Development of small-angle scattering pair distribution function analysis techniques and application to nanoparticles assemblies

Chia-Hao Liu

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

ABSTRACT

Development of small-angle scattering pair distribution function analysis techniques and application to nanoparticles assemblies

Chia-Hao Liu

With the improvement in synthesis method, a variety of nanoparticles (NPs) with nearly uniform distribution in size and morphology are now available to scientists. This progress opens a new opportunity of assembling these high quality nanoparticles into metamaterial - nanoparticle assemblies (NPAs). The properties of NPA depend on the interactions between constituent NPs, therefore NPA offer a distinct advantage in designing material properties that are not available in the bulk phase (crystal) or discrete phase (nanoparticle). Novel application of NPA in modern devices, such as solar cells and field effect transistors, had also been demonstrated. The spatial arrangements of NPs is the key factor to their interactions, therefore, it is crucial to characterize the structure of NPA quantitatively. The technique of diffraction plays an unique role for characterizing NPA structure, as it not only offers the structural type, which may also be obtained from imagine technique, but also yields structural information in three-dimension, such interparticle distance and the range of structural coherence of the packing order. Traditionally, the diffraction analysis is based on crystallography and is carried out in reciprocal space. However, it is known that the local structure is overlooked in this kind of crystallographic analysis, which places a challenge for have a comprehensive understanding of the NPA structure.

The pair distribution function (PDF) analysis, which is powerful in probing local structures for atomic systems, serves as a promising tool for characterizing NPA structure. However, the approach of using PDF analysis for NPA structure characterization has barely been explored. In this thesis, I will present the methodological developments of the PDF technique. Starting from presenting a machine-learning-assisted approach for predicting the space group of its structure from the PDF, I will be focusing on the aspect of accelerating the structure modeling steps with PDF. Next, the development of pair distribution function analysis in small-angle scattering domain sasPDF will be introduced, including software package PDFgetX3 which is aiming to facilitate the extraction of PDF from small-angle scattering data quickly. The approach of sasPDF is validated against three representative structures across different levels of structural order. Finally, the example of applying sasPDF method to identify the jamming transition signature in polymer-ligated NPA is introduced, followed by another example of discovering multiply-twinned structure from the reprogramming of DNA-ligated NPA.

Table of Contents

List of Figures iv

List of Tables ix

Acknowledgments xi

1 Application of nanoparticles and their assemblies 1

1.1 Introduction 1

1.1.1 Nanoparticle assemblies (NPA) 2

1.1.2 DNA-ligated nanoparticle assemblies 3

1.1.3 Polymer-ligated nanoparticle assemblies 3

1.1.4 Properties of NPA-based materials 4

1.2 Current status of structural characterization techniques for NPs and NPAs . 5

1.3 Outline of this thesis 6

2 Pair-distribution function (PDF) technique 7

2.1 Introduction 7

2.2 PDF theory 8

2.3 Data collection for PDF 11

2.4 Data analysis for PDF 13

2.4.1 Structure modeling 13

3 Machine learning approach to determine the space group of a structure from the atomic PDF 15
3.1 Introduction 15
3.2 Machine Learning experiments 18
3.2.1 Space Group Determination based on Logistic Regression (LR) model 23
3.2.2 Space group determination based on convolutional neural network (CNN) 27
3.3 Results and discussion 31
3.3.1 Space group determination on calculated PDFs 31
3.3.2 Space Group Determination on Experimental PDFs 33
3.4 Conclusion 35
3.5 Appendix 37
3.5.1 Logistic Regression and Elastic Net Regularizations 37
3.5.2 Robustness of the CNN model 39

4 sasPDF: pair distribution function analysis of nanoparticle superlattice assemblies from small-angle-scattering data 41
4.1 Introduction 41
4.2 Samples 43
4.3 sasPDF method 43
4.4 Software 51
4.5 PDF method 53
4.6 Application to representative structures 55
4.7 Appendix 61
4.7.1 Illustration of of data acquisition strategy 61

5 Applications of sasPDF method on nanoparticle assemblies 67

5.1 A structural signature for jamming in polymer-ligated nanoparticle assemblies 67
  5.1.1 introduction 67
  5.1.2 Experiment 68
  5.1.3 Method 70
  5.1.4 Results 71
  5.1.5 Conclusion 76
  5.1.6 Appendix 76
5.2 Multiply twinned structure in DNA-ligated Au nanoparticle assemblies . . . 77
  5.2.1 Results 77
  5.2.2 Conclusion 78

Bibliography **78**

List of Figures

2.1 Schematic of the RA-PDF experimental setup.K and K' is the wave vector for the incident and scattered x-ray beam respectively. Q is the momentum transfer vector, which is showing in the inset. 12

3.1 Example of (a) normalized PDF $\mathbf{X}$ and (b) its quadratic form $\mathbf{X}^2$ of compound $Li_{18}Ta_6O_{24}$ (space group $P2/c$). 22

3.2 Accuracy in determining space group when top-i predictions are considered (A_i). Inset shows the first discrete differences ($\Delta A_i = A_i - A_{i-1}$) when i predictions are considered. Blue represents the result of the logistic regression model with $\mathbf{X}^2$ and red is the result from the convolutional neural network model. 25

3.3 The ratio of correctly classified structures v.s. space group number from (a) logistic regression model (LR) with quadratic feature $\mathbf{X}^2$ and (b) convolutional neural network (CNN) model. Marker size reflects the relative frequency of space group in the training set. Markers are color coded with corresponding crystal systems (triclinic (dark blue), monoclinic (orange), orthorhombic (green), tetragonal (blue), trigonal (grey), hexagonal (yellow) and cubic (dark red). 26

3.4 Schematic of our convolutional neural network (CNN) architecture. 28

3.5 Accuracy of the CNN model on the training set (blue), the testing set (red) and the optimization loss against the testing set (green) with respect to number of epochs during the training step. 30

3.6 The confusion matrix of our CNN model. The row labels indicate the correct space group and the column labels the space group returned by the model. An ideal model would result in a confusion matrix with all diagonal values being 1 and all off-diagonal values being zero. The numbers in parentheses are the space-group number. 34

4.1 Example of the 1D diffraction pattern $I_m(Q)$ from the Cu_2S NPA sample. The data were collected with the spot exposure time and scan exposure time reported in the text. The inset shows the corresponding 2D diffraction image. The horizontal stripes in the image are from the dead zone between panels of the detector. The diagonal line is the beam-stop holder. 45

4.2 Illustration of the interactive interface for tuning the process parameters in the PDFgetS3 program. 52

4.3 Measured (a) scattering intensity $I_m(Q)$ (grey) and form factor $P(Q)$ (blue), (b) reduced total structure function $F(Q)$ (red) and (c) PDF (open circle) of Au NPA. In (c), the PDF calculated from body-center cubic (bcc) model is shown in red and the difference between the measured PDF and the bcc model is plotted in green with an offset. 56

4.4 Measured PDF (open circle) of a Cu_2S NPA sample with the best fit PDF from the fcc model (red line). The Difference curve between the data aAs a result, nd model is plotted offset below in green. The inset shows the region of the first four nearest neighbor peaks of the PDF along with the best-fit fcc model. 59

4.1 (a) Reduced structure functions $F(Q)$ and (b) PDFs $G(r)$ of the SiO_2 NPA sample with different scan exposure times. Blue is from data with 1 s scan exposure time and red is from data with 30 s scan exposure time. In both panels, data are plotted with a small offset for ease of viewing. In both cases the form factor was measured with an scan exposure time of 600 s. 62

4.2 (a) Reduced structure functions $F(Q)$ and (b) PDFs $G(r)$ of the SiO_2 NPA sample processed with form factor $P(Q)$ from different scan exposure times. Blue is made with a form-factor measured for 30 s and red is with a form factor collected for 600 s. In both cases the scan exposure time for the NPA sample was 600 s. In both panels, data are plotted with a small offset for ease of viewing. 63

4.3 (a) Reduced structure functions $F(Q)$ and (b) PDFs $G(r)$ of the SiO_2 NPA sample. Blue is from data collected at Columbia University using a SAXSLAB (Amherst, MA) instrument with a 2-hour (7200 s) scan exposure time for both $I(Q)$ and $P(Q)$ measurements. Red is from data collected at beamline 11-BM, NSLS-II with 30 s scan exposure time for both $I_m(Q)$ and $P(Q)$ measurements. 64

4.4 (a) Form factor signal from Cu_2S NPs. Blue is the raw data collected at an in-house instrument and red is the data smoothed by applying a Savitzky-Golay filter with window size 13 and fitted polymer order 2. (b) reduced structure functions, $F(Q)$, and (c) PDFs, $G(r)$ from the Cu_2S NPA sample. In both panel, blue represents the data processed with raw form factor signal and red represents the data processed with smoothed form factor signal. Curves are offset from each other slightly for ease of view. 65

4.5 Semi-quantitative structural analysis on Cu_2S NPA sample. 66

5.1 Measured PDFs of, from top to bottom, H-31, H-41, H-62, H-80, H-106, H-129 samples. 71

5.2 Measured PDF (open circle) of H-31 sample and calculated PDFs (solid lines) from (a) fcc, (b) hcp, (c) icosahedral (d) damped sine-wave models. In each panel, the line in dark red is the PDF calculated from the corresponding model with optimum parameters. From (a) to (c), the line in grey is the PDF calculated from the same model but with small ADPs. In (d), the line in grey is the PDF calculated from the undamped sine-wave model. Dashed lines indicate maxima of the sharper PDFs in each panel. 72

5.3 PDFs of, from top to bottom, H-31, H-41, H-62, H-80, H-106, H-129 plotted on a renormalized r-axis, r/λ, where λ is the refined wavelength of the best-fit damped sine-wave model. 73

5.4 Measured PDFs (open circles), full-r fit (grey) and high-r fit (red) of (a) H-41, (b) H-80, and (c) H-129 samples. The difference between two models (brown) is plotted below in each panel. 79

5.5 Hard-sphere parameter, ξ_h, for medium (blue) and high (red) graft density samples. The shaded area is the region of M_n where an anomalous enhancement in gas permeability was previously reported. This enhancement is reproduced in our samples as shown in the inset where the permeability ratio P_ϕ/P_b is plotted from samples with graft densities $\Sigma = 0.43$ chains/nm^2 (blue) and $\Sigma = 0.66$ chains/nm^2 (red) similar to the ones in the x-ray experiments. The horizontal dashed line in the inset is $P_\phi/P_b = 1$ for reference. 80

5.6 Measured PDFs of, from top to bottom, M-29, M-41, M-65, M-78, M-101, M-132 samples. 81

5.7 PDFs of, from top to bottom, M-29, M-41, M-65, M-78, M-101, M-132 plotted on a renormalized r-axis, r/λ, where λ is the refined wavelength of the best-fit damped sine-wave model. 82

5.8 Measured PDFs (open circles), full-r fit (grey) and high-r fit (red) of (a) M-41, (b) M-78, and (c) M-132 samples. The difference between two models (brown) is plotted below in each panel. 83

5.9 Measured PDFs from the fcc-bcc phase transition. From bottom to top, each PDF corresponds to data collected at 0, 40, 80, 120, 160, 220, 280, 360, 480 and 800 minutes after the extra DNA strands was added. 84

5.10 Scatter plot of agreement factors (Rw) of fcc model (red) and bcc model (blue) vs data collected at different reaction time. 85

5.11 Measured PDF (blue) at reaction time = 800 mins and PDF from best-fit fcc model (red). The difference (green) is plotted with an offset for the ease of reading. 86

5.12 Scatter plot of agreement factors (R_{w}) for decahedron (green), octahedron (red) and icosahedron (blue) fit to the PDF collected at reaction time = 800 mins, plotted as a function of the number of particles per model. The agreement factor from crystalline model (fcc) to the same PDF is labeled in a dashed line. 87

5.13 Measured PDF (blue) at reaction time = 800 mins and PDF from best-fit decahedron cluster model (red). The difference curve (green) is plotted with an offset for the ease of reading. The shaded area of difference curve labels the improvement of decahedron cluster from fcc model. 88

List of Tables

3.1 Space group and corresponding number of entries considered in this study. . 18

3.2 Parameters used to calculate PDFs from atomic structures. All parameters follow the same definitions as in [53]. 21

3.3 Top-6 space-group predictions from the CNN model on experimental PDFs. Bold-faced prediction is the most probable space group from existing literatures listed in the Refs. column. More than one predictions are highlighted when these space groups are regarded as highly similar in literatures. Details about these cases will be discussed in the text. The Note column specifies if the PDF is from a crystalline (C) or nanocrystalline (NC) sample. The experimental data were collected under various instrumental conditions which are not identical to the training set and experimental data were measured at the room temperature, unless otherwise specified. Dagger is used to label the data that the CNN model fails to predict the correct space group. 36

S1 Accuracies of CNN model with different sets of hyper parameters. Accuracy is abbreviated as accu. in the table. The last row specifies the optimum set of hyperparmeters for our final CNN model. 40

1 Nanoparticle assemblies (NPA) considered in this study. Building block indicates the NP and surfactant linkers used to build the assemblies. D is the particle diameter (one standard deviation in parentheses) estimated from TEM images and reported in the original publications listed in the Ref. column. Beamline is the x-ray beamline where the SAXS data were measured (see text for details). PMA is Poly(methyl acrylate) and DDT is dodecanethiol. 44

2 Refined parameters for NPA samples. Model column specifies the structural model used to fit the measured PDF. a is the lattice constant of the unit cell, PDP stands for particle displacement parameters, which is an indication of the uncertainty in position of the nanoparticles. r_{damp} is the standard deviation of the Gaussian damping function defined in Eq. 12. Scale is a constant factor being multiplied to the calculated PDF. R_w is the residual-function, commonly used as a measure for the goodness of fit. 57

1 Polymer-grafted silica NP samples. M_n is the molecular weight of the grafted chain in kg/mol and Σ is the polymer graft density on the surface of the nanoparticles in chains/nm^2. 69

Acknowledgments

The graduate study is a long journey. I would like to start by thanking my academic advisor Simon Billinge who is the best advisor I could ever ask for. Without his guidance, I would be nowhere near to where I am standing now. From scientific reasoning to communicating, Simon demonstrates how a scientist should behave. It is such a great honor and privilege to be able to learn so closely with him over the years. I also want to thank Simon for his endless patience for letting me explore research ideas and giving me advices when I needed the most during my graduate study.

I also want to thank my group members including Max Terban, Soham Banerjee, Kirsten Jensen, Zurab Guguchia, Long Yang, Chris Wright, Anton Kovyakh, Elizabeth Culbertson, Songsheng Tao, Yevgeny Rakita, Ben Frandsen, Pavol Juhas, Ran Gu for contributing intellectual discussions and laughters in both appropriate and inappropriate ways. I want to thank my collaborators throughout the years, who always enrich me with their knowledge from different domains, including Yunzhe Tao, Eileen Buenning, Ji Xu, Mayank Jhalaria, Paul Todd and Alison Wustrow. Of course, their advisors, Prof. Qiang Du, Prof. Sanat K. Kumar, Prof. Daniel J. Hsu and Prof. James R. Nielson.

Finally, I would like to thank my family - my parents Tony and Helen for their unfailing supports and love, my sister Winny for being there whenever I needed her and my dog Sparky who always brings me happiness (and his toys) and teaches me what is unconditional love.

To my family, Winny, Helen, Tony and Sparky.

Chapter 1

Application of nanoparticles and their assemblies

1.1 Introduction

Nanoparticles (NPs) are generally regarded as objects in the size of 1 to 100 nanometers. In the last few decades, research about NPs growths exponentially as on this length scale exciting phenomena such as superparamagnetism in magnetic NPs [15; 121], carrier multiplication in semiconductor NPs [130; 119] and tunable band gap [32; 91] emerge due to quantum mechanical effects. Those properties are attributed to a wide range of factors, such as size, morphology, chemical composition and surface chemistries of the NPs [177; 99]. With the advent of high degrees of control over nanoparticle synthesis, narrow size-distribution and extensive tunability in its morphology, chemical, electronic and magnetic properties had been reported [129; 77; 47]. Since then, attention start turning to integrating NPs with modern applications, ranging from light emitting [156; 81], energy harvesting [131; 206], to biomedical sensing [72; 154] devices, which were reported to have lower cost and higher efficiency than the traditional devices based on bulk crystals [159; 2].

1.1.1 Nanoparticle assemblies (NPA)

The progress in synthesizing nearly uniform NPs brings up attention of assembling them as metamaterials; nanoparticle assemblies, NPAs. The research of assembling microscopic objects into ordered 2D and 3D structures can be traced back to the mid 1980s, when researchers studied the super structures formed by colloidal polystyrene particles between two smooth glass boundaries [143; 188]. At that time, external confinements were still required for the assemblies of ordered structure and the yield was still low. In the early 1990s, pioneered work on the formations of ordered 2D networks of Au [61] and Ag_2S [125] NPs was reported. In the mid 1990s, after the seminal work on assembling CdSe NPs into 2D and 3D ordered structures along with a control in lattice constant for the NPAs formed [127], research about NPAs has grown exponentially. In the following years, ordered assemblies from TiO_2 [33], Ag [195; 123], SiO2 [1] NPs has been reported. More recently, diversified structures from NPAs based on binary NPs [166] or anisotropic NPs [126; 124] had also been observed.

NPAs can directly form in its colloidal solution, or on a substrate after evaporation or dewetting [144; 108; 39]. The NP and ligand attached (or "linkers" in some literature) are the building block of the NPA. The formation of NPA depends closely on the interactions between its building blocks and environmental factors of the synthesis process [122; 124]. So far, a wide range of interactions, such as van der Waals force, Coulomb force due to surface charge or electric dipoles between NPs and hard-sphere repulsions between the ligand, had been reported to be the driving force for the assembly of different NPs. [103; 180; 179] In addition, environmental factors like capillary force [39], ambient temperature [207] and external magnetic field [144] had also been reported as the key to the formation of certain NPA systems. For most of the systems, ligands are non-biological molecules that can be covalently bonded to the NPs of interest, while providing functionality at the opposite end [145]. Work had shown ligands can not only stabilize the growth of NPs but also guide the formation of NPAs [44; 108; 28] For example, the structure of Au NPAs is altered between body-centered cubic (bcc) and hexagonal closest packed (hcp) structure depending on the ratio ligand length and particle radius [28].

1.1.2 DNA-ligated nanoparticle assemblies

Instead of non-biological ligands that are commonly seen in NPA systems, a seminal work in the late 1990s demonstrated the realization of 3D ordered assemblies of Au NPs with thiol-modified DNA as the ligands [123]. A good control in the size and morphology of the assemblies formed had been predicted as the interaction between DNA sequence were well understood and the length of discrete DNA sequences can be specifically tailored [3; 112]. Work had also shown the formation DNA-based NPAs is reversible [123]. There has been a rich body of literature about synthesizing DNA-based NPAs with different levels of control in morphologies and structure types [133; 6; 112]. Recently, progress in synthesizing ordered assemblies that depend merely on the DNA ligands but not the NP had also been reported and structures that are not previously accessible (including noncrystalline phase) had been observed [8; 181]. The programmability of DNA-based NPAs make this technique a promising approach for synthesizing and engineering artificial materials.

1.1.3 Polymer-ligated nanoparticle assemblies

Soft molecules such as polymers can also be used as the ligands for NPAs. Work had shown the ordered structures in 2D and 3D, which are commonly observed from NPAs, can be achieved with polymer-ligated Au NPAs system [205; 204]. Polymeric ligands offer a control of the effective size through changes to the polymers molecular weight, chemical nature, architecture, persistence length and surrounding solvent [200]. Rather diversified morphologies, ranging from stings (1D), sheets (2D) and spherical clusters (3D), had been observed in SiO2 NPAs with polystyrene ligands, by simply changing the length and density of the polymer ligand [1]. Furthermore, by carefully designing the morphology of the ligands, anisotropic structures can be formed with spherical constituent NPs [1]. In addition to the advantage of programmability, the polymer-ligated NPAs are also suitable for industrial-scale applications such as filled rubbers and membranes for gas separations [101].

1.1.4 Properties of NPA-based materials

The properties of NPA-based materials, such as mechanical [1], optical [173] electrical [128] and magnetic [174] properties, had been shown to be highly tunable. Applications of NPA-based devices such as solar cells and field effect transistors have been demonstrated [178; 164; 177]. It is known that the overall properties of NPA-based materials are based on the interactions between constituent particles [145; 112]. This allows NPA-based materials to achieve unique properties that are not observed in its discrete phase (NP) or in its bulk phase. The spatial arrangements of constituent particles play an important role for the interaction between NPs. For example, by changing the separation of adjacent Ag NPs, the plasmonic frequency of the DNA-ligated Ag NPA was shifted across the spectrum of visible lights [201] and the magnetic properties such as the remanent magnetization and coercive field can be tuned by varying the interpaticle distance of dodecanediol-capped Fe_3O_4 NPAs [198]. Given such rich tunablility in terms of material properties and the great potential for device application, it is then crucial to characterize the structures of NPAs quantitatively if their properties are to be optimized.

1.2 Current status of structural characterization techniques for NPs and NPAs

Scattering and electron microscopy (EM) have been the major techniques for studying the structure of NP and NPAs [128; 179; 54]. In particular, the technique of transmission electron microscopy (TEM) is commonly used as it directly yields high-resolution images of the NPs and NPAs. However, for both the case of NPs and NPAs, it is necessary to either analyze the images manually [192; 171] or match observed images with patterns algorithmically generated from known structures [166; 104] to obtain quantitative structural information about the sample . This approach can yield the structure types [104; 208] but does not typically result in the kind of quantitative 3D structural information we are used to obtaining for atomic structures of crystals, including accurate inter-particle vectors and distributions of inter-particle distances, or the range of structural coherence of the packing order. It is desirable to explore scattering approaches that can yield structural information in 3D.

Depending on the measured range of scattering vector Q, scattering data can be categorized into wide-angle and small-angle regimes. The wide-angle x-ray (WAXS) scattering data is typically collected at the range of $Q \geq 0.1$ $\mathring{\mathrm{A}}^{-1}$. The information encoded in this range corresponds to the inter-atomic distances present in a material, which is usually at the length scale of angstroms to nanometers. The technique of wild-angle scattering is an invaluable method for studying the crystalline structure of NPs [14; 155; 157]. On the other hands, the small-angle x-ray scattering (SAXS) data is usually collected at $Q < 0.1$ $\mathring{\mathrm{A}}^{-1}$. In this range, the scattering data yields information about the material on nano- to micrometer scales. The technique of small-angle scattering started as a tool for studying the intrinsic shape, size distributions and scattering density of NPs on these scales [63; 191; 17] and it was later used to study the particle arrangement in the NPAs as the correlation peaks appear in the small-angle scattering data resembling atomic-scale interference peaks (Diffuse scattering and Bragg peaks) when NPs aggregate, yield information about particle packing [128; 133].

Although developments of modeling diffraction patterns from WAXS and SAXS data had been reported [169; 98; 197], the analysis in reciprocal space, which is based on crystallography, is less favorable when the structure is only short-range ordered [24; 34; 92], as is the case for NPs and NPAs [163; 113] For materials that are only ordered in short-range, structural information can quantitatively extracted by the atomic pair distribution function (PDF) analysis [149; 50; 210; 92]. However, there has been barely attempt for extending the powerful PDF technique to the small-angle scattering data for characterizing the structure of NPAs.

1.3 Outline of this thesis

This thesis will be centered around the method developments for different aspects of PDF analysis. This thesis will be structured as follows: In Chapter 2, basic theory, along with an overview on the data collection and data analysis steps, of the PDF technique will be reviewed. In Chapter 3, an approach of using machine learning method to assist structure solution with PDFs will be presented. In Chapter 4, the PDF method in small-angle regime sasPDF and its software implementation PDFgetS3 will be introduced, followed by Chapter 5, which is about the application of sasPDF method to systems with different levels of structure order.

Chapter 2

Pair-distribution function (PDF) technique

2.1 Introduction

There are various techniques for characterizing the atomic structure in a material. The technique of powder diffraction has been heavily used in this aspect. For a powder diffraction experiment, the sample is illuminated by a beam of neutron or x-ray. The incident beam is then diffracted by the scatterers in the sample (nuclei for neutron beam and atom for x-ray) and a 2D diffraction pattern which can be further reduced to a 1D spectrum [23]. The 1D powder diffraction pattern contains signal from diffuse scattering and Bragg diffraction [194]. In the conventional crystallographic analysis, material structure is determined solely by the information encoded in Bragg peaks [194], where their positions yield the symmetry information and lattice constants of the structure and their intensities yield information about the arrangement of atoms in the structure [138]. However, the diffuse scattering signal, which provides the information of local structure is often ignored. As a result, in crystallographic analysis which assumes periodicity in atomic arrangement, only the average structure (long-range) is characterized but not the local structure (short-ranged). The later is commonly observed in finite objects such as NPs and NPAs [24; 92]. Therefore it is desired to have a analysis approach, which both Bragg diffraction and diffuse scattering are considered ("total scattering").

The pair distribution function (PDF) is the Fourier transform of total scattering structure factor [50]. This technique was firstly used for studying the structures of amorphous materials, such as glasses and liquids [49; 193]. Recently, it has been applied to study structure of disordered crystalline materials and nanomaterials [142; 117; 30]. By considering both Bragg (long range) and diffuse (short range) scattering, PDF is favorable for obtaining a comprehensive understanding of the structure in question. In this chapter, we will briefly review the theoretical and experimental aspects of PDF technique, followed by the discussion on the data analysis approach.

2.2 PDF theory

To derive the formalism of atomic PDF, we will start from the scattering of an atom m. In the kinematical limit, the scattering amplitude is [22]

$$\Psi_m(\mathbf{Q}) = f_m(\mathbf{Q}) \exp\left[i\mathbf{Q}\cdot\mathbf{r_m}\right] \tag{2.1}$$

where $\mathbf{Q}$ is the scattering vector, namely the difference between incoming $\mathbf{K}_i$ and scattered beam $\mathbf{K}_s$, $\mathbf{Q} = (\mathbf{K}_s - \mathbf{K}_i)$. $\mathbf{r_m}$ and $f_m(\mathbf{Q})$ is the position and atomic form factor of the atom respectively. For an atom with volume V and electron density as a function of position $\rho(\mathbf{r_m})$, the atomic form factor is defined as [65]

$$f_m(\mathbf{Q}) = \int_V \rho(\mathbf{r_m}) \exp\left(i\mathbf{Q}\cdot\mathbf{r}\right) \mathrm{d}\mathbf{r}, \tag{2.2}$$

Consider a unit cell with N_s atoms, the coherent scattering intensity $I_c(Q)$ is [50; 65]

$$I_c(\mathbf{Q}) = \sum_{m=1}^{N_s}\sum_{n=1}^{N_s} \Psi_m^*(\mathbf{Q})\Psi_n(\mathbf{Q}) \tag{2.3}$$

$$= \sum_{m=1}^{N_s}\sum_{n=1}^{N_s} f_m^*(\mathbf{Q}) f_n(\mathbf{Q}) \exp\left[i\mathbf{Q}\cdot(\mathbf{r}_m - \mathbf{r}_n)\right], \tag{2.4}$$

where $f_m(\mathbf{Q})$ and $\mathbf{r}_m$ are the atomic form factor amplitude and position of m-th atom in the unit cell, respectively.

