | - | - | 5 ± 2 |
| Dispersity 1 (Å) | - | - | 1.30 ± 0.07 | - | - | 0 ± 0 |
| Dispersity 2 (Å) | - | - | - | - | - | 0.2 ± 0.2 |
| $R_{wp}$ (%) | 31.82 | 8.30 | 1.61 | 26.97 | 2.17 | 1.57 |

## Comparison of a two-phase model using the $[Bi_{22}O_{26}]$ cluster and the $[Bi_{38}O_{45}]$ cluster, with the one-phase model of the two clusters

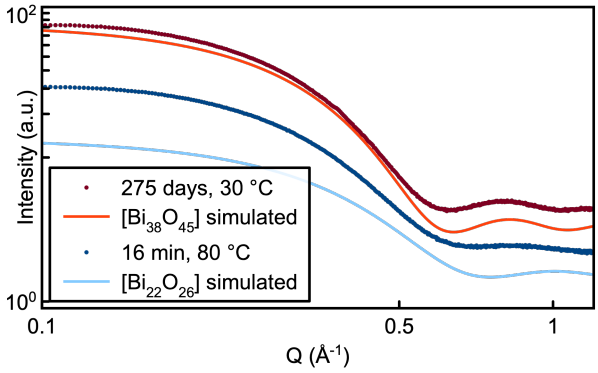



**Figure S6: The datasets obtained at different stages of the reaction plotted with the simulated data of the $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ cluster. The dataset from the sample left at room temperature for 275 days is very similar to the $[Bi_{38}O_{45}]$ cluster, while the dataset for the sample heated to $T = 80$ °C for $t = 16$ minutes is similar to a combination of both $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$, with $[Bi_{22}O_{26}]$ as the majority species. Adapted from Anker et al.[33]**

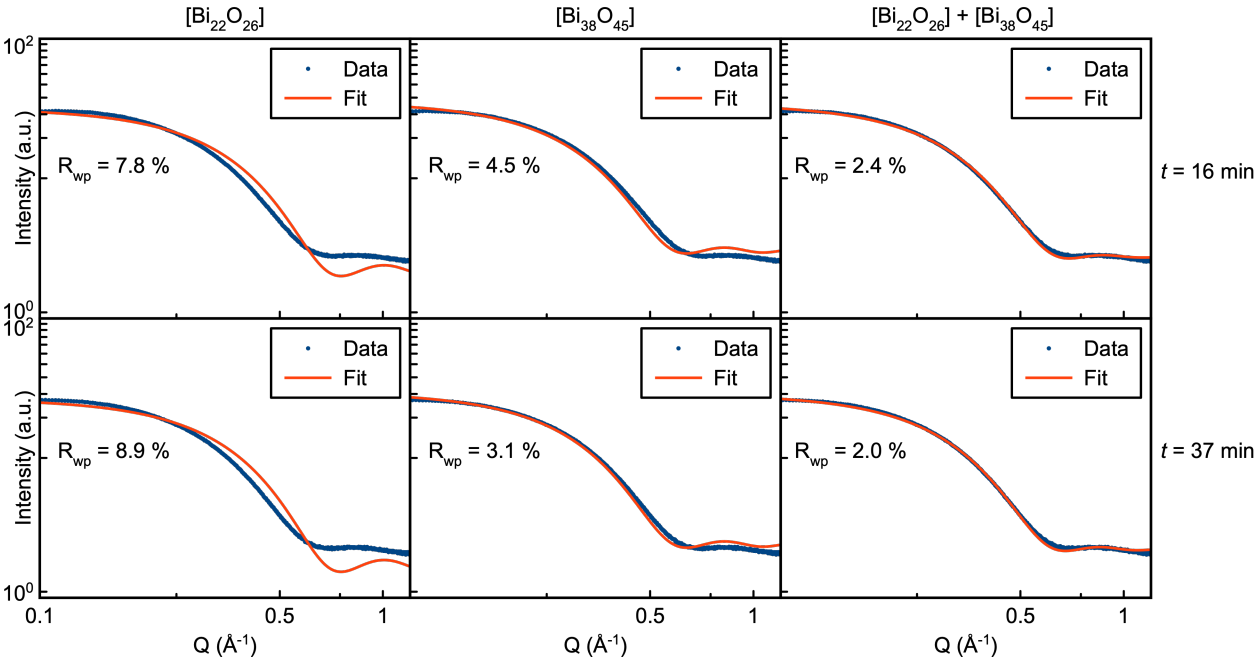



**Figure S7: The models including a combination of both the $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ structures fit the *in situ* SAXS data obtained at $T = 80$ °C after $t = 16$ and 37 min better than either the individual $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ clusters. Adapted from Anker et al.[33]**

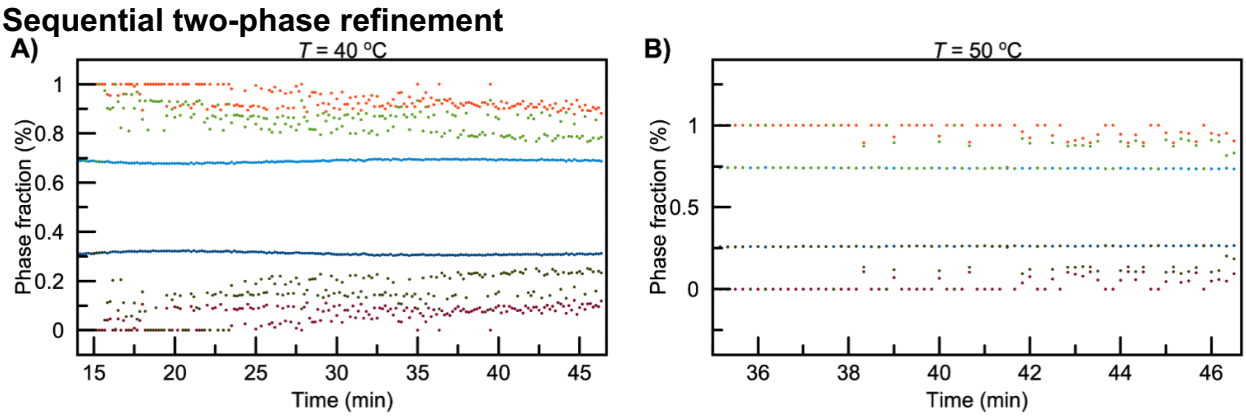



**Figure S8: Phase fraction of the $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ structures during the *in situ* experiments is plotted versus time of the reaction for the $T = 50$ °C and $T = 40$ °C *in situ* PDF & SAXS measurement and by complex modelling, which shows that the $[Bi_{22}O_{26}]$ cluster is growing to the $[Bi_{38}O_{45}]$ cluster. Blue: SAXS, Red: PDF, Green: Complex Modelling. Adapted from Anker et al.[33]**

## Cluster growth from the crystalline $[Bi_6O_4(OH)_4(NO_3)_6(H_2O)_2]\cdot(H_2O)$ and $[Bi_6O_5(OH)_3(NO_3)_5]\cdot(H_2O)_3$ dissolved in DMSO

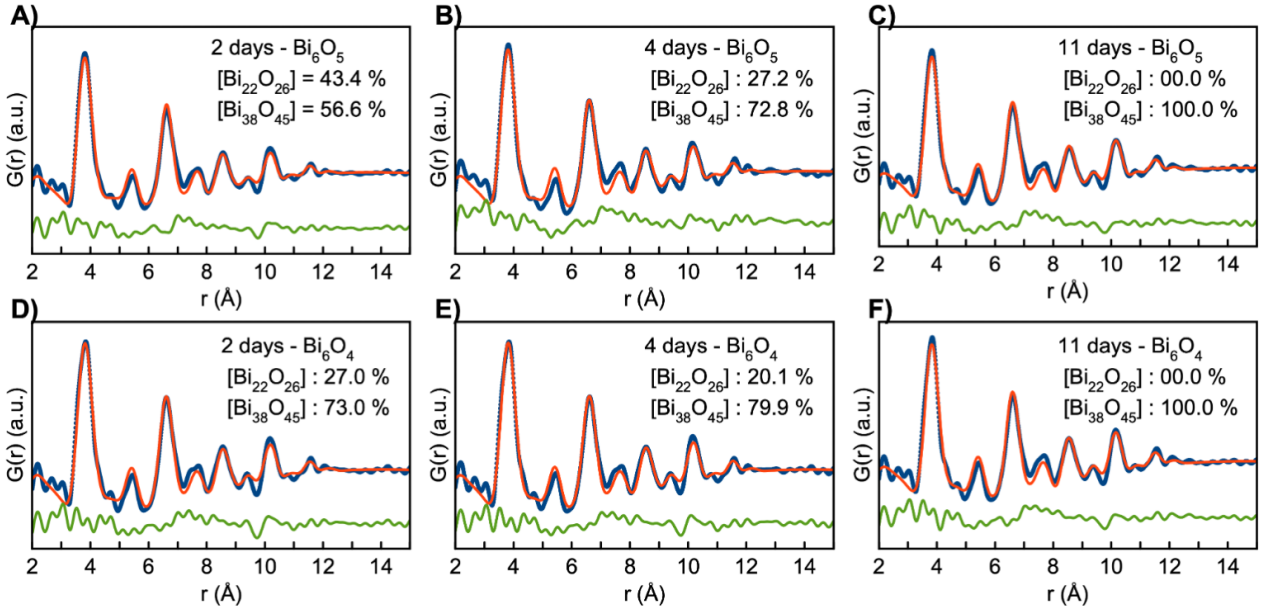



**Figure S9: The two-phase refinements with the $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ cluster of the A–C) $[Bi_6O_5(OH)_3(NO_3)_5]\cdot(H_2O)_3$ crystal mixed with DMSO and left undisturbed for $t = 2$-, 4- and 11 days at $T = 25$ °C. D–F) the $[Bi_6O_4(OH)_4(NO_3)_6(H_2O)_2]\cdot(H_2O)$ crystal mixed with DMSO and left undisturbed for $t = 2$-, 4- and 11 days at $T = 25$ °C. Adapted from Anker et al.[33]**

# Appendix III: Supplementary Information for Chapter 6

### Goodness-of-fit parameter $R_{wp}$ value

The goodness–of–fit parameter is the value that is minimized during a refinement procedure. The $R_{wp}$ value is the most common goodness-of-fit parameter used in PDF analysis. It is given by

**Equation S1**

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

Where $G_{obs}$ and $G_{calc}$ are the observed and calculated PDFs, P is the parameters that are refined in the model.

### Simulation parameters for the β-$Bi_2O_3$ crystal

**Table S8: Simulated parameters for all the examples provided in the paper.**

| | **2x2x3 unit cells of β-$Bi_2O_3$** |
|---|---|
| r-range (Å) | 2 – 15 |
| r-step | 0.01 |
| $Q_{min}$ ($Å^{-1}$) | 0.7 |
| $Q_{max}$ ($Å^{-1}$) | 15.5 |
| $Q_{damp}$ ($Å^{-1}$) | 0.04 |
| ADP ($Å^2$) | 1.35 |

### Fitting variables for ML-MotEx

**Table S9: Instrumental- and fitting parameters for the 4 examples provided in the paper. The initial guesses of the parameters are written inside the parentheses, while the fitting range is written outside the parentheses.**

| | **β-$Bi_2O_3$** |
|---|---|
| **PDF** | |
| r-range (Å) | 2 – 15 |
| r-step | 0.01 |
| $Q_{min}$ ($Å^{-1}$) | 0.7 |
| $Q_{max}$ ($Å^{-1}$) | 15.5 |

| | |
|---|---|
| $Q_{damp}$ (Å$^{-1}$) | 0.04 |
| $Q_{broad}$ (Å$^{-1}$) | - |
| Oxygen threshold (Å) | 2.5 |
| PDF scaling factor | 0 – 1.5 (0.9) |
| Delta2 (Å$^2$) | - |
| $ADP_{C/Mo/w/Bi}$ (Å$^2$) | 0.01 – 3 (1.35) |
| $ADP_O$ (Å$^2$) | 2 |
| Isotropic expansion | 0.98 – 1.02 (1.00) |
| **SAXS** | |
| Q-range (Å$^{-1}$) | 0.09 – 2.7 |
| $R_{max}$ (Å) | 100 Å |
| SAXS scaling factor | 0 – $10^{-5}$ ($4*10^{-6}$) |
| Background | 0 – 4 (2.4) |

## Cookie-Cutter algorithm

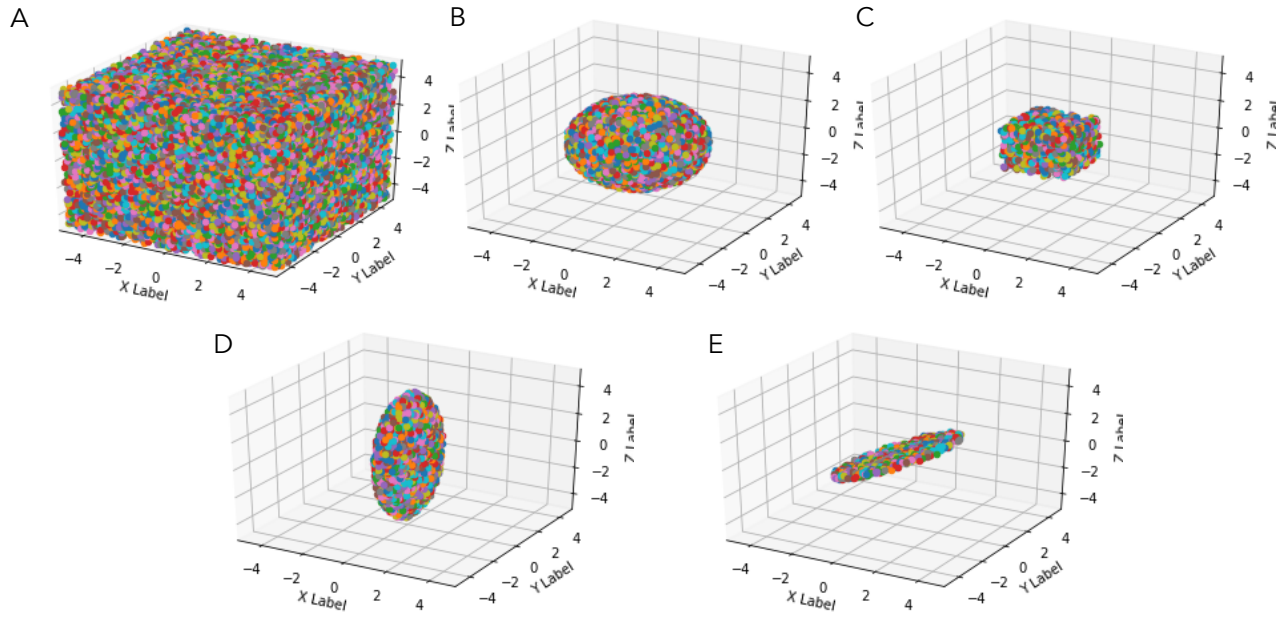



**Figure S10: 4 examples of shapes that can be cut out with the cookie-cutter tool. A) A box with atoms 0.1 Å apart. All atoms outside B) the sphere, C) the cube, D) the ellipsoid, E) the disk, have been removed.**

