crystals

MDPI

Article

The Sensitivity of the Pair-Angle Distribution Function to Protein Structure

Patrick Adams †, Jack Binns †, Tamar L. Greaves and Andrew V. Martin *

School of Science, RMIT University, Melbourne, VIC 3000, Australia; S3826109@student.rmit.edu.au (P.A.); jack.binns@rmit.edu.au (J.B.); tamar.greaves@rmit.edu.au (T.L.G.)

* Correspondence: andrew.martin@rmit.edu.au

† These authors contributed equally to this work.

Received: 15 July 2020; Accepted: 17 August 2020; Published: 20 August 2020

Abstract: The continued development of X-ray free-electron lasers and serial crystallography techniques has opened up new experimental frontiers. Nanoscale dynamical processes such as crystal growth can now be probed at unprecedented time and spatial resolutions. Pair-angle distribution function (PADF) analysis is a correlation-based technique that has the potential to extend the limits of current serial crystallography experiments, by relaxing the requirements for crystal order, size and number density per exposure. However, unlike traditional crystallographic methods, the PADF technique does not recover the electron density directly. Instead it encodes substantial information about local three-dimensional structure in the form of three- and four-body correlations. It is not yet known how protein structure maps into the many-body PADF correlations. In this paper, we explore the relationship between the PADF and protein conformation. We calculate correlations in reciprocal and real space for model systems exhibiting increasing degrees of order and secondary structural complexity, from disordered polypeptides, single alpha helices, helix bundles and finally a folded 100 kilodalton protein. These models systems inform us about the distinctive angular correlations generated by bonding, polypeptide chains, secondary structure and tertiary structure. They further indicate the potential to use angular correlations as a sensitive measure of conformation change that is complementary to existing structural analysis techniques.

Keywords: pair-angle distribution function; X-ray cross-correlation analysis; proteins; structural analysis; X-ray crystallography; serial crystallography; X-ray free-electron laser

1. Introduction

Proteins, as sophisticated macromolecular machines, are integral to complex biochemical processes such as DNA transcription, respiration, cell signalling and molecular transport. A detailed understanding of human physiology and disease at the molecular level requires an intricate knowledge of how proteins function [1–4]. Capturing the functional motions of proteins in laboratory experiments is extremely challenging, and there is currently no way to observe the structure of a single protein at high temporal and spatial resolution as it performs its function. Instead, proteins are studied as ensembles, which imposes limitations on what information can be obtained experimentally. The most dominant technique in structural biology is X-ray crystallography [5–7], where protein crystals enhance the scattering signal at the expense of introducing artefacts due to crystal packing [8,9]. Non-crystalline ensembles are studied by a range of techniques including small-angle X-ray/neutron scattering (SAXS/SANS), nuclear magnetic resonance (NMR) or cryo-electron microscopy (cryo-EM). Each of these techniques has limitations, either in the preparation of the sample (e.g., crystallisation or solvation), limits on the size of the subject protein molecules or in the structural information provided.

For example, protein NMR is capable of providing highly detailed atomic and dynamical information, although it is limited to small proteins or subunits [10,11].

Not all proteins can be characterised by a single experimental technique. For example, membrane proteins comprise up to 50% of small-molecule drug targets [2], but less than 10% of the protein structures in the Protein Data Bank (PDB) [5]. Located in cell membranes, they do not readily crystallise due to a high degree of flexibility and the presence of both hydrophilic and hydrophobic surfaces which hinders organisation into well-ordered three-dimensional lattices [12,13]. Despite advances in membrane-based crystallisation and micro-focus synchrotron beams [14,15], membrane proteins remain very challenging targets for X-ray crystallography experiments.

Crystallography requires well-ordered crystals, a minimum crystal size, and ideally a single crystal per exposure. The crystalline order must persist to high resolution to enable accurate fitting of atomic models, which works best with data approaching 2 Å resolution. For synchrotron sources, the minimum crystal size to achieve this resolution is approximately 5–30 micron, depending on the sample and the beamline [16]. In turn, these requirements mean very few studies of chemically driven time-resolved protein crystallography are currently feasible, as the diffusion of the triggering molecule into the crystal is impacted by crystal size. As a result, such time-resolved structural studies are often limited to SAXS experiments on functional assemblies [17].