If the scattering from a sample is isotropic, for example, it is an untextured powder or a liquid with no anisotropy, the observed scattering intensity will depend only on the magnitude of $\mathbf{Q}$, $|\mathbf{Q}| = Q$ and not its direction in space. The observed scattering intensity in this case will depend on the orientationally averaged $I_c(\mathbf{Q})$,

$$I_c(Q) = \left\langle \sum_{m=1}^{N_s}\sum_{n=1}^{N_s} f_m^*(\mathbf{Q}) f_n(\mathbf{Q}) \exp\left[i\mathbf{Q}\cdot(\mathbf{r}_m - \mathbf{r}_n)\right] \right\rangle, \tag{2.5}$$

where $\langle\cdot\rangle$ denotes the orientational average. In situations where there is the electron density of the scatterer is uncorrelated with the structure Eq. 2.5 may be further arranged as [65; 50]

$$I_c(Q) = N_s \left\langle f^2(Q) \right\rangle + \sum_{m=1}^{N_s}\sum_{n\neq m}^{N_s} \left\langle f_m^*(Q)\right\rangle \left\langle f_n(Q)\right\rangle \left\langle \exp\left[i\mathbf{Q}\cdot(\mathbf{r}_m - \mathbf{r}_n)\right]\right\rangle. \tag{2.6}$$

Since the PDF is a Fourier transform of the reduced structure function $F(Q) = Q\left[S(Q) - 1\right]$, we will start deriving the definition of $S(Q)$. From the Faber-Ziman formalism [51], the structure function $S(Q)$ is defined as

$$S(Q) = \frac{I_c(Q)}{N_s\langle f(Q)\rangle^2} - \frac{\langle f^2(Q)\rangle - \langle f(Q)\rangle^2}{\langle f(Q)\rangle^2}, \tag{2.7}$$

We note that if we assume the atomic form factor exhibits no orientational preference and plug in $\langle f^2(Q)\rangle = \langle f(Q)\rangle^2$, Eq. 6 becomes

$$I_c(Q) = N_s\langle f^2(Q)\rangle S(Q). \tag{2.8}$$

This expression is equivalent to representing the atoms as points at the position of their scattering center, convoluted with their electron distributions. The resulting structure function, $S(Q)$, yields the arrangement of scatterers in the sample. To express Eq. 5 in a similar fashion as the Faber-Ziman formalism, we first normalize Eq. 5 with total number of scatters N_s, subtract $\langle f(Q)\rangle^2$ and normalize it with $\langle f^2(Q)\rangle$ and we arrive

$$S(Q) - 1 = \frac{I_c(Q)}{N_s\langle f(Q)\rangle^2} - \frac{\langle f^2(Q)\rangle}{\langle f(Q)\rangle^2} \tag{2.9}$$

$$= \frac{1}{N_s\langle f(Q)\rangle^2}\sum_{m=1}^{N_s}\sum_{n\neq m}^{N_s}\langle f_m^*(Q)\rangle\langle f_n(Q)\rangle\langle\exp\left[i\mathbf{Q}\cdot(\mathbf{r}_m - \mathbf{r}_n)\right]\rangle\,. \tag{2.10}$$

The orientational average of the exponential term in Eq. 2.9 can be further evaluated if the scattering is isotropic, which is the case of powder diffraction [50]

$$\langle\exp\left[i\mathbf{Q}\cdot(\mathbf{r}_m - \mathbf{r}_n)\right]\rangle = \frac{\sin\left(Qr_{mn}\right)}{Qr_{mn}}, \tag{2.11}$$

where $r_{mn} = |\mathbf{r_m} - \mathbf{r_n}|$. Plugging Eq. 2.11 back to Eq. 2.9 and substitute in the definition of PDF $\mathcal{F}(r)$ [50]

$$\mathcal{F}(r) = \frac{2}{\pi}\int_0^\infty Q\left[S(Q) - 1\right]\sin\left(Qr\right)\mathrm{d}Q \tag{2.12}$$

$$= \frac{2}{\pi}\int_0^\infty \frac{1}{N_s r_{mn}\langle f(Q)\rangle^2}\sum_{m=1}^{N_s}\sum_{n\neq m}^{N_s}\langle f_m^*(Q)\rangle\langle f_n(Q)\rangle\sin\left(Qr_{mn}\right)\sin\left(Qr\right)\mathrm{d}Q. \tag{2.13}$$

Since sine function forms a orthogonal basis, the integration over Q in Eq. 2.13 results in delta-functions [22]

$$\int_0^\infty \sin\left(Qr_{mn}\right)\sin\left(Qr\right)\mathrm{d}Q = \frac{\pi}{2}\left[\delta(r - r_{mn}) - \delta(r + r_{mn})\right]. \tag{2.14}$$

By constraining our attention on the positive axis, Eq. 2.13 can be rewritten as

$$\mathcal{F}(r) = \frac{1}{rN_s\langle f(Q)\rangle^2}\sum_{m=1}^{N_s}\sum_{n\neq m}^{N_s}\langle f_m^*(Q)\rangle\langle f_n(Q)\rangle\,\delta(r - r_{mn}). \tag{2.15}$$

We note that above formalism can be extended to systems with multiple types of atoms. In this case, the $\langle f(Q)\rangle^2$ term should be treated as the sample-averaged, squared form factors from all atoms in the unit cell [50].

The PDF is purely based on the inter-atomic distances r_{mn} presented in the structure and no periodicity is assumed, which makes it is a general tool to study both crystalline and nanomaterials. The information in PDF is in the real space, which yields an intuitive way to interpret the information directly from the spectrum or to construct structure model for testing hypothesis.

2.3 Data collection for PDF

In the modern x-ray diffraction experiments, the rapid-acquisition PDF (RA-PDF) setup is commonly used as it can significantly shorten the data collection time [38]. In this setup, the sample is mounted perpendicular to the incident x-ray beam, with a large 2D area detector placed behind. The detector is usually located close to the sample so that the momentum transfer Q is maximized (Fig. 2.1). In practice, the sample may not be exactly perpendicular to the area detector and the oblique incidence, along with detector geometry, can be calibrated by measuring the standard materials, such as Ni or CeO_2, and comparing the position of measured Debye-Scherrer rings with known results. Once the calibration is done, each pixel on the area detector can be assigned to a certain Q value and the 2D diffraction image can be reduced into 1D diffraction pattern by correcting experimental factors, like the electronic noise from the detector, background scattering, multiple scattering etc., and performing azimuthal averaging along Q-values. There are several industry-standard softwares packages like FIT2D, PYFAI, that provide the capability of calibration and integration. Once the diffraction pattern is obtained, there are software tools for carrying

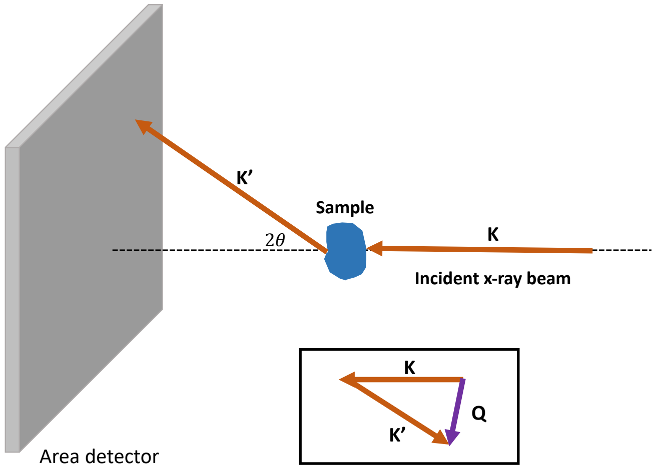


Figure 2.1: Schematic of the RA-PDF experimental setup.K and K' is the wave vector for the incident and scattered x-ray beam respectively. Q is the momentum transfer vector, which is showing in the inset.

out the transformation of PDF from x-ray scattering data PDFgetX3 [87] and neutron scattering PDFgetN [90] in a fast and reliable way.

2.4 Data analysis for PDF

Because of physical constraints in the experiment, only the scattering intensities from the interval $[Q_{\mathrm{min}}, Q_{\mathrm{max}}]$ are accessible. So the PDF measured $G(r)$ is in fact

$$\begin{aligned} G(r) &= \mathcal{F}(r) - \frac{2}{\pi}\left\{\int_0^{Q_{\mathrm{min}}} + \int_{Q_{\mathrm{max}}}^{\infty}\right\} F(Q)\sin{(Qr)}\,\mathrm{d}Q && (2.16) \\ &= \mathcal{F}(r) - \frac{2}{\pi}\int_0^{Q_{\mathrm{min}}} F(Q)\sin{(Qr)}\,\mathrm{d}Q. && (2.17) \end{aligned}$$

The contribution from the interval $[Q_{\mathrm{max}}, \infty]$ is dropped because work had shown the errors introduced by the high Q signal is minimal for high quality experiment [182]. The contribution from the interval $[0, Q_{\mathrm{min}}]$ is originated from the small-angle scattering signal, yielding a baseline which is a straight line for a bulk material and a function for a nanomaterial, depending its morphologies and size [52]. The delta-functions in Eq. 2.15 also broaden into Gaussian peaks in the measured signal to account for the thermal motion of atoms [50].

2.4.1 Structure modeling

The structural information encoded in a measured PDF can be extracted directly by analyzing the peaks. For a given peak, its position yields average separation of the atomic-pair in question, its integrated intensity (area under the peak) gives the coordination number of the atomic pair and its width and shape give the probability distribution of atomic position.

Though a good amount of information can be extracted with model-independent approaches, "structure modeling" is probably the most common approach as it yields fully quantitative information about the structure in question. Work had been devoted to understanding the structural and experimental factors that may appear in the measured PDF [182; 141; 83], therefore it is possible to compare a calculated PDF directly with the measured one. By varying the parameter values so that the difference between the measured and calculated PDF is minimized, it is then possible to draw inference about the structure in question based on the best-fit parameters. This process can be formulated as an optimization problem

$$\arg\min_{\theta} \left\| G(r) - G_{calc}(r;\, \theta) \right\|_2^2 , \tag{2.18}$$

where $G(r)$, $G_{calc}(r;\, \theta)$ stands for measured and calculated PDF respectively, θ is a vector of length p, where p is the total number of parameters considered in the model, and $\|\cdot\|_2$ is L-2 norm. There are script-based [89] and GUI-based [148; 53] programs for carrying the structure modeling step for PDF.

Chapter 3

Machine learning approach to determine the space group of a structure from the atomic PDF

3.1 Introduction

Crystallography is used to determine crystal structures from diffraction patterns [60], including patterns from powdered samples [138]. The analysis of single crystal diffraction is the most direct approach for solving crystal structures. However, powder diffraction becomes the best option when single crystals with desirable size and quality are not available.

A crystallographic structure solution makes heavy use of symmetry information to succeed. The first step is to determine the unit cell and space group of the underlying structure. Information about this is contained in the positions (and characteristic absences) of Bragg peaks in the diffraction pattern. This process of determining the unit cell and space group of the structure is know as "indexing" the pattern [60]. Indexing is inherently challenging for powder diffraction due to the loss of explicit directional information in the pattern, which is the result of projecting the data from three-dimensions into a one-dimensional pattern [48; 120]. However, there are a number of different algorithms available that work well in different situations [190; 41; 29; 5] Once the unit cell information is determined, an investigation on systematic absences of diffraction peaks is carried out to identify the space group. Various methods in determining space group information, based on either statistical or brute-force searches, have been used [132; 116; 4; 42].

The problem is even more difficult when the structural correlations only extend on nanometer length-scales as crystallography breaks down [24]. In this case progress can be made using atomic pair distribution function (PDF) methods for structure refinements [149; 50; 36; 210; 92]. PDFs may also be used for studying structures of bulk materials.

There has been some success in using PDF for structure solution [86; 25; 88; 40]. However, a major challenge for PDF structure solution is that, unlike powder diffraction case, a peak in the PDF simply indicates a characteristic distance existing in the structure but no overall information about the underlying unit cell [50]. Therefore, the symmetry information can not be inferred by the traditional indexing protocols that are predicated on the crystallography. However, to date there has not been a theory for identifying the space group directly given the PDF. Being able to determine the symmetry information based on the PDF will lead to more possibilities of solving structures from a wider class of materials.

Recently, machine learning (ML) has emerged as a powerful tool in different fields, such as in image classification [100] and speech recognition [73]. Moreover, ML models even outperform a human in cases such as image classifications [70] and the game of Go [168]. ML provides an platform of exploring the predictive relationship between the input and output of a problem, given a considerable amount of data is supplied for a ML model to “learn”. We know that the symmetry information is present in the powder diffraction pattern, and that the PDF is simply a Fourier transform of that pattern. We therefore reason that the symmetry information survives in the PDF though we do not know explicitly how it is encoded. We can qualitatively deduce that a higher symmetry structure, such as cubic, will produce a lower density of PDF peaks than a lower symmetry structure such as tetragonal. However, to date, there has not been a theory for identifying the space group directly, given the PDF. Here we attempt to see whether a ML algorithm can be trained to recognise the space group of the underlying structure, given a PDF as input. We note a recent paper that describes an attempt to determine the space group from powder diffraction pattern [137]. In this case a promising accuracy of 81 % was obtained in determining space group on simulated data, but the convolutional neural network (CNN) model they used was not able to determine space group from experimental data selected in their work.

To prepare data for training a ML model, we compute PDFs from 45 space groups, totaling $101,802$ structures, deposited in the Inorganic Crystal Structure Database (ICSD) [18]. The space groups chosen were the most heavily represented, accounting for more than 80% of known inorganic compounds [187].

The first ML model we tried was logistic regression (LR), which is a rather simple ML model. Although quite successful, we explored a more sophisticated ML model, a convolutional neural network (CNN). The CNN model outperforms the LR model by 15 %, reaching an accuracy of 91.9 % for obtaining the correct space group in the top-6 predicted results on the testing set. In particular, the CNN showed a significant improvement over LR in classifying challenging cases such as structures with lower symmetry.

The CNN model is also tested on experimental PDFs where the underlying structures are known but the data are subject to experimental noise and collected under various instrumental conditions. High accuracy in determining space groups from experimental PDFs was also demonstrated.

3.2 Machine Learning experiments

Machine learning (ML) is centered around the idea of exploring the predictive but oftentimes implicit relationship between inputs and outputs of a problem. By feeding considerable amount of input and output pairs (training set) to a learning algorithm, we hope to arrive at a prediction model which is a good approximation to the underlying relationship between the inputs and outputs. If the exact form of the output is available, either discrete or continuous, before the training step, the problem is categorized as “supervised learning” under the context of ML. The space-group determination problem discussed in this paper also falls into the supervised learning category. In the language of ML, the inputs are often denoted as “features” of the data and the outputs are usually called the “labels”. Both inputs and outputs could be a scalar or a vector. After learning the prediction model is then tested against a set of input and output pairs which have not seen by the training algorithm (the so-called testing set) in order to independently validate the performance of the prediction model.

In the context of the space group determination problem, the input that we want to interrogate is PDF data. We can select any feature or features from the data, for example, the feature we choose could be the PDF itself. The label is the space group of the structure that gave rise to the PDF. The database we will use to train our model is a pool of known structures. Strictly, we choose all the known structures from 45 most heavily represented space groups in the ICSD, which accounts for 80 % of known inorganic compounds [187]. These were further pruned to remove duplicate entries (same composition *and* same structure). The space groups considered and the number of unique structures in each space group are reproduced in Table 3.1.

Table 3.1: Space group and corresponding number of entries considered in this study.

Space group (order)	# of entries
$P\bar{1}(2)$	4615
$P2_1(4)$	581
$Cc(9)$	489
$P2_1/m(11)$	1247
$C2/m(12)$	3529
$P2/c(13)$	442
$P2_1/c(14)$	7392
$C2/c(15)$	3704
$P2_12_12_1(19)$	701
$Pna2_1(33)$	743
$Cmc2_1(36)$	525
$Pmmm(47)$	646
$Pbam(55)$	745
$Pnnm(58)$	477
$Pbcn(60)$	478
$Pbca(61)$	853
$Pnma(62)$	6930
$Cmcm(63)$	2249
$Cmca(64)$	575
$Cmmm(65)$	513
$Immm(71)$	754
$I4/m(87)$	569
$I4_1/a(88)$	397
$I\bar{4}2d(122)$	373

$P4/mmm(123)$	1729
$P4/nmm(129)$	1376
$P4_2/mnm(136)$	870
$I4/mmm(139)$	4028
$I4/mcm(140)$	1026
$I4_1/amd(141)$	700
$R\bar{3}(148)$	1186
$R3m(160)$	482
$P\bar{3}m1(164)$	1005
$R\bar{3}m(166)$	2810
$R\bar{3}c(167)$	1390
$P6_3/m(176)$	1289
$P6_3mc(186)$	849
$P6/mmm(191)$	3232
$P6_3/mmc(194)$	3971
$Pa\bar{3}(205)$	447
$F\bar{4}3m(216)$	2893
$Pm\bar{3}m(221)$	2933
$Fm\bar{3}m(225)$	4860
$Fd\bar{3}m(227)$	4382
$Ia\bar{3}d(230)$	455
total	101,802

We then computed the PDF from each of $101,802$ structures. The parameters capturing finite Q-range and instrumental conditions, are reproduced in Table 3.2. Those parameters

Table 3.2: Parameters used to calculate PDFs from atomic structures. All parameters follow the same definitions as in [53].

Parameter	Value
r_{min} (Å)	1.5
r_{max} (Å)	30.0
Q_{max} (Å^{-1})	0.5
Q_{min} (Å^{-1})	23.0
r_{grid} (Å)	$\frac{\pi}{Q_{\text{max}}}$
ADP (Å^2)	0.008
Q_{damp} (Å^{-1})	0.04
Q_{broad} (Å^{-1})	0.01

are chosen such that they are close to the values that is practically attainable at most synchrotron facilities. With the r-grid and r-range reported in Table 3.2, each computed PDF is a 209×1 vector, denoted $\mathbf{G}$. Depending on the atom types in the compounds, the amplitude of the PDF may vary drastically, which is inherently problematic for most ML algorithms [79] To avoid this problem, we determine a normalized form of each $\mathbf{G}$, $\mathbf{X}$ defined according to

$$\mathbf{X} = \frac{\mathbf{G} - \min(\mathbf{G})}{\max(\mathbf{G}) - \min(\mathbf{G})}, \tag{3.1}$$

where $\min(\cdot)$ and $\max(\cdot)$ mean taking the minimum and maximum value of $\mathbf{G}$ respectively. Since $\min(G)$ is always a negative number for the reduced PDF, $G(r)$, that we compute from the structure models, this definition results in the value of $\mathbf{X}$ always ranging between 0 and 1. An example of $\mathbf{X}$ from $\text{Li}_{18}\text{Ta}_6\text{O}_{24}$ (sapce group $P2/c$) is shown in Fig. 3.1(a).

For our learning experiments, we randomly select 80% of the data entries from each space

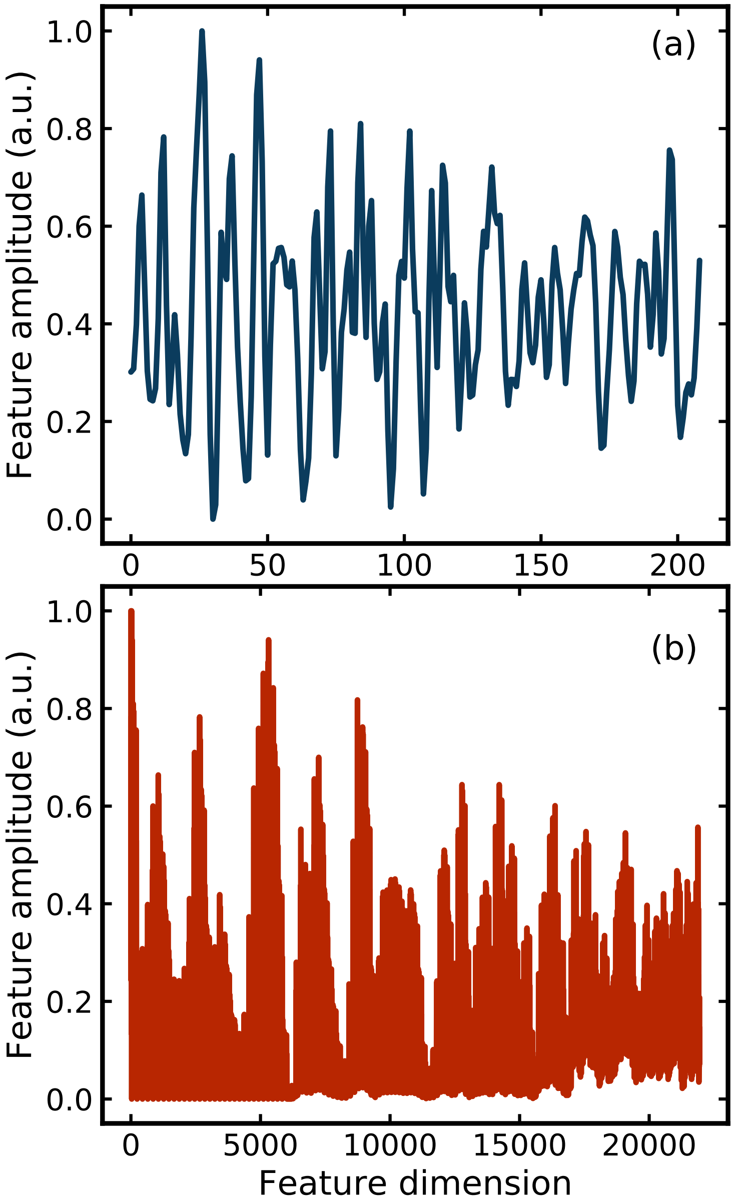


Figure 3.1: Example of (a) normalized PDF $\mathbf{X}$ and (b) its quadratic form $\mathbf{X}^2$ of compound $Li_{18}Ta_6O_{24}$ (space group $P2/c$).

group as the training set and reserve the remaining 20% of data entries as the testing set.

All learning experiments were carried out on one or multiple computation nodes of Habanero shared high performance cluster (HPC) at Columbia University. Each computation node consists of 24 cores of CPUs (Intel Xeon Processor E5-2650 v4), 128GB memory and 2 GPUs (Nvidia K80 GPUs).

3.2.1 Space Group Determination based on Logistic Regression (LR) model

We start our learning experiment with a rather simple model, logistic regression (LR). In the setup of the LR model the probability of a given feature being classified as a particular space groups is parametrized by a "logistic function" [69]. Forty-five space groups are considered in our study, therefore there are the same number of logistic functions, each with a set of parameters left to be determined. Since the space group label is known for each data in the training set, the learning algorithm is then used to find an optimized set of parameters to each of the forty-five logistic functions such that the overall probability of determining the correct space group on all training data is maximized. As a common practice, we also include "regularization" [69] to reduce overfitting in the trained model. The regularization scheme chosen in our implementation is "elastic net" which is known for encouraging sparse selections on strongly correlated variables [211]. Two hyperparameters α and Λ are introduced under the context of our regularization scheme. The explicit definition of these two parameters is presented in the Appendix section. Our LR model is implemented through SCIKIT-LEARN [140]. The optimum α, Λ for our LR model is determined by cross-validation [69] in the training stage.

The best LR model with $\mathbf{X}$ as the input yields an accuracy of 20 % at $(\alpha, \Lambda) =$

$(10^{-5}, 0.75)$. This result is better than a random guess from 45 space groups (2 %) but is still far from satisfactory. We reason that the symmetry information depends not on the absolute value of the PDF peak positions, which depend on specifics of the chemistry, but on their relative positions. This information may be more apparent in an autocorrelation of the PDF with itself, which is a quadratic feature in ML language. Our quadratic feature, $\mathbf{X}^2$, is defined as

$$\mathbf{X}^2 = \{\mathrm{X}_i\mathrm{X}_j \mid i, j = 1, 2, \ldots d,\ j > i\} \tag{3.2}$$

where d is the dimension of $\mathbf{X}$ and $\mathbf{X}^2$ is a vector of dimension $\frac{d(d-1)}{2} \times 1$. An example of the quadratic feature from $\mathrm{Li}_{18}\mathrm{Ta}_6\mathrm{O}_{24}$ (space group $P2/c$) is shown in Fig. 3.1(b).

The best LR model with $\mathbf{X}^2$ as the input yields an accuracy of 44.5 % at $(\alpha, \Lambda) = (10^{-5}, 1.0)$. This is much better than for the linear feature, but still quite low. However, the goal of space-group determination problem is to find the right space group, not necessarily to have it returned in the top position in a rank ordered list of suggestions. We therefore define alternative accuracy (A_6) that allow the correct space group to appear at any position in the top-6 space groups returned by the model. The values of A_i ($i = 1, 2, \ldots 6$) and their first discrete differences $\Delta A_i = A_i - A_{i-1}$ ($i = 2, 3, \ldots, 6$) of our best LR model are shown in Fig. 3.2. We observed a more than 10 % improvement in the alternative accuracy after considering top-2 predictions from the LR model (ΔA_2) and the improvement (ΔA_i) diminishes monotonically when more predictions are considered, as expected. Top-6 estimate is yielding a good accuracy (77 %) and this is still a small enough number of space groups that could be tested manually in any structure determination.

The ratio of correctly classified structures vs. space group number is shown Fig. 3.3(a). The space group numbering follows standard convention [67]. Higher space group number means a more symmetric structure and we find, in general, the LR model yields a decent performance in predicting space groups from structures with high symmetry but it performs

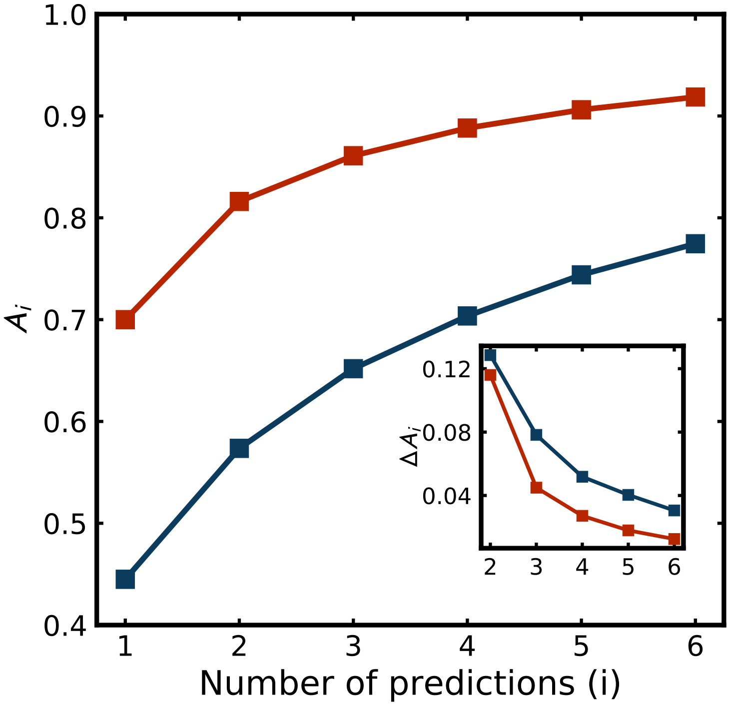


Figure 3.2: Accuracy in determining space group when top-i predictions are considered (A_i). Inset shows the first discrete differences ($\Delta A_i = A_i - A_{i-1}$) when i predictions are considered. Blue represents the result of the logistic regression model with $\mathbf{X}^2$ and red is the result from the convolutional neural network model.