**Table S10: Parameter thresholds for sampling geometries with spherical-, cubic-, ellipsoidal- or a disk shape. The parameters are sampled from a uniform distribution in this range.**

| **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 |

## 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:[97]

**Equation S2**

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

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

## Analysis of the PDF fits of data from the [$Bi_{38}O_{45}$] cluster

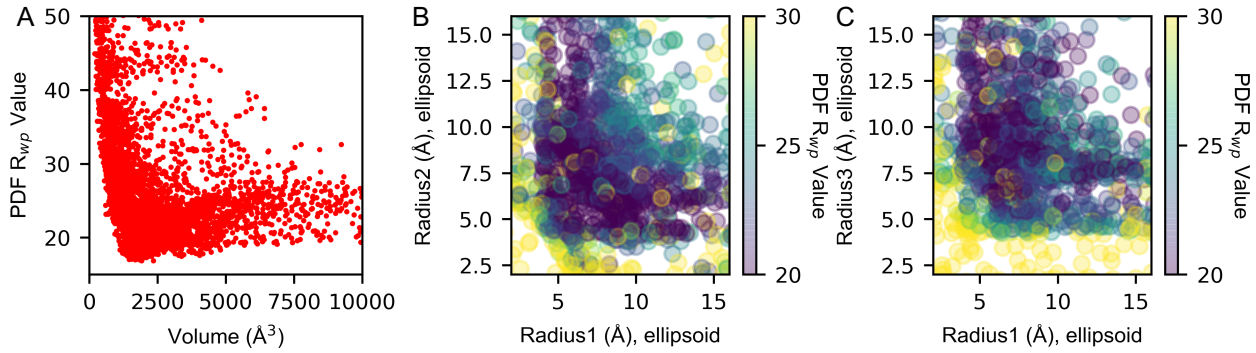



**Figure S11: Overview of the $R_{wp}$ values of the PDF fits versus important structural parameters. A) The $R_{wp}$ values versus the volume of the structure, B) The $R_{wp}$ value versus radius1 and radius2 in the ellipsoidal shaped structures and C) The $R_{wp}$ value versus radius1 and radius3 in the ellipsoidal shaped structures.**

## 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 S11: Refinement details the best fitting cluster extracted with brute-force modelling and ML-MotEx to the background subtracted from the [$Bi_{38}O_{45}$] cluster.**

**Fits of the PDF and SAXS data using the [$Bi_{38}O_{45}$] cluster as model**

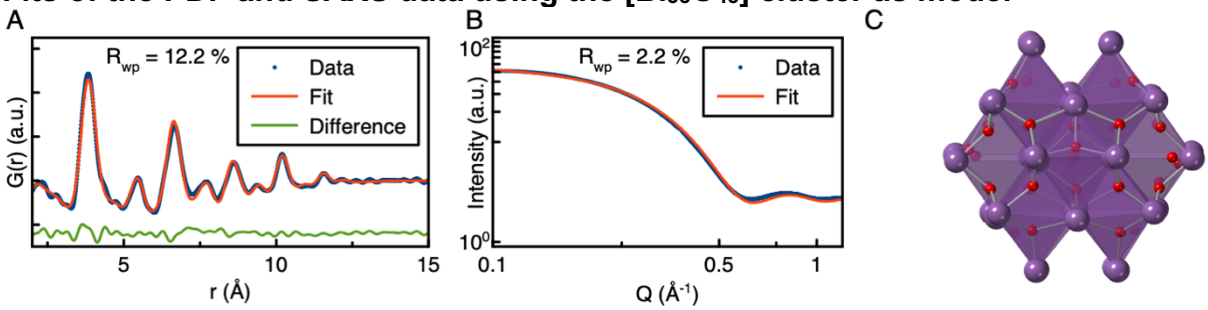



**Figure S12: Fit of the A) PDF data, B) SAXS data, using the C) [$Bi_{38}O_{45}$] cluster.**

| Example system | PDF | SAXS |
|---|---|---|
| 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) | - |
| Q-range (Å$^{-1}$) | - | 0.09 – 1.2 |
| $R_{max}$ (Å) | - | 100 |
| SAXS scaling factor | - | 0 – ∞ ($4.1 \cdot 10^{-6}$) |
| Background | - | 0 – ∞ (2.4) |

**Table S12: Refinement details the fit of the [$Bi_{38}O_{45}$] cluster to the background subtracted PDF and SAXS 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 parentheses.**

## ML-MotEx does not perform well on SAXS data due to the missing atomistic information compared to PDF data

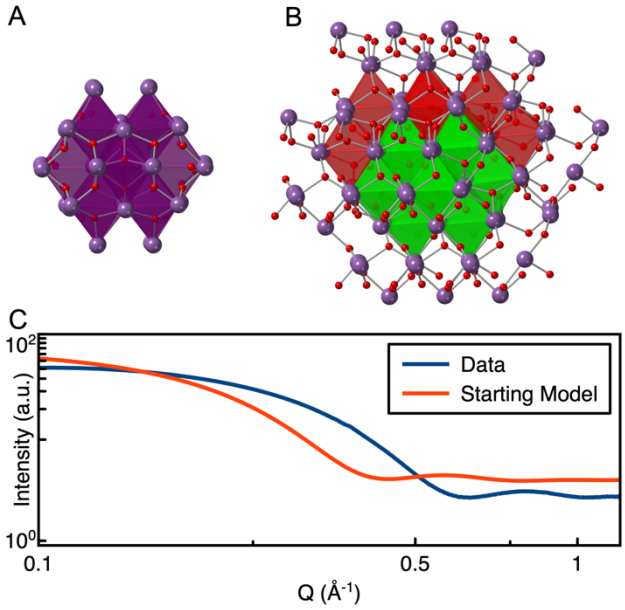



**Figure S13: A) The [$Bi_{38}O_{45}$] cluster B) The β-$Bi_2O_3$ structure with highlighted polyhedra in red and green, where the green corresponds to the [$Bi_{38}O_{45}$] cluster. Some polyhedra are drawn as balls and stick instead, to enhance visibility. C) Comparison of the experimental SAXS of the [$Bi_{38}O_{45}$] cluster and simulated data of a 2x2x3 unit cell of β-$Bi_2O_3$. The simulation parameters mimic typical values of a PDF dataset and can be seen in Table S8.**

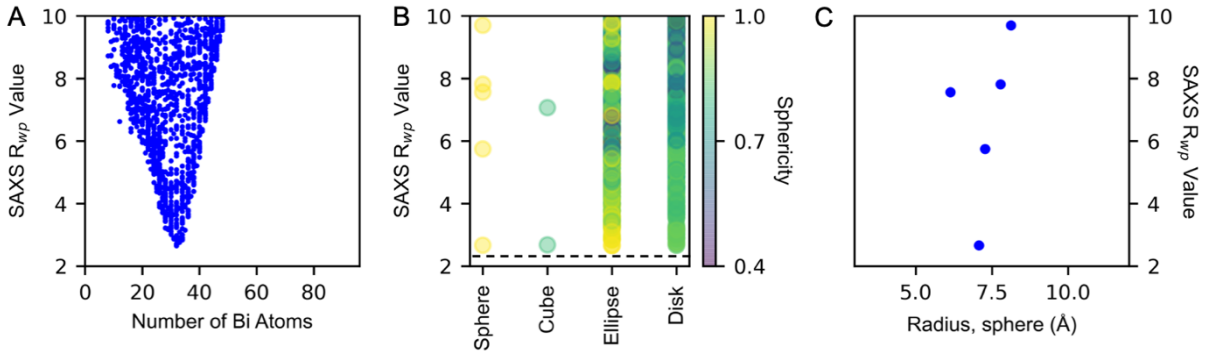



**Figure S14: Overview of the $R_{wp}$ values of the SAXS fits versus important structural parameters. A) The $R_{wp}$ values versus the number of bismuth atoms present in the structure, B) The $R_{wp}$ value versus the shape and**

**sphericity of the structures and C) The $R_{wp}$ value versus the spherical radius of the structures made with a spherical shape.**

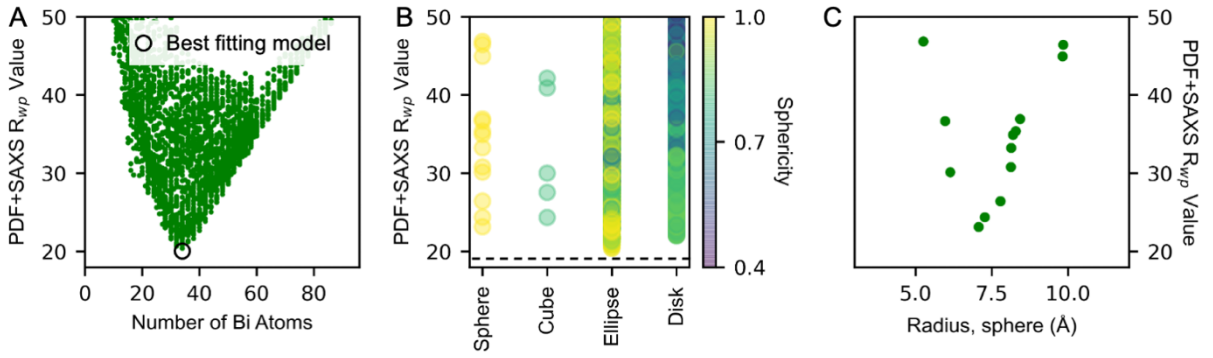



**Figure S15: Overview of the $R_{wp}$ values of the PDF fits + $R_{wp}$ values of the SAXS fits versus important structural parameters. A) The $R_{wp}$ values versus the number of bismuth atoms present in the structure, B) The $R_{wp}$ value versus the shape and sphericity of the structures and C) The $R_{wp}$ value versus the spherical radius of the structures made with a spherical shape.**

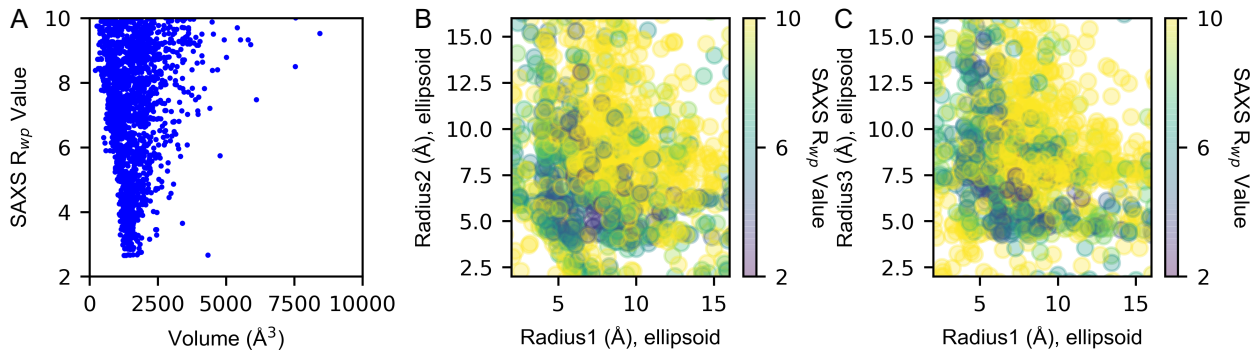



**Figure S16: Overview of the $R_{wp}$ values of the SAXS fits versus important structural parameters. A) The $R_{wp}$ values versus the volume of the structure, B) The $R_{wp}$ value versus radius1 and radius2 in the ellipsoidal shaped structures and C) The $R_{wp}$ value versus radius1 and radius3 in the ellipsoidal shaped structures.**

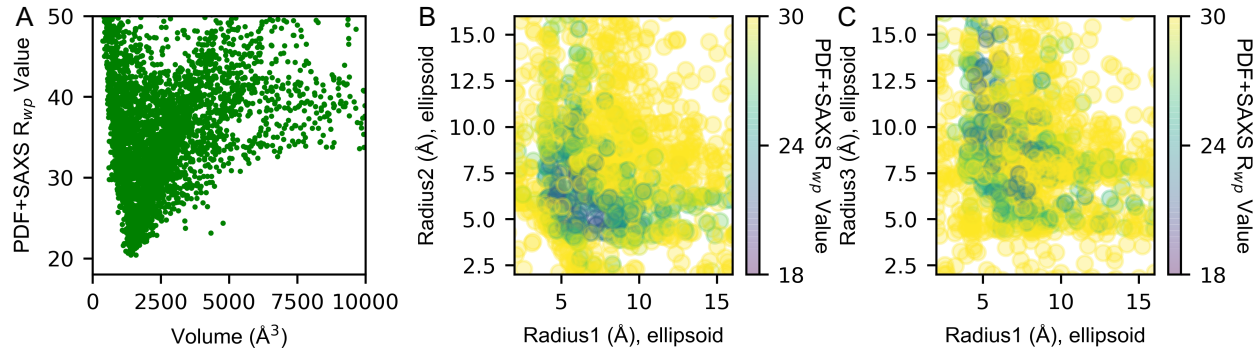



**Figure S17: Overview of the $R_{wp}$ values of the PDF fits + the $R_{wp}$ values of the SAXS fits versus important structural parameters. A) The $R_{wp}$ values versus the volume of the structure, B) The $R_{wp}$ value versus radius1 and radius2 in the ellipsoidal shaped structures and C) The $R_{wp}$ value versus radius1 and radius3 in the ellipsoidal shaped structures.**