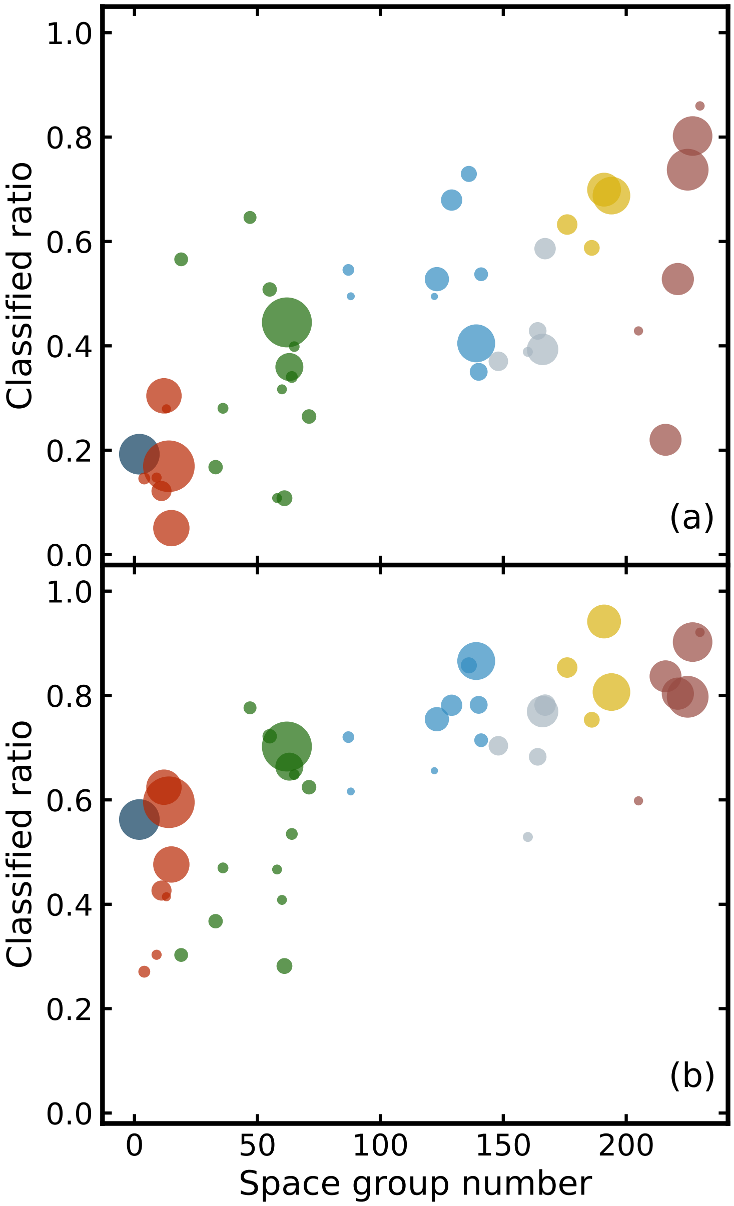


Figure 3.3: The ratio of correctly classified structures v.s. space group number from (a) logistic regression model (LR) with quadratic feature $\mathbf{X}^2$ and (b) convolutional neural network (CNN) model. Marker size reflects the relative frequency of space group in the training set. Markers are color coded with corresponding crystal systems (triclinic (dark blue), monoclinic (orange), orthorhombic (green), tetragonal (blue), trigonal (grey), hexagonal (yellow) and cubic (dark red).

poorly on classifying low symmetry structures.

3.2.2 Space group determination based on convolutional neural network (CNN)

The result from the linear ML model (LR) is promising, prompting us to move to a more sophisticated deep learning model. Deep learning models [106; 64] have been successfully applied to various fields, ranging from computer vision [71; 100; 150], natural language processing [10; 175; 94] to material science [152; 209]. In particular, we sought to use a convolutional neural network (CNN) [105].

The performance of a CNN depends on the overall architecture as well as the choice of hyperparameters such as the size of kernels, number of channels at each convolutional layer, the pooling size and the dimension of the fully-connected (FC) layer [64]. However there is no well-established protocol for selecting these parameters, which is a largely trial-and-error effort for any given problem. We build our CNN by tuning hyperparameters and validating the performance on the testing data, which is just 20% of the total data.

The resulting CNN built for the space-group determination problem is illustrated in Fig. 3.4. The input PDF is a 1D signal sequence of dimension $209 \times 1 \times 1$. We first apply a convolution layer of 256 channels with kernel size 32×1 to extract the first set of feature maps [105] of dimension $209 \times 1 \times 256$. It has been shown that applying a nonlinear activation function to each output improves not only the ability for a model to learn complex decision rules but also the numerical stability during the optimization step [106]. We chose rectified linear unit (ReLU) [43] as our activation function for the network. After the first convolution layer, we apply 64-channel kernel of size 32×1 to the first feature map and generate the second set of feature maps of dimension $209 \times 1 \times 64$. Similar to the first convolution layer,

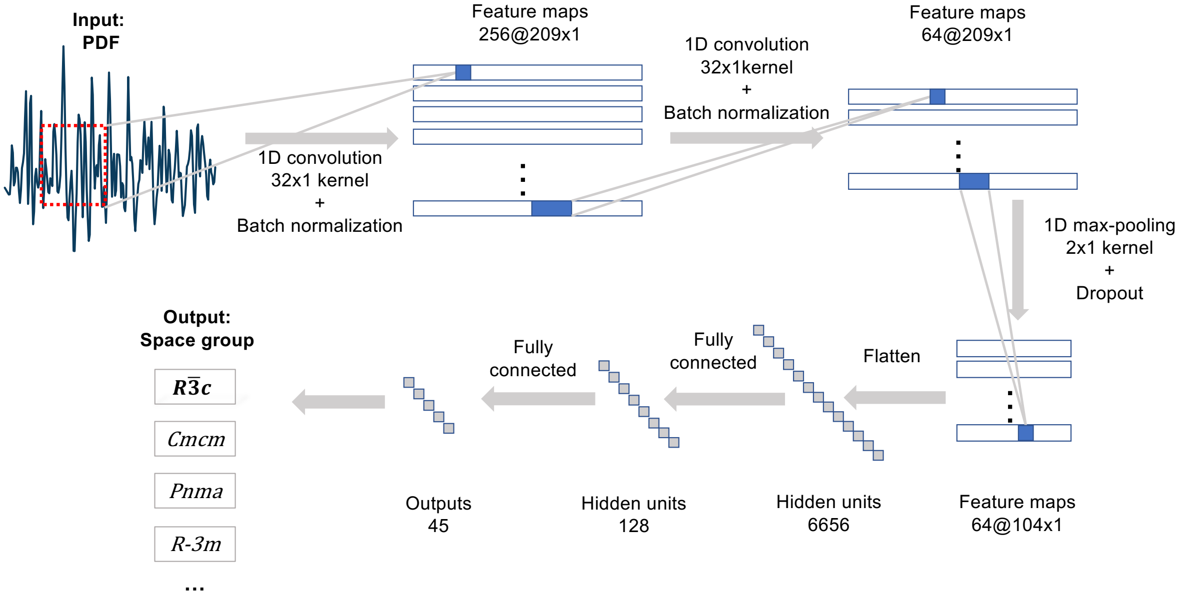


Figure 3.4: Schematic of our convolutional neural network (CNN) architecture.

the second feature map is also activated by ReLU. This is followed by a max-pooling layer [82] of size 2, which is applied to reduce overfitting. After the subsampling process in the max-pooling layer, the output is of size $104 \times 1 \times 64$ and it is then flattened to size of 6556×1 before two fully-connected layers of size 128 and 45 are applied. The first FC layer is used to further reduce the dimensionality of output from the max-pooling layer and it is activated with ReLu. The second FC layer is activated with softmax function [64] to output the probability of the input PDF being one of the 45 space groups considered in our study.

Categorical cross entropy loss [27] is used for training our model. It is apparent from Table 3.1 that the number of data entries in each space group are not evenly distributed, varying from 373 ($I\bar{4}2d$) to 7392 ($P2_1/c$) per space group. We would like to avoid the possibility of obtaining a neural network that is biased towards space groups with abundant data entries. To mitigate the effect of the unbalanced data set, loss from each training sample is multiplied by a class weight [95] which is the inverse of the ratio between the number of data entries from the same space group label in the training sample and the size of entire training set. We then use Adaptive Moment Estimation (Adam) [96] as the stochastic optimization method to train our model with a mini-batch size of 64. During the training step, we follow the same protocol outlined in [71] to perform the weight initialization [70] and batch normalization [78]. A dropout strategy [170] is also applied in the pooling layer to reduce over-fitting in our neural network. The parameters in the CNN model are iteratively updated through the stochastic gradient descent method (Adam).

Learning rate is a parameter that affects how drastically the parameters are updated at each iteration. A small learning rate is preferable when the parameters are close to some set of optimal values and vice versa. Therefore, an appropriate schedule of learning rate is crucial for training a model. Our training starts with a learning rate of 0.1, and the value is reduced by a factor of 10 at epochs 81 and 122. With the learning rate schedule described, the optimization loss against the testing set, along with the prediction accuracy on the training and testing sets, are plotted with respect to the number of epochs in Fig. 3.5. Our training is terminated after 164 epochs when the training accuracy, testing accuracy and optimization loss all plateau, meaning no significant improvement to the model would be gained with further updates to the parameters. Our CNN model is implemented with Keras [37] and trained on a single Nvidia Tesla K80 GPU.

Under the architecture and training protocol discussed above, our best CNN model yields an accuracy of 70.0 % from top-1 prediction and 91.9 % from top-6 predictions, which outperforms the LR model by 15 %. Similarly, from Fig. 3.2, we observe a more than 10 % improvement in the alternative accuracy after considering top-2 predictions (ΔA_2) in the CNN model and the improvement (ΔA_i) decreases monotonically, even in a more drastic trend than the case of LR model, when more predictions are considered.

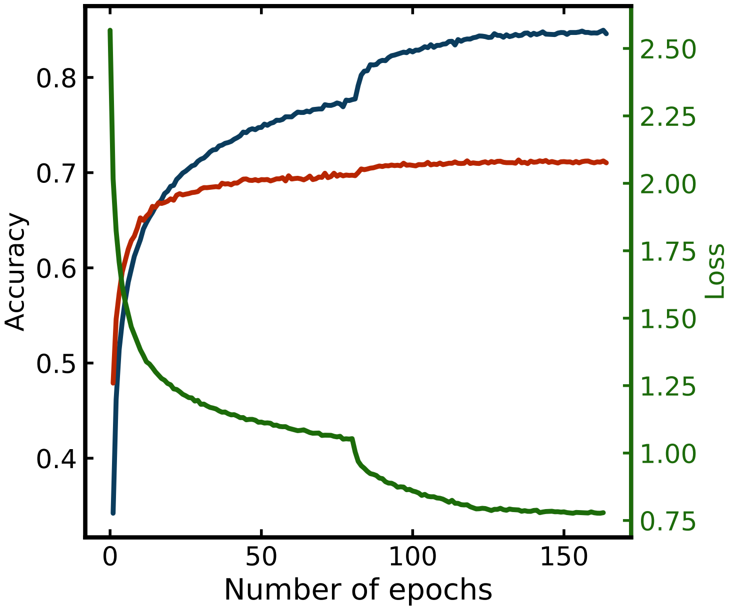


Figure 3.5: Accuracy of the CNN model on the training set (blue), the testing set (red) and the optimization loss against the testing set (green) with respect to number of epochs during the training step.

3.3 Results and discussion

3.3.1 Space group determination on calculated PDFs

The main result of the work is that, for the CNN model and defining success that the correct space-group is found in the top-6 choices, we achieve greater than 90% success rate (the correct space-group is returned in the top position 70 % of the time) when just the normalized PDF is given to the ML model. This success rate is much greater than random guessing and suggests that this approach may be a practically useful way of getting space-group information from PDFs. Below we explore in greater detail the performance of the CNN, including analyzing how it fails when it gets the answer wrong.

In general, it is fair to expect a ML model to achieve a higher accuracy on a space group with abundant training samples. However, from Fig. 3.3, it is clear that the LR model even fails to identify well represented space groups across all space group numbers. On the other hand, a positive correlation between the size of training data and the classification ratio is observed in the CNN model. Furthermore, except for space group $Ia\bar{3}d$ which is the most symmetric space group, the classification ratios on the rarely seen groups are lower than the well represented groups in our CNN model. However, the main result is that the CNN performs significantly better than the LR model for all space groups, especially on structures with lower symmetry. There is an overall trend towards increase in the prediction ability as the symmetry increases, and there are outliers, but there seems to be a trend that the CNN model is better at predicting space groups for more highly populated space groups.

The confusion matrix [172] is a common tool to assess the performance of a ML model. The confusion matrix, $\mathbf{M}$, is an N by N matrix, where N is the number of labels in the dataset. The rows of $\mathbf{M}$ identify the true label (correct answer) and the columns of $\mathbf{M}$ mean the label predicted by the model. The numbers in the matrix are the proportion of results in each category. For example, the diagonal elements indicate the proportion of outcomes where the correct label was predicted in each case, and the matrix element in the $Fd\bar{3}m$ row and the $F\bar{4}3m$ column (value 0.05) is the proportion of PDFs from an $Fd\bar{3}m$ space group structure that were incorrectly classified as being in space group $F\bar{4}3m$. For an ideal prediction model, the diagonal elements of the confusion matrix should be 1.0 and all off-diagonal elements would be zero. The confusion matrix from our CNN model is shown Fig. 3.6.

We observe "teardrop" patterns in the columns of $P\bar{1}$, $P2_1/c$ and $Pnma$, meaning the CNN model tends to incorrectly assign a wide range of space groups into these groups. On the surface, this behavior is worrying but the confusions actually correspond to the real group-subgroup relation which has been known and tabulated in literature [7; 31; 67]. For the case of $P\bar{1}$, the major confusion groups ($P2_1/c$, $C2/c$ and $P2/c$) are in fact minimal non-isomorphic supergroups of $P\bar{1}$. Moreover, $P2_12_12_1$ shares the same subgroup ($P2_1$) with $P2_1/c$ and $Pbca$ is a supergroup of $P2_12_12_1$ while $Pbcn$ is a supergroup of $P2_1/c$. Similar reasoning can be applied to the case of $P2_1/c$ and $Pnma$ as well. The statistical model appears to be picking up some real underlying mathematical relationships.

We also investigate the cases with low classification accuracy (low value in diagonal elements) from the CNN model. $P2_1$ is the group with the lowest accuracy (27 %) among all labels. The similar group-subgroup reasoning also holds for this case as well. $P2_1/c$ (32 % error rate) is, again, a supergroup of $P2_1$ and $C2/c$ (10 % error rate) is a supergroup of $P2_1/c$. The same reasoning holds for other confusion cases and we will not explicitly iterate through it here, but this suggests that these closely group/sub-group related space groups should also be considered whenever the CNN model returns another one in the series. It is possible to train a different CNN model which focuses on disambiguating space groups that are closely related by the group/sub-group relationship. However, we did not implement this kind of hierarchical model in our study.

3.3.2 Space Group Determination on Experimental PDFs

The CNN model is used to determine the space group of 15 experimental PDFs and the results are reported in Table 3.3. For each experimental PDF, structures are known from previous studies which are also referenced in the table. Both crystalline (C) or nanocrystalline (NC) samples with a wide range of structural symmetries are covered in this set of experimental PDFs. It is worth noting that the sizes of the NC samples chosen are roughly equal to or larger than 10 nm, at which size in our measurements the PDF signal from the NC material falls off roughly at the same rate as that from crystalline PDFs in the training set. Every experimental PDF is subject to experimental noise and collected under various instrumental conditions that result in aberrations to the PDF that are not identical to parameter values used to generate our training set (Table 3.2). It is therefore expected that the CNN classifier will work less well than on the testing set. From Table 3.3, it is clear that the CNN model yields an overall satisfactory result in determining space groups from experimental data with the space group from 12 out of 15 test cases properly identified in the top-6 predictions.

Here we comment on the performance of the CNN. In the cases of $IrTe_2$ at 10 K, the material has been reported in the literature in both $C2/m$ and $P\bar{1}$ space groups [118; 183], and it is not clear which is correct. The CNN returned both space groups in the top six. Furthermore, for data from the same sample at room temperature, the CNN model identifies not only the correct space group ($P\bar{3}m1$), but also the space groups that the structure will occupy below the low-temperature symmetry lowering transition ($C2/m$, $P\bar{1}$). For the case of $BaTiO_3$ nanoparticles, the CNN model identifies two space groups that are considered in the literature to yield rather equivalent explanatory power ($R3m$, $P4/mmm$) [102; 136]. It

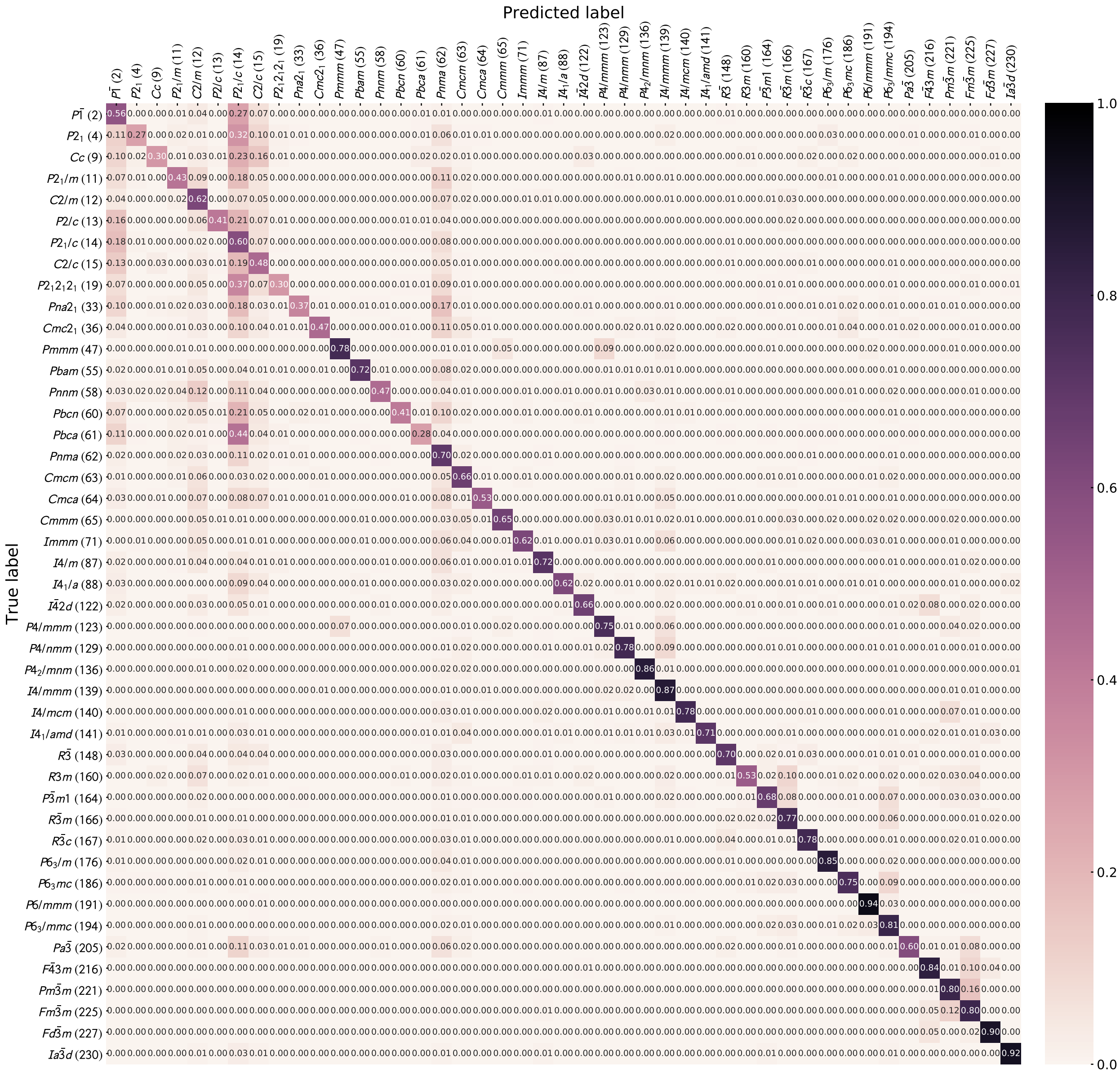


Figure 3.6: The confusion matrix of our CNN model. The row labels indicate the correct space group and the column labels the space group returned by the model. An ideal model would result in a confusion matrix with all diagonal values being 1 and all off-diagonal values being zero. The numbers in parentheses are the space-group number.

is encouraging that the CNN appears to be getting the physics right in these cases.

Investigating the failing cases from the CNN model (entries with a dagger in Table 3.3) also reveals insights about the decision rules learned by the model. Sr_2IrO_4, was firstly identified as a perovskite structure with space group $I4/mmm$ [153], but later work pointed out that a lower symmetry group $I4_1/acd$ is more appropriate due to correlated rotations of the corner-shared IrO_6 octahedra about the c-axis [76; 167]. There is a long-wavelength modulation of the rotations along the c-axis resulting a supercell with a five-times expansion along that direction ($a = 5.496$ Å, $c = 25.793$ Å). The PDF will not be sensitive to such a long-wavelength superlattice modulation which may explain why the model does not identify a space group close to the $I4_1/acd$ space group, reflecting additional symmetry breaking due to the supermodulation. It is not completely clear what the space group would be for the rotated octahedra without the supermodulation, so we are not sure if this space group is among the top-6 that the model found.

Somewhat surprisingly the CNN fails to find the right space group for wurtzite CdSe, which is a very simple structure, but rather finds space groups with low symmetries. One possible reason is that we know there is a high degree of stacking faulting in the bulk CdSe sample that was measured. This was best modelled as a phase mixture of wurtzite (space group $P6_3mc$) and zinc-blende (space group $F\bar{4}3m$) [117]. The prediction of low symmetry groups might reflect the fact the underlying structure can not be described with a single space group.

3.4 Conclusion

We demonstrate an application of machine learning (ML) to determine the space group directly from an atomic pair distribution function (PDF). We also present a convolutional

Table 3.3: Top-6 space-group predictions from the CNN model on experimental PDFs. Bold-faced prediction is the most probable space group from existing literatures listed in the Refs. column. More than one predictions are highlighted when these space groups are regarded as highly similar in literatures. Details about these cases will be discussed in the text. The Note column specifies if the PDF is from a crystalline (C) or nanocrystalline (NC) sample. The experimental data were collected under various instrumental conditions which are not identical to the training set and experimental data were measured at the room temperature, unless otherwise specified. Dagger is used to label the data that the CNN model fails to predict the correct space group.

Sample	1st	2nd	3rd	4th	5th	6th	Refs.	Note
Ni	$\boldsymbol{Fm\bar{3}m}$	$Pm\bar{3}m$	$Fd\bar{3}m$	$F\bar{4}3m$	$P4/mmm$	$P6_3/mmc$	[135]	C
Fe_3O_4	$\boldsymbol{Fd\bar{3}m}$	$I4_1/amd$	$R\bar{3}m$	$Fm\bar{3}m$	$F\bar{4}3m$	$P6_3/mmc$	[56]	C
CeO_2	$\boldsymbol{Fm\bar{3}m}$	$Fd\bar{3}m$	$Pm\bar{3}m$	$F\bar{4}3m$	$Pa\bar{3}$	$P4/mmm$	[199]	C
$Sr_2IrO_4^\dagger$	$Fm\bar{3}m$	$P6/mmm$	$P6_3/mmc$	$Pm\bar{3}m$	$Fd\bar{3}m$	$R\bar{3}m$	[76; 167]	C
$CuIr_2S_4$	$\boldsymbol{Fd\bar{3}m}$	$Fm\bar{3}m$	$F\bar{4}3m$	$R\bar{3}m$	$Pm\bar{3}m$	$R3m$	[58]	C
$CdSe^\dagger$	$P2_1/c$	$P\bar{1}$	$C2/c$	$Pnma$	$Pna2_1$	$P2_12_12_1$	[117]	C
$IrTe_2$	$C2/m$	$\boldsymbol{P\bar{3}m1}$	$P2_1/c$	$P\bar{1}$	$P2_1/m$	$C2/c$	[118; 202]	C
$IrTe_2$@10K	$\boldsymbol{C2/m}$	$P6_3/mmc$	$P6/mmm$	$P4/mmm$	$\boldsymbol{P\bar{1}}$	$P2_1/c$	[118; 183]	C
Ti_4O_7	$\boldsymbol{P\bar{1}}$	$C2/c$	$P2_1/c$	$C2/m$	$Pnnm$	$P4_2/mnm$	[114]	C
$MAPbI_3$@130K	$P\bar{1}$	$P2_1/c$	$C2/c$	$P2_12_12_1$	$\boldsymbol{Pnma}$	$Pna2_1$	[176]	C
$MoSe_2$	$\boldsymbol{P6_3/mmc}$	$R3m$	$R\bar{3}m$	$P6_3mc$	$P4/mmm$	$Fd\bar{3}m$	[80]	C
TiO_2(anatase)	$\boldsymbol{I4_1/amd}$	$C2/m$	$P2_1/m$	$C2/c$	$P\bar{1}$	$P2_1/c$	[74]	NC
TiO_2(rutile)	$\boldsymbol{P4_2/mnm}$	$C2/m$	$P2_1/c$	$P\bar{1}$	$P2_1/m$	$Pnma$	[13]	NC
$Si^\dagger$	$P6_3mc$	$I\bar{4}2d$	$R3m$	$C2/c$	$P\bar{1}$	$Pbca$	[160]	NC
$BaTiO_3$	$\boldsymbol{R3m}$	$\boldsymbol{P4/mmm}$	$C2/m$	$P6_3/mmc$	$Pnma$	$Cmcm$	[102; 136]]	NC

neural network (CNN) model which yields a promising accuracy (91.9 %) from the top-6 predictions when it is evaluated against the testing data. Interestingly, the trained CNN model appears to capture decision rules that agree with the mathematical (group-subgroup) relationships between space groups. The trained CNN model is tested against 15 experimental PDFs, including crystalline and nanocrystalline samples. Space groups from 12 of these experimental data were successfully found in the top-6 predictions by the CNN model. This shows great promise for preliminary, model-independent assessment of PDF data from well ordered crystalline or nanocrystalline materials.

3.5 Appendix

3.5.1 Logistic Regression and Elastic Net Regularizations

Consider a dataset with total M structures and K distinct space-group labels. Each structure has a space group and we denote the space group of m-th structure as k_m where $k_m \in \{1, 2, \dots K\}$, our complete set of space groups. In the setup of LR model, the probability of a feature x_m of dimension d, which is a computable from m-th structure, belongs to a specific space group k_m is parametrized as

$$\Pr(k_m|x_m, \beta^{k_m}) = \frac{\exp\left(\beta_0^{k_m} + \sum\limits_{i=1}^{d} \beta_i^{k_m} x_{m,i}\right)}{1 + \exp\left(\beta_0^{k_m} + \sum\limits_{i=1}^{d} \beta_i^{k_m} x_{m,i}\right)}, \tag{S1}$$

where $\beta^{k_m} = \{\beta_0^{k_m}, \beta_1^{k_m}, \dots, \beta_d^{k_m}\}$ is a set of parameters to be determined. The index k_m runs from 1 to 45 which corresponds total number of space groups considered in our study. Since the space group k and feature x are both known for the training data, the learning algorithm is then used to find a optimized set of $\beta = \{\beta^{k_m} : k_m = 1, 2, \dots, K\}$ which maximizes the overall probability in determining correct space group $\Pr(k_m|x_m, \beta^{k_m})$ on all

M training data.

For each of the M structures, there will be a binary result for classification; Either the space group label is correctly classified or not. This process can be regarded as M independent Bernoulli trials. The probability function for a single Bernoulli trial is expressed as

$$f(k_m|x_m,\boldsymbol{\beta}^{k_m}) = \left[\Pr(k_m|x_m,\boldsymbol{\beta}^{k_m})\right]^{\gamma_m} \left[1-\Pr(k_m|x_m,\boldsymbol{\beta}^{k_m})\right]^{1-\gamma_m}, \tag{S2}$$

where γ is an indicator. $\gamma_m = 1$ if the space-group label k_m is correctly predicted and $\gamma_m = 0$ if the prediction is wrong. Since each classification are independent, the joint probability function for M classifications on the space-group label, $f_M(\boldsymbol{K}|x,\boldsymbol{\beta})$, is written as

$$f_M(\mathbf{K}|\mathbf{x},\boldsymbol{\beta}) = \prod_{m=1}^{M} f(k_m|x_m,\boldsymbol{\beta}^{k_m}), \tag{S3}$$

where $\mathbf{K} = \{k_m\}$ and $\mathbf{x} = \{x_m\}$. Furthermore, since both the label and features are known in the training set, Eq. S3 is just a function of β,

$$L(\beta) = f_M(\mathbf{K}|\mathbf{x},\boldsymbol{\beta}) \tag{S4}$$

Logarithm is a monotonic transformation. Taking logarithm of Eq. S4 does not change the original behavior of the function and it improves the numerical stability as the product of probabilities is turned into sum of logarithm of probabilities and extreme values from the product can still be computed numerically. We therefore arrive the "log-likelihood" function

$$l(\beta) = \log(L(\beta)) \tag{S5}$$

It is common to include "regularization" [69] for reducing overfitting in the model. The regularization scheme chosen in our implementation is "elastic net" which is known for encouraging sparse selections on strongly correlated variables [211]. The explicit definitions of the log-likelihood function with elastic regularization is written as

$$l_t(\beta) = l(\beta) + \alpha\Big(\Lambda \left\|\beta\right\|_1 + (1-\Lambda) \left\|\beta\right\|_2^2\Big), \tag{S6}$$

where $\left\|\cdot\right\|$ and $\left\|\cdot\right\|_2^2$ stands for L1 and L2 norm [75] respectively. Two hyperparameters α and Λ are introduced under this regularization scheme. α is a hyperparameter that determines the overall "strength" of the regularization and Λ governs the relative ratio between L1 and L2 regularization [211]. Detailed steps on optimizing Eq. S6 is beyond the scope of this paper, but they are available in most of standard ML reviews [69; 27].

3.5.2 Robustness of the CNN model

The classification accuracies from CNN models with different sets of hyperparameters, such as number of filters, kernel size and pooling size, are reproduced in Table S1. The classification accuracy only vary modestly across different sets of hyperparameters and this implies the robustness of our CNN architecture. We determined the desired architecture of our CNN model based on the classification accuracy on the testing set and the learning curves (loss, training accuracy and testing accuracy) reported in Fig. 3.5.

Table S1: Accuracies of CNN model with different sets of hyper parameters. Accuracy is abbreviated as accu. in the table. The last row specifies the optimum set of hyperparmeters for our final CNN model.

# filters	kernel size	# hidden units	# ensembles	Top-1 accu. (%)	Top-6 accu. (%)
128, 32	24	128	2	64.1	90.7
256, 64	24	128	2	68.6	91.6
64, 64	24	128	2	67.4	91.1
128, 64	32	128	2	69.0	91.7
128, 64	16	128	2	66.6	91.3
128, 64	24	256	2	69.2	91.6
128, 64	24	64	2	66.4	91.2
128, 64	24	128	1	65.7	91.1
128, 64	24	128	3	68.2	91.6
256, 64	**32**	**128**	**3**	**70.0**	**91.9**

Chapter 4

sasPDF: pair distribution function analysis of nanoparticle superlattice assemblies from small-angle-scattering data

4.1 Introduction

With the advent of high degrees of control over nanoparticle synthesis [129; 77; 47] attention is turning to assembling superlattices of them as metamaterials [28; 35] and applications of nanoparticle assemblies (NPA) based devices such as solar cells and field effect transistors have been demonstrated [178; 164; 177]. It is crucial to study the structures of these NPAs if their properties are to be optimized. For example, it has been shown that the mechanical [1], optical [201], electrical [189] and magnetic [174] properties can be further engineered by controlling the spatial arrangement of the constituents in the NPA.

CHAPTER 4. SASPDF: PAIR DISTRIBUTION FUNCTION ANALYSIS OF NANOPARTICLE SUPERLATTICE ASSEMBLIES FROM SMALL-ANGLE-SCATTERING DATA

Getting detailed quantitative structural information from NPAs, especially in 3D, is a challenging and largely unsolved problem. Small angle scattering and electron microscopy (EM) have been the major techniques for studying the structure of NPAs [128; 179]. The technique of TEM yields high-resolution images of NPAs. To obtain quantitative structural information it is necessary to either analyze the images manually [192] or match observed images with patterns that are algorithmically generated from known structures [166]. This approach can yield the structure types [208] but does not typically result in the kind of quantitative 3D structural information we are used to obtaining for atomic structures of crystals, including accurate inter-particle vectors and distributions of inter-particle distances, or the range of structural coherence of the packing order. It is desirable to explore scattering approaches that can yield that kind of information.

The technique of small-angle x-ray or neutron scattering (SAS) has been an important tool to study objects that have sizes from nano- to micrometer length-scales [186; 63; 66; 97], such as large nanocrystals [147] and biological molecules [97], yielding information about the intrinsic shape, size distributions and scattering density of objects on these scales [63; 16; 139; 191; 17].

When these nanoscale objects aggregate, correlation peaks appear in the SAS data resembling atomic-scale interference peaks (Diffuse scattering and Bragg peaks), but encoding information about particle packing rather than atomic packing [128; 133]. Obtaining structural information about the NPAs from these correlation peaks appears to be a promising approach. Although the recent developments in SAS modeling demonstrates the ability to account for phase, morphology and orientations of NPs in a lattice [197; 111], fitting the SAS data with robust structural models to obtain quantitative information about the structure has barely been explored [112]

On the other hand, the atomic pair distribution function (PDF) analysis of x-ray and neutron powder diffraction has proven to be a powerful tool for characterizing local order in materials, and for extracting quantitative structural information [149; 50; 210; 92] when the atoms are not long-range ordered, as is the case in nanoparticles. Here we extend PDF analysis to handle correlation peaks in the small angle scattering data, allowing us to study the arrangement of particles in nanoparticle assemblies using the same quantitative modeling tools that are available for studying the atomic arrangements in nanoparticles themselves. We describe the extension of the PDF equations in the small-angle scattering (SAS) regime and describe the data collection protocol for optimum data quality. We also present the PDFgetS3 software package that can be readily used to extract the PDF from small-angle scattering data. We then apply the sasPDF method to investigate structures of some representative NPA samples with different levels of structural order.

4.2 Samples

To test the method we obtained SAS data from the samples listed in Table 1. Synthesis details of these NPA samples can be found in the references listed in the table.

4.3 sasPDF method

The data were collected using a standard SAXS setup at an x-ray synchrotron source, with a 2D area detector mounted perpendicular to the beam in transmission geometry. Both the Cu_2S NPA and the SiO_2 NPA samples were measured at beamline 11-BM at the National Synchrotron Light Source-II (NSLS-II). The Cu_2S NPA powders were sealed between two rectangular Kapton tapes with a circular deposited area of diameter about 3 mm and thickness about 0.2 mm. The SiO_2 NPA formed a circular, free-standing stable film of diameter about 5 mm and thickness about 1 mm which was mounted perpendicular to the beam and

Table 1: Nanoparticle assemblies (NPA) considered in this study. Building block indicates the NP and surfactant linkers used to build the assemblies. D is the particle diameter (one standard deviation in parentheses) estimated from TEM images and reported in the original publications listed in the Ref. column. Beamline is the x-ray beamline where the SAXS data were measured (see text for details). PMA is Poly(methyl acrylate) and DDT is dodecanethiol.

Sample	Building block	D (nm)	Beamline	Ref.
Au NPA	DNA-capped Au NP	11.4(1.0)	X21	[133]
Cu_2S NPA	DDT-capped Cu_2S NP	16.1(1.3)	11-BM	[68]
SiO_2 NPA	PMA-capped SiO_2 NP	14(4)	11-BM	[21]

no further sealing was carried out. The scattering data of the Cu_2S NPA and SiO_2 NPA samples were collected with a Pilatus 2M (Dectris, Switzerland) detector with a sample-detector distance 2.02 m using an x-ray wavelength of 0.918 Å. An example of the diffraction image from the Cu_2S NPA is shown in the inset of Fig. 4.1. The scattering from these samples is isotropic as the sample consists of powders of randomly oriented NPA crystallites, and the 2D diffraction images can be reduced to a 1D diffraction pattern, $I_m(Q)$, by performing an azimuthal integration around rings of constant scattering angle on the detector. This was done using PYFAI [93]. This requires a calibration of the experiment geometry described below, but the integrated 1D pattern from the 2D diffraction image is shown in Fig. 4.1. The relative positions and intensities of sharp peaks in the $I_m(Q)$ originate from the Debye-Scherrer rings in the 2D image. We need to use a data acquisition strategy that mitigates effects of x-ray beam-damage to the sample. The linkers that connect nanoparticles in the assemblies play a crucial role for the NPA structure formed but are susceptible to degrada-

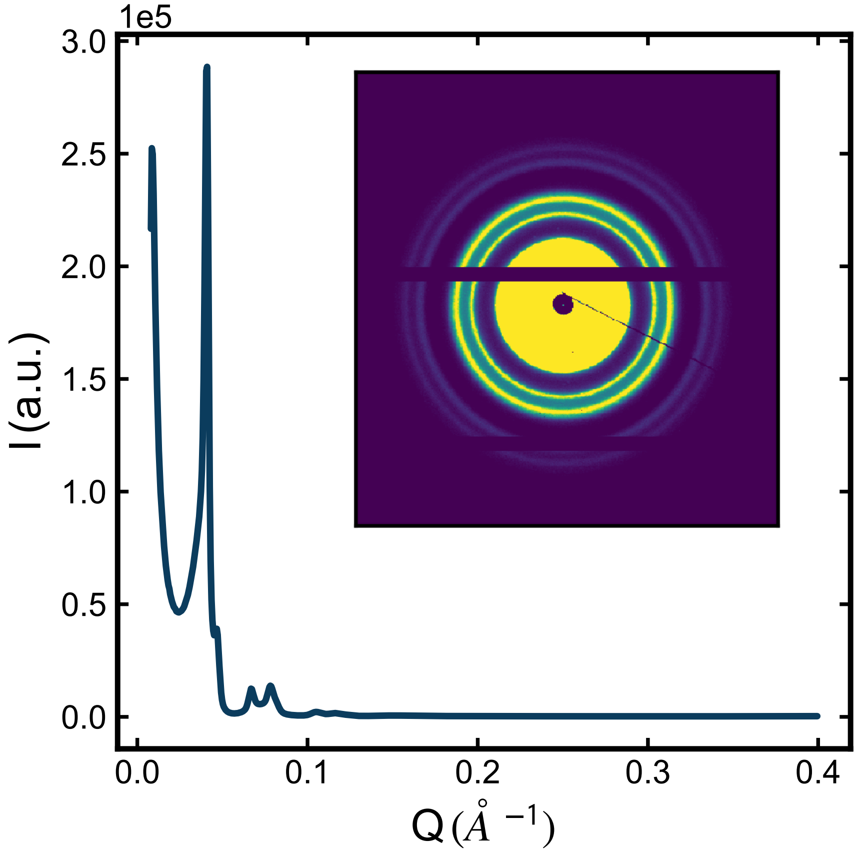


Fig. 4.1: Example of the 1D diffraction pattern $I_m(Q)$ from the Cu_2S NPA sample. The data were collected with the spot exposure time and scan exposure time reported in the text. The inset shows the corresponding 2D diffraction image. The horizontal stripes in the image are from the dead zone between panels of the detector. The diagonal line is the beam-stop holder.

tion in the intense x-ray beam that may result in changes in the NPA structure. To describe the strategy we separate the concepts of a "spot exposure time" and the "scan exposure time". The latter is the total integrated exposure time to obtain a dataset with sufficient statistics. The former is the length of time that any spot on the sample is exposed. The scan exposure then consists of multiple spot exposures, where the sample is translated after each spot exposure so that a fresh region of sample is exposed. For ease of experimentation we would like to determine a spot exposure time that is as long as possible whilst ensuring that the sample has not degraded significantly during that exposure. We have found that the maximum safe spot exposure time depends on the nature of the NPA sample, as well as experimental conditions such as x-ray energy, flux and sample temperature. It therefore requires a trial-and-error approach to determine it. To choose the optimal spot exposure time we locate the beam on a fixed spot of the sample and take a sequence of short exposures, monitoring for significant changes in the intensity of the strongest correlation peak in $I_m(Q)$. The spot exposure time determined this way for our experimental setup was 30 s for both Cu_2S NPA and SiO_2 NPA samples and the scan exposure time was 5 minutes (30 s, 10 spots) for the Cu_2S NPA sample and 10 minutes (30 s, 20 spots) for the SiO_2 NPA sample.

The scan exposure time is estimated based on an assessment of noise in the PDF given a desired Q_{max}, but it depends sensitively on the counting statistics in the high-Q region of the diffraction pattern, which is easiest to assess by looking in the high-Q region of the reduced structure function $F(Q)$. For illustration purposes, the effect of scan exposure time on the $F(Q)$ (and the resulting PDF) is illustrated in Fig. 4.1 of Appendix section.

For the calibration of the experimental geometry, such as sample-detector distance and detector tilting we use the calibration capability in the Python package PYFAI [93]. We first measured silver behenate (AgBh) [62] as a well characterized calibration sample. The d-spacing of the calibration sample, the x-ray wavelength and the pixel dimensions of the detector are known, which allows the geometric parameters to be refined in PYFAI. We found that selecting the strongest few rings (even just two or three work well) in the image allowed a stable refinement of the calibration parameters.

Finally, in this study we also consider legacy data from measurements carried out previously [133]. The data of the Au NPA sample were collected at beamline X21 at the National Synchrotron Light Source (NSLS) from a sample loaded into a wax-sealed 1 mm diameter quartz capillary. The scattering data were collected with a MarCCD (Rayonix, USA) area detector using an x-ray wavelength of 1.55 Å. Details of the measurements are reported in [133].

The PDF, denoted $G(r)$, is a truncated sine Fourier transform of the reduced structure function $F(Q) = Q\left[(S(Q) - 1)\right]$ [50]

$$G(r) = \frac{2}{\pi} \int_{Q_{\min}}^{Q_{\max}} F(Q) \sin(Qr)\, \mathrm{d}Q. \tag{1}$$

Since $F(Q)$ can be easily computed once $S(Q)$ is available, we will first focus on describing the precise definition of $S(Q)$ and its relation to the measured diffraction pattern $I_m(Q)$. The measured intensity, $I_m(Q)$, depends on experimental details such as the flux, and beam size of the x-ray source, the data collection time and the sample density. From the point of developing the SASPDF formalism, we will focus on the coherent scattering intensity $I_c(Q)$ [50] which is obtained after correcting $I_m(Q)$ for the experimental factors as we describe below.

The coherent scattering intensity $I_c(Q)$ from a unit cell with N_s atoms is [50; 65]

$$I_c(\mathbf{Q}) = \sum_{m=1}^{N_s} \sum_{n=1}^{N_s} f_m^*(\mathbf{Q}) f_n(\mathbf{Q}) \exp\left[i\mathbf{Q} \cdot (\mathbf{r}_m - \mathbf{r}_n)\right], \tag{2}$$

where $\mathbf{Q}$ is the scattering vector, $f_m(\mathbf{Q})$ and $\mathbf{r}_m$ are the atomic form factor amplitude and position of m-th atom in the unit cell, respectively.

If the scattering from a sample is isotropic, for example, it is an untextured powder or a liquid with no anisotropy, the observed scattering intensity will depend only on the magnitude of $\mathbf{Q}$, $|\mathbf{Q}| = Q$ and not its direction in space. The observed scattering intensity in this case will depend on the orientationally averaged $I_c(\mathbf{Q})$,

$$I_c(Q) = \left\langle \sum_{m=1}^{N_s} \sum_{n=1}^{N_s} f_m^*(\mathbf{Q}) f_n(\mathbf{Q}) \exp\left[i\mathbf{Q} \cdot (\mathbf{r}_m - \mathbf{r}_n)\right] \right\rangle, \tag{3}$$

where $\langle \cdot \rangle$ denotes the orientational average.

This formalism is readily extended to the case where the scattering objects are not atoms, but are some other finite-sized object, for example, nanoparticles. In this case, the atomic form-factor would be replaced with the form-factor for the scattering objects in question. The form factor $f(\mathbf{Q})$ for a generalized scatterer, with volume V and its electron density as a function of position $\rho(\mathbf{r})$ is [65]

$$f(\mathbf{Q}) = \int_V \left[\rho(\mathbf{r}) - \rho_0\right] \exp\left(i\mathbf{Q} \cdot \mathbf{r}\right) \mathrm{d}\mathbf{r}, \tag{4}$$

where ρ_0 is the average electron density of the ambient environment of the scatterers.

In situations where there is only one type of scatterer we pull the form factors out of the sum, and if the electron density of the scatterer is approximately spherical Eq. 2.5 may be further simplified to [65; 50]

$$I_c(Q) = N_s \left\langle f^2(Q) \right\rangle + \left\langle f(Q) \right\rangle^2 \left\langle \sum_{m=1}^{N_s} \sum_{n \neq m}^{N_s} \exp\left[i\mathbf{Q} \cdot (\mathbf{r}_m - \mathbf{r}_n)\right] \right\rangle. \tag{5}$$

Following the Faber-Ziman formalism [51],

$$S(Q) = \frac{I_c(Q)}{N_s \langle f(Q) \rangle^2} - \frac{\langle f^2(Q) \rangle - \langle f(Q) \rangle^2}{\langle f(Q) \rangle^2}, \tag{6}$$

we plug in $\langle f^2(Q) \rangle = \langle f(Q) \rangle^2$ and Eq. 6 becomes

$$I_c(Q) = N_s \langle f^2(Q) \rangle S(Q). \tag{7}$$

This expression is equivalent to representing the scatterers as points at the position of their scattering center, convoluted with their electron distributions. The resulting structure function, $S(Q)$, yields the arrangement of scatterers in the sample. This expression is often expressed in the SAS literature as [65]

$$S(Q) = \frac{I_c(Q)}{N_s P(Q)}. \tag{8}$$

Where $P(Q)$ is equivalent to $\langle f^2(Q) \rangle$ [65], the orientational average of the square of the form-factor. We note that, as with the atomic PDF, the above analysis can be generalized to the cases of scattering from multiple types of scatterers [98; 197; 165] and in the SAS case approximate corrections for asphericity of the electron density [85; 162; 207], may be applied.

To determine $S(Q)$ we need to have $P(Q)$. $P(Q)$ can be computed from a given electron density, or directly measured. For the case of a NPA sample, the precise scattering properties of the NP ensemble in the sample, including any polydispersity or distribution of geometric shapes, are not always known, therefore it is best to measure the form factor directly, as described below. In general we do not know N_s and all of the experimental factors (for example, the incident flux, multiple scattering and so on). The algorithm [26] that is widely used for carrying out corrections for these effects in the atomic PDF literature [87] is also suitable for the SAS data. It takes advantage of our knowledge of the asymptotic behavior of the $S(Q)$ function to obtain an *ad hoc* but robust estimation of $S(Q)$ from the measured $I_m(Q)$. This is described in detail in [87]. The resulting scale of the PDF is not well determined, but when fitting models to the data this is not a problem [141], and in practice it gives close to a correct scale for high quality measurements. Here we show that we can take the same approach to obtain the PDF from the measured SAS data here.

In the test experiments we describe here, in each case the form factor of the nanoparticles was obtained from a measurement. The NPs are suspended in solvent at a sufficient dilution to avoid significant inter-particle correlations. The SAS signal of the dilute NP solution is measured with good statistics over the same range of Q as the measurement of the nanoparticle assemblies themselves, and ideally on the same instrument. The signal of the solvent and its holder is also measured and then subtracted from the SAS signal of the dilute NP solution to obtain the correct particle form factor signal. We emphasize that it is important to measure exactly the same batch of NPs to have an accurate form factor for the NPA sample considered.

A form factor measured with high statistics is crucial as the signal in $I_c(Q)$ is weak in the high-Q region and noise from the $P(Q)$ measurement can be significant in this region. Fig. 4.2 shows the effect on $F(Q)$ (and the resulting PDF), when processed using $P(Q)$ from different scan exposure times. It is clear that the statistics of the form-factor measurement has a significant effect on the results. In cases where any signal in $P(Q)$ does not change rapidly it may be smoothed to reduce the effects of limited statistics, at the cost of possibly introducing bias if the smoothing is not done ideally. This will be particularly relevant when the nanoparticles are not monodisperse, as is somewhat common.

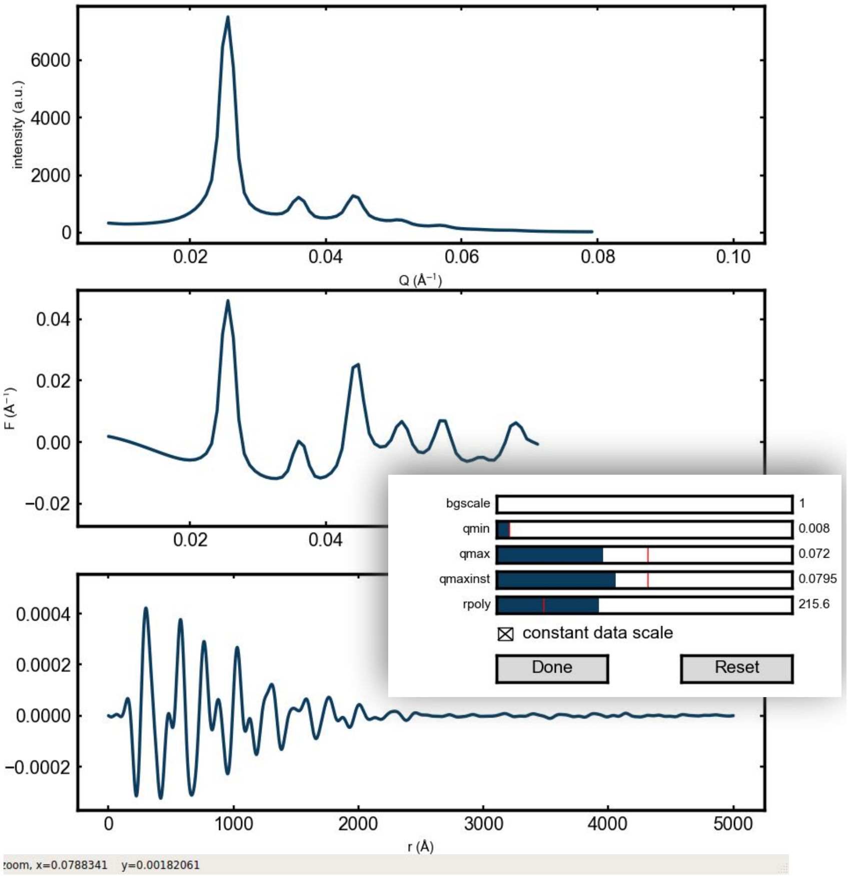
The experimental PDF $G(r)$ is then obtained via the Fourier transformation, Eq. 1. The success of the sasPDF method depends heavily on the good statistics (high signal-to-noise ratio) throughout the entire diffraction pattern $I_c(Q)$ and the form factor $P(Q)$, as important information about the structure may reside in the high-Q region where the signal intensity is weak. It is recommended to use intense radiation sources such as synchrotrons. A comparison in data quality from an in-house instrument and a synchrotron source is shown in Fig. 4.3 of Appendix section.

4.4 Software

To facilitate the sasPDF method, we implemented a PDFgetS3 software program for extracting the sasPDF from experimental data. Information about obtaining the software is on the diffpy organization website (https://www.diffpy.org). The software is currently supported in Python 2 (2.7) and Python 3 (3.4 and above). It requires a license and is free for researchers conducting open academic research, but other uses require a paid license.

The PDFgetS3 program takes in a measured diffraction pattern $I_m(Q)$ and a form factor, $P(Q)$, as the inputs and applies a series of operations such as subtraction of experimental effects and form factor normalization and outputs the PDF, $G(r)$. If the square of the orientationally averaged form-factor $\langle f(Q)\rangle^2$ is available, both $P(Q)$ and $\langle f(Q)\rangle^2$ can be specified in the program, and the $S(Q)$ will be computed based on Eq. 6 which accounts for the anisotropy of scatterers in the material. Processing parameters used in PDFgetS3 operations, such as the form-factor file, the Q-range of the Fourier transformation on $F(Q)$ and the r-grid of the output $G(r)$, can be set in a configuration file in the same way detailed in [87]. Similar to PDFgetX3, an interactive window for tuning these processing parameters, is also available in PDFgetS3. An illustration of such interactive interface is shown in the Fig. 4.2. Sliders for each processing parameter allow the user to inspect the effect on the output PDF data immediately.

Once the optimal processing parameters are determined based on the quality of the PDF, those parameter values will be stored as part of the metadata in the output $G(r)$ file. The final values of Q_{min} and Q_{max} should be used when calculating PDF from a structure model, as these parameters contribute to the ripples in the PDF [141]. Full details on how to use the program is available on the diffpy organization website.



bgscale 1
qmin 0.008
qmax 0.072
qmaxinst 0.0795
rpoly 215.6
☒ constant data scale

zoom, x=0.0788341 y=0.00182061

Fig. 4.2: Illustration of the interactive interface for tuning the process parameters in the PDFgetS3 program.

4.5 PDF method

The PDF gives the scaled probability of finding two scatterers in a material a distance r apart [50]. For a macroscopic object with N scatterers, the atomic pair density, $\rho(r)$, and $G(r)$ can be calculated from a known structure model using

$$\rho(r) = \frac{1}{4\pi r^2 N} \sum_m \sum_{n \neq m} \frac{f_m(Q) f_n^*(Q)}{\langle f(Q) \rangle_{s.a.}^2} \delta(r - r_{mn}), \tag{9}$$

and

$$G(r) = 4\pi r \left[\rho(r) - \rho_0\right]. \tag{10}$$

Here, ρ_0 is the number density of scatters in the object. $f_m(Q) = \langle f_m(\mathbf{Q}) \rangle$ is the orientationally averaged form-factor of the m-th scatterer. $\langle f(Q) \rangle_{s.a.} = \sum_{m=1}^{N} (\frac{N_m}{N}) f_m(Q)$ denotes the sample average of $f(Q)$ over all scatterers in the material, where N_m is the number of scatterers that are of the same kind as the m-th scatter. Finally, r_{mn} is the distance between the m-th and n-th scatterer. We use Eq. 10 to fit the PDF generated from a structure model to a PDF determined from experiment.

PDF modeling, where it is carried out, is performed by adjusting parameters of the structure model, such as the lattice constants, positions of scatterers and particle displacement parameters (PDPs), to maximize the agreement between the theoretical and an experimental PDF. In practice, the delta functions in Eq. 10 are Gaussian-broadened to account for thermal motion of the scatterers and the equation is modified with a damping factor to account for instrument resolution effects. The modeling of SASPDF can be done seamlessly with tools developed in the atomic PDF field, with parameter values scaled accordingly. We outline the modeling procedure using PDFGUI [53], which is widely used to model the atomic PDF. In PDFGUI, the nanoparticle arrangements can simply be treated analogously as atomic structures, with a unit cell and fractional coordinates, but the lattice constants reflect the size of the NPA, which is usually at the order of 100 nm = 1000 Å. The atomic displacement parameters (ADPs) defined in PDFGUI can be directly mapped to the particle displacement parameters (PDPs) in the SASPDF case and, empirically, we find the PDP values are roughly four to five orders of magnitude larger than the values of its counterpart on the atomic scale, therefore starting values of 500 Å^2 are reasonable. These will be adjusted to the best-fit values during the refinement. The PDF peak intensity depends on the scattering length of relevant particle, which in the case of x-rays scattering off atoms, is the atomic number of the atom. For the SASPDF case we do not know explicitly how to scale the scattering strength of the particles, but for systems with a single scatterer, this constitutes an arbitrary scale factor that we neglect.

The measured SASPDF signal falls off with increasing r. The damping may originate from various factors, for example, the instrumental Q-space resolution [50] and finite range of order in the superlattice assembly. In PDFGUI there is a a Gaussian damping function $B(r)$,

$$B(r) = \exp\left[-\frac{(rQ_{damp})^2}{2}\right]. \tag{11}$$

We define a r_{damp} parameter

$$r_{damp} = \frac{1}{Q_{damp}}, \tag{12}$$

which is the distance where about one third of the SASPDF signal disappears completely. It is also possible to generalize the modeling process to the case of a customized damping function and non-crystallographic structure with Diffpy-CMI [89], which is a highly flexible PDF modeling program. In the following section, we use PDFGUI for modeling data from more ordered samples (Au NPA and Cu_2S NPA) and Diffpy-CMI for modeling data from a disordered sample (SiO_2 NPA).

4.6 Application to representative structures

To illustrate the SASPDF method we have applied it to some representative nanoparticle assemblies from the literature [133; 68; 21]. The first example is from DNA templated gold nanoparticle superlattices, originally reported in [133]. The measured intensity, $I_m(Q)$, the reduced total structure function $F(Q) = Q[S(Q) - 1]$, and the PDF $G(r)$ are shown in Fig. 4.3(a), (b) and (c), respectively. It is clear that the data corrections and normalizations to get $F(Q)$ result in a more prominent signal in the high-Q regime of the scattering data, and a highly structured PDF after the Fourier transform (Fig. 4.3(c)).

The PDF signal dies off around 350 nm, which puts a lower bound on the size of the NPA. The first peak in the PDF is located at 30.07 nm which corresponds to the nearest inter-particle distance in the assembly. This distance is expected because the shortest inter-particle distance can be approximated as the average size of Au NPs (11.4 nm) plus the average surface-to-surface distance (d_{ss}) between nearest neighbor NPs (18 nm) [133]. Peaks beyond the nearest neighbor give an indication of characteristic inter-particle distances in the assembly and codify the 3D arrangement of the nanoparticles in space.