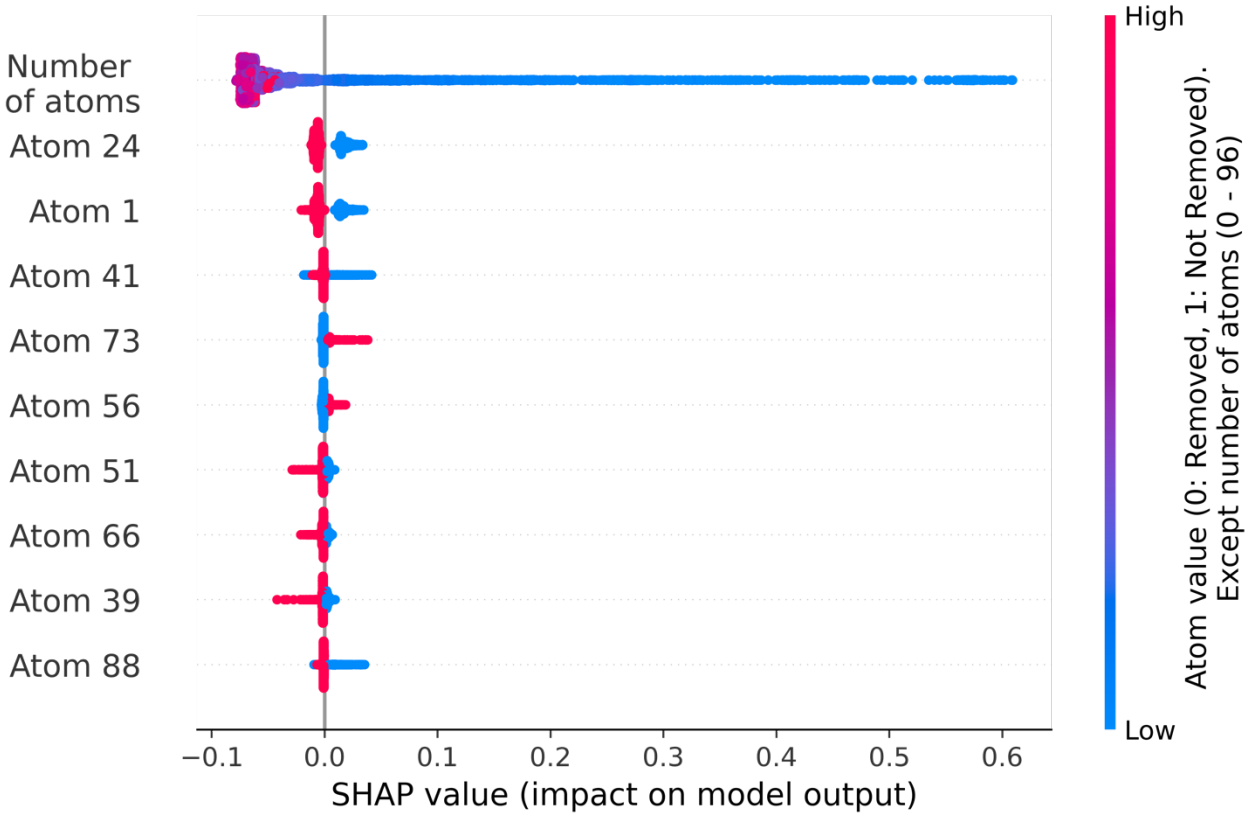



**Figure S18: SHAP summary plot from ML-MotEx utilized on PDF data from the [$Bi_{38}O_{45}$] cluster.**

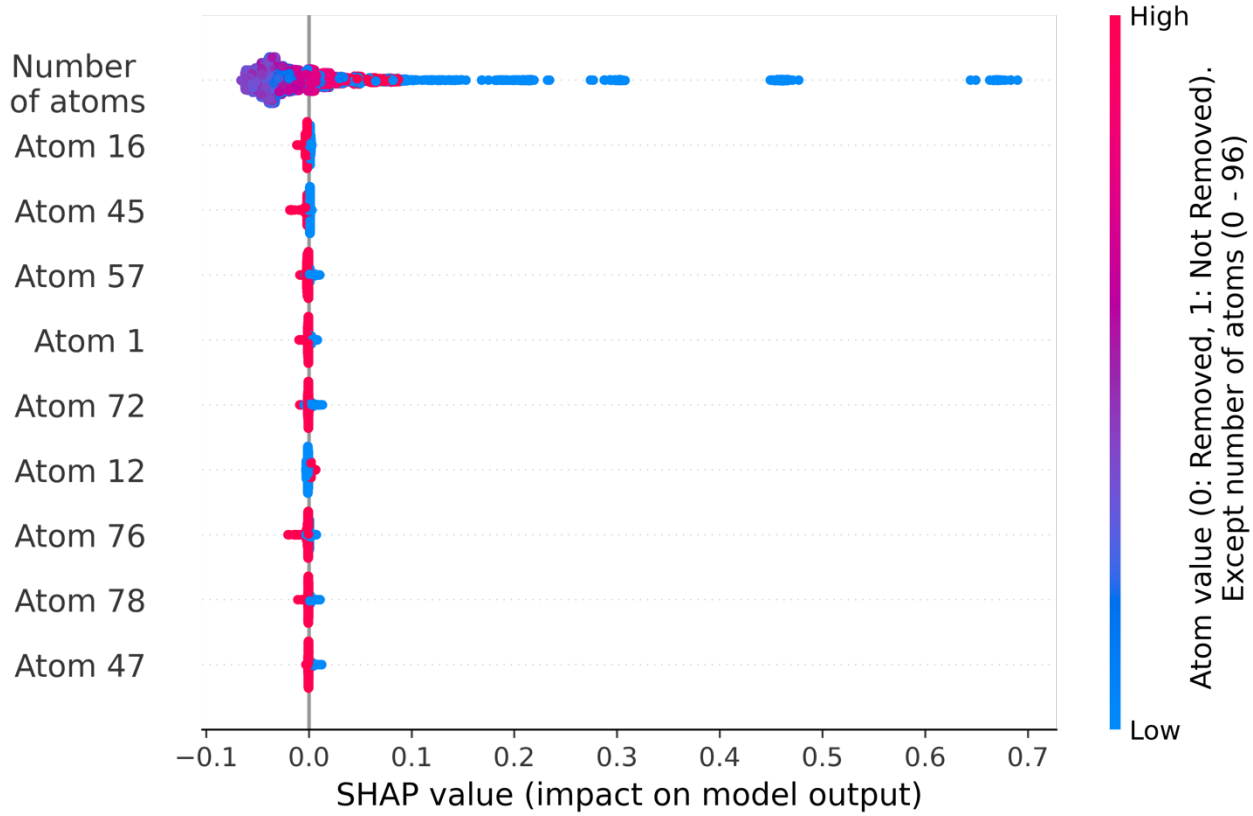



**Figure S19: SHAP summary plot from ML-MotEx utilized on SAXS data from the [$Bi_{38}O_{45}$] cluster.**

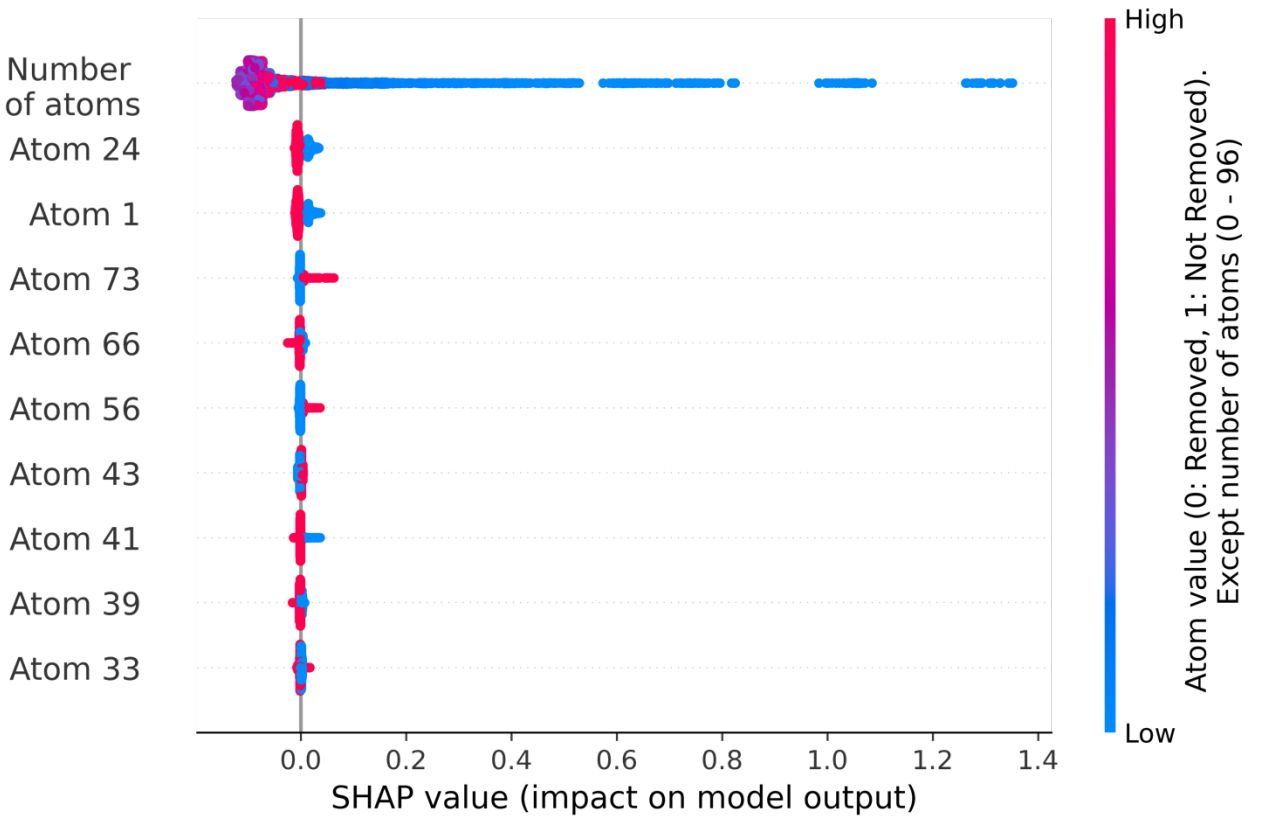



**Figure S20: SHAP summary plot from ML-MotEx utilized on combined PDF and SAXS data from the [$Bi_{38}O_{45}$] cluster.**

While ML-MotEx carves out a physical closed-shell structure (yellow atoms) without 'dangling' atoms when it has been trained on PDF data, this is not the case when it is trained on SAXS data. We argue that ML-MotEx is a better approach for PDF modelling since PDF contains information about the atomic arrangement, while the SAXS data is better suited to understand size and shape.


**Figure S21: Visualisation of the most preferred atoms (yellow) to keep in the structure by ML-MotEx and the atoms that should be removed (purple).**

In order to understand how ML-MotEx orders the importance of the atoms to yield a low $R_{wp}$ of the fit, the atoms were ordered with respect to their SHAP value. The first atom was the one with the lowest SHAP value (pushes the $R_{wp}$ value down) and the last atom was the atom with the highest SHAP value (pushes the $R_{wp}$ value up). Afterwards, they were iteratively included in the model with the one with the lowest SHAP value first and fitted to the data. The $R_{wp}$ values of the fits from this process is illustrated in Figure S21 for the SAXS and PDF data individually as well as combined. From Figure S21, the best fitting model for the PDF $R_{wp}$ value is when the 37 atoms with the lowest SHAP values are kept in the model and the rest of the atoms are removed. This model is shown and fitted to the PDF and SAXS data in Figure S23. Fitting details are given in Table S13. The same routine is made for SAXS data illustrated in Figure S24 and Table S14 and for combined analysis of PDF and SAXS data, where the $R_{wp}$ values are summed, in Figure S25 and Table S15. The results show that the PDF want to keep an 'closed shell physical structure', while the SAXS data just generate a structure that has the right size and shape that fits the SAXS data well. We therefore conclude that the ML-MotEx approach is better suited for techniques as PDF with atomistic information compared to SAXS data with information about the particle morphology.

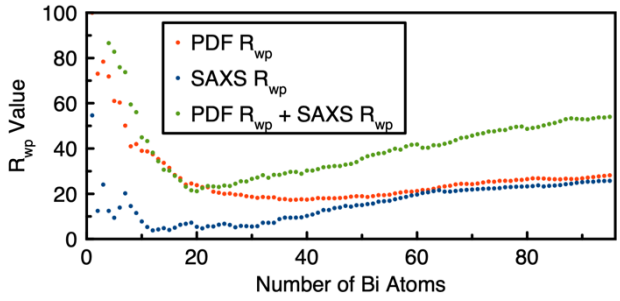



**Figure S22: The $R_{wp}$ values of the fits made by including various numbers of atoms in the structure starting from the atom with the lowest SHAP value (pushes the $R_{wp}$ value most down) after ML-MotEx has been applied on the PDF-, SAXS- and combined dataset.**

*Best fitting structure found with ML-MotEx on the PDF data from the [$Bi_{38}O_{45}$] cluster*

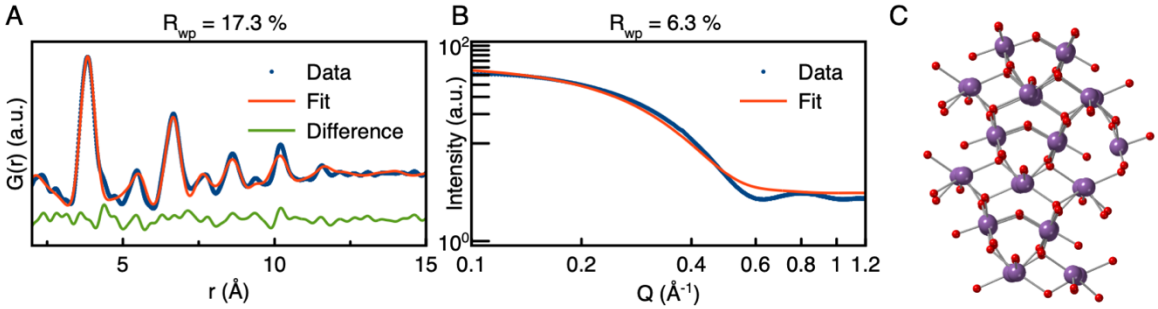



**Figure S23: Fits of the cut out of the 37 most preferred bismuth atoms of the PDF model, A) PDF fit, B) SAXS fit, C) The structure of the 37 most preferred Bi atoms and all oxygen atoms within 2.5 Å to a Bi atom.**

| Example system | PDF | SAXS |
|---|---|---|
| r-range (Å) | 2 – 15 | - |
| Scale | 0 – 10 (0.86) | - |
| $ADP_{Bi}$ (Å$^2$) | 0 – 4 (1.7) | - |
| $ADP_O$ (Å$^2$) | Fixed to 1 | - |
| Isotropic expansion | 0.95 – 1.05 (0.98) | - |
| Q-range (Å$^{-1}$) | - | 0.09 – 1.2 |
| $R_{max}$ (Å) | - | 100 |
| SAXS scaling factor | - | 0 – ∞ ($4.3 \cdot 10^{-6}$) |
| Background | - | 0 – ∞ (2.8) |

**Table S13: Refinement details the fit of the [$Bi_{37}O_{85}$] cluster extracted with ML-MotEx with the use of PDF data to the background subtracted PDF and SAXS 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 parentheses.**

*Best fitting structure found with ML-MotEx on the SAXS data from the [$Bi_{38}O_{45}$] cluster*