A semi-quantitative interpretation of conventional powder diffraction data suggested the Au NPA forms a body-centered cubic (bcc) structure [133]. We therefore test the bcc model against the measured PDF. The fit is shown in Fig. 4.3(c) and the refined parameters are reproduced in Table 2. The agreement between the bcc model and the measured data is good. We refine a lattice parameter that is ∼3 % smaller than the value reported from the semi-quantitative analysis. Additionally, the PDF gives information about the disorder in the system in the form of the crystallite size (∼350 nm) and the particle displacement parameter (PDP), the nanoparticle assembly version of the atomic displacement parameter (ADP) in atomic systems. The PDF derived crystallite size is drastically smaller than the

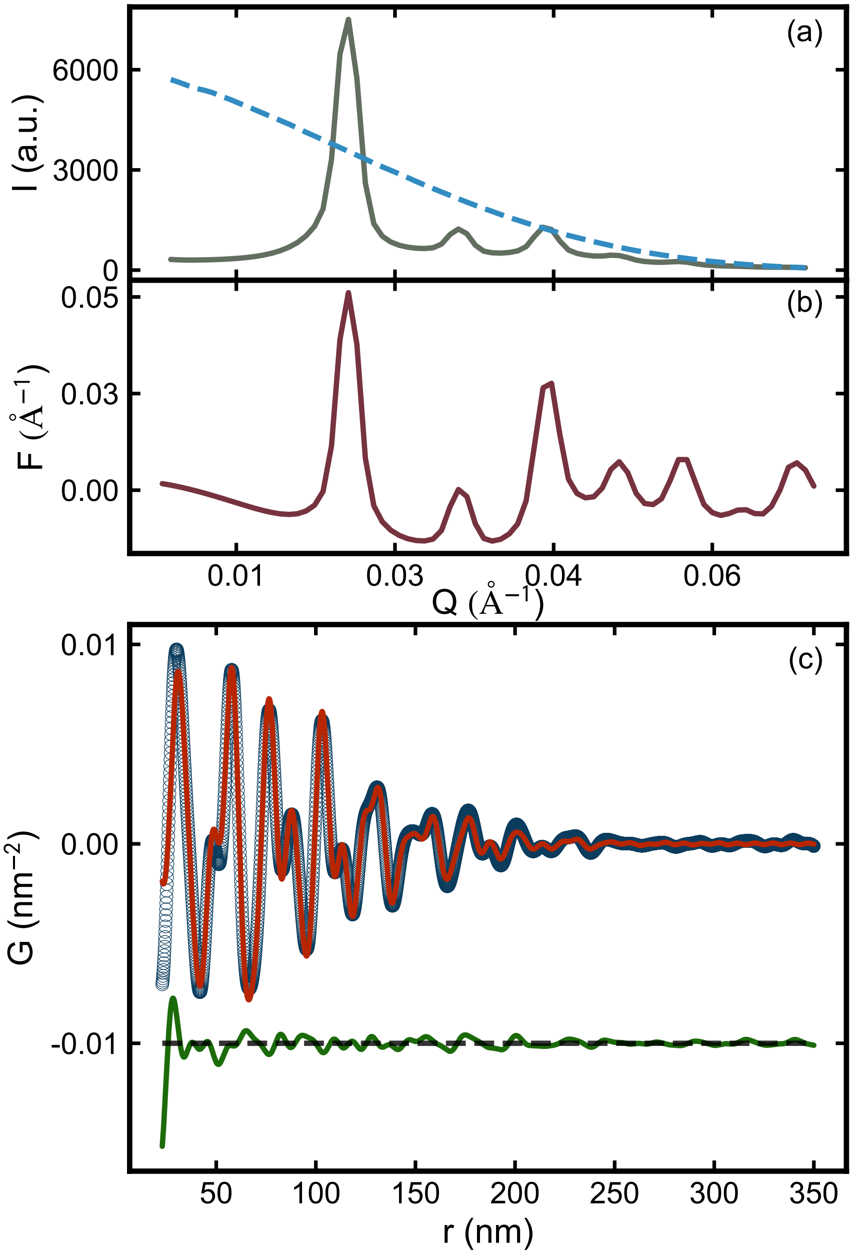


Fig. 4.3: Measured (a) scattering intensity $I_m(Q)$ (grey) and form factor $P(Q)$ (blue), (b) reduced total structure function $F(Q)$ (red) and (c) PDF (open circle) of Au NPA. In (c), the PDF calculated from body-center cubic (bcc) model is shown in red and the difference between the measured PDF and the bcc model is plotted in green with an offset.

Table 2: Refined parameters for NPA samples. Model column specifies the structural model used to fit the measured PDF. a is the lattice constant of the unit cell, PDP stands for particle displacement parameters, which is an indication of the uncertainty in position of the nanoparticles. r_{damp} is the standard deviation of the Gaussian damping function defined in Eq. 12. Scale is a constant factor being multiplied to the calculated PDF. R_w is the residual-function, commonly used as a measure for the goodness of fit.

	Au NPA	Cu_2S NPA
Model	bcc	fcc
a (nm)	34.73	26.55
PDPs (nm^2)	4.78	0.253
r_{damp} (nm)	83.3	61.4
Scale	0.537	0.361
R_w	0.172	0.221

value (∼500 nm) estimated from the FWHM of the first correlation peak [133] and it is clear by visual inspection of the PDF that the ∼500 nm value is an overestimate. These results suggest that even in the case where it is straightforward to infer the geometry of underlying assembly using qualitative and semi-quantitative means there is an advantage to carrying out a more fully quantitative SASPDF analysis.

Next we consider a dataset from a dodecanethiol (DDT)-capped Cu_2S NPA [68]. In this case the form factor is measured on an in-house Cu K_α instrument. This was necessary in the current case because the instability of the nanoparticles in suspension prevented a good measurement to be made at the synchrotron. As a result the form factor measurement was somewhat noisy (Fig. 4.4(a), blue curve) and we elected to smooth it by applying a Savitzky-Golay filter [134]. The smoothing parameters of window size and polynomial order were selected as 13 and 2, respectively, based on a trial and error approach optimized to result in a good smoothing without changing the shape of the signal. The smoothed curve is shown in Fig. 4.4(a). It is worth noting that in general, a smoothing process may start failing when the signal-to-ratio in the data exceeds a certain threshold, and so good starting data is always desirable. A conventional semi-quantitative analysis on diffraction data from the sample collected on an in-house Cu K_α instrument is shown in Fig. 4.5. It suggests the NPA forms a face-centered cubic (fcc) structure with an inter-particle distance of 18.8 nm. The SAS PDF obtained from the same NPA sample is shown in Fig. 4.4. It clearly shows that peaks die out at around 300 nm, which again signifies the crystallite size of the assembly. The first peak of the measured PDF is at 18.5 nm, corresponding to the inter-particle distance in the NPA. This value is about 1.6 % smaller than the value estimated from the semiquantitative analysis.

The best-fit PDF of a close-packed face-centered cubic (fcc) structural model is shown in red in the figure and refined structural parameters are presented in Table 2. The fcc model

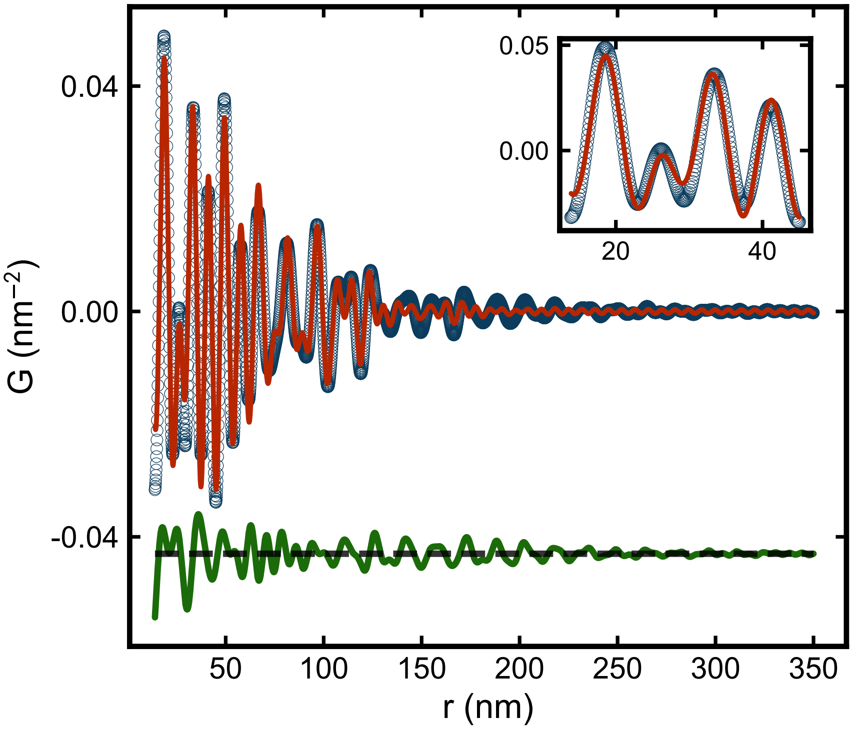


Fig. 4.4: Measured PDF (open circle) of a Cu$_2$S NPA sample with the best fit PDF from the fcc model (red line). The Difference curve between the data aAs a result, nd model is plotted offset below in green. The inset shows the region of the first four nearest neighbor peaks of the PDF along with the best-fit fcc model.

yields a rather good agreement with the measured PDF of Cu_2S NPA in the short-range (up to ∼130 nm). Interestingly, the refined lattice parameter of this cubic model is 26.55 nm, from which we can calculate an average inter-particle spacing of 18.78 nm, which is much closer to the value estimated from the in-house data than directly extracting the position of the first peak in the PDF. The first peak in the PDF calculated from the model lines up with that from the data at 18.5 nm, which means that the position of the peak, as extracted from the peak maximum, underestimates the actual inter-particle distance by $\sim 1.5\%$, which may be due to the sloping background in the $G(r)$ function [50]. Quatntiative modeling is always preferred for obtaining the most precise determination of inter-particle distance.

The region of the first four nearest-neighbor peaks in the PDF, together with the fit, is shown in the inset to Fig. 4.4. A close investigation of this region shows subtle shifts in peak positions between the measured PDF and the refined fcc model. At around 26 nm (second peak), the peak from refined model is shifted to higher-r compared to the measured data, while at around 33 nm (third peak), the relative shift in peak position is towards the low-r direction. These discrepancies suggest the NPA structure is more complicated than a simple fcc structure and may reflect the presence of internal twined defects, for example [12]. Furthermore, it is clear that signal persists in the measured PDF in the high-r region that is not captured by the single-phase damped fcc model. There is clearly more to learn about the structure of the NPA by finding improved structural models and fitting them to the PDF, though this is beyond the scope of the current paper.

It is worth noting that the refined PDP value of DDT-capped Cu_2S NPA is significantly smaller than that of the DNA-templated Au NPA described above. A small PDP means the positional disorder of the NPs is small which would be expected with shorter, more rigid, linkers between the particles. The inter-particle distance (18.8 nm) can be decomposed into the sum of the average particle diameter (16.1 nm) and the particle-surface to particle-surface distance $d_{ss} = 2.7$ nm. Based on the chemistry the linker would have length 1.7 nm in the fully stretched out state, which would result in a maximal $d_{ss} = 3.4$ nm if the linkers were stretched out and oriented radially. Half the observed surface-surface distance, $d_{ss}/2 = 1.4$ nm. This result is reasonable, suggesting the linkers are either not straight, or not radial, or possibly partially interleaved. Nonetheless, this shorter linker would be expected to be more rigid and therefore consistent with our observation of a smaller PDP value from the sasPDF analysis.

4.7 Appendix

4.7.1 Illustration of of data acquisition strategy

In this section, important effects related to the data quality are illustrated. In general, for a successful sasPDF experiment, it is crucial to achieve a high signal-to-noise ratio throughout the entire Q-range for both the form factor and sample measurements. Figs. S4.1 and S4.2 show the effect of insufficient counting statistics in the sample and form factor measurements, respectively. Fig. S4.3 compares the data quality from an in-house instrument and a synchrotron source. Finally, Fig. S4.4 shows the remedial effect of smoothing data from in-house measured form factor with insufficient statistics. The proper remedy is to measure with sufficient statistics in the first place.

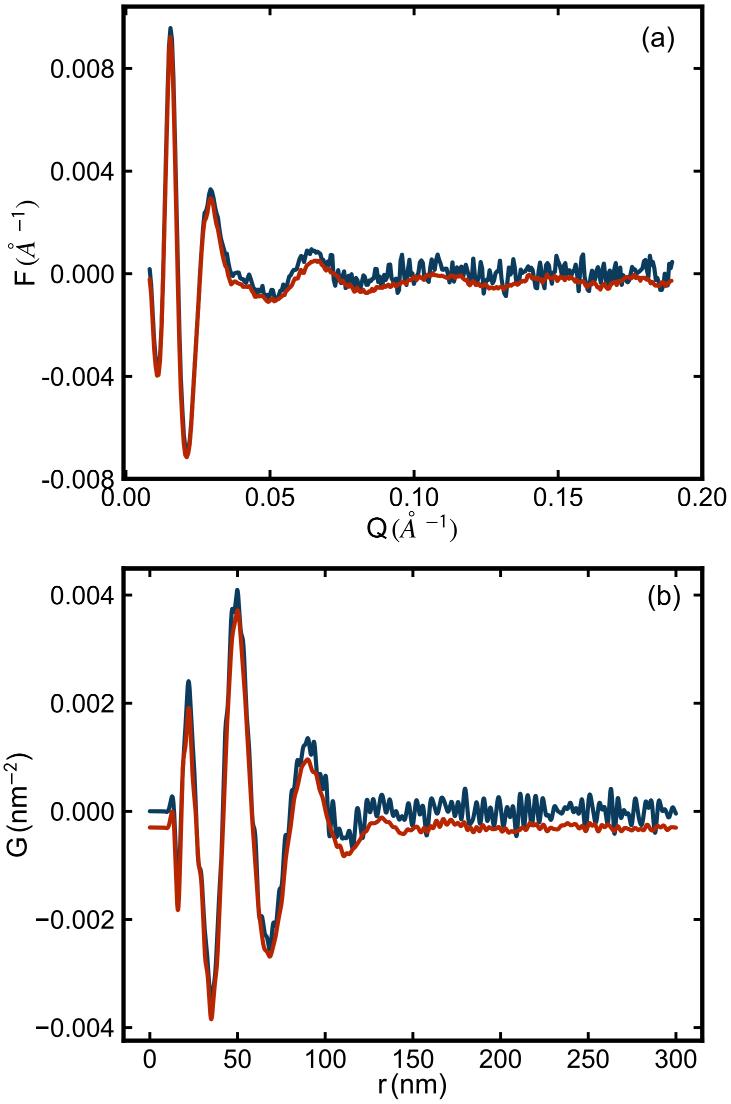


Fig. S4.1: (a) Reduced structure functions $F(Q)$ and (b) PDFs $G(r)$ of the SiO_2 NPA sample with different scan exposure times. Blue is from data with 1 s scan exposure time and red is from data with 30 s scan exposure time. In both panels, data are plotted with a small offset for ease of viewing. In both cases the form factor was measured with an scan exposure time of 600 s.

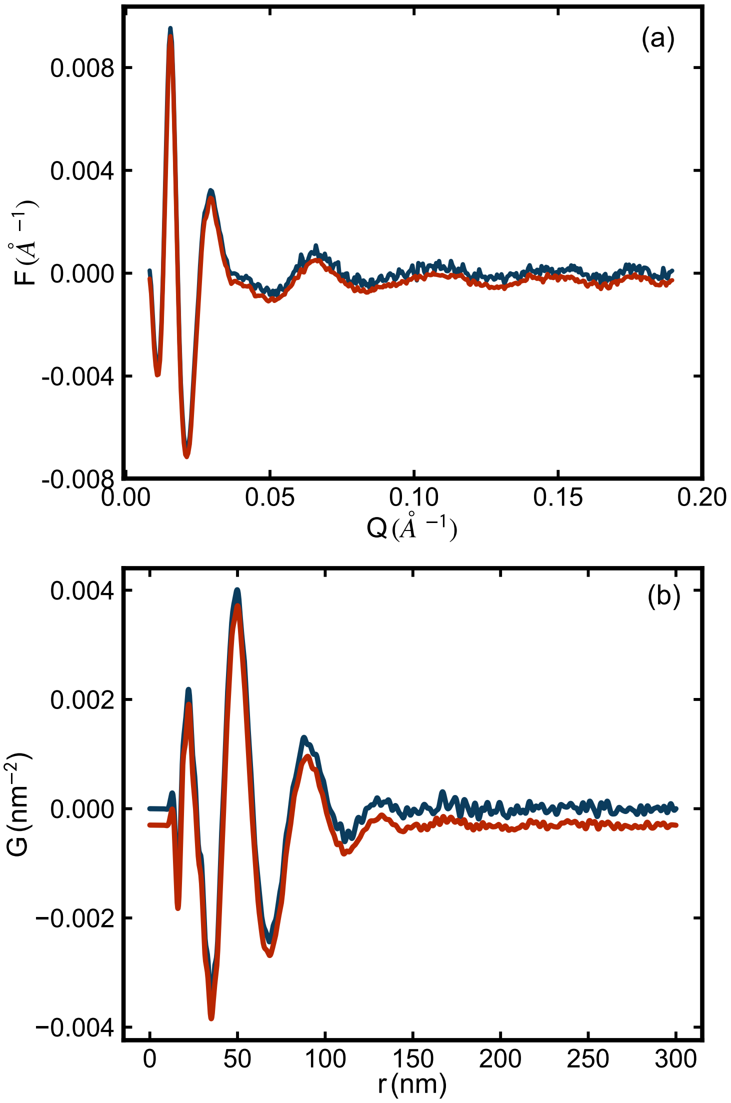


Fig. S4.2: (a) Reduced structure functions $F(Q)$ and (b) PDFs $G(r)$ of the SiO_2 NPA sample processed with form factor $P(Q)$ from different scan exposure times. Blue is made with a form-factor measured for 30 s and red is with a form factor collected for 600 s. In both cases the scan exposure time for the NPA sample was 600 s. In both panels, data are plotted with a small offset for ease of viewing.

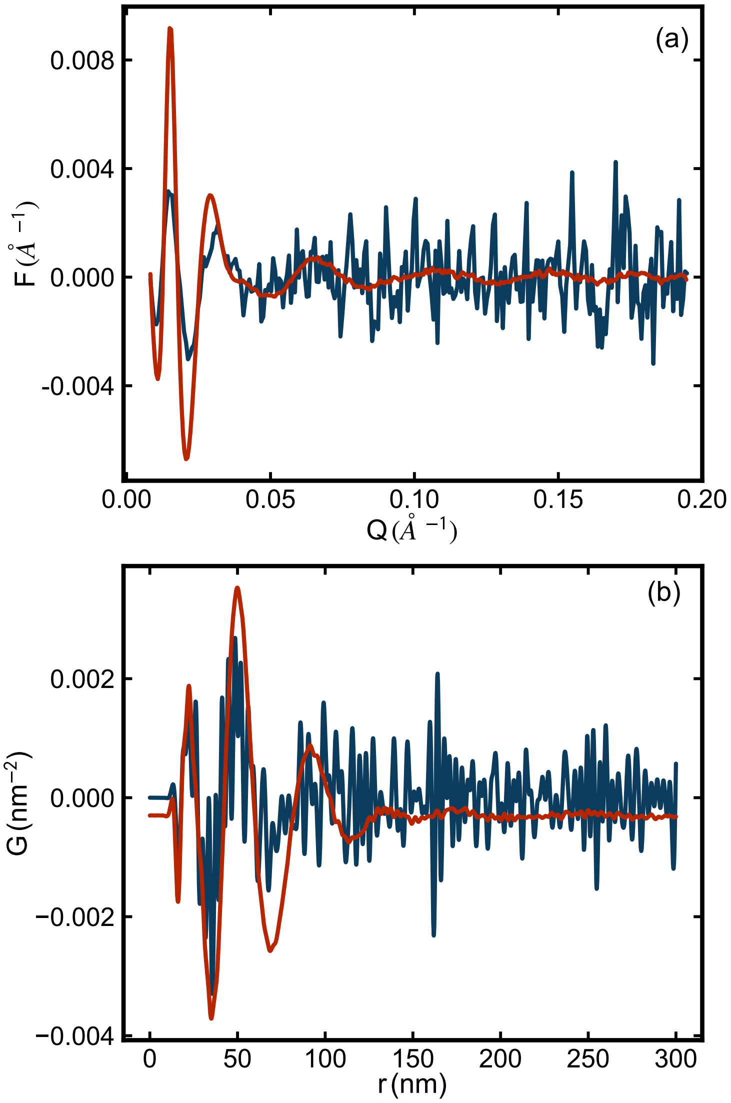


Fig. S4.3: (a) Reduced structure functions $F(Q)$ and (b) PDFs $G(r)$ of the SiO_2 NPA sample. Blue is from data collected at Columbia University using a SAXSLAB (Amherst, MA) instrument with a 2-hour (7200 s) scan exposure time for both $I(Q)$ and $P(Q)$ measurements. Red is from data collected at beamline 11-BM, NSLS-II with 30 s scan exposure time for both $I_m(Q)$ and $P(Q)$ measurements.

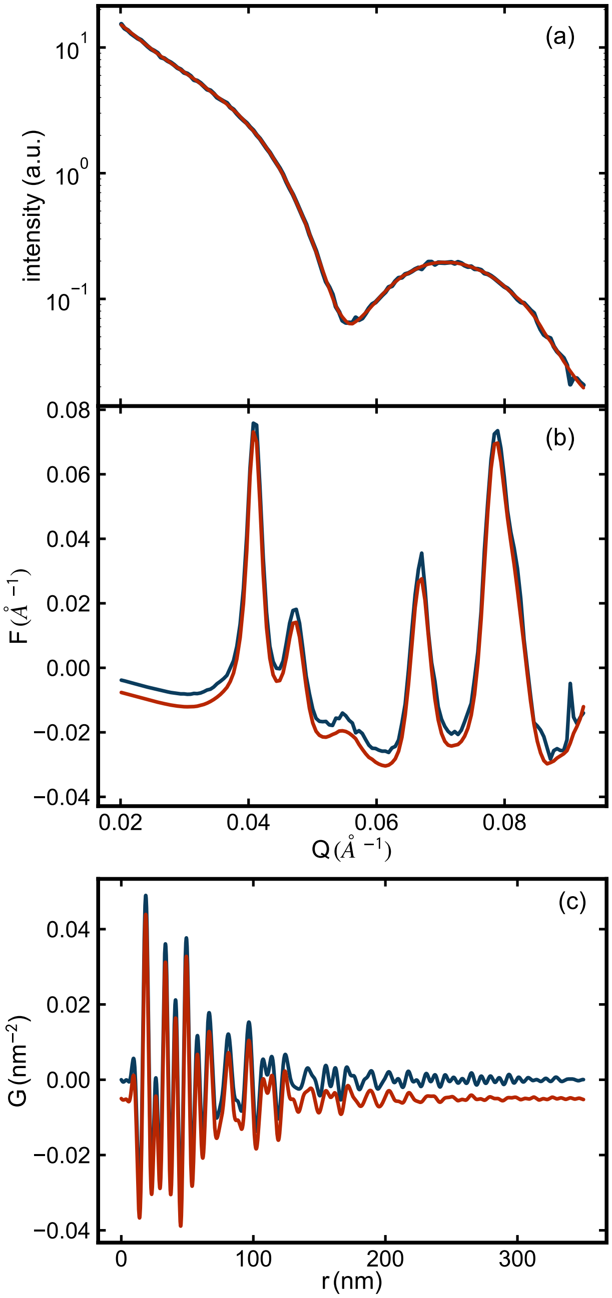


Fig. S4.4: (a) Form factor signal from Cu_2S NPs. Blue is the raw data collected at an in-house instrument and red is the data smoothed by applying a Savitzky-Golay filter with window size 13 and fitted polymer order 2. (b) reduced structure functions, $F(Q)$, and (c) PDFs, $G(r)$ from the Cu_2S NPA sample. In both panel, blue represents the data processed with raw form factor signal and red represents the data processed with smoothed form factor signal. Curves are offset from each other slightly for ease of view.

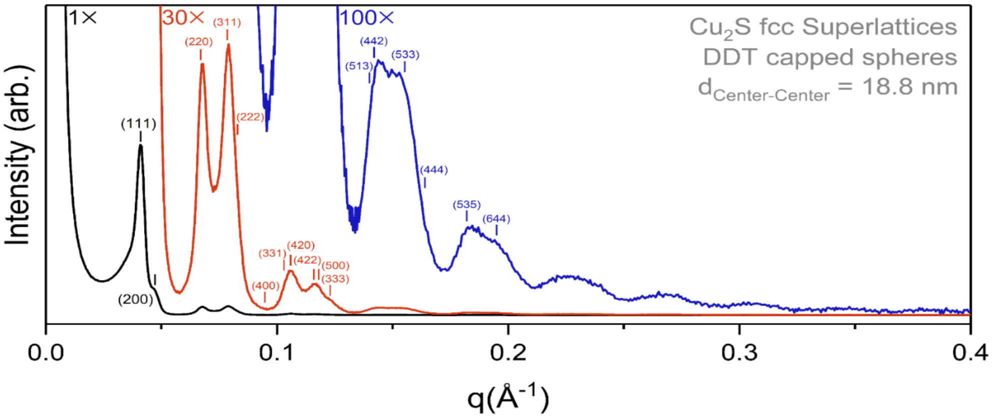


Fig. S4.5: Semi-quantitative structural analysis on Cu_2S NPA sample.

Chapter 5

Applications of sasPDF method on nanoparticle assemblies

5.1 A structural signature for jamming in polymer-ligated nanoparticle assemblies

5.1.1 introduction

There has been considerable interest in the jamming transition, especially as it relates to unusual properties of materials such as foams, toothpaste etc [109; 185; 20]. A jammed state is defined as one that is microscopically disordered and can support weight with only elastic deformation. While much work has focused on the dynamics of these materials [45; 46], there has been continuing interest in obtaining a structural signature of this transition.

In this context we study the systems of matrix-free polymer grafted nanoparticles (PGNs) which show enhanced gas transport and a suppression in physical aging relative to the neat polymer [158; 146; 151]. Grafting the polymers onto the surface of the inorganic nanoparticles circumvents the challenge of reproducibly obtaining uniformly dispersed nanoparticles [107]. Previous work had shown that at fixed grafting density (a specific case is 0.47 chains/nm^2), the gas permeation of these materials is always higher than that of the pure polymer. The permeation displays a maximum as a function of the graft chain length, in this case at a molecular weight in the vicinity of $\approx$ 90 kDa. In addition, independent linear oscillatory shear rheology shows that this maximum corresponds to a jamming-unjamming "transition" as a function of chain molecular weight [84]. Here we show through the analysis of x-ray scattering that there is a structural signature of this proposed jamming transition. As we reduce the chain length of the grafts we find that the first peak of the pair distribution function between NP centers shows a significant narrowing, while leaving the total number of neighbors effectively unchanged. This picture, which is consistent with the idea that force chains are a signature of jamming, suggest that this dynamic transition is associated with static signatures.

5.1.2 Experiment

The samples we consider are spherical silica NP cores (14 ± 4 nm) grafted with poly(methyl acrylate) (PMA) chains at a fixed grafting density of $\approx$ 0.47 chains/nm^2 (medium) and $\approx$ 0.66 chains/nm^2 (high). The NPA formed a circular, free-standing stable film of diameter about 5 mm. Details of synthesis is reported in [21]. The chain length of the grafted chains are varied systematically in a series of experiments and the details information of measured samples is reported in Table 1. The samples were measured at the DUBBLE beam line (BM26) at the European synchrotron radiation facility (ESRF) and at the Complex Materials Scattering (CMS, 11-BM) beamline of the National Synchrotron Light Source II (NSLS-II) at Brookhaven National Laboratory (BNL). Both experiments were conducted using the rapid acquisition PDF approach [38]. Films of the polymer-ligated NPs were supported

Table 1: Polymer-grafted silica NP samples. M_n is the molecular weight of the grafted chain in kg/mol and Σ is the polymer graft density on the surface of the nanoparticles in chains/nm^2.

Sample	Σ	M_n	Sample	Σ	M_n
H-31	0.66	31	M-29	0.47	29
H-41	0.66	41	M-41	0.47	41
H-62	0.66	62	M-65	0.47	65
H-80	0.66	80	M-78	0.47	78
H-106	0.66	106	M-101	0.47	101
H-129	0.66	129	M-132	0.47	132

by a bracket with the surface of films perpendicular to incident x-ray beam. At DUBBLE beamline, area detecor in use was Pilatus 1M (Dectris, Switzerland) and sample-detector distance was 2.37 m with x-ray wavelength of 0.979 Å. At CMS beamline, area detecor in use was Pilatus 2M (Dectris, Switzerland) and sample-detector distance was at 2.02 m with x-ray wavelength of 0.918 Å. Both setups were chosen such that the maximumly accessible momentum transfer (Q_{max}) is about 0.15 Å^{-1}. The "spot exposure time", which is the length of time that any spot on the sample is exposed, was set to 30 s for both experiments at CMS and at BM26. This value was determined by locating the beam on a fixed spot of the sample and taking a sequence of short exposures, while ensuring there is no significant changes in the intensity of the strongest correlation peak in the diffraction pattern. Fro desired data statistics, 20 images, collected with spot exposure time, were summed together (accounting for "scan exposure time"). Detailed discussion about the affects of these two parameters to the data quality is presented in greater detail elsewhere [110].

These systems have amorphous arrangements of the nanoparticles [21], which is advantageous for studying the jamming transition as it eliminates effects due to specific inter-particle contacts and correlations coming from the packing; However, it also presents challenges for detailed study of the structure. To obtain a quantitative analysis of these systems we have extended the application of atomic pair distribution function (PDF) analysis, which is the structural approach of choice for studying atomic liquids and amorphous materials [59; 203; 196; 19], to the small-angle scattering regime (SASPDF). This required a significant development of the methods and software, which will be presented in detail in a separate paper [110], but is summarized in the Method section below.

5.1.3 Method

The PDF is experimentally accessible as the Fourier transform of the properly corrected and normalized diffraction intensity from an isotropically scattering sample such as a uniform crystalline powder or an amorphous material or liquid. It yields the probability of finding a neighboring scattering object (atoms in regular PDF) at a distance r away from another object. For the case of our polymer-ligated NPs, the scattering contrast between the NP cores and grafted polymer – thus the $G(r)$ obtained – corresponds to the ensemble average of the center-to-center separation between the NPs. This is accomplished by dividing the coherent scattering signal by the form factor of the particles, which we measure directly. The PDFs from high graft density samples are shown in Fig. 5.1. The PDF yields a measure of the probability of finding a scattering object, in this case a silica nanoparticle, at some distance-r away from another one. The PDF can be computed by locating a particle at the origin, moving our in the radial direction from that particle and counting the density of particles at r away from the particle at the center. The PDF can be understood as a histogram of inter-atomic distances [50].

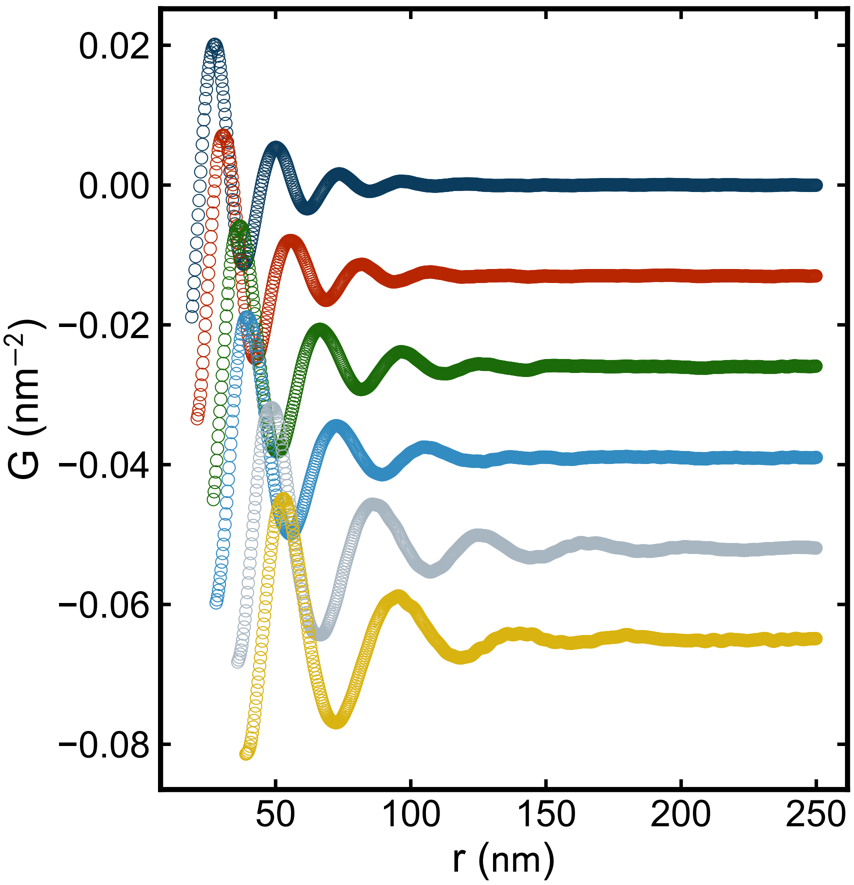


Fig. 5.1: Measured PDFs of, from top to bottom, H-31, H-41, H-62, H-80, H-106, H-129 samples.

5.1.4 Results

We find that the polymer-ligated NPs arrange in an isotropic packing about a central particle and there is no evidence of close-packing such as face-centered cubic (fcc) or icosahedral structures for all graft polymer lengths and grafting densities that we studied. This is evidenced by the fact that the PDF signal is well fit by a single-frequency sine-wave (Fig. 5.2), which is evidence that the packing of particles around the central particle is isotropic in space. If it were not, for example if there were a tendency towards fcc packing, the inter-particle distances would be different in different directions, and multiple Fourier components

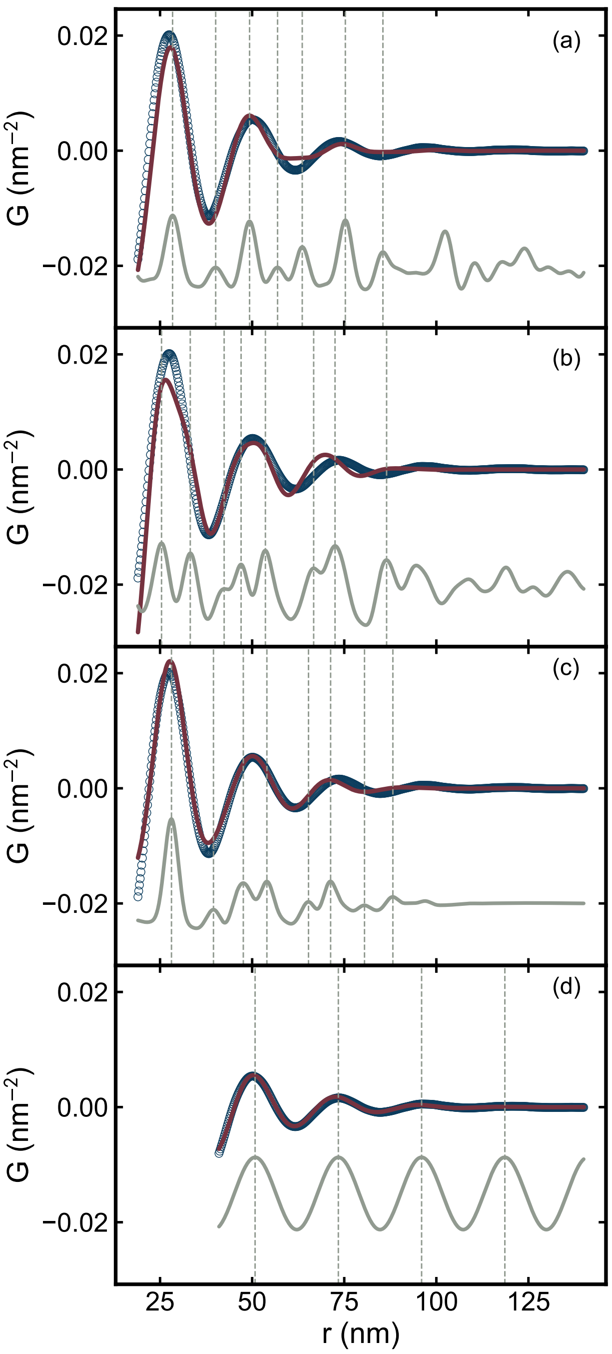


Fig. 5.2: Measured PDF (open circle) of H-31 sample and calculated PDFs (solid lines) from (a) fcc, (b) hcp, (c) icosahedral (d) damped sine-wave models. In each panel, the line in dark red is the PDF calculated from the corresponding model with optimum parameters. From (a) to (c), the line in grey is the PDF calculated from the same model but with small ADPs. In (d), the line in grey is the PDF calculated from the undamped sine-wave model. Dashed lines indicate maxima of the sharper PDFs in each panel.

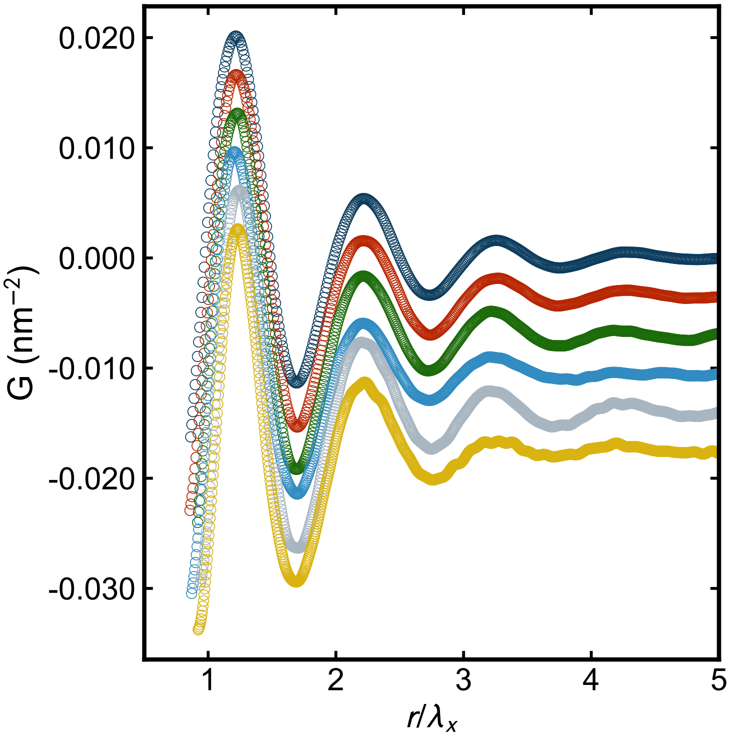


Fig. 5.3: PDFs of, from top to bottom, H-31, H-41, H-62, H-80, H-106, H-129 plotted on a renormalized r-axis, r/λ, where λ is the refined wavelength of the best-fit damped sine-wave model.

would be needed to explain the measured PDF. Therefore, the nanoparticle assemblies, at all graft polymer lengths, are in structural terms much closer to a random liquid or amorphous material with no directional packing.

It is evident from Fig. 5.1 that the average particle-particle separation grows with grafting polymer length, as one might expect. However, for all molecular weights, the PDFs are self-similar: scaling the r-axis by the mean inter-particle separation results in all the curves collapsing onto each other (Fig. 5.3.) There is no appearance of order such as fcc in the local packing for shorter grafting polymers, which might be a structural signature of jamming and correlate with the enhanced gas separation properties of the materials.

However, close inspection of the scaling figure indicates that small changes are evident in the width of the first peak. The width of PDF peaks indicates the level of disorder of the system [50]. Liquids and amorphous materials that are dense random packings of hard spheres have nearest neighbor shells that are significantly sharper than higher-neighbor shells [57; 55; 184; 185], meaning the motions of nearest neighbors are highly correlated [83]. This scenario is indeed true for the the low M_n samples, but interestingly, less so for the high M_n samples. To investigate the change of the 1st peak width across samples, we define an order parameter ξ, which is presented in the following way. We investigate the correlation of nearest neighbors by testing a partial sine-wave model , that is fitted at higher-r range, against the first neighbor peak in the measured PDFs (high-r fit.) The partial sine-wave model describes the first neighbor peak well for longest chain length sample (M-132) (red in Fig. 5.4(c)), signifying the motions of nearest neighbors in the sample are not highly correlated. However, the partial sine-wave model does not describe the first neighbor peak well for the short chain length sample (M-41) (red in Fig. 5.4(a)). A sample with an intermediate graft-polymer chain length appears to have behavior in between these two (red in Fig. 5.4(b)). The first peak in the data is much sharper than the damped sine-wave peak, as we would expect for something behaving like a hard-sphere random packed solid. The different graft polymer length samples have well dispersed but randomly packed structures. This observation indicates that the nature of the spheres is crossing over from more hard-sphere behavior to soft spheres. In the former case, we presume that there is little inter-penetration of the grafted molecules of neighboring silica spheres, whereas in the latter, there is a greater degree of inter-penetration. The behavior of the NPA is crossing over from hard to soft on going from sample H-41 to H-129. Similar behavior is observed for medium graft density sample as well and results are shown in Figs. 5.6 to 5.8.

To explore this crossover in greater detail, we consider the whole series of graft-polymer lengths. To quantify this behavior we define a "hard-sphere parameter" that is a measure of the degree of hard-sphere behavior. When we fit the damped sine-waves, the wavelength and damping factor are varied to give the best agreement, or lowest R_w. In Fig. 5.4, we exclude the first peak when carrying out the damped sine wave fitting and we used the same refined parameters to plot the sine-wave all the way to below the first peak. We call this the high-r fit. It is also possible to carry out the fit including the first peak. We call this the full-r fit. Because the nearest neighbor peak in the PDF is the strongest feature, the full-r fit parameters are heavily weighted towards fitting this peak well. As we discuss above, we therefore expect the high-r and full-r fits to be quite different for the hard-sphere case, but to become much more similar for a soft-sphere model. We can use R_w as a measure of this. We therefore define our hard-sphere parameter, ξ_h as

$$\xi_h = \frac{R_w^h(M_n) - R_w^f(M_n)}{R_w^h(M_{n0}) - R_w^f(M_{n0})}, \tag{1}$$

where $R_w^h(M_n)$ is the R_w from the high-r fit for the sample with polymer graft length M_n, and $R_w^f(M_n)$ is the full-r fit equivalent. This parameter will be large when the system is behaving as a hard-sphere system, and will become zero in the soft-sphere limit and M_{n0} is the molecular weight of the shortest polymers in the sample. By normalizing it to the value for our smallest polymer chain lengths we give it the characteristic of an order parameter, that crosses between 1 and 0. The PDFs obtained by the full-r fits are shown in Fig. 5.4 as the solid grey lines, showing the much better fit to the first peak in these fits. The hard-sphere parameter ξ_h for the high graft density samples is shown as the red-dashed line in Fig. 5.5. The ξ_h crosses over smoothly from large to small with increasing M_n reaching close to zero at around $M_n = 110$ kg/mol. By this point the nanoparticle assembly is behaving like a soft-sphere system. This cross-over is close to the region where the dynamic jamming transition has been observed [84] in a similar system and the jamming transition therefore appears to be the loss of collectivity/coherent dynamics of the near-neighbor shell. We also plot on

Fig. 5.5 a shaded region which corresponds to the region of an anomalous enhancement in gas permeability has been reported for similar membranes [21]. The values shown here are our own measurements of gas permeability from samples similar to those measured in the x-ray measurements. Enhancement in permeability ratio is defined as the permeability of the target gas in the composite membrane, P_ϕ, normalized to the permeability of that gas in a membrane of the pure polymer, denoted P_b. There is an enhancement in permeability ratio P_ϕ/P_b of CO_2 in the intermediate M_n region, which coincides with the region where the NPA crosses over from hard-sphere to soft-sphere behavior.

5.1.5 Conclusion

By applying pair distribution function analysis to small-angle x-ray scattering data from polymer-ligated nanoparticle assemblies, we identify a structural signature of jamming transition, which is associated with the change of the first peak width in the PDF. The identified region agree well with the region from dynamical characterization tools. In addition, the jamming transition region also maps to the region where enhancement in gas separation was reported. The jamming transition can be understood as the cross-over from hard- to soft-sphere behavior in the system, which leads to the loss of collectivity dynamics of the near-neighbor shell.

5.1.6 Appendix

In this section, we present similar analysis results from the medium graft density samples.

5.2 Multiply twinned structure in DNA-ligated Au nanoparticle assemblies

In this section, we will briefly discuss a legacy data of DNA-ligated Au nanoparticle assemblies published in [207]. Previous work had shown this system can be gradually transformed from the body-centered cubic (bcc) phase into face-centered cubic (fcc) phase after inputting specific DNA strands. The entire reaction was reported to take about 800 minutes and the diffraction patterns were collected throughout the process. From the analysis reported before (which was done in reciprocal space), the transformation from fcc to bcc phase was due to the nucleation and growth of fcc embryos within the bcc starting phase and no intermediate phases were involved.

5.2.1 Results

To start our analysis, we first transform the diffraction patterns collected throughout the reaction into pair distribution functions (PDFs) (Fig. 5.9). We observer a gradual change in the signal between two end members. We first fit the bcc and fcc model throughout the data. We find that for the data collected at earlier reaction time, the agreement factor (R_w) [50] from the bcc fits (blue in Fig. 5.10) is better, however it gradually degrades and after reaction time = 280 mins., for fcc fits (red in Fig. 5.10) takes over the bcc fits. However even for the end member (reaction time = 800 mins), which was reported to be pure fcc phase, the R_w value of best fit fcc model is still at the higher end (0.3) and considerable residual is presented from the fit (green in Fig. 5.11), which implies the fcc model might be far from the correct structure model. To have a clearer idea about the underlying structure of the end member (reaction time = 800 mins), we employ the "cluster-mining" approach [11], which is based on fitting the measured data in a highly constrained manner, against an algorithmically generated pool of candidate structures. This approach had been reported to be promising in finding and evaluating structural models of small metallic nanoparticles.

In our cluster-mining approach, we consider candidate structures from three motifs, octahedron, decahedron and icosahedron and the scatter plot of R_{w} for the candidate structures, along with the R_{w} value from our previous fcc fit is shown in Fig. 5.12. It is clear that the cluster-mining approach discovers a series of decahedral structures (green pentagons in Fig. 5.12) that are more optimal than the fcc structure model we considered before. The fit from the best decahedron structure (number of particles = 192) is shown in Fig. 5.13. Based on the residual, we find the best-fit decahedron model indeed remove significant portion of misfit presented in the fcc model, however a further investigation is needed for solving the structure entirely.

5.2.2 Conclusion

Multiply-twinned structures have been observed and extensively studied in fcc metal such as Au, Ag, Pt and Pd [115; 161] and similar structure was also observed in PdSe and DNA-ligated NPA [163; 9]. The identification of non-fcc model at the end member implies the reprogramming process of DNA-ligated Au NPA might be mapped to the same crystallization process observed in other NPA or crystals, which offers a new angle for understanding the reprogramming process.

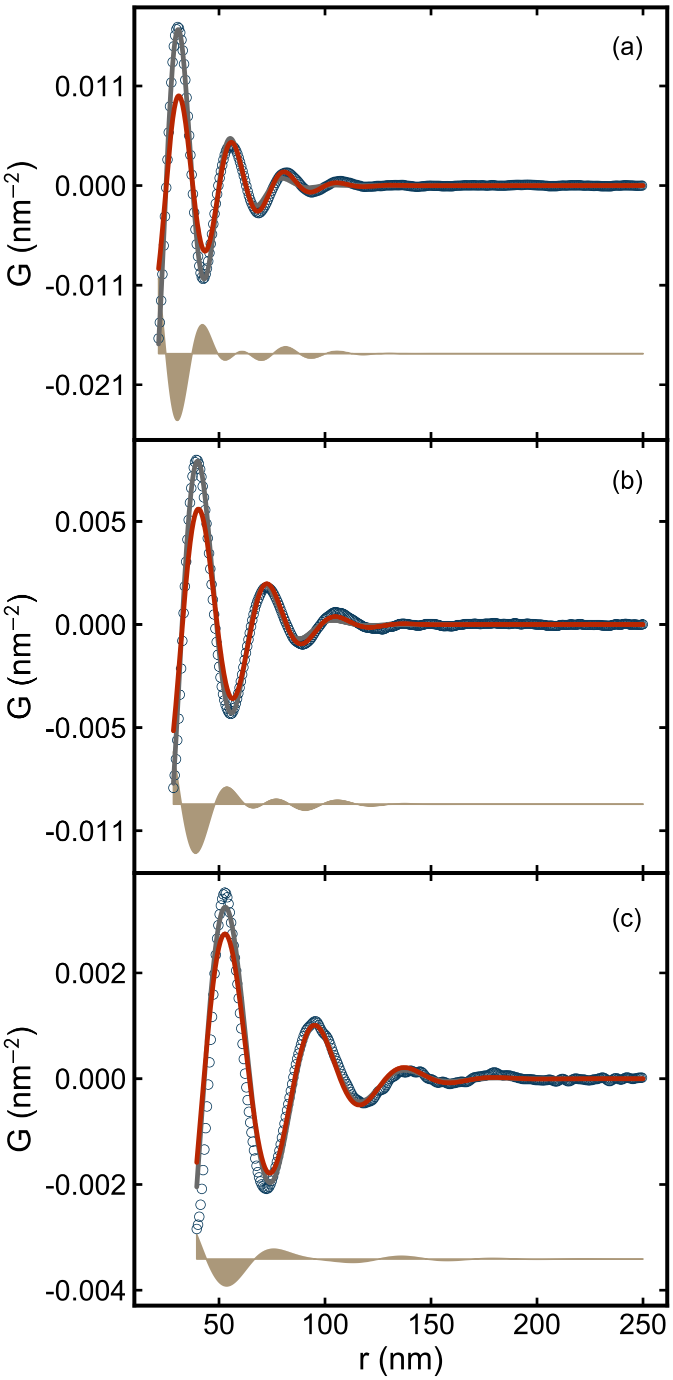


Fig. 5.4: Measured PDFs (open circles), full-r fit (grey) and high-r fit (red) of (a) H-41, (b) H-80, and (c) H-129 samples. The difference between two models (brown) is plotted below in each panel.

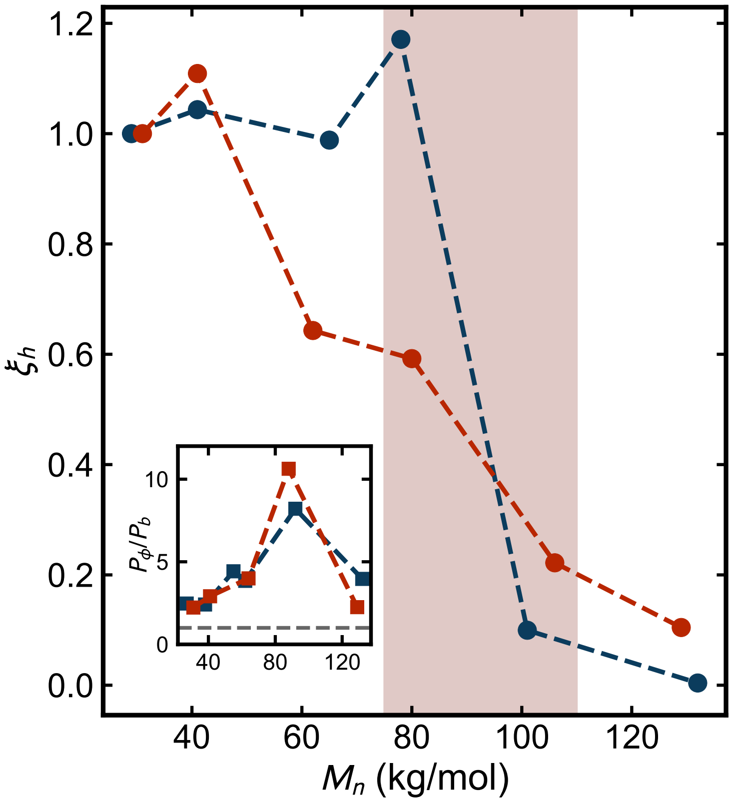


Fig. 5.5: Hard-sphere parameter, ξ_h, for medium (blue) and high (red) graft density samples. The shaded area is the region of M_n where an anomalous enhancement in gas permeability was previously reported. This enhancement is reproduced in our samples as shown in the inset where the permeability ratio P_ϕ/P_b is plotted from samples with graft densities $\Sigma =$ 0.43 chains/nm^2 (blue) and $\Sigma = 0.66$ chains/nm^2 (red) similar to the ones in the x-ray experiments. The horizontal dashed line in the inset is $P_\phi/P_b = 1$ for reference.

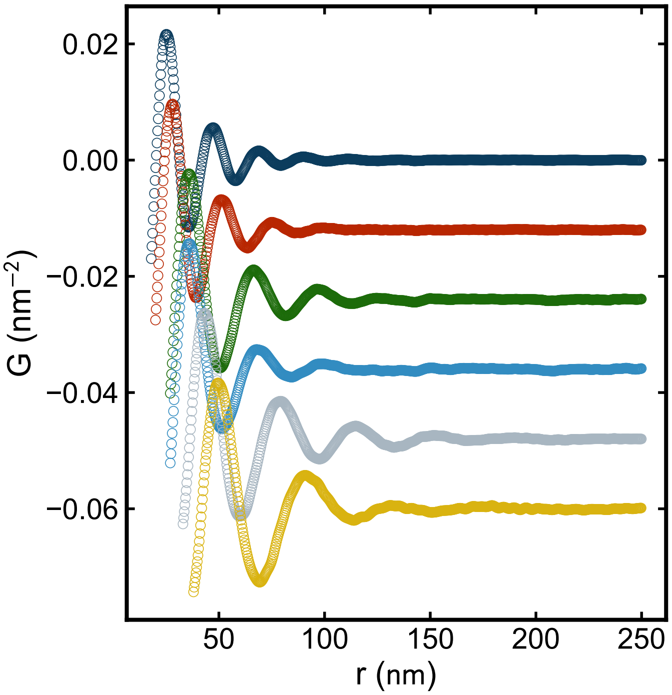


Fig. 5.6: Measured PDFs of, from top to bottom, M-29, M-41, M-65, M-78, M-101, M-132 samples.

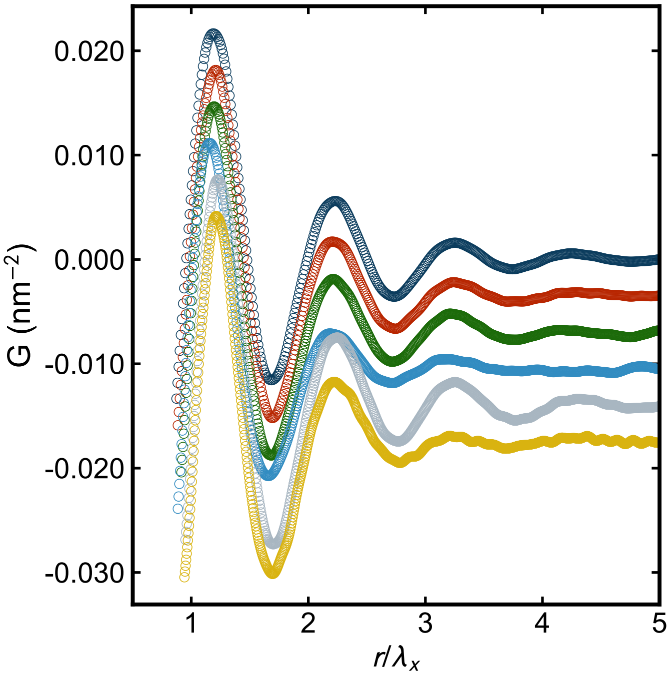


Fig. 5.7: PDFs of, from top to bottom, M-29, M-41, M-65, M-78, M-101, M-132 plotted on a renormalized r-axis, r/λ, where λ is the refined wavelength of the best-fit damped sine-wave model.