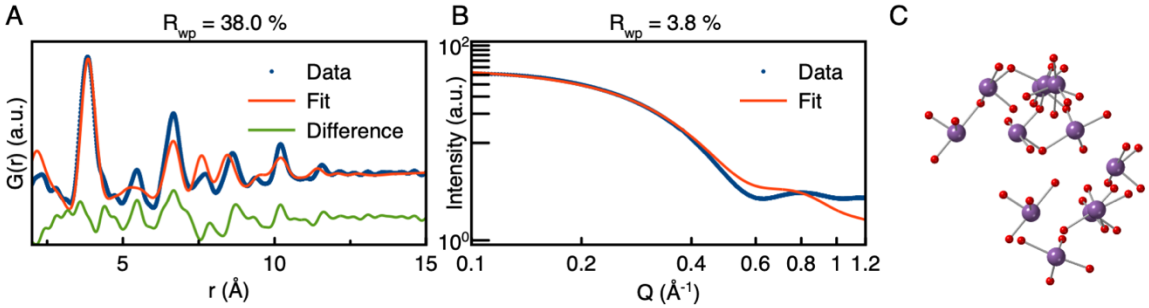



**Figure S24: Fits of the cut out of the 12 most preferred bismuth atoms of the SAXS model, A) PDF fit, B) SAXS fit, C) The structure of the 12 most preferred Bi atoms and all oxygen atoms within 2.5 Å to a Bi atom.**

| Example system | PDF | SAXS |
|---|---|---|
| r-range (Å) | 2 – 15 | - |
| Scale | 0 – 10 (1.80) | - |
| $ADP_{Bi}$ ($Å^2$) | 0 – 4 (1.8) | - |
| $ADP_O$ ($Å^2$) | Fixed to 1 | - |
| Isotropic expansion | 0.95 – 1.05 (0.99) | - |
| Q-range ($Å^{-1}$) | - | 0.09 – 1.2 |
| $R_{max}$ (Å) | - | 100 |
| SAXS scaling factor | - | 0 – ∞ ($3.3 \cdot 10^{-6}$) |
| Background | - | 0 – ∞ (0.0) |

**Table S14: Refinement details the fit of the [$Bi_{12}O_{40}$] cluster extracted with ML-MotEx with the use of SAXS data to the background subtracted PDF and SAXS 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 parentheses.**

*Best fitting structure found with ML-MotEx on the combined PDF and SAXS data from the [$Bi_{38}O_{45}$] cluster*

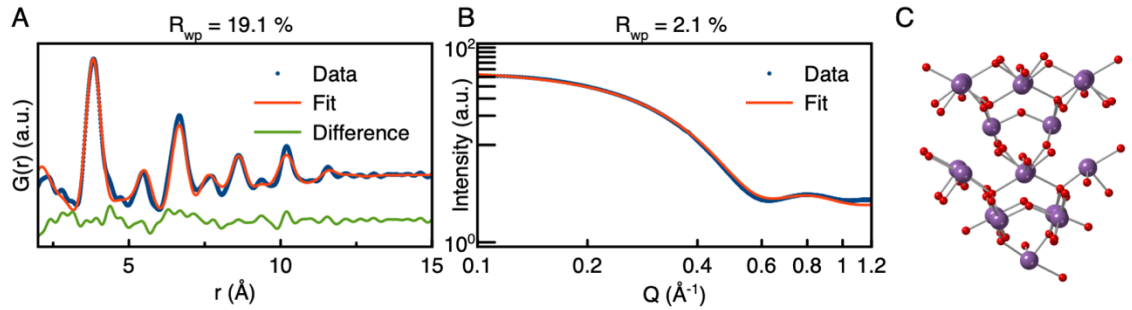



**Figure S25: Fits of the cut out of the most preferred 20 bismuth atoms of the combined PDF+SAXS model, A) PDF fit, B) SAXS fit, C) The structure of the 20 most preferred Bi atoms and all oxygen atoms within 2.5 Å to a Bi atom.**

| Example system | PDF | SAXS |
|---|---|---|
| r-range (Å) | 2 – 15 | - |
| Scale | 0 – 10 (1.29) | - |
| $ADP_{Bi}$ (Å$^2$) | 0 – 4 (1.9) | - |
| $ADP_O$ (Å$^2$) | Fixed to 1 | - |
| Isotropic expansion | 0.95 – 1.05 (0.97) | - |
| Q-range (Å$^{-1}$) | - | 0.09 – 1.2 |
| $R_{max}$ (Å) | - | 100 |
| SAXS scaling factor | - | 0 – ∞ ($1.3 \cdot 10^{-6}$) |
| Background | - | 0 – ∞ (1.6) |

**Table S15: Refinement details the fit of the [$Bi_{20}O_{56}$] cluster extracted with ML-MotEx with the use of combined PDF and SAXS data to the background subtracted PDF and SAXS 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 parentheses.**

# Appendix IV: Supplementary Information for Chapter 7

**Table S16: The number of metal/oxygen atoms in known POM structures in solution,[85] which were used as a chemical restrain of which clusters were included in the database. M is all transition metals, post-transition metals and lanthanoids.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $MO_4$ | $M_2O_7$ | $M_3O_9$ | $M_3O_{10}$ | $M_4O_{12}$ | $M_4O_{13}$ | $M_5O_{14}$ | |
| | $M_5O_{15}$ | $M_{10}O_{28}$ | $MO_2$ | $M_{14}O_{42}$ | $M_{15}O_{36}$ | $MO_{10}$ | $MO_7$ |
| | $M_2O_{10}$ | $M_6O_{19}$ | $M_7O_{22}$ | $M_{24}O_{72}$ | $M_7O_{24}$ | $M_8O_{26}$ | |
| | $M_8O_{28}$ | $M_{36}O_{112}$ | $M_{18}O_{56}$ | $M_2O_5$ | $M_2O_4$ | $MO_3$ | |
| | $M_8O_{27}$ | $M_{10}O_{34}$ | $M_5O_{23}$ | $M_6O_{26}$ | $M_{12}O_{40}$ | $M_{11}O_{39}$ | |
| | $M_9O_{34}$ | $M_{18}O_{62}$ | $M_9O_{33}$ | $M_{19}O_{67}$ | $M_9O_{37}$ | $M_6O_{18}$ | |
| | $M_{10}O_{38}$ | $M_6O_{20}$ | $M_{10}O_{32}$ | $M_{11}O_{38}$ | $M_{11}O_{40}$ | $M_5O_{19}$ | |
| | $M_6O_{22}$ | $M_{22}O_{74}$ | $M_{34}O_{116}$ | $M_{10}O_{36}$ | $M_{19}O_{69}$ | $M_{20}O_{70}$ | |
| | $M_{21}O_{71}$ | $M_{12}O_{42}$ | $M_{17}O_{61}$ | $M_{15}O_{56}$ | $M_{12}O_{48}$ | $M_{48}O_{184}$ | |

**Table S17: Instrumental parameters for PDFs used to determine if two clusters are similar. The isotropic atomic displacement parameters (ADP) have been set high to emphasize the general trends in the PDF and not the disorder.**

| |
|---|
| $R_{Range}$ = 0 – 30 Å |
| $R_{step}$ = 0.1 Å |
| $Q_{min}$ = 0.2 $Å^{-1}$ |
| $Q_{max}$ = 30 $Å^{-1}$ |
| $Q_{damp}$ = 0.01 $Å^{-1}$ |
| ADP = 1 $Å^2$ |

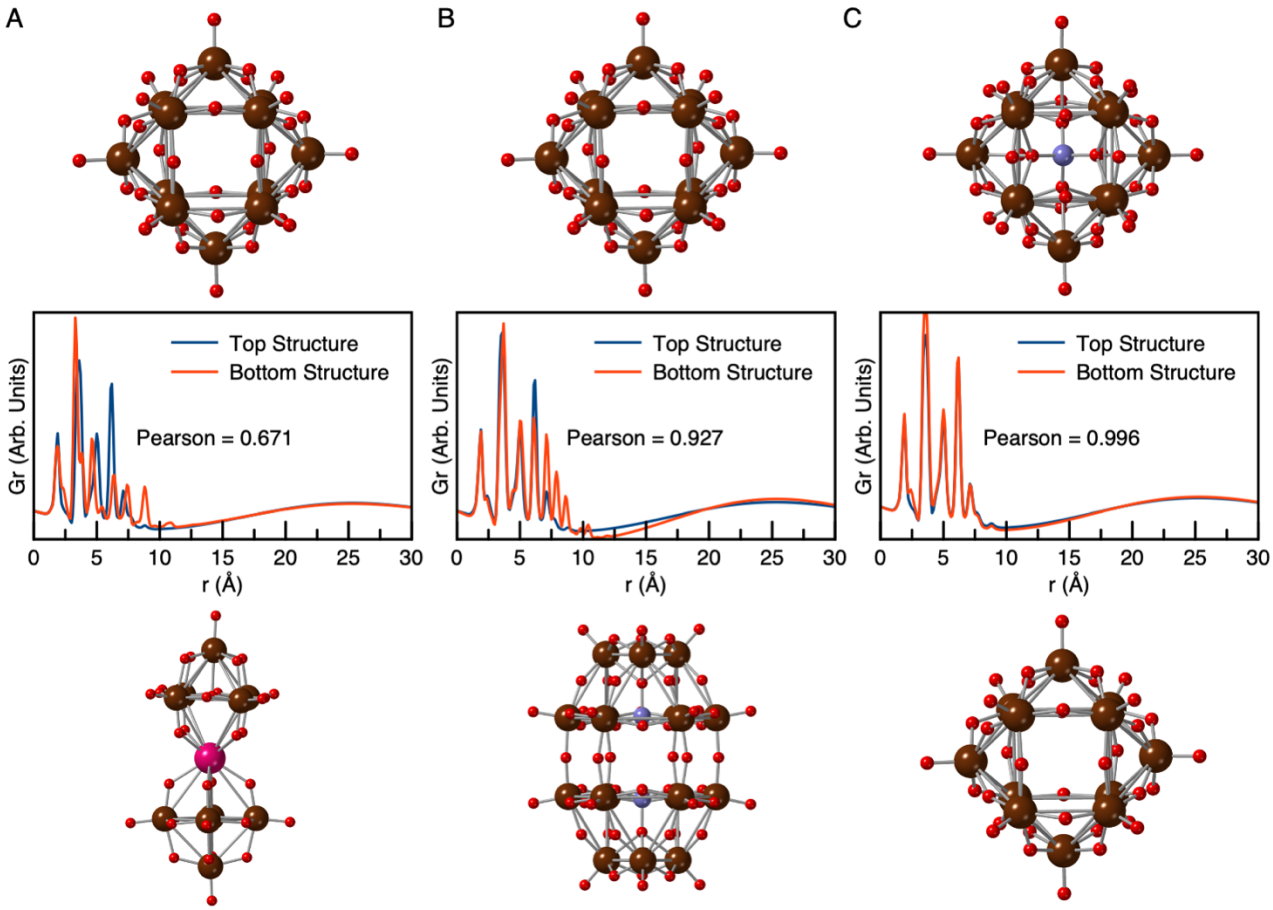



**Figure S26: Three examples of clusters with their corresponding simulated PDFs and the PDFs Pearson correlation value. Simulation parameters are given in Table S17.**

**Table S18: Parameter range for the simulated datasets. The simulated parameters were determined using Latin hypercube sampling.[84] The parameters are further described in Diffpy-CMI.[48]**

| PDF | SAXS |
|---|---|
| $R_{Range}$ = 0 – 10 Å | $Q_{Range}$ = 0 – 2 Å$^{-1}$ |
| $R_{step}$ = 0.1 Å | Background = 0 – 0.01 |
| $Q_{min}$ = 0 – 2 Å$^{-1}$ | DebyeSumRmax = $10^7$ Å |
| $Q_{max}$ = 14 – 28 Å$^{-1}$ | Gaussian Noise RMS = 0 – 0.01 |
| $Q_{damp}$ = 0.01 – 0.04 Å$^{-1}$ | |
| ADP = 0 - 2 Å$^2$ | |

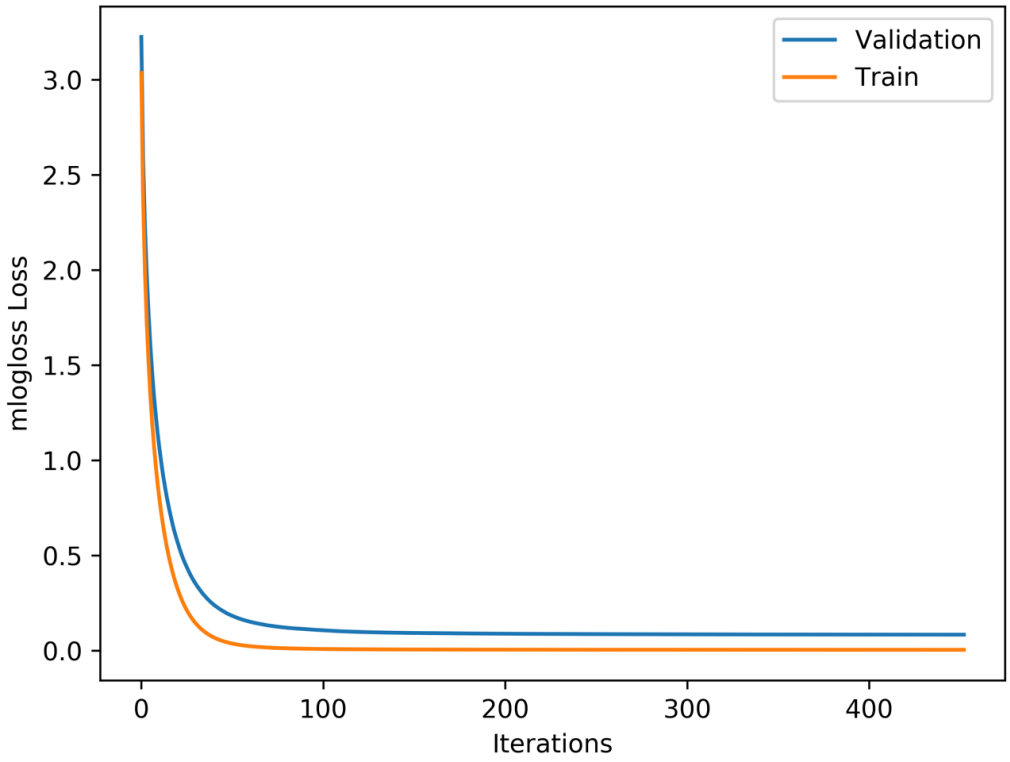



**Figure S27: Loss curve for the training of the GBDT algorithm to predict which POM cluster a PDF match. The training loss converges to zero, while the validation loss is slightly higher (overfitting). This means that the model learns to predict perfectly on previously seen datasets, while it has small errors on unknown data.**

**Table 19: The mean and standard deviation based on five iterations where the model has been trained on different simulated PDFs and predictions have been done on the same test set.**

| | **2** | **3** | **5** | **8** | **98** |
|---|---|---|---|---|---|
| **xPDF** | 43.84 ± 0.09 | 40.32 ± 0.66 | 53.95 ± 0.00 | 71.38 ± 0.09 | 93.95 ± 0.16 |
| **nPDF** | 46.95 ±0.55 | 39.37 ± 0.27 | 54.00 ± 0.36 | 73.27 ± 0.31 | 94.27 ± 0.23 |
| **ePDF** | 45.10 ± 0.33 | 41.67 ± 0.51 | 50.93 ± 0.27 | 74.27 ± 0.00 | 95.98 ± 0.17 |
| **xSAXS** | 29.12 ± 3.32 | 48.67 ± 0.44 | 60.95 ± 0.00 | 70.56 ± 0.11 | 93.59 ± 0.11 |
| **xPDF + xSAXS** | 55.12 ± 0.22 | 55.76 ± 0.00 | 75.17 ± 0.00 | 83.88 ± 0.18 | 97.02 ± 0.17 |
| **xPDF + nPDF + xSAXS** | 65.55 ± 0.11 | 65.1 ± 0.11 | 77.47 ± 0.09 | 86.68 ± 0.00 | 97.52 ± 0.32 |

# Appendix V: Published paper

The following pages contain the article published based on the results in this Thesis.