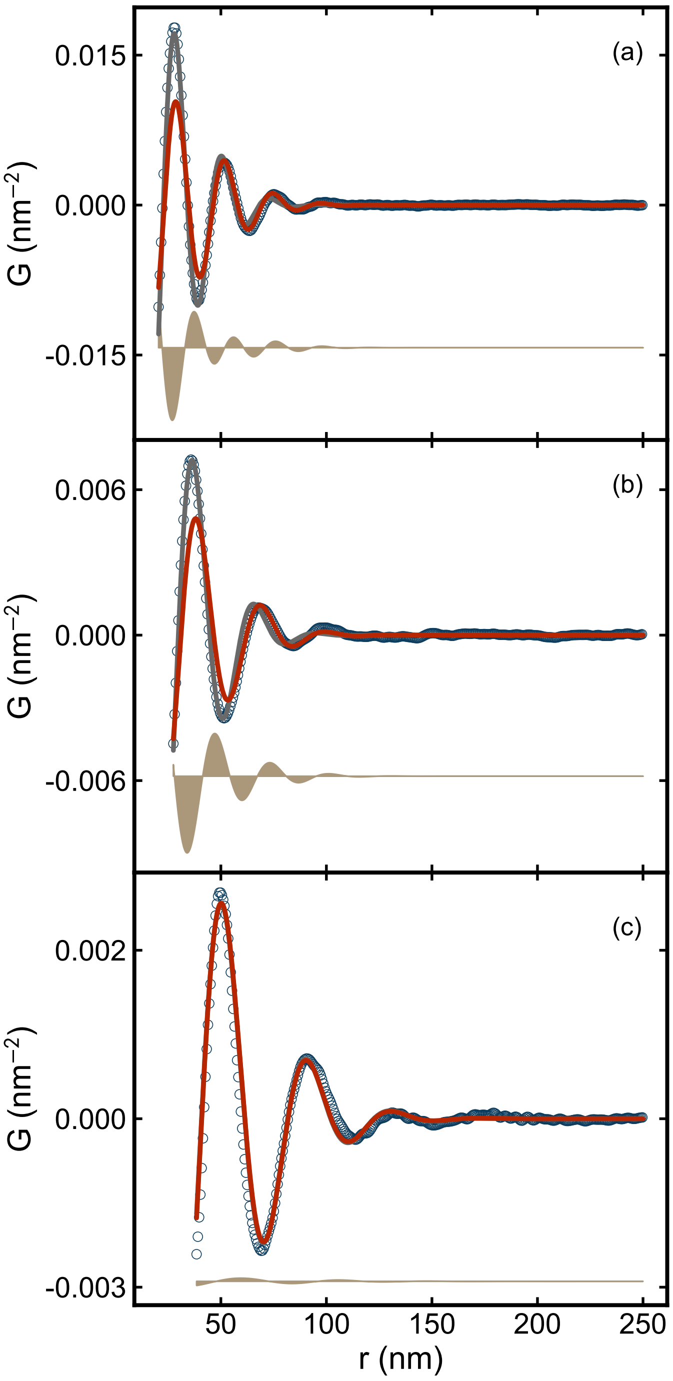


Fig. 5.8: Measured PDFs (open circles), full-r fit (grey) and high-r fit (red) of (a) M-41, (b) M-78, and (c) M-132 samples. The difference between two models (brown) is plotted below in each panel.

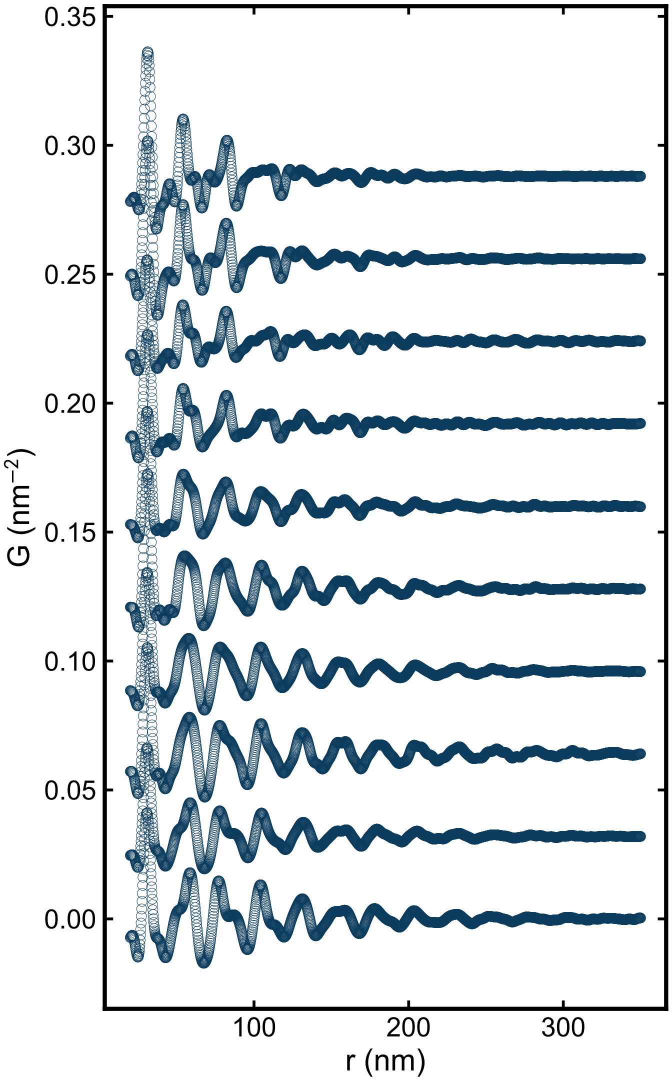


Fig. 5.9: Measured PDFs from the fcc-bcc phase transition. From bottom to top, each PDF corresponds to data collected at 0, 40, 80, 120, 160, 220, 280, 360, 480 and 800 minutes after the extra DNA strands was added.

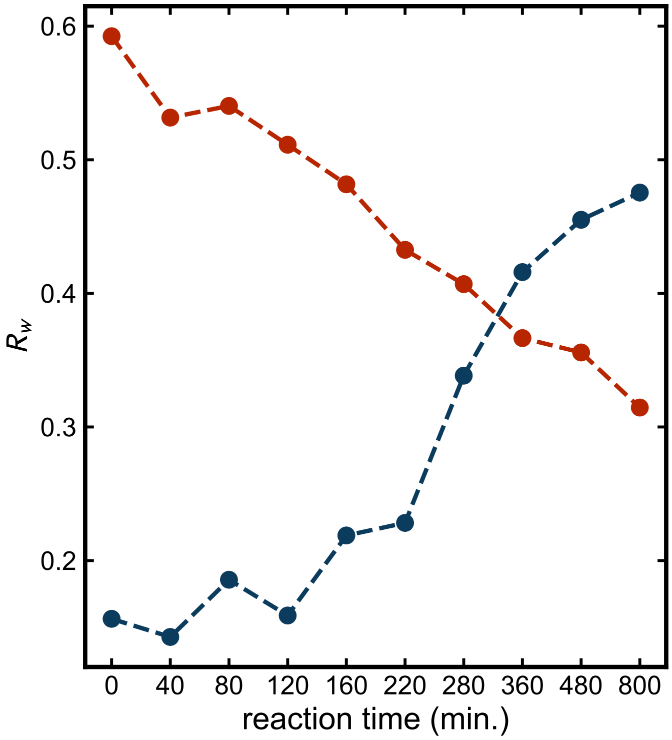


Fig. 5.10: Scatter plot of agreement factors (Rw) of fcc model (red) and bcc model (blue) vs data collected at different reaction time.

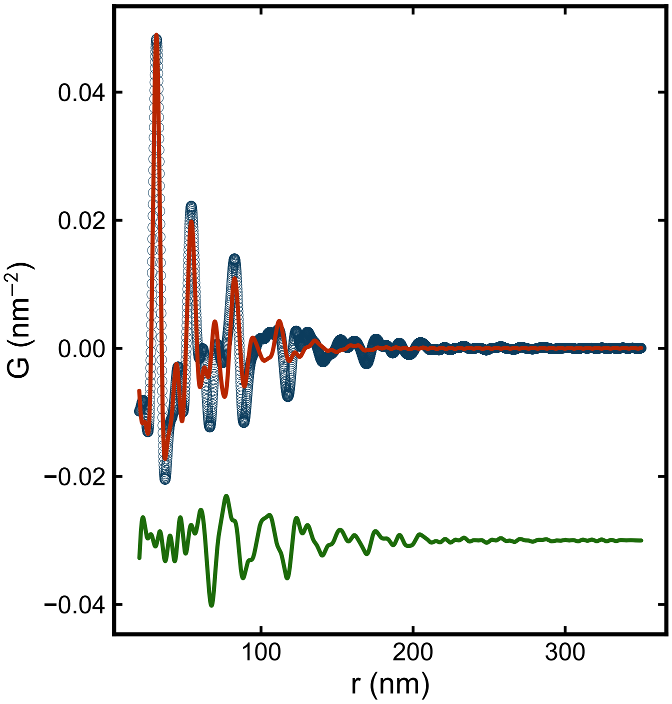


Fig. 5.11: Measured PDF (blue) at reaction time = 800 mins and PDF from best-fit fcc model (red). The difference (green) is plotted with an offset for the ease of reading.

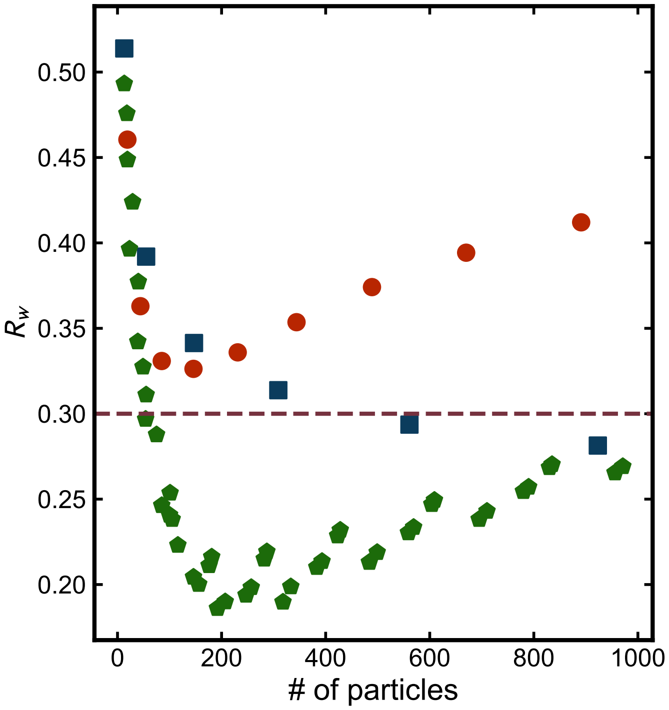


Fig. 5.12: Scatter plot of agreement factors (R_w) for decahedron (green), octahedron (red) and icosahedron (blue) fit to the PDF collected at reaction time = 800 mins, plotted as a function of the number of particles per model. The agreement factor from crystalline model (fcc) to the same PDF is labeled in a dashed line.

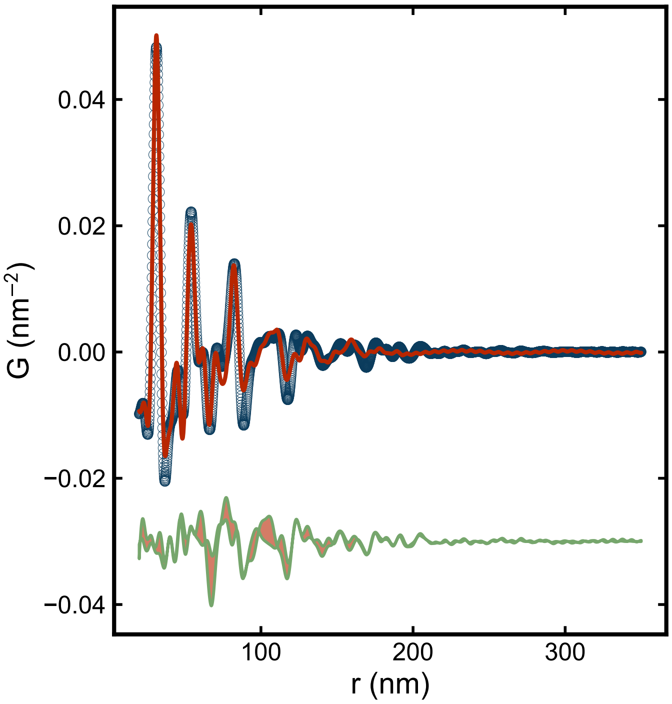


Fig. 5.13: Measured PDF (blue) at reaction time = 800 mins and PDF from best-fit decahedron cluster model (red). The difference curve (green) is plotted with an offset for the ease of reading. The shaded area of difference curve labels the improvement of decahedron cluster from fcc model.

Bibliography

[1] Pinar Akcora, Hongjun Liu, Sanat K. Kumar, Joseph Moll, Yu Li, Brian C. Benicewicz, Linda S. Schadler, Devrim Acehan, Athanassios Z. Panagiotopoulos, Victor Pryamitsyn, Venkat Ganesan, Jan Ilavsky, Pappanan Thiyagarajan, Ralph H. Colby, and Jack F. Douglas. Anisotropic self-assembly of spherical polymer-grafted nanoparticles. 8(4):354–359.

[2] Dmitry Aldakov, Aurlie Lefranois, and Peter Reiss. Ternary and quaternary metal chalcogenide nanocrystals : Synthesis, properties and applications. 1(24):3756–3776.

[3] A. Paul Alivisatos, Kai P. Johnsson, Xiaogang Peng, Troy E. Wilson, Colin J. Loweth, Marcel P. Bruchez, and Peter G. Schultz. Organization of 'nanocrystal molecules' using DNA. 382(6592):609–611.

[4] A. Altomare, M. Camalli, C. Cuocci, C. Giacovazzo, A. Moliterni, and R. Rizzi. EXPO2009: Structure solution by powder data in direct and reciprocal space. *J Appl Cryst*, 42(6):1197–1202, December 2009.

[5] A. Altomare, G. Campi, C. Cuocci, L. Eriksson, C. Giacovazzo, A. Moliterni, R. Rizzi, and P.-E. Werner. Advances in powder diffraction pattern indexing: N-TREOR09. *J Appl Cryst*, 42(5):768–775, October 2009.

[6] Ebbe S. Andersen, Mingdong Dong, Morten M. Nielsen, Kasper Jahn, Ramesh Subramani, Wael Mamdouh, Monika M. Golas, Bjoern Sander, Holger Stark, Cristiano L. P. Oliveira, Jan Skov Pedersen, Victoria Birkedal, Flemming Besenbacher, Kurt V. Gothelf, and Jrgen Kjems. Self-assembly of a nanoscale DNA box with a controllable lid. 459(7243):73–76.

[7] E. Ascher, V. Gramlich, and H. Wondratschek. Korrekturen zu den Angaben Untergruppen' in den Raumgruppen der Internationalen Tabellen zu Bestimmung von Kristallstrukturen (1935), Band I. Corrections to the sections Untergruppen' of the

space groups in Internationale Tabellen zur Bestimmung von Kristallstrukturen (1935), Vol. I. *Acta Cryst B*, 25(10):2154–2156, October 1969.

[8] Evelyn Auyeung, Joshua I. Cutler, Robert J. Macfarlane, Matthew R. Jones, Jinsong Wu, George Liu, Ke Zhang, Kyle D. Osberg, and Chad A. Mirkin. Synthetically programmable nanoparticle superlattices using a hollow three-dimensional spacer approach. 7(1):24–28.

[9] Evelyn Auyeung, Ting I. N. G. Li, Andrew J. Senesi, Abrin L. Schmucker, Bridget C. Pals, Monica Olvera de la Cruz, and Chad A. Mirkin. DNA-mediated nanoparticle crystallization into Wulff polyhedra. 505(7481):73–77.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv14090473 Cs Stat*, Sep. Purif. Technol. 2014.

[11] S. Banerjee, C.-H. Liu, K. M. Jensen, P. Juhs, J. D. Lee, M. Tofanelli, C. J. Ackerson, C. B. Murray, and S. J. L. Billinge. Cluster-mining: An approach for determining core structures of metallic nanoparticles from atomic pair distribution function data. 76(1):24–31.

[12] Soham Banerjee, Chia-Hao Liu, Kirsten M. O. Jensen, Pavol Juhás, Jennifer D. Lee, Christopher J. Ackerson, Christopher B. Murray, and Simon J. L. Billinge. Cluster-mining: An approach for determining core structures of metallic nanoparticles from atomic pair distribution function data. 2019. arXiv:1901.08754 [cond-mat.mtrl-sci].

[13] W. H. Baur and A. A. Khan. Rutile-type compounds. IV. SiO2, GeO2 and a comparison with other rutile-type structures. *Acta Cryst B*, 27(11):2133–2139, November 1971.

[14] M. G. Bawendi, A. R. Kortan, M. L. Steigerwald, and L. E. Brus. Xray structural characterization of larger CdSe semiconductor clusters. 91(11):7282–7290.

[15] C. P. Bean and J. D. Livingston. Superparamagnetism. 30(4):S120–S129.

[16] G. Beaucage. Approximations Leading to a Unified Exponential/Power-Law Approach to Small-Angle Scattering. 28(6):717–728.

[17] G. Beaucage, H. K. Kammler, and S. E. Pratsinis. Particle size distributions from small-angle scattering using global scattering functions. 37(4):523–535.

[18] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch. New developments in the Inorganic Crystal Structure Database (ICSD): Accessibility in support of materials research and design. *Acta Cryst B*, 58(3):364–369, June 2002.

[19] C. J. Benmore. A Review of High-Energy X-Ray Diffraction from Glasses and Liquids.

[20] Ludovic Berthier and Giulio Biroli. Theoretical perspective on the glass transition and amorphous materials. 83(2):587–645.

[21] Connor R. Bilchak, Eileen Buenning, Makoto Asai, Kai Zhang, Christopher J. Durning, Sanat K. Kumar, Yucheng Huang, Brian C. Benicewicz, David W. Gidley, Shiwang Cheng, Alexei P. Sokolov, Matteo Minelli, and Ferruccio Doghieri. Polymer-Grafted Nanoparticle Membranes with Controllable Free Volume. 50(18):7111–7120.

[22] S. J. L. Billinge. Nanometre-scale structure from powder diffraction: Total scattering and atomic pair distribution function analysis.

[23] S. J. L. Billinge. Local structure from total scattering and atomic pair distribution function (pdf) analysis. In Robert E. Dinnebier and Simon J. L. Billinge, editors, *Powder diffraction: theory and practice*, pages 464 – 493, London, England, 2008. Royal Society of Chemistry.

[24] S. J. L. Billinge and I. Levin. The problem with determining atomic structure at the nanoscale. *Science*, 316:561–565, 2007.

[25] Simon J. L. Billinge, Philip M. Duxbury, Douglas S. Gonalves, Carlile Lavor, and Antonio Mucherino. Recent results on assigned and unassigned distance geometry with applications to protein molecules and nanostructures. *Ann. Oper. Res.*, pages 1–43, 2018. to be published.

[26] Simon J. L. Billinge and Christopher L. Farrow. Towards a robust *ad-hoc* data correction approach that yields reliable atomic pair distribution functions from powder diffraction data. *J. Phys: Condens. Mat.*, 25:454202, 2013.

[27] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., 2006.

[28] Michael A. Boles, Michael Engel, and Dmitri V. Talapin. Self-Assembly of Colloidal Nanocrystals: From Intricate Structures to Functional Materials. 116(18):11220–11289.

[29] A. Boultif and D. Louër. Powder pattern indexing with the dichotomy method. *J Appl Cryst*, 37(5):724–731, October 2004.

[30] Emil S. Božin, Christos D. Malliakas, Petros Souvatzis, Thomas Proffen, Nicola A. Spaldin, Mercouri G. Kanatzidis, and Simon J. L. Billinge. Entropically stabilized local dipole formation in lead chalcogenides. *Science*, 330:1660, 2010.

[31] L. L. Boyle and J. E. Lawrenson. Klassengleichen supergroup–subgroup relationships between the space groups. *Acta Cryst A*, 28(6):489–493, November 1972.

[32] Louis Brus. Electronic wave functions in semiconductor clusters: Experiment and theory. 90(12):2555–2560.

[33] Shelly D. Burnside, Valery Shklover, Christophe Barb, Pascal Comte, Francine Arendse, Keith Brooks, and Michael Grtzel. Self-Organization of TiO2 Nanoparticles in Thin Films. 10(9):2419–2425.

[34] Julyan H. E. Cartwright and Alan L. Mackay. Beyond crystals: The dialectic of materials and information. 370(1969):2807–2822.

[35] Ji-Hyuk Choi, Han Wang, Soong Ju Oh, Taejong Paik, Pil Sung, Jinwoo Sung, Xingchen Ye, Tianshuo Zhao, Benjamin T. Diroll, Christopher B. Murray, and Cherie R. Kagan. Exploiting the colloidal nanocrystal library to construct electronic devices. 352(6282):205–208.

[36] Joshua J. Choi, Xiaohao Yang, Zachariah M. Norman, Simon J. L. Billinge, and Jonathan S. Owen. Structure of Methylammonium Lead Iodide Within Mesoporous Titanium Dioxide: Active Material in High-Performance Perovskite Solar Cells. *Nano Lett.*, 14(1):127–133, January 2014.

[37] François Chollet et al. Keras. `https://keras.io`, 2015.

[38] Peter J. Chupas, Xiangyun Qiu, J. C. Hanson, P. L. Lee, Clare P. Grey, and Simon J. L. Billinge. Rapid acquisition pair distribution function analysis (RA-PDF). *J. Appl. Crystallogr.*, 36:1342–1347, 2003.

[39] Jacob. W. Ciszek, Ling Huang, Stefan Tsonchev, YuHuang Wang, Kenneth R. Shull, Mark A. Ratner, George C. Schatz, and Chad A. Mirkin. Assembly of Nanorods into Designer Superstructures: The Role of Templating, Capillary Forces, Adhesion, and Polymer Hydration. 4(1):259–266.

[40] Matthew J. Cliffe, Martin T. Dove, D. A. Drabold, and Andrew L. Goodwin. Structure determination of disordered materials from diffraction data. *Phys. Rev. Lett.*, 104(12):125501, 2010.

[41] A. A. Coelho. Indexing of powder diffraction patterns by iterative use of singular value decomposition. *J Appl Cryst*, 36(1):86–95, February 2003.

[42] A. A. Coelho. An indexing algorithm independent of peak position extraction for X-ray powder diffraction patterns. *J Appl Cryst*, 50(5):1323–1330, October 2017.

[43] G. E. Dahl, T. N. Sainath, and G. E. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8609–8613, May 2013.

[44] Marie-Christine Daniel and Didier Astruc. Gold Nanoparticles: Assembly, Supramolecular Chemistry, Quantum-Size-Related Properties, and Applications toward Biology, Catalysis, and Nanotechnology. 104(1):293–346.

[45] G. D'Anna and G. Gremaud. The jamming route to the glass state in weakly perturbed granular media. 413(6854):407–409.

[46] O. Dauchot, G. Marty, and G. Biroli. Dynamical Heterogeneity Close to the Jamming Transition in a Sheared Granular Material. 95(26):265701.

[47] Celso de Mello Doneg, Peter Liljeroth, and Daniel Vanmaekelbergh. Physicochemical Evaluation of the Hot-Injection Method, a Synthesis Route for Monodisperse Nanocrystals. 1(12):1152–1162.

[48] P. M. de Wolff. On the determination of unit-cell dimensions from powder diffraction patterns. *Acta Cryst*, 10(9):590–595, Sep. Purif. Technol. 1957.

[49] P. Debye and H. Menke. The determination of the inner structure of liquids by x-ray means. *Physik. Z.*, 31:797–8, 1930.

[50] T. Egami and S. J. L. Billinge. *Underneath the Bragg peaks: structural analysis of complex materials*. Elsevier, Amsterdam, 2nd edition, 2012.

[51] T. E. Faber and J. M. Ziman. A theory of the electrical properties of liquid metals iii. the resistivity of binary alloys. *Philos. Mag.*, 11:153–157, 1965.

[52] C. L. Farrow and S. J. L. Billinge. Relationship between the atomic pair distribution function and small angle scattering: implications for modeling of nanoparticles. *Acta Crystallogr. A*, 65(3):232–239, 2009.

[53] C. L. Farrow, P. Juhás, Jiwu Liu, D. Bryndin, E. S. Božin, J. Bloch, Th. Proffen, and S. J. L. Billinge. PDFfit2 and PDFgui: Computer programs for studying nanostructure in crystals. *J. Phys: Condens. Mat.*, 19:335219, 2007.

[54] Riccardo Ferrando, Julius Jellinek, and Roy L. Johnston. Nanoalloys: From Theory to Applications of Alloy Clusters and Nanoparticles. 108(3):845–910.

[55] J. L. Finney, A. Hallbrucker, I. Kohl, A. K. Soper, and D. T. Bowron. Structures of High and Low Density Amorphous Ice by Neutron Diffraction. 88(22):225503.

[56] M. E. Fleet. The structure of magnetite. *Acta Cryst B*, 37(4):917–920, Am. Pharm. Rev. 1981.

[57] J. Fortner and J. S. Lannin. Radial distribution functions of amorphous silicon. 39(8):5527–5530.

[58] Takao Furubayashi, Takehiko Matsumoto, Takatsugu Hagino, and Shoichi Nagata. Structural and Magnetic Studies of Metal-Insulator Transition in Thiospinel CuIr 2S 4. *J. Phys. Soc. Jpn.*, 63(9):3333–3339, Sep. Purif. Technol. 1994.

[59] K. Furukawa. The radial distribution curves of liquids by diffraction methods. 25(1):395.

[60] Carmelo Giacovazzo. *Direct Phasing in Crystallography: Fundamentals and Applications.* International Union of Crystallography, Chester, England : Oxford ; New York, 1 edition edition, February 1999.

[61] Michael Giersig and Paul Mulvaney. Preparation of ordered colloid monolayers by electrophoretic deposition. 9(12):3408–3413.

[62] R. Gilles, U. Keiderling, and A. Wiedenmann. Silver behenate powder as a possible low-angle calibration standard for small-angle neutron scattering. 31(6):957–959.

[63] O. Glatter. A new method for the evaluation of small-angle scattering data. 10(5):415–421.

[64] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, November 2016.

[65] A. Guinier. *X-ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies.* W.H. Freeman, San Francisco, 1963.

[66] Andr Guinier. *X-Ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies.* Courier Corporation.

[67] Theo Hahn, editor. *International Tables for Crystallography, Volume A: Space Group Symmetry.* Springer, Dordrecht, 5th edition edition, Am. Pharm. Rev. 2002.

[68] Wei Han, Luoxin Yi, Nan Zhao, Aiwei Tang, Mingyuan Gao, and Zhiyong Tang. Synthesis and Shape-Tailoring of Copper Sulfide/Indium Sulfide-Based Nanocrystals. 130(39):13152–13161.

[69] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition, 2009.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity Mappings in Deep Residual Networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 630–645. Springer International Publishing, 2016.

[72] Rudolf Hergt, Silvio Dutz, Robert Mller, and Matthias Zeisberger. Magnetic particle hyperthermia: Nanoparticle magnetism and materials development for cancer therapy. 18(38):S2919–S2934.

[73] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.*, 29(6):82–97, November 2012.

[74] M. Horn, C. F. Schwerdtfeger, and E. P. Meagher. Refinement of the structure of anatase at several temperatures*. *Zeitschrift fur Kristallographie*, 136:273–281, November 1972.

[75] Roger A. Horn. *Matrix Analysis: Second Edition*. Cambridge University Press, New York, NY, 2 edition edition, October 2012.

[76] Q. Huang, J. L. Soubeyroux, O. Chmaissem, I. Natali Sora, A. Santoro, R. J. Cava, J. J. Krajewski, and W. F. Peck. Neutron Powder Diffraction Study of the Crystal Structures of Sr2RuO4 and Sr2IrO4 at Room Temperature and at 10 K. *Journal of Solid State Chemistry*, 112(2):355–361, October 1994.

[77] Taeghwan Hyeon, Su Seong Lee, Jongnam Park, Yunhee Chung, and Hyon Bin Na. Synthesis of Highly Crystalline and Monodisperse Maghemite Nanocrystallites without a Size-Selection Process. 123(51):12798–12801.

[78] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv150203167 Cs*, February 2015.

[79] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013.

[80] P. B. James and M. T. Lavik. The crystal structure of MoSe2. *Acta Cryst*, 16(11):1183–1183, November 1963.

[81] Eunjoo Jang, Shinae Jun, Hyosook Jang, Jungeun Lim, Byungki Kim, and Younghwan Kim. White-Light-Emitting Diodes with Quantum Dot Color Converters for Display Backlights. 22(28):3076–3080.

[82] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, Sep. Purif. Technol. 2009.

[83] I. K. Jeong, R. H. Heffner, M. J. Graf, and S. J. L. Billinge. Lattice dynamics and correlated atomic motion from the atomic pair distribution function. *Phys. Rev. B*, 67:104301, 2003.

[84] Mayank Jhalaria, Eileen Buenning, Yucheng Huang, Madhusudan Tyagi, Reiner Zorn, Michaela Zamponi, Victoria Garca-Sakai, Jacques Jestin, Brian C. Benicewicz, and Sanat K. Kumar. Accelerated Local Dynamics in Matrix-Free Polymer Grafted Nanoparticles. 123(15):158003.

[85] Matthew R. Jones, Robert J. Macfarlane, Byeongdu Lee, Jian Zhang, Kaylie L. Young, Andrew J. Senesi, and Chad A. Mirkin. DNA-nanoparticle superlattices formed from anisotropic building blocks. 9(11):913–917.

[86] P. Juhás, D. M. Cherba, P. M. Duxbury, W. F. Punch, and S. J. L. Billinge. Ab initio determination of solid-state nanostructure. *Nature*, 440(7084):655–658, 2006.

[87] P. Juhás, T. Davis, C. L. Farrow, and S. J. L. Billinge. PDFgetX3: A rapid and highly automatable program for processing powder diffraction data into total scattering pair distribution functions. *J. Appl. Crystallogr.*, 46:560–566, 2013.

[88] P. Juhás, L. Granlund, S. R. Gujarathi, P. M. Duxbury, and S. J. L. Billinge. Crystal structure solution from experimentally determined atomic pair distribution functions. *J. Appl. Crystallogr.*, 42(3):623–629, Jun 2010.

[89] Pavol Juhás, Christopher L. Farrow, Xiaohao Yang, Kevin R. Knox, and Simon J. L. Billinge. Complex modeling: a strategy and software program for combining multiple information sources to solve ill-posed structure and nanostructure inverse problems. *Acta Crystallogr. A*, 71(6):562–568, Nov 2015.

[90] Pavol Juhs, Jaap N. Louwen, Lambert van Eijck, Eelco T. C. Vogt, and Simon J. L. Billinge. *PDFgetN3*: atomic pair distribution functions from neutron powder diffraction data using *ad hoc* corrections. *J. Appl. Crystallogr.*, 51(5):1492–1497, Oct 2018.

[91] Cherie R. Kagan, Efrat Lifshitz, Edward H. Sargent, and Dmitri V. Talapin. Building devices from colloidal quantum dots. 353(6302).

[92] David A Keen and Andrew L Goodwin. The crystallography of correlated disorder. *Nature*, 521(7552):303–309, 2015.

[93] J. Kieffer, G. Ashiotis, A. Deschildre, Z. Nawaz, J. P. Wright, D. Karkoulis, and F. E. Picca. The fast azimuthal integration python library: pyFAI. *J. Appl. Crystallogr.*, 48:510–519, 2015.

[94] Yoon Kim. Convolutional Neural Networks for Sentence Classification. *ArXiv14085882 Cs*, August 2014.

[95] Gary King and Langche Zeng. Logistic Regression in Rare Events Data. *Polit. Anal.*, 9(2):137–163, 2001/ed.

[96] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs*, December 2014.

[97] Michel H. J. Koch, Patrice Vachette, and Dmitri I. Svergun. Small-angle scattering: A view on the properties, structures and structural changes of biological macromolecules in solution. 36(2):147–227.

[98] Michael Kotlarchyk and SowHsin Chen. Analysis of small angle neutron scattering spectra from polydisperse interacting colloids. 79(5):2461–2469.

[99] Maksym V. Kovalenko, Liberato Manna, Andreu Cabot, Zeger Hens, Dmitri V. Talapin, Cherie R. Kagan, Victor I. Klimov, Andrey L. Rogach, Peter Reiss, Delia J. Milliron, Philippe Guyot-Sionnnest, Gerasimos Konstantatos, Wolfgang J. Parak, Taeghwan Hyeon, Brian A. Korgel, Christopher B. Murray, and Wolfgang Heiss. Prospects of Nanoscience with Nanocrystals. 9(2):1012–1057.

[100] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[101] Sanat K. Kumar, Brian C. Benicewicz, Richard A. Vaia, and Karen I. Winey. 50th Anniversary Perspective: Are Polymer Nanocomposites Practical for Applications? 50(3):714–731.

[102] G. H. Kwei, A. C. Lawson, S. J. L. Billinge, and S.-W. Cheong. Structures of the ferroelectric phases of barium titanate. *J. Phys. Chem.*, 97:2368, 1993.

[103] Y. Lalatonne, J. Richardi, and M. P. Pileni. Van der Waals versus dipolar forces controlling mesoscopic organizations of magnetic nanocrystals. 3(2):121–125.

[104] Karel Lambert, Bram De Geyter, Iwan Moreels, and Zeger Hens. PbTe—CdTe Core—Shell Particles by Cation Exchange, a HR-TEM study. 21(5):778–780.

[105] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, November 1998.

[106] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[107] Chunzhao Li, Junwon Han, Chang Y. Ryu, and Brian C. Benicewicz. A Versatile Method To Prepare RAFT Agent Anchored Substrates and the Preparation of PMMA Grafted Nanoparticles. 39(9):3175–3183.

[108] Sen Li, HongZhe Wang, WeiWei Xu, HongLei Si, XiaoJun Tao, Shiyun Lou, Zuliang Du, and Lin Song Li. Synthesis and assembly of monodisperse spherical Cu2S nanocrystals. 330(2):483–487.

[109] Andrea J. Liu and Sidney R. Nagel. Jamming is not just cool any more. 396(6706):21–22.

[110] Chia-Hao Liu, Eric Janke, Ruipen Li, Pavol Juhás, Oleg Gang, Dimitri V. Talapin, and Simon J. L. Billinge. saspdf: pair distribution function analysis of nanoparticle assemblies from small-angle-scattering data, 2019.

[111] Fang Lu, Thi Vo, Yugang Zhang, Alex Frenkel, Kevin G. Yager, Sanat Kumar, and Oleg Gang. Unusual packing of soft-shelled nanocubes. 5(5):eaaw2399.

[112] Robert J. Macfarlane, Byeongdu Lee, Matthew R. Jones, Nadine Harris, George C. Schatz, and Chad A. Mirkin. Nanoparticle Superlattice Engineering with DNA. 334(6053):204–208.

[113] Giulia Fulvia Mancini, Tatiana Latychevskaia, Francesco Pennacchio, Javier Reguera, Francesco Stellacci, and Fabrizio Carbone. Order/Disorder Dynamics in a Dodecanethiol-Capped Gold Nanoparticles Supracrystal by Small-Angle Ultrafast Electron Diffraction. 16(4):2705–2713.

[114] M. Marezio and P. D. Dernier. The crystal structure of Ti4O7, a member of the homologous series TinO2n-1. *Journal of Solid State Chemistry*, 3(3):340–348, August 1971.

[115] L. D. Marks and David J. Smith. High resolution studies of small particles of gold and silver: I. Multiply-twinned particles. 54(3):425–432.

[116] A. J. Markvardsen, K. Shankland, W. I. F. David, J. C. Johnston, R. M. Ibberson, M. Tucker, H. Nowell, and T. Griffin. ExtSym: A program to aid space-group determination from powder diffraction data. *J Appl Cryst*, 41(6):1177–1181, December 2008.

[117] A. S. Masadeh, E. S. Božin, C. L. Farrow, G. Paglia, P. Juhás, A. Karkamkar, M. G. Kanatzidis, and S. J. L. Billinge. Quantitative size-dependent structure and strain determination of CdSe nanoparticles using atomic pair distribution function analysis. *Phys. Rev. B*, 76:115413, 2007.

[118] Nobuhiro Matsumoto, Kouji Taniguchi, Ryo Endoh, Hideaki Takano, and Shoichi Nagata. Resistance and Susceptibility Anomalies in IrTe2 and CuIr2Te4. *Journal of Low Temperature Physics*, 117(5):1129–1133, December 1999.

[119] John A. McGuire, Milan Sykora, Jin Joo, Jeffrey M. Pietryga, and Victor I. Klimov. Apparent Versus True Carrier Multiplication Yields in Semiconductor Nanocrystals. 10(6):2049–2057.

[120] A. D. Mighell and A. Santoro. Geometrical ambiguities in the indexing of powder patterns. *J Appl Cryst*, 8(3):372–374, June 1975.

[121] Maria Mikhaylova, Do Kyung Kim, Natalia Bobrysheva, Mikhail Osmolowsky, Valentin Semenov, Thomas Tsakalakos, and Mamoun Muhammed. Superparamagnetism of Magnetite Nanoparticles: Dependence on Surface Modification. 20(6):2472–2477.

[122] Younjin Min, Mustafa Akbulut, Kai Kristiansen, Yuval Golan, and Jacob Israelachvili. The role of interparticle and external forces in nanoparticle assembly. 7(7):527–538.

[123] Chad A. Mirkin, Robert L. Letsinger, Robert C. Mucic, and James J. Storhoff. A DNA-based method for rationally assembling nanoparticles into macroscopic materials. 382(6592):607–609.

[124] Karol Miszta, Joost de Graaf, Giovanni Bertoni, Dirk Dorfs, Rosaria Brescia, Sergio Marras, Luca Ceseracciu, Roberto Cingolani, Ren van Roij, Marjolein Dijkstra, and Liberato Manna. Hierarchical self-assembly of suspended branched colloidal nanocrystals into superlattice structures. 10(11):872–876.

[125] L. Motte, F. Billoudet, and M. P. Pileni. Self-Assembled Monolayer of Nanosized Particles Differing by Their Sizes. 99(44):16425–16429.

[126] Catherine J. Murphy, Tapan K. Sau, Anand M. Gole, Christopher J. Orendorff, Jinxin Gao, Linfeng Gou, Simona E. Hunyadi, and Tan Li. Anisotropic Metal Nanoparticles: Synthesis, Assembly, and Optical Applications. 109(29):13857–13870.

[127] C. B. Murray, C. R. Kagan, and M. G. Bawendi. Self-Organization of CdSe Nanocrystallites into Three-Dimensional Quantum Dot Superlattices. 270(5240):1335–1338.

[128] C. B. Murray, C. R. Kagan, and M. G. Bawendi. Synthesis and Characterization of Monodisperse Nanocrystals and Close-Packed Nanocrystal Assemblies. 30(1):545–610.

[129] C. B. Murray, D. J. Norris, and M. G. Bawendi. Synthesis and characterization of nearly monodisperse CdE (E = sulfur, selenium, tellurium) semiconductor nanocrystallites. 115(19):8706–8715.

[130] Gautham Nair and Moungi G. Bawendi. Carrier multiplication yields of CdSe and CdTe nanocrystals by transient photoluminescence spectroscopy. 76(8):081304.

[131] Keisuke Nakayama, Katsuaki Tanabe, and Harry A. Atwater. Plasmonic nanoparticle enhanced light absorption in GaAs solar cells. 93(12):121904.

[132] M. A. Neumann. X-Cell: A novel indexing algorithm for routine tasks and difficult cases. *J Appl Cryst*, 36(2):356–365, Am. Pharm. Rev. 2003.

[133] Dmytro Nykypanchuk, Mathew M. Maye, Daniel van der Lelie, and Oleg Gang. DNA-guided crystallization of colloidal nanoparticles. 451(7178):549–552.

[134] Sophocles J. Orfanidis. *Introduction to Signal Processing*. Prentice Hall.

[135] E.a. Owen and E.l. Yates. LXVI. X-ray measurement of the thermal expansion of pure nickel. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 21(142):809–819, Am. Pharm. Rev. 1936.

[136] Katharine Page, Thomas Proffen, Markus Niederberger, and Ram Seshadri. Probing Local Dipoles and Ligand Structure in BaTiO3 Nanoparticles. *Chem. Mater.*, 22(15):4386–4391, August 2010.

[137] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, and K.-S. Sohn. Classification of crystal structure using a convolutional neural network. *IUCrJ*, 4(4):486–494, July 2017.

[138] Vitalij K. Pecharsky and Peter Y. Zavalij. *Fundamentals of Powder Diffraction and Structural Characterization of Materials*. Springer, New York, USA, 2005.

[139] Jan Skov Pedersen. Analysis of small-angle scattering data from colloids and polymer solutions: Modeling and least-squares fitting. 70:171–210.

[140] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12:2825, October 2011.

[141] P. F. Peterson, E. S. Božin, Th. Proffen, and S. J. L. Billinge. Improved measures of quality for atomic pair distribution functions. *J. Appl. Crystallogr.*, 36:53, 2003.

[142] V. Petkov, S. J. L. Billinge, P. Larson, S. D. Mahanti, T. Vogt, K. K. Rangan, and M. G. Kanatzidis. Structure of nanocrystalline materials using atomic pair distribution function analysis: study of $\mathrm{LiMoS_2}$. *Phys. Rev. B*, 65:092105, 2002.

[143] Pawel Pieranski, L. Strzelecki, and B. Pansu. Thin Colloidal Crystals. 50(12):900–903.

[144] M. P. Pileni. Nanocrystal Self-Assemblies: Fabrication and Collective Properties. 105(17):3358–3371.

[145] M.-P. Pileni. Self-Assembly of Inorganic Nanocrystals: Fabrication and Collective Intrinsic Properties. 40(8):685–693.

[146] Paul Podsiadlo, Amit K. Kaushik, Ellen M. Arruda, Anthony M. Waas, Bong Sup Shim, Jiadi Xu, Himabindu Nandivada, Benjamin G. Pumplin, Joerg Lahann, Ayyalusamy Ramamoorthy, and Nicholas A. Kotov. Ultrastrong and Stiff Layered Polymer Nanocomposites. 318(5847):80–83.

[147] Jrg Polte, Robert Erler, Andreas F. Thnemann, Sergey Sokolov, T. Torsten Ahner, Klaus Rademann, Franziska Emmerling, and Ralph Kraehnert. Nucleation and Growth of Gold Nanoparticles Studied via in situ Small Angle X-ray Scattering at Millisecond Time Resolution. 4(2):1076–1082.

[148] Th. Proffen and S. J. L. Billinge. PDFFIT, a program for full profile structural refinement of the atomic pair distribution function. *J. Appl. Crystallogr.*, 32:572–575, 1999.

[149] Thomas Proffen, Katharine L. Page, Sylvia E. McLain, Bjorn Clausen, Timothy W. Darling, James A. TenCate, Seung-Yub Lee, and Ersan Ustundag. Atomic pair distribution function analysis of materials containing crystalline and amorphous phases. *Z. Kristallogr.*, 220(12):1002–1008, 2005.

[150] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ArXiv151106434 Cs*, November 2015.

[151] N. Ramesh, P. K. Davis, J. M. Zielinski, R. P. Danner, and J. L. Duda. Application of free-volume theory to self diffusion of solvents in polymers below the glass transition temperature: A review. 49(23):1629–1644.

[152] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: Recent applications and prospects. *Npj Comput. Mater.*, 3(1):54, December 2017.

[153] John J. Randall, Lewis Katz, and Roland Ward. The Preparation of a Strontium-Iridium Oxide Sr2IrO41,2. *J. Am. Chem. Soc.*, 79(2):266–267, January 1957.

[154] L. Harivardhan Reddy, Jos L. Arias, Julien Nicolas, and Patrick Couvreur. Magnetic Nanoparticles: Design and Characterization, Toxicity and Biocompatibility, Pharmaceutical and Biomedical Applications. 112(11):5818–5878.

[155] Franz X. Redl, Charles T. Black, Georgia C. Papaefthymiou, Robert L. Sandstrom, Ming Yin, Hao Zeng, Christopher B. Murray, and Stephen P. O'Brien. Magnetic,

Electronic, and Structural Characterization of Nonstoichiometric Iron Oxides at the Nanoscale. 126(44):14583–14599.

[156] Ute Resch-Genger, Markus Grabolle, Sara Cavaliere-Jaricot, Roland Nitschke, and Thomas Nann. Quantum dots versus organic dyes as fluorescent labels. 5(9):763–775.

[157] Katherine P. Rice, Aaron E. Saunders, and Mark P. Stoykovich. Seed-Mediated Growth of Shape-Controlled Wurtzite CdSe Nanocrystals: Platelets, Cubes, and Rods. 135(17):6669–6676.

[158] Perla Rittigstein and John M. Torkelson. Polymernanoparticle interfacial interactions in polymer nanocomposites: Confinement effects on glass transition temperature and suppression of physical aging. 44(20):2935–2943.

[159] Aurora Rizzo, Yanqin Li, Stefan Kudera, Fabio Della Sala, Marco Zanella, Wolfgang J. Parak, Roberto Cingolani, Liberato Manna, and Giuseppe Gigli. Blue light emitting diodes based on fluorescent CdSeZnS nanocrystals. 90(5):051106.

[160] Parham Rohani, Soham Banerjee, Souroush Ashrafi-Asl, Mohammad Malekzadeh, Reza Shahbazian-Yassar, Simon J. L. Billinge, and Mark T. Swihart. Synthesis and properties of boron-hyperdoped silicon nanoparticles. *Adv. Funct. Mater.*, 2018. Published.

[161] N. M. Rosengaard and H. L. Skriver. Calculated stacking-fault energies of elemental metals. 47(19):12865–12873.

[162] Michael B. Ross, Jessie C. Ku, Victoria M. Vaccarezza, George C. Schatz, and Chad A. Mirkin. Nanoscale form dictates mesoscale function in plasmonic DNAnanoparticle superlattices. 10(5):453–458.

[163] Sara M. Rupich, Elena V. Shevchenko, Maryna I. Bodnarchuk, Byeongdu Lee, and Dmitri V. Talapin. Size-Dependent Multiple Twinning in Nanocrystal Superlattices. 132(1):289–296.

[164] Edward H. Sargent. Solar Cells, Photodetectors, and Optical Sources from Infrared Colloidal Quantum Dots. 20(20):3958–3964.

[165] A. J. Senesi and B. Lee. Small-angle scattering of particle assemblies. 48(4):1172–1182.

[166] Elena V. Shevchenko, Dmitri V. Talapin, Nicholas A. Kotov, Stephen O'Brien, and Christopher B. Murray. Structural diversity in binary nanoparticle superlattices. 439(7072):55–59.

[167] Tetsuo Shimura, Yoshiyuki Inaguma, Tetsuro Nakamura, Mitsuru Itoh, and Yukio Morii. Structure and magnetic properties of ${\mathrm{Sr}}_{2\mathrm{\ensuremath{-}}\mathit{x}}$${\mathit{A}}_{\mathit{x}}$${\mathrm{IrO}}_{4}$ (A=Ca and Ba). *Phys. Rev. B*, 52(13):9143–9146, October 1995.

[168] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, October 2017.

[169] Robert L. Snyder, Jaroslav Fiala, Hans J. Bunge, Hans Joachim Bunge, and International Union of Crystallography. *Defect and Microstructure Analysis by Diffraction*. Oxford University Press.

[170] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*, 15(1):1929–1958, January 2014.

[171] Ivana Srnov-loufov, Frantiek Lednick, Antonn Gemperle, and Juliana Gemperlov. CoreShell (Ag)Au Bimetallic Nanoparticles: Analysis of Transmission Electron Microscopy Images. 16(25):9928–9935.

[172] Stephen V. Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1):77–89, October 1997.

[173] James J. Storhoff, Anne A. Lazarides, Robert C. Mucic, Chad A. Mirkin, Robert L. Letsinger, and George C. Schatz. What Controls the Optical Properties of DNA-Linked Gold Nanoparticle Assemblies? 122(19):4640–4650.

[174] Shouheng Sun, C. B. Murray, Dieter Weller, Liesl Folks, and Andreas Moser. Monodisperse FePt Nanoparticles and Ferromagnetic FePt Nanocrystal Superlattices. 287(5460):1989–1992.

[175] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[176] I. P. Swainson, R. P. Hammond, C. Soullière, O. Knop, and W. Massa. Phase transitions in the perovskite methylammonium lead bromide, CH3ND3PbBr3. *Journal of Solid State Chemistry*, 176(1):97–104, November 2003.

[177] Dmitri V. Talapin, Jong-Soo Lee, Maksym V. Kovalenko, and Elena V. Shevchenko. Prospects of Colloidal Nanocrystals for Electronic and Optoelectronic Applications. 110(1):389–458.

[178] Dmitri V. Talapin and Christopher B. Murray. PbSe Nanocrystal Solids for n- and p-Channel Thin Film Field-Effect Transistors. 310(5745):86–89.

[179] Dmitri V. Talapin, Elena V. Shevchenko, Maryna I. Bodnarchuk, Xingchen Ye, Jun Chen, and Christopher B. Murray. Quasicrystalline order in self-assembled binary nanoparticle superlattices. 461(7266):964–967.

[180] Dmitri V. Talapin, Elena V. Shevchenko, Christopher B. Murray, Alexey V. Titov, and Petr Krl. DipoleDipole Interactions in Nanoparticle Superlattices. 7(5):1213–1219.

[181] Ye Tian, Yugang Zhang, Tong Wang, Huolin L. Xin, Huilin Li, and Oleg Gang. Lattice engineering through nanoparticleDNA frameworks. 15(6):654–661.

[182] B. H. Toby and T. Egami. Accuracy of pair distribution function analysis applied to crystalline and noncrystalline materials. *Acta Crystallogr. A*, 48(3):336–46, 1992.

[183] Tatsuya Toriyama, Masao Kobori, Takehisa Konishi, Yukinori Ohta, Kunihisa Sugimoto, Jungeun Kim, Akihiko Fujiwara, Sunseng Pyon, Kazutaka Kudo, and Minoru Nohara. Switching of Conducting Planes by Partial Dimer Formation in IrTe2. *J. Phys. Soc. Jpn.*, 83(3):033701, February 2014.

[184] S. Torquato and F. H. Stillinger. Controlling the Short-Range Order and Packing Densities of Many-Particle Systems. 106(33):8354–8359.

[185] S. Torquato and F. H. Stillinger. Jammed hard-particle packings: From Kepler to Bernal and beyond. 82(3):2633–2672.

[186] John Turkevich and Harry Hopkins Hubbell. Low Angle X-Ray Diffraction of Colloidal Gold and Carbon Black1a. 73(1):1–7.

[187] V. S. Urusov and T. N. Nadezhina. Frequency distribution and selection of space groups in inorganic crystal chemistry. *J Struct Chem*, 50(1):22–37, December 2009.

[188] David H. Van Winkle and C. A. Murray. Layering transitions in colloidal crystals as observed by diffraction and direct-lattice imaging. 34(1):562–573.

[189] Danil Vanmaekelbergh and Peter Liljeroth. Electron-conducting quantum dot solids: Novel materials based on colloidal semiconductor nanocrystals. 34(4):299–312.

[190] J. W. Visser. A fully automatic program for finding the unit cell from powder data. *J Appl Cryst*, 2(3):89–95, August 1969.

[191] Vladimir V. Volkov and Dmitri I. Svergun. Uniqueness of ab initio shape determination in small-angle scattering. 36(3-1):860–864.

[192] Z. L. Wang. Transmission Electron Microscopy of Shape-Controlled Nanocrystals and Their Assemblies. 104(6):1153–1175.

[193] B. E. Warren. *x-ray Diffraction*. Addison-Wesley, New York.

[194] B. E. Warren. *X-ray Diffraction*. Dover, New York, 1990.

[195] Robert L. Whetten, Marat N. Shafigullin, Joseph T. Khoury, T. Gregory Schaaff, Igor Vezmar, Marcos M. Alvarez, and Angus Wilkinson. Crystal Structures of Molecular Gold Nanocrystal Arrays. 32(5):397–406.

[196] Adrian C. Wright. Diffraction studies of glass structure. 123(1):129–148.

[197] K. G. Yager, Y. Zhang, F. Lu, and O. Gang. Periodic lattices of arbitrary nano-objects: Modeling and applications for self-assembled systems. 47(1):118–129.

[198] Tianzhong Yang, Chengmin Shen, Zian Li, Huairuo Zhang, Congwen Xiao, Shutang Chen, Zhichuan Xu, Dongxia Shi, Jianqi Li, and Hongjun Gao. Highly Ordered Self-Assembly with Large Area of Fe3O4 Nanoparticles and the Magnetic Properties. 109(49):23233–23236.

[199] Masatomo Yashima and Syuuhei Kobayashi. Positional disorder of oxygen ions in ceria at high temperatures. *Appl. Phys. Lett.*, 84(4):526–528, January 2004.

[200] Xingchen Ye, Chenhui Zhu, Peter Ercius, Shilpa N. Raja, Bo He, Matthew R. Jones, Matthew R. Hauwiller, Yi Liu, Ting Xu, and A. Paul Alivisatos. Structural diversity in binary superlattices self-assembled from polymer-grafted nanocrystals. 6(1):1–10.

[201] Kaylie L. Young, Michael B. Ross, Martin G. Blaber, Matthew Rycenga, Matthew R. Jones, Chuan Zhang, Andrew J. Senesi, Byeongdu Lee, George C. Schatz, and Chad A.

Mirkin. Using DNA to Design Plasmonic Metamaterials with Tunable Optical Properties. 26(4):653–659.

[202] Runze Yu, Soham Banerjee, H. C. Lei, Ryan Sinclair, Milinda Abeykoon, H. D. Zhou, Cedomir Petrovic, Zurab Guguchia, and Emil Bozin. Absence of local fluctuating dimers in superconducting $\mathrm{Ir}_{1-x}(\mathrm{Pt,Rh})_x\mathrm{Te}_2$. *Phys. Rev. B*, 97(17):174515, 2018.

[203] Gilles Zerah and JeanPierre Hansen. Selfconsistent integral equations for fluid pair distribution functions: Another attempt. 84(4):2336–2343.

[204] Honghu Zhang, Wenjie Wang, Mufit Akinc, Surya Mallapragada, Alex Travesset, and David Vaknin. Assembling and ordering polymer-grafted nanoparticles in three dimensions. 9(25):8710–8715.

[205] Jianyuan Zhang, Peter J. Santos, Paul A. Gabrys, Sangho Lee, Caroline Liu, and Robert J. Macfarlane. Self-Assembling Nanocomposite Tectons. 138(50):16228–16231.

[206] Qifeng Zhang, Evan Uchaker, Stephanie L. Candelaria, and Guozhong Cao. Nanomaterials for energy conversion and storage. 42(7):3127–3171.

[207] Yugang Zhang, Suchetan Pal, Babji Srinivasan, Thi Vo, Sanat Kumar, and Oleg Gang. Selective transformations between nanoparticle superlattices via the reprogramming of DNA-mediated interactions. 14(8):840–847.

[208] Zhongbin Zhuang, Qing Peng, Boce Zhang, and Yadong Li. Controllable Synthesis of Cu2S Nanocrystals and Their Assembly into a Superlattice. 130(32):10482–10483.

[209] Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M. Ghiringhelli. Insightful classification of crystal structures using deep learning. *Nat. Commun.*, 9(1):2775, July 2018.

[210] Mirijam Zobel, Reinhard B. Neder, and Simon A. J. Kimber. Universal solvent restructuring induced by colloidal nanoparticles. *Science*, 347(6219):292–294, 2015.

[211] Hui Zou and Trevor Hastie. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.