GDCh Research Articles Angewandte International Edition Chemie

Check for updates

VIP Cluster Compounds Very Important Paper

How to cite:
International Edition: doi.org/10.1002/anie.202103641
German Edition: doi.org/10.1002/ange.202103641

## Structural Changes during the Growth of Atomically Precise Metal Oxido Nanoclusters from Combined Pair Distribution Function and Small-Angle X-ray Scattering Analysis

*Andy S. Anker, Troels Lindahl Christiansen, Marcus Weber, Martin Schmiele, Erik Brok, Emil T. S. Kjær, Pavol Juhás, Rico Thomas, Michael Mehring,* and Kirsten M. Ø. Jensen**

**Abstract:** *The combination of in situ pair distribution function (PDF) analysis and small-angle X-ray scattering (SAXS) enables analysis of the formation mechanism of metal oxido nanoclusters and cluster–solvent interactions as they take place. Herein, we demonstrate the method for the formation of clusters with a $[Bi_{38}O_{45}]$ core. Upon dissolution of crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ in DMSO, an intermediate rapidly forms, which slowly grows to stable $[Bi_{38}O_{45}]$ clusters. To identify the intermediate, we developed an automated modeling method, where smaller $[Bi_xO_y]$ structures based on the $[Bi_{38}O_{45}]$ framework are tested against the data. $[Bi_{22}O_{26}]$ was identified as the main intermediate species, illustrating how combined PDF and SAXS analysis is a powerful tool to gain insight into nucleation on an atomic scale. PDF also provides information on the interaction between nanoclusters and solvent, which is shown to depend on the nature of the ligands on the cluster surface.*

### Introduction

The design of new functional materials relies on understanding the fundamental chemical reactions that govern material formation and growth. In inorganic and materials chemistry, we are still challenged in describing these processes on an atomic level,[1] as studies of nucleation and growth phenomena are challenging. In recent years, atomically precise metal oxido nanoclusters of titanium,[2] cerium[3] and bismuth have been intensively studied due to their wide range of applications, for example, for photocatalysis,[2] oxygen storage,[3] in medicine,[4] as radiopaque materials[4b,5] or as building blocks for advanced catalysts.[6] Apart from their technological relevance understanding the solution chemistry of such metal oxido clusters on an atomic and molecular scale can open new opportunities for synthesizing nanoscale metal oxides in a controlled manner.[6,7]

We have in recent years focused on the chemistry of bismuth oxido clusters and have synthesized a range of differently sized molecules and nanoclusters.[4a,8] These nanoclusters are generally composed of a bismuth oxido core $[Bi_xO_y]^{z+}$ of varying nuclearity and charge, and anionic ligands (e.g. nitrate, carboxylates, aryl sulfonates, silanolates) for charge compensation. In the following, we will refer to the structures of the nanoclusters as $[Bi_xO_y]$ neglecting charge and ligand shell. The nanoclusters can be considered built from $[Bi_6O_8]$ units which share edges (as defined from the octahedra formed from Bi atoms, Figure 1 A,B) to form larger clusters. These clusters represent cut-outs of the fluorite-type structure which constitutes the basis of δ- and β-$Bi_2O_3$.[6,9] For example, the $[Bi_{22}O_{26}]$ cluster[8e] (Figure 1 C) is composed of 6 of these basic motifs and the $[Bi_{38}O_{45}]$ cluster[4a] (Figure 1 D) of 13. Similar structures are seen in metal oxido clusters of other large cations such as $[Ce_{22}O_{24}]$,[10] $[Ce_{38}O_{54}]$,[3a] $[U_{38}O_{56}]$,[11] and $[Pu_{38}O_{56}]$.[12] The $[M_6O_x]$ building block, where M is a metal, is therefore anticipated to play an important role in the cluster growth, however, an understanding of the mechanisms controlling this process in solution is still missing. In situ investigations have shed light on cluster reactions in solution, and the cluster growth from $[Bi_6O_8]$ to $[Bi_{38}O_{45}]$ has previously been investigated with electrospray ionization mass spectrometry (ESI-MS).[13] A variety of bismuth oxido clusters were observed during the reaction, but $[Bi_6O_8]$, $[Bi_{22}O_{27}]$ and $[Bi_{38}O_{45}]$ clusters were

[*] A. S. Anker, Dr. T. L. Christiansen, E. T. S. Kjær, Dr. K. M. Ø. Jensen
Department of Chemistry and Nano-Science Center
University of Copenhagen
Universitetsparken 5, 2100 Copenhagen E (Denmark)
E-mail: kirsten@chem.ku.dk
Dr. M. Weber, R. Thomas, Prof. Dr. M. Mehring
Fakultät für Naturwissenschaften, Institut für Chemie
Professur Koordinationschemie, Technische Universität Chemnitz
Strasse der Nationen 62, 09111 Chemnitz (Germany)
and
Center for Materials, Architectures and Integration of Nanomembranes (MAIN)
Rosenbergstrasse 6, 09126 Chemnitz (Germany)
E-mail: michael.mehring@chemie.tu-chemnitz.de
Dr. M. Schmiele, Dr. E. Brok
Niels Bohr Institute and Nano-Science Center
University of Copenhagen
Universitetsparken 5, 2100 Copenhagen E (Denmark)
Dr. P. Juhás
Computational Science Initiative, Brookhaven National Laboratory
98 Rochester Street, Upton, NY 11973 (USA)
Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.202103641.

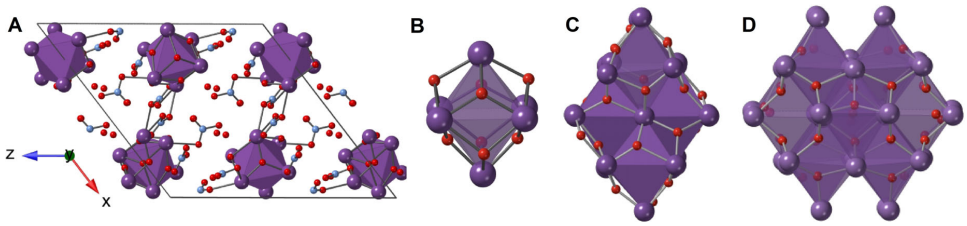



***Figure 1.*** A) Structure of crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$.[8a] Hydrogen atoms have been omitted for clarity. B) An octahedral $[Bi_6O_8]$ unit. C) The $[Bi_{22}O_{26}]$[8e] cluster structure. D) The $[Bi_{38}O_{45}]$[4a] cluster structure. In all structures, bismuth is shown in purple, oxygen in red, and nitrogen in blue.

identified as particularly stable, which agrees well with other cluster studies.[3a,9a,11,12,14] From these experiments, Sattler et al.[13] proposed that cluster growth happens through di- or oligomerization reactions and rearrangement processes, with cluster growth starting from hexanuclear metal oxido clusters. Based on the ESI-MS findings, as well as on knowledge on crystal structures of polynuclear clusters (containing 6–10 Bi atoms), the reaction is expected to proceed through intermediate species formed by addition of bismuth containing fragments of dissociated $[Bi_6O_8]$ units to the cluster core.[15] However, while mass spectrometry gives information about the atomic composition of the species present during growth, the evolution of the atomic structure in solution is yet to be unraveled.

The molecular structures of many bismuth oxido nanoclusters are well known, as these can be characterized with single crystal diffraction. The structures of the clusters shown in Figure 1 have all been determined from single crystal X-ray diffraction studies. However, characterizing cluster structures directly in solution, as is needed for in situ studies, is more challenging. Extended X-ray absorption fine structure (EXAFS) has been used to characterize the atomic structure of some small clusters[14a,16] and limited structural information can be obtained from Nuclear Magnetic Resonance (NMR) spectroscopy.[16b] However, these methods only provide information on the very local atomic range. Here, we use in situ X-ray total scattering (TS) and pair distribution function analysis[17] (PDF) to follow the formation of $[Bi_{38}O_{45}]$ from $[Bi_6O_8]$ as a model reaction for nanocluster formation and growth. As the Fourier transform of the TS signal, the PDF represents a histogram of all interatomic distances in a sample, and thus provide atomic structural information from the local to the global range. PDF is an excellent technique for in situ studies of materials formation in solution,[18] and has been applied to study a few large cluster systems previously[12a 16b] in order to characterize their solution structure, but not in situ to follow cluster reactions. While PDF analysis offers information on the atomic structure of matter, it is less sensitive to for example, cluster or particle size and shape. A way to complement the structural information available in the PDF is to combine the technique with small-angle X-ray scattering (SAXS), which provides information on the size, morphology and size dispersion of clusters or particles.[19]

We here use PDF and SAXS to study the formation of the $[Bi_{38}O_{45}]$ cluster starting from $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$[8a] in dimethyl sulfoxide (DMSO) and obtain detailed information on the reaction pathway from octahedral $[Bi_6O_8]$ units to $[Bi_{38}O_{45}]$ nanoclusters. To analyze the data, we have developed an automated modeling approach[20] to identify intermediates and the reaction pathway. This automated approach, where all relevant smaller cluster structures, based on the $[Bi_{38}O_{45}]$ framework, is tested against the experimental PDF, allows us to readily identify the size and structure evolution during the reaction. Our results clearly identify $[Bi_{22}O_y]$ as a metastable intermediate structure in the reaction, and we do not observe $[Bi_6O_8]$ units as a reaction intermediate on the time scale of our experiments. We also demonstrate how PDF can be used to characterize restructuring effects of the solvent at the surface of nanoclusters and show that the cluster–solvent interaction affects the $[Bi_{38}O_{45}]$ clusters themselves, as it introduces disorder in their structure. This effect is highly dependent on the presence and type of the ligand covering the clusters. The methods and modeling tools developed here, using automated and combined SAXS and PDF analysis thus provides insights into cluster chemistry in solution.

## Results and Discussion

Figure 2A shows in situ X-ray total scattering data obtained during the formation of $[Bi_{38}O_{45}]$ from $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ dissolved in DMSO. The experiment was done at $T = 80\,°C$, and data sets obtained from similar experiments performed at lower temperatures ($T =$ 30 °C–60 °C) can be found in Figure S1. At the beginning of the reaction, Bragg peaks from the crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ phase (Figure 2D) is seen, as crystals at this point are still in suspension in DMSO. After approximately 7 min, the crystals have fully dissolved, and only diffuse scattering is seen. We therefore turn to PDF for further structural analysis, and the corresponding PDFs are plotted in Figure 2B. The PDFs are the Fourier transformation of the TS data, where the signal from DMSO has been subtracted, as illustrated in Figure S2. In the beginning of the experiment (Figure 2E), the PDF shows the presence of long-range order from the crystalline starting material, and as expected, the data can be fitted with the $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ structure. Towards the end of the experiments the long-range order disappears, and the PDFs now show peaks up to only about 12 Å.

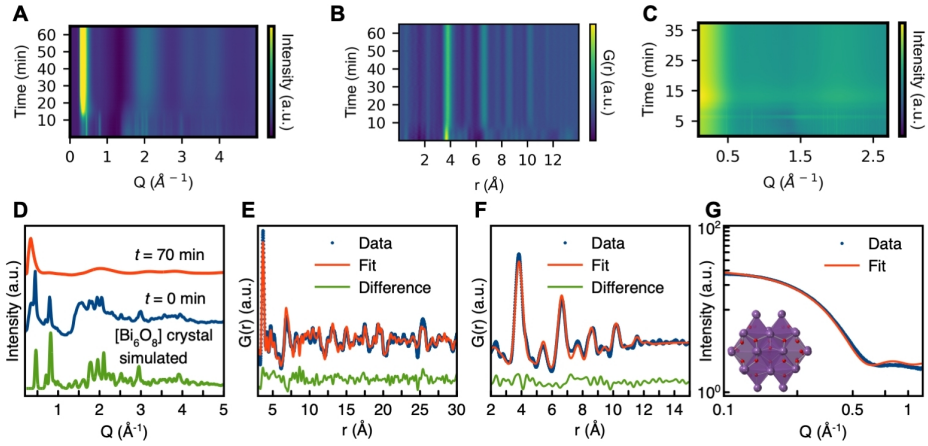



***Figure 2.*** A) Background subtracted in situ X-ray TS data obtained during the experiment performed at $T = 80\,°C$ plotted as a function of time. While Bragg peaks are seen at the beginning of the reaction, only diffuse features are seen after dissolution. B) Corresponding PDFs plotted as a function of time. C) Background subtracted SAXS data obtained during a similar experiment plotted as a function of time. The intensity is shown on a logarithmic scale. D) Simulated X-ray scattering data of the $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3\,H_2O$ crystal structure, compared to background subtracted data from $t = 0$ min and $t = 70$ min, $T = 80\,°C$. E) Fit of the $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3\,H_2O$ crystal structure to the PDF obtained from the first frame in the experiment at $t = 0$ min. $R_{wp} = 29.0\,\%$. F) Fit of the $[Bi_{38}O_{45}]$ cluster structure to the PDF obtained at the end of the experiment at $t = 70$ min; $R_{wp} = 16.6\,\%$. G) Fit of the $[Bi_{38}O_{45}]$ cluster structure to the SAXS data obtained at the end of the experiment at $t = 37$ min; $R_{wp} = 3.1\,\%$. The SAXS data are plotted in a log–log plot.

In Figure 2F, the PDF obtained at the end of the experiment is fitted with the $[Bi_{38}O_{45}]$ model, confirming the product of the reaction. The cluster model (Figure 1D) was obtained from the structure of crystalline $[Bi_{38}O_{45}(NO_3)_{20}(DMSO)_{28}](NO_3)_4\cdot 4\,DMSO$,[4a] where an isolated $[Bi_{38}O_{45}]$ cluster was cut out of the crystal structure. The modeling (described in detail in Table S1) was set up so that the basic atomic structure was kept fixed during the refinement. The excellent fit illustrates that the structure of the $[Bi_{38}O_{45}]$ cluster in solution is similar (if not identical) to that seen in the crystalline phase, as also confirmed from DLS measurements (see Figure S3).

The presence of the $[Bi_{38}O_{45}]$ cluster at the end of the experiment is further supported by SAXS experiments. In situ SAXS data from a similar experiment are shown in Figure 2C. Bragg peaks from the crystalline $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3\,H_2O$ phase can again be identified at the beginning of the reaction (Figure 2C) but disappear after approximately 10 min. Figure 2G shows the SAXS data obtained at the end of the experiment fitted with the $[Bi_{38}O_{45}]$ cluster using the Debye equation as described in detail in Table S2. Again, the fits show excellent agreement between data and model.

Having established that we can use PDF and SAXS to confirm the starting point and end point of the reaction, we now move to characterize the structures seen during the cluster growth and identify intermediate species. Figure 3A shows selected PDFs from the process, clearly illustrating that a smaller cluster than $[Bi_{38}O_{45}]$ is present at the beginning of the reaction after dissolution. Initially, we had expected to see the presence of isolated $[Bi_6O_8]$ clusters in solution, since it is described as particularly stable.[9a,13] The calculated PDF from a $[Bi_6O_8]$ cluster is shown in Figure 3A showing only a few peaks from the small structure. Surprisingly, we cannot

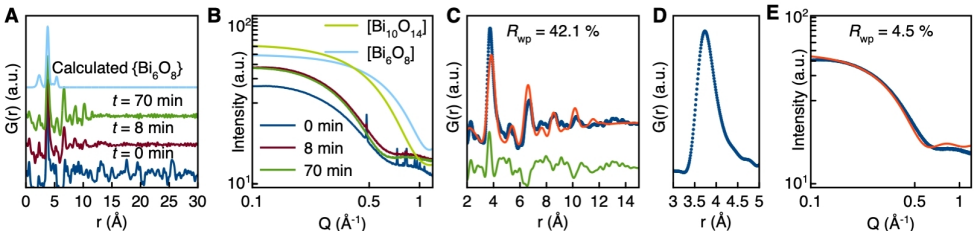



***Figure 3.*** A, B) Comparison of experimental PDFs (A) and SAXS data (log–log plot) (B) with calculated patterns from an isolated $[Bi_6O_8]$ unit and a $[Bi_{10}O_{14}]$ "dimer". C) Fit of the $[Bi_{38}O_{45}]$ cluster structure to the PDF obtained after the dissolution of the crystalline starting material; $T = 80\,°C$, $t = 8$ min. D) Asymmetric peak at 3.8 Å in the PDF. E) A log–log plot of the fit of the $[Bi_{38}O_{45}]$ cluster structure to the SAXS data obtained after the dissolution of the crystalline precursor; $T = 80\,°C$, $t = 16$ min.

identify a PDF similar to this at any point in our in situ data from any reaction temperature, as peaks to higher $r$ values are seen immediately after dissolution has occurred. This indicates that the small $[Bi_6O_8]$ clusters are not stable in DMSO on the time scales and experimental conditions of the in situ X-ray scattering experiments. This observation is confirmed by SAXS (Figure 3B), where isolated, monodisperse $[Bi_6O_8]$ clusters are not seen at any point in our data. We also do not observe the presence of for example, $[Bi_{10}O_{14}]$ structures (two edge-sharing $[Bi_6O_8]$ units) or any smaller fragments. While we cannot rule out that such structures are present in very small concentrations in the samples, they are not main species.

$[Bi_6O_8]$ was observed as a very dominant species in in situ ESI-MS experiments.[13] However, the conditions used for the ESI-MS measurements were different from the current scattering studies, as the hydrolysis reactions in the ESI-MS experiments were followed at room temperature, at much lower Bi concentration (0.05 mM–1 mM for ESI-MS vs. 230 mM for PDF/SAXS), and in a different solvent, namely a mixture of DMSO and dichloromethane, $CH_2Cl_2$. These highly different conditions may explain the differences in the species observed. However, ESI-MS is also more sensitive to minor species than both PDF and SAXS, although they will only be detected if they can easily be charged. The two experiments are thus not in contradiction but provide information for different conditions and different time scales: While the ESI-MS studies show that $[Bi_6O_8]$ species form rapidly after dissolution, their absence from our in situ X-ray scattering data indicate very low solubility and thus low concentration of these species and fast condensation after dissolution in DMSO. Both experiments show the formation of $[Bi_{22}O_{26}]$ clusters as quite stable intermediate and $[Bi_{38}O_{45}]$ as final products.

We can now analyze the structure of the intermediate species. From the similar PDF peak positions between the intermediate structure ($t = 8$ min) and the final cluster ($t =$ 70 min) seen in Figure 3A, it is evident that the intermediate is structurally related to the $[Bi_{38}O_{45}]$ cluster, as many of the main peaks show up in both PDFs. However, the peaks representing the largest interatomic distances are not present in the PDF from the intermediate cluster, which also appears to be significantly more disordered, as seen from the asymmetric broadening of the first Bi···Bi peak at $r = 3.8$ Å in Figure 3D. Figure 3C shows a fit of the $[Bi_{38}O_{45}]$ structure to the PDF from the intermediate cluster. In Figure 3E, the SAXS fit of the $[Bi_{38}O_{45}]$ cluster to data representing the intermediate confirm that the cluster is smaller than the final $[Bi_{38}O_{45}]$.

To characterize the cluster size evolution during the reaction, we first analyze the in situ SAXS data with a form factor analysis. Details on the modeling and examples of fits are given in Table S3 and Figure S4. The results (Figures S5 and S6) show that the average cluster size increases during the reaction, and the clusters present are of a higher size dispersity at the beginning of the reaction than at the end, where the presence of stable $[Bi_{38}O_{45}]$ clusters results in a very low size dispersity.

The observation that the intermediate cluster is smaller but structurally close to the $[Bi_{38}O_{45}]$ structure agrees well with the chemistry of bismuth oxido clusters discussed above, as several different sizes of bismuth oxido clusters, all built up from $[Bi_6O_8]$ units can be synthesized. We therefore attempted to fit other known cluster structures, with sizes between $[Bi_6O_8]$ to $[Bi_{38}O_{45}]$, to the intermediate PDF as shown in Figure S7. The best cluster candidates were the $[Bi_{18}O_{36}]$ and $[Bi_{22}O_{38}]$ structures, agreeing well with observations from in situ ESI-MS analysis, molecular dynamics simulations and quantum chemical calculations.[13,15a,b] However, to look further into the time-dependent structural changes, and to probe a larger structural space, we developed an automated method, where all smaller cluster structures based on the $[Bi_{38}O_{45}]$ framework were tested against the experimental PDF. Automated modeling methods for PDF analysis have recently been proposed by Banerjee et al.[21] for analysis of metal nanoparticles. In a recent study, we developed a method, where structural motifs in molybdenum oxide structures were identified by fitting automatically generated fragments of know polyoxometalate structures to PDFs.[20] We here apply an extension of this latter method to in situ data in order to identify intermediates and reaction pathways.

The principle is illustrated in Figure 4. First, the $[Bi_{38}O_{45}]$ model is fit to the PDF obtained from the last frame in the experimental series. It was thereafter tested whether a smaller version of the structure, i.e., with some Bi and O atoms removed would result in a better fit to the data, as described in more detail in the Supporting Information, section I.

This method can give us an overview of the best fitting structure (built up from edge-sharing $[Bi_6O_8]$ units) for every frame obtained during the reaction. Figure 5A–C shows the results from the automated modeling described above, where the refined number of Bi atoms is plotted as a function of time for the in situ experiments done at $T = 80$ °C, 60 °C, and 30 °C. Results from experiments done at other temperatures are given in Figure S8. Figure 5A–C firstly show that larger structures appear to be stabilized at higher reaction temperatures, and that the clusters grow with time. In Figure 5A ($T =$ 80 °C) plateaus at approximately 32 Bi atoms and 28 Bi atoms are identified, labelled as I and II. The same plateau at 28 Bi atoms is seen for the data obtained at $T = 60$ °C (Figure 5B), whereas data from $T = 30$ °C (Figure 5C) show plateaus at 25 and 20 Bi atoms (III and IV). The occurrence of clusters of a given size is illustrated in Figure 5D, which shows a histogram of the cluster structures identified in all three datasets. This shows that a large proportion of the identified clusters have between 20 and 25 Bi atoms. In Figure 6E–H, we show the structures I–IV along with fits to selected data. While these structures fit the data well, they appear unphysical, as "dangling" Bi atoms are seen. However, when looking closer, we identify that they share a stable backbone of 22 Bi atoms. This is particularly clear for cluster III and IV, as seen in Figure 6A, where the Bi positions of these are compared to the $[Bi_{22}O_{26}]$ structure, and fits of the $[Bi_{22}O_{26}]$ cluster to the same frames (Figure 6B,C) are reasonable.

The results from the automated modeling thus indicate that $[Bi_{22}O_{26}]$ is an important intermediate in the reaction. We also saw other frequent sizes, such as $Bi_{28}$ and $Bi_{32}$. However, these structures do not appear as closed-shell, physical structures, and are unlikely to be stable. The SAXS form

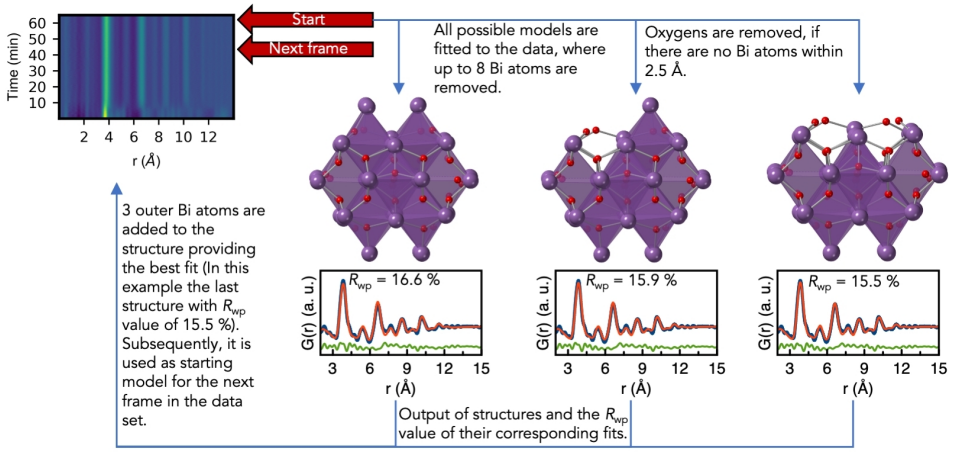



***Figure 4.*** Sketch of the automated modeling process. The last frame of the in situ dataset is modeled with the $[Bi_{38}O_{45}]$ cluster structure and all structures in which up to 8 Bi atoms from the outer shell have been removed together with nonbonding oxygen atoms. The best fitting structure (with three Bi atoms added) is used as a starting point for the fitting of the second-last frame. This process is repeated for all frames in the reaction.

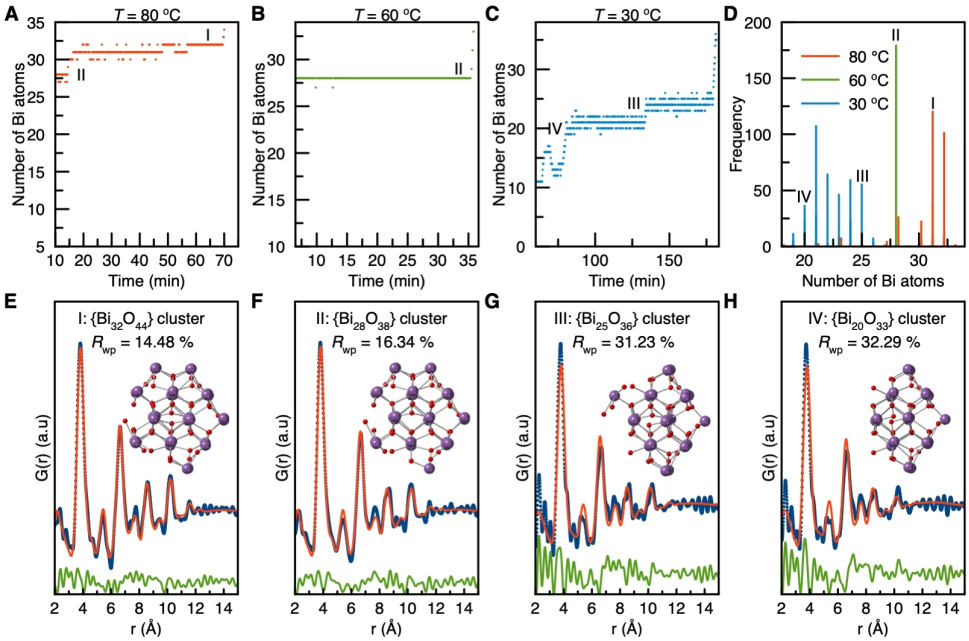



***Figure 5.*** A–C) The number of bismuth atoms in the best fitting structure for each frame is plotted versus reaction time for the experiments done at $T = 80$, 60, and 30 °C. D) Histogram representing the number of Bi atoms in the best fitting structure through the reaction. Frequency is the number of times the specific cluster size is present during the reaction. E–H) Clusters and fits that have been obtained in the automated modeling process.

factor analysis described above (Figures S5 and S6) furthermore showed that before the reaction finished, a larger size dispersion is seen, thus indicating that multiple cluster sizes can be present during the reaction. We therefore tried fitting both the PDF and SAXS data with a two-structure model with the $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ clusters. The refined parameters are given in Tables S4–7 and the fits to both SAXS and PDF data are shown in Figure S9, where they are compared to the

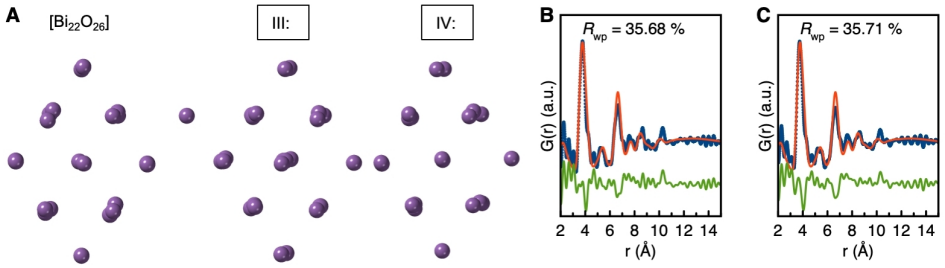



***Figure 6.*** A) Comparison of the Bi positions in the $[Bi_{22}O_{26}]$ cluster with those of structures III and IV, found from the automated modeling method. B, C) Fit to the in situ PDF ($T = 30\,°C$) obtained at $t = 152$ min (B) and $t = 103$ min (C) using the $[Bi_{22}O_{26}]$ cluster as the model.

results from the automated modeling. In section K in the Supporting Information, we compare simulated SAXS data of the $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ clusters to data at two different time points in the reaction and compare two-phase SAXS refinements to fits using individual $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ clusters. Generally, the two-structure models provide similar fit qualities or better.

Having established that the main species present throughout the reaction are $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$, we can sequentially analyze the data to extract phase fractions of the two clusters as a function of time. The results from this analysis are shown in Figure 7 (see Figure S12 for results from other reaction temperatures). Results from SAXS is shown in blue, and PDF in red, illustrating how the $[Bi_{38}O_{45}]$ fraction increases with function of time and temperature. Analysis of SAXS and PDF data independently show similar phase fractions of the two clusters in the high temperature cases, and the amount of the $[Bi_{38}O_{45}]$ cluster steadily increases. We therefore attempted a combined analysis method, that is, using the so-called complex modeling approach.[19b–d,22] Complex modeling using SAXS and PDF analysis has been exemplified for CdS nanoparticles, where Farrow et al.[19b] showed that by combining the two techniques, a much more robust description of particle size and shape could be obtained. Here, we have in a similar way used the program Diffpy-CMI for complex modeling of PDF and SAXS data. However, we have applied the Debye equation to calculate the scattering pattern in both the SAXS regime and PDF regime, as described in detail in the Supporting Information, sections L and M. The results of the combined modeling are shown in green in Figure 7 and generally agree well with the individual SAXS and PDF analysis, but with more certainty on the refined parameters as seen from the less scattered data points.

While the results from the experiments done at higher temperatures ($T = 50\,°C$–$80\,°C$) generally show clear trends and a good agreement between the SAXS and PDF results, larger discrepancies are observed for the data seen from experiments done at lower temperatures, illustrated for $T = 30\,°C$ in Figure 7C. Here, the PDF results are somewhat scattered, while the SAXS refinements appear to give a clearer trend. The dispersity analysis illustrated in Figures S5 and S6 furthermore showed a larger dispersion of sizes for low temperature experiments, and it may be that small fractions of other cluster structures or fragments are present for longer times during these slower reactions.

The data indicate that in none of the experiments, a full reaction to $[Bi_{38}O_{45}]$ takes place, as some $[Bi_{22}O_{26}]$ is present at the end of the experiment. The refined fraction of the final $[Bi_{38}O_{45}]$ cluster is approximately 60 % at the end of the in situ experiment at $T = 80\,°C$, and for the reactions done at lower temperatures, the $[Bi_{22}O_{26}]$ cluster remains the majority species in the solution. However, PDF and SAXS data collected from samples kept up to 275 days at room temperature show excellent fits with a single phase $[Bi_{38}O_{45}]$ model (Figures S16 and S18).

Combined, the results from the automated PDF analysis and the complex modeling of SAXS and PDF data allow us to establish the overall reaction route. When $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$ is dissolved in DMSO, larger clusters quickly form, and we never observe free $[Bi_6O_8]$ (or smaller structure fragments) in our data at a time resolution

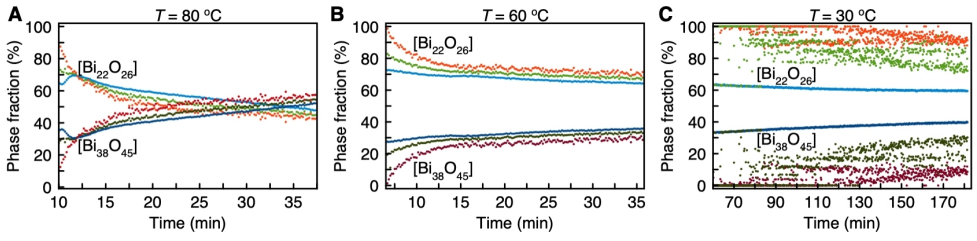



***Figure 7.*** Phase fractions of the $[Bi_{22}O_{26}]$ and $[Bi_{38}O_{45}]$ clusters, plotted as a function of time for experiments at A) $T = 80\,°C$, B) 60 °C, and C) 30 °C. Results from PDF are plotted in red, from SAXS in blue, and from complex modeling in green.

of 4 s. This process most likely takes place through condensation reactions and rearrangement processes, as also suggested by Sattler et al.[13] The main intermediate species we observe is $[Bi_{22}O_{26}]$, as concluded from both SAXS and PDF data. This cluster is quite stable and co-exists with $[Bi_{38}O_{45}]$ in the time range investigated with in situ experiments. The local structure changes taking place during dissolution are discussed further in the Supporting Information, section O. The experiments discussed so far were all done from $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3H_2O$. Nearly identical results were obtained when the starting crystalline material was changed to $[Bi_6O_4(OH)_4(NO_3)_6(H_2O)_2]\cdot(H_2O)$ (see Figure S18). This indicates that the results presented are general for compounds based on a hexanuclear bismuth oxido core $[Bi_6O_8]$ dissolved in DMSO.

Using SAXS and PDF, we have identified the main intermediate species in the process. Nevertheless, some questions regarding the cluster growth remain unanswered from the current study. As discussed above, it has previously been suggested that cluster growth takes place through oligomerization reactions involving fragments of $[Bi_6O_8]$ units as key species. Our data and time resolution do not allow us to identify these short-lived species or the specific mechanism. Several processes may take place in the reaction: for example, elusive, larger intermediates assembled from $[Bi_6O_8]$ units may form, which could collapse or partly dissociate to result in the $[Bi_{22}O_{26}]$ structure. Dissociation of $[Bi_6O_8]$ fragments could also take place before oligomerization to lead stepwise to $[Bi_{22}O_{26}]$ and finally to $[Bi_{38}O_{45}]$. This hypothesis is supported by ESI-MS studies of the hydrolysis of bismuth carboxylates, which revealed intermediates between $Bi_3$- and $Bi_{13}$-oxido clusters. Independent analytic techniques to support their formation in solution are missing, but several single crystal X-ray structures of other polynuclear bismuth oxido clusters including at least one $[Bi_6O_8]$ unit with various ligands have been reported.[9a,15c] The time resolution for the scattering experiments currently possible limits our insight into the fast processes in solution; for example, molecular dynamic simulations have suggested that the growth from $[Bi_6O_8]$ units take place on the nanosecond to microsecond time scale.[15a] In the future, new possibilities for ultrafast, high flux X-ray scattering studies may open for studies of such processes.[23] If one could chemically stabilize a series of species formed on the way to the $[Bi_{22}O_{26}]$ clusters, new insight could also be obtained.

### *Ligand Exchange: From $[Bi_{38}O_{45}(NO_3)_{20}(DMSO)_\delta]$ to $[Bi_{38}O_{45}(OMc)_{24}(DMSO)_\delta]$*

PDF analysis also allows us to study cluster–solvent interactions and any structural changes that may take place during exchange of ligands in the clusters. Figure 8A shows PDFs obtained when heating a solution of $[Bi_{38}O_{45}(NO_3)_{20}(DMSO)_{28}](NO_3)_4\cdot 4DMSO$ in DMSO with sodium methacrylate. In this process, nitrate ligands are substituted with methacrylate[5,9a] making the clusters soluble in organic solvents. After 7 min of the reaction, significant changes appear. Figure 8B shows selected PDFs obtained over the course of the reaction. At the beginning, both the peaks at $r \approx 3.8$ Å and $r \approx 6.7$ Å have shoulders which disappear through the ligand exchange, and all of the peaks become narrower and more symmetric as the reaction progresses.

We start by modeling the first frame of the reaction with the $[Bi_{38}O_{45}]$ cluster structure, Figure 9A. While this model fits all main PDF peaks, significant misfits are seen, and the difference curve is similar to an exponentially dampened sinusoidal function. This behavior has previously been described by Zobel et al., who related it to solvent restructuring effects on the surface of nanoparticles.[24] Figure 9B shows the fit if a dampened sinusoidal function is added to the $[Bi_{38}O_{45}]$ model. This significantly improves the fit to the data, suggesting a strong interaction between the cluster (or its ligands) and the DMSO solvent. A PDF obtained from pure DMSO is shown in Figure S19 confirming that the oscillation does not arise from the solvent itself.

At the end of the reaction, the data (Figure 9D) is very well described by the $[Bi_{38}O_{45}]$ cluster, giving a $R_{wp}$ value of 12.2%. Here, the exponential dampened sinusoidal does not contribute to the fit as seen in Figure 9E. Figure 9C (bottom) shows the $R_{wp}$ values of the fits during the reaction. This behavior indicates that the structuring of the solvent dominates at the beginning of the reaction but when the ligand is exchanged to methacrylate, the effect vanishes. This is also clear when plotting the contribution of the dampened sinusoidal oscillation through the reaction, Figure 9F.

The ligand exchange, and the solvent interactions also appear to affect the atomic structure of the bismuth oxido clusters. Figure 9C (top) shows the refined ADPs of the Bi atoms through the reaction. This parameter relates to the width of the PDF peaks, and decreases through the reaction,

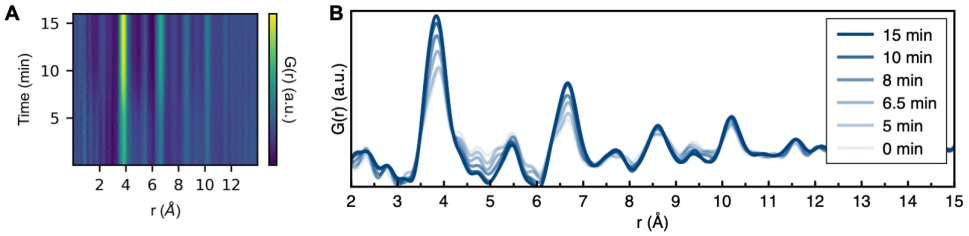



***Figure 8.*** A) PDFs obtained from in situ X-ray TS data collected during the ligand-exchange experiment performed at $T = 80$ °C, plotted as a function of time. B) Selected PDFs from the experiment.

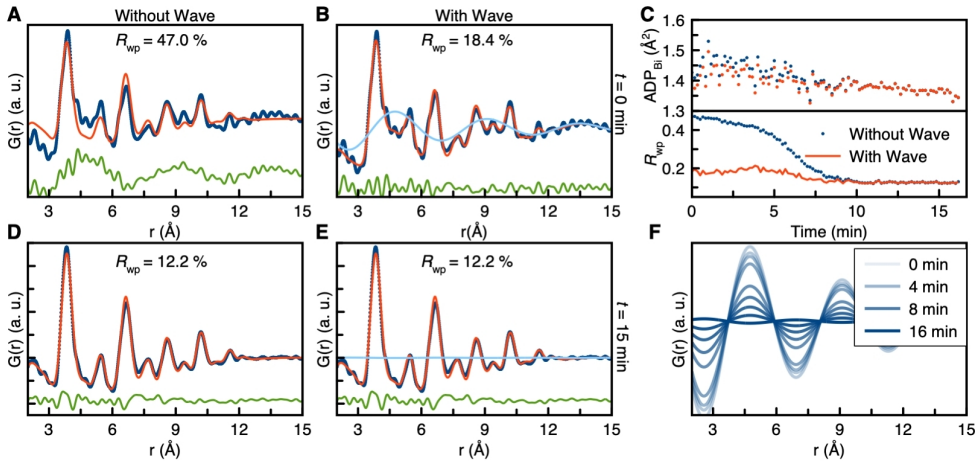



***Figure 9.*** Representative fit of a PDF obtained during ligand exchange, obtained $t = 0$ min into the ligand-exchange reaction A) without including solvent restructuring effects (wave), B) including solvent restructuring effects (wave). C) $R_{wp}$ values and ADP values of Bi obtained when modeling the time-resolved PDFs, obtained during ligand exchange, with the $[Bi_{38}O_{45}]$ model. D) Representative fit of a PDF obtained during ligand exchange, obtained $t = 15$ min into the reaction without including solvent restructuring effects and E) when including solvent restructuring effects at the surface of the nanocluster. F) The contribution of the solvent restructuring effects during the ligand-exchange reaction.

showing that the ligand-exchanged clusters with methacrylate are more ordered. The PDFs obtained from the beginning of the reaction furthermore show highly asymmetric peaks, as seen most clearly from the shoulder peak at about 4.2 Å. This asymmetry is likely to be an effect of the strong interaction between the cluster and the DMSO solvent. Modeling details are given in Table S8,9.

The solvent restructuring in the beginning of the reaction is likely due to partial dissociation of nitrate ligands, which results in a cationic cluster, making the coordination to DMSO strong. This can result in some distortion of the $[Bi_{38}O_{45}]$ core as seen from PDF peaks asymmetry. When the nitrate ligand is substituted with methacrylate, the strong bonding between methacrylate and the $[Bi_{38}O_{45}]$ clusters hamper the coordination of DMSO to the nanoclusters, and the effect in the PDF vanishes. This is also reflected when considering the crystal structures of $[Bi_{38}O_{45}(NO_3)_{20}(DMSO)_{28}](NO_3)_4{\cdot}4DMSO$[4a] and $[Bi_{38}O_{45}(OMc)_{24}(DMSO)_9]{\cdot}2DMSO{\cdot}7H_2O$:[5] For the nitrate-coordinated cluster, 28 DMSO molecules coordinate to the bismuth oxido core, while this is only 9 for the crystal structure containing methacrylate.

A sketch of how the solvent restructures at the surface of the $[Bi_{38}O_{45}]$ clusters is illustrated in Figure 10. The DMSO molecules coordinate to the nanocluster through strong bonding to the surface of the bismuth oxido cluster with directions perpendicular to the surface of the $[Bi_{38}O_{45}]$ clusters. This results in regions with alternating low and high electron density due to packing of solvent molecules. Collectively, this gives rise to a sinusoidal oscillation contributing to the PDF. Since the solvent effect is strongest close to the surface of the $[Bi_{38}O_{45}]$ clusters and it slowly transforms into a bulk solvent structure of DMSO, the sinusoidal oscillation exponentially decays. However, we observe solvent effects of up to $r = 30$ Å.

## Conclusion

In situ PDF and SAXS have been used to follow the reaction from a suspension of $[Bi_6O_5(OH)_3(NO_3)_5]{\cdot}3H_2O$ in DMSO to $[Bi_{38}O_{45}]$ clusters in solution (Figure 11). The structural modeling was done using the Debye equation, and we applied combined modeling of the PDF and SAXS data. To follow the cluster growth taking place during the reaction, we introduced an automated modeling method, where intermediate species were identified by iteratively removing atoms from the final $[Bi_{38}O_{45}]$ structure. This analysis showed that starting from $[Bi_6O_8]$ a $[Bi_{22}O_{26}]$ intermediate very quickly forms and slowly transforms to $[Bi_{38}O_{45}]$. The formation of the $[Bi_{22}O_{26}]$ structure unexpectedly proceeded immediately after dissolution of the crystalline starting material, meaning that the time resolution of the X-ray scattering experiments does not allow us to identify any short-lived building blocks that may form before the larger clusters are seen. Questions regarding e.g., oligomerization reactions and cluster fragmentation thus remain open, but the combination of SAXS and PDF provided new insight into the cluster growth process of bismuth oxido species including information on the time scale at which processes occur.

We furthermore demonstrate that due to nitrate ligand dissociation, a strong interaction between DMSO and the bismuth oxido clusters induce restructuring of solvent molecules around the bismuth oxido clusters. This interaction leads to disorder of the atomic structure in the bismuth oxido clusters themselves, as seen from the width and asymmetry of PDF peaks. When methacrylate is introduced as a ligand in

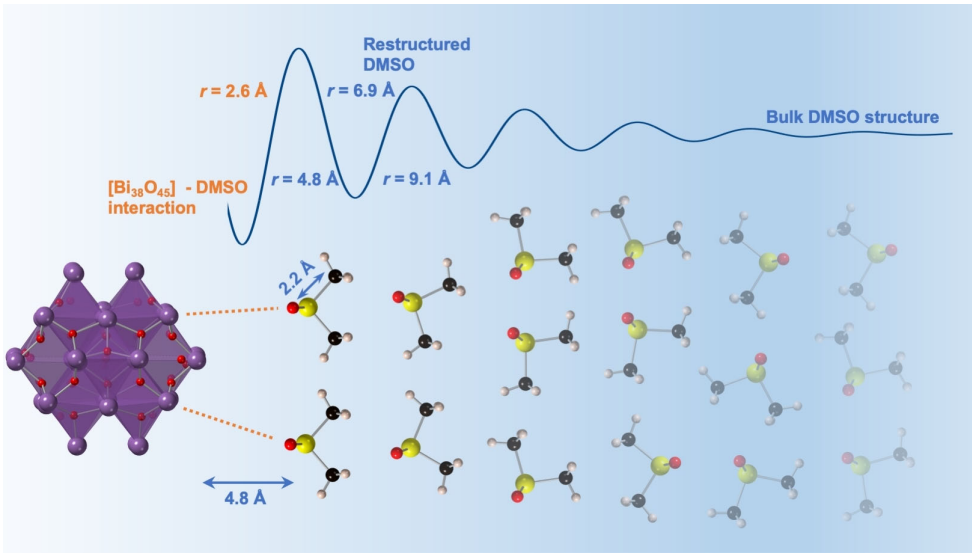



***Figure 10.*** Illustration of how the surface of the $[Bi_{38}O_{45}]$ cluster, with nitrate as the ligand, induces a solvent restructuring effect up to $r = 30$ Å. Interactions between the bismuth oxido clusters and the DMSO molecules induce an ordering of the solvent molecules near the $[Bi_{38}O_{45}]$ cluster surface, which gives rise to an exponentially dampened sinusoidal oscillation in the PDF. Carbon atoms are shown in black, sulfur in yellow, and hydrogen in beige.

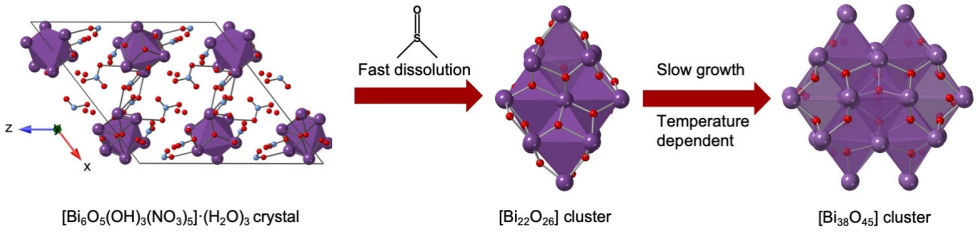



***Figure 11.*** The $[Bi_6O_5(OH)_3(NO_3)_5]\cdot 3\,H_2O$ crystals dissolve in DMSO to form a $[Bi_{22}O_{26}]$ cluster as an intermediate before forming atomically precise $[Bi_{38}O_{45}]$ clusters.

the $[Bi_{38}O_{45}]$ cluster, it screens the cluster from strong solvent interactions, consequently leaving the $[Bi_{38}O_{45}]$ core more ordered and closer to that observed in the crystalline structure. However, the $[Bi_{38}O_{45}]$ core structure stays intact during the ligand substitution reaction, which demonstrates that such atomically precise metal oxido nanoclusters can serve as model structures to study ligand effects at the surface of metal oxide nanoparticles.

The chemical knowledge built up in this study is likely transferable to the growth of other metal oxido clusters, at least to metal oxides of the fluorite type such as uranium, cerium and plutonium oxido nanoclusters, which also show a stable nuclearity of 38 metal atoms in their cluster core. The insights that can be obtained from combined in situ X-ray scattering studies furthermore open for understanding material formation in solution. For example, SAXS and PDF studies may allow the identification of prenucleation clusters, and provide a much deeper understanding of the fundamental processes involved in nucleation. Studies of atomically monodisperse metal oxido cluster structures, as done here, can furthermore provide knowledge of metal oxide chemistry, and it is likely that the metastable $[Bi_{22}O_{26}]$ cluster observed here may be an important species in bismuth oxide formation. Understanding the solution chemistry of metal oxido clusters on an atomic and molecular scale can thus open new opportunities for synthesizing nanoscale metal oxides in a controlled manner.

## Acknowledgements

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 804066). We furthermore thank DANSCATT (supported by the Danish Agency for Science and

Higher Education) for support. A.S.A. acknowledges the Siemens Foundation for support for his thesis project. We acknowledge DESY (Hamburg, Germany), a member of the Helmholtz Association HGF, for the provision of experimental facilities. Parts of this research were carried out at beamline P02.1 and P07 at Petra III, and we thank Martin Etter, Jozef Bednarcik and Ann-Christin Dippel for assistance in using the beamline. We acknowledge SOLEIL for provision of synchrotron radiation facilities, and we thank Thomas Bizien for assistance in using beamline SWING. We acknowledge the European Synchrotron Radiation Facility for provision of synchrotron radiation facilities, and we thank Peter Boesecke for assistance in using beamline ID02. We thank the Deutsche Forschungsgemeinschaft (DFG SPP1415: Crystalline Nonequilibrium Phases—Preparation, Characterization and in situ Studies of Formation Mechanisms; ME 2284/3-2) for support. M.S. and E.B. acknowledge the Innovation Fund Denmark (IFD) as part of project Linking Industry to Neutrons and X-rays (LINX) under File No. 5152-00005B for partly funding this project. Open access funding enabled and organized by Projekt DEAL.

### *Conflict of Interest*

The authors declare no conflict of interest.

**Keywords:** bismuth · in situ studies · nanostructures · pair distribution function · small-angle X-ray scattering

[1] E. D. Bøjesen, B. B. Iversen, *CrystEngComm* **2016**, *18*, 8332 – 8353.
[2] a) P. D. Matthews, T. C. King, D. S. Wright, *Chem. Commun.* **2014**, *50*, 12815 – 12823; b) P. Coppens, Y. Chen, E. Trzop, *Chem. Rev.* **2014**, *114*, 9645 – 9661; c) W.-H. Fang, L. Zhang, J. Zhang, *J. Am. Chem. Soc.* **2016**, *138*, 7480 – 7483; d) S. Yang, H.-C. Su, J.-L. Hou, W. Luo, D.-H. Zou, Q.-Y. Zhu, J. Dai, *Dalton Trans.* **2017**, *46*, 9639 – 9645; e) P. D. Matthews, N. Li, H.-K. Luo, D. S. Wright, *Chem. Eur. J.* **2016**, *22*, 4632 – 4633.
[3] a) K. J. Mitchell, K. A. Abboud, G. Christou, *Nat. Commun.* **2017**, *8*, 1445; b) B. Russell-Webster, K. A. Abboud, G. Christou, *Chem. Commun.* **2020**, *56*, 5382 – 5385.
[4] a) L. Miersch, T. Rüffer, M. Schlesinger, H. Lang, M. Mehring, *Inorg. Chem.* **2012**, *51*, 9376 – 9384; b) F. Y. Du, J. M. Lou, R. Jiang, Z. Z. Fang, X. F. Zhao, Y. Y. Niu, S. Q. Zou, M. M. Zhang, A. H. Gong, C. Y. Wu, *Int. J. Nanomed.* **2017**, *12*, 5973 – 5992.
[5] L. Miersch, T. Rüffer, M. Mehring, *Chem. Commun.* **2011**, *47*, 6353 – 6355.
[6] M. Schlesinger, S. Schulze, M. Hietschold, M. Mehring, *Dalton Trans.* **2013**, *42*, 1047 – 1056.
[7] M. Weber, M. Schlesinger, M. Mehring, *Cryst. Growth Des.* **2016**, *16*, 5678 – 5688.
[8] a) F. Lazarini, *Acta Crystallogr.* **1978**, *34*, 3169 – 3173; b) D. L. Rogow, H. Fei, D. P. Brennan, M. Ikehata, P. Y. Zavalij, A. G. Oliver, S. R. J. Oliver, *Inorg. Chem.* **2010**, *49*, 5619 – 5624; c) M. Mehring, S. Paalasmaa, M. Schürmann, *Eur. J. Inorg. Chem.* **2005**, 4891 – 4901; d) M. Mehring, D. Mansfeld, S. Paalasmaa, M. Schürmann, *Chem. Eur. J.* **2006**, *12*, 1767 – 1781; e) D. Mansfeld, M. Mehring, M. Schürmann, *Angew. Chem. Int. Ed.* **2005**, *44*, 245 – 249; *Angew. Chem.* **2005**, *117*, 250 – 254; f) X. Kou, X. Wang, D. Mendoza-Espinosa, L. N. Zakharov, A. L. Rheingold, W. H. Watson, K. A. Brien, L. K. Jayarathna, T. A. Hanna, *Inorg. Chem.* **2009**, *48*, 11002 – 11016; g) M. Weber, T. Rüffer, F. Speck, F. Göhler, D. P. Weimann, C. A. Schalley, T. Seyller, H. Lang, M. Mehring, *Inorg. Chem.* **2020**, *59*, 3353 – 3366.
[9] a) M. Mehring in *Clusters—Contemporary Insight in Structure and Bonding* (Ed.: S. Dehnen), Springer International Publishing, Cham, **2017**, pp. 201 – 268; b) M. Mehring, *Coord. Chem. Rev.* **2007**, *251*, 974 – 1006.
[10] I. L. Malaestean, A. Ellern, S. Baca, P. Kogerler, *Chem. Commun.* **2012**, *48*, 1499 – 1501.
[11] C. Falaise, C. Volkringer, J. F. Vigier, A. Beaurain, P. Roussel, P. Rabu, T. Loiseau, *J. Am. Chem. Soc.* **2013**, *135*, 15678 – 15681.
[12] a) L. Soderholm, P. M. Almond, S. Skanthakumar, R. E. Wilson, P. C. Burns, *Angew. Chem. Int. Ed.* **2008**, *47*, 298 – 302; *Angew. Chem.* **2008**, *120*, 304 – 308; b) R. E. Wilson, S. Skanthakumar, L. Soderholm, *Angew. Chem. Int. Ed.* **2011**, *50*, 11234 – 11237; *Angew. Chem.* **2011**, *123*, 11430 – 11433.
[13] D. Sattler, M. Schlesinger, M. Mehring, C. A. Schalley, *ChemPlusChem* **2013**, *78*, 1005 – 1014.
[14] a) K. Takao, S. Takao, A. C. Scheinost, G. Bernhard, C. Hennig, *Inorg. Chem.* **2012**, *51*, 1336 – 1344; b) S. Takao, K. Takao, W. Kraus, F. Emmerling, A. C. Scheinost, G. Bernhard, C. Hennig, *Eur. J. Inorg. Chem.* **2009**, 4771 – 4775; c) S. L. Estes, M. R. Antonio, L. Soderholm, *J. Phys. Chem. C* **2016**, *120*, 5810 – 5818.
[15] a) M. Walther, D. Zahn, *Eur. J. Inorg. Chem.* **2015**, 1178 – 1181; b) M. Walther, D. Zahn, *Chem. Phys. Lett.* **2018**, *691*, 87 – 90; c) M. Schlesinger, A. Pathak, S. Richter, D. Sattler, A. Seifert, T. Rüffer, P. C. Andrews, C. A. Schalley, H. Lang, M. Mehring, *Eur. J. Inorg. Chem.* **2014**, 4218 – 4227.
[16] a) C. Hennig, A. Ikeda-Ohno, W. Kraus, S. Weiss, P. Pattison, H. Emerich, P. M. Abdala, A. C. Scheinost, *Inorg. Chem.* **2013**, *52*, 11734 – 11743; b) Y.-J. Hu, K. E. Knope, S. Skanthakumar, L. Soderholm, *Eur. J. Inorg. Chem.* **2013**, 4159 – 4163; c) C. Hennig, S. Takao, K. Takao, S. Weiss, W. Kraus, F. Emmerling, A. C. Scheinost, *Dalton Trans.* **2012**, *41*, 12818 – 12823.
[17] S. J. L. Billinge, M. G. Kanatzidis, *Chem. Commun.* **2004**, 749 – 760.
[18] K. M. Ø. Jensen, C. Tyrsted, M. Bremholm, B. B. Iversen, *ChemSusChem* **2014**, *7*, 1594 – 1611.
[19] a) S. J. L. Billinge, I. Levin, *Science* **2007**, *316*, 561 – 565; b) C. Farrow, C. Shi, P. Juhás, X. Peng, S. J. L. Billinge, *J. Appl. Crystallogr.* **2014**, *47*, 561 – 565; c) P. Juhás, C. L. Farrow, X. Yang, K. R. Knox, S. J. L. Billinge, *Acta Crystallogr. Sect. A* **2015**, *71*, 562 – 568; d) C. L. Farrow, S. J. L. Billinge, *Acta Crystallogr. Sect. A* **2009**, *65*, 232 – 239.
[20] T. L. Christiansen, E. T. S. Kjær, A. Kovyakh, M. L. Röderen, M. Høj, T. Vosch, K. M. Ø. Jensen, *J. Appl. Crystallogr.* **2020**, *53*, 148 – 158.
[21] S. Banerjee, C.-H. Liu, J. D. Lee, A. Kovyakh, V. Grasmik, O. Prymak, C. Koenigsmann, H. Liu, L. Wang, A. M. M. Abeykoon, S. S. Wong, M. Epple, C. B. Murray, S. J. L. Billinge, *J. Phys. Chem. C* **2018**, *122*, 29498 – 29506.
[22] a) M. G. Tucker, D. A. Keen, M. T. Dove, A. L. Goodwin, Q. Hui, *J. Phys. Condens. Matter* **2007**, *19*, 335218; b) V. Krayzman, I. Levin, M. G. Tucker, *J. Appl. Crystallogr.* **2008**, *41*, 705 – 714.
[23] P. Fromme, *Nat. Chem. Biol.* **2015**, *11*, 895 – 899.
[24] a) M. Zobel, R. B. Neder, S. A. J. Kimber, *Science* **2015**, *347*, 292 – 294; b) S. L. J. Thomä, S. W. Krauss, M. Eckardt, P. Chater, M. Zobel, *Nat. Commun.* **2019**, *10*, 995.

Manuscript received: March 14, 2021
Accepted manuscript online: May 30, 2021
Version of record online: ■■ ■■, ■■■■

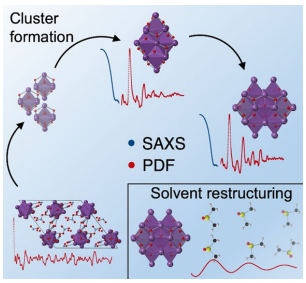
  

# Research Articles

## Cluster Compounds

A. S. Anker, T. L. Christiansen, M. Weber, M. Schmiele, E. Brok, E. T. S. Kjær, P. Juhás, R. Thomas, M. Mehring,* K. M. Ø. Jensen* ———— ▮▮▮▮–▮▮▮▮

Structural Changes during the Growth of Atomically Precise Metal Oxido Nanoclusters from Combined Pair Distribution Function and Small-Angle X-ray Scattering Analysis




The development of tools for combined, automated analysis of pair distribution function (PDF) and small-angle X-ray scattering (SAXS) data enabled the identification of intermediate species in the formation of atomically precise $[Bi_{38}O_{45}]$ nanoclusters (see picture). By studying cluster–solvent interactions it was also found that the solvent can affect the metal oxido structure.