The development of serial crystallography techniques has permitted the use of micron-size protein crystals and, when combined with X-ray free-electron lasers, significantly improved the time-resolution of experiments [18]. The method of delivering micron-sized crystals to an X-ray beam using micro-fluidic streams was developed to replenish the sample in destructive XFEL experiments [19]. By exploiting the femtosecond duration of XFEL pulses, the serial technique enabled the dose limits of conventional protein crystallography to be massively exceeded [19].

The serial technique has also proved advantageous at synchrotron sources for studying membrane protein crystals at room temperature [20,21] and for time-resolved studies [22]. Room temperature studies are achieved at synchrotrons with serial crystal delivery by spreading the radiation dose over an ensemble of crystals, reducing radiation damage per crystal [20]. Serial methods have also been developed for proteins in solution to study time-resolved mixing experiments at synchrotrons [23] and light-induced protein dynamics at X-ray free-electron lasers [24,25].

Serial diffraction experiments have driven the development of powerful new data analysis methods. One of these new approaches is X-ray cross-correlation analysis (XCCA), primarily developed for single particle imaging [26–28] and disordered materials [29–31]. This approach analyses the correlations between the diffracted intensities at different scattering vectors, averaged over many particles in random orientations. This technique has been successfully applied to a diverse range of materials, for instance, revealing structural distortions in viral capsids [32] and transient precursor structures formed during nanocrystal self-assembly [33]. XCCA has also been used to study liquid crystal thin films, determining the extension of the pair distribution function into two dimensions [34]. Progress in this area is summarised in recent reviews [35,36].

From a statistical point of view, the intensity correlations can be mapped into real space to recover a sum of three- and four-atom correlations known as the pair-angle distribution function (PADF) [37]. This can been seen as a natural extension to higher dimensions of the two-body statistical information contained in powder diffraction and small-angle X-ray scattering (SAXS). The PADF contains bond angle information and, at larger distances, its angular structure provides a "fingerprint" that can identify local molecular structures.

Unlike conventional crystallography, XCCA and PADF analysis can be applied when there are tens or hundreds of crystals in the beam per exposure. These methods lie in between conventional crystallography that requires a single crystal (or a few domains at most) per exposure and powder diffraction where there are a very large number of crystals and the scattering becomes isotropic. A second potential advantage of PADF and XCCA analysis for protein crystals is that these methods impose less stringent requirements on crystallinity and crystal size than conventional crystallographic

experiments. Removing these requirements allows PADF experiments to be applied to ensembles of nanocrystals, rather than large, high-quality single crystals. As compared to single crystals, the diffusion times of chemical reagents through such ensembles are reduced, opening the way for chemically triggered time-resolved experiments or for studies of dynamical process such as crystal nucleation and growth.

The disadvantage of the PADF technique for protein crystal data is that in its current form it does not recover electron density, only correlation functions that characterise structural statistics. These correlation functions are in a sense analogous to Patterson maps which encode interatomic vectors in crystallography. Therefore, the sensitivity of these correlations to protein structure needs to be understood in order to determine their potential benefits for studying protein crystals. Protein structures have been recovered from simulated single particle data using XCCA [38], suggesting there may be suitable extensions of XCCA to structure determination from protein crystals in the future. Here, we present the analysis of multi-atom PADF statistics [37] as a first step toward more advanced XCCA-based structure determination from crystals.

Our goal here is to explore the sensitivity of multi-atom PADF statistics to protein structure using atomic models and structure factors from the PDB. First, we calculate the PADF directly from atomic models to understand how protein structure is represented in real-space angular correlations. By using the example of a disordered polypeptide we observe the angular structure of the intramolecular bonding in the peptide chain. Then, we gradually increase the structural complexity by studying an alpha helix, a small alpha helix bundle and finally a protein with approximately 30 helices. The angular peak positions in the PADF distributions can act like “fingerprints" to identify these different types of structural order. Second, we then study how much structural PADF information would be accessed in experiments by generating intensity correlations and PADF plots from the experimentally derived structure factors. We also investigate how the finite radial and angular resolution in PADF measurements impacts the accessibility of the structural information.

2. Background on the PADF Technique

The starting point for PADF analysis are the correlations of intensities for each pair of detector pixels. An angular average of these q-space intensity correlations is taken around the beam axis and an ensemble average is taken over all the diffraction patterns as follows,

$$C(q,q',\theta) = \frac{1}{N}\sum_{i=1}^{N}\int I_i(q,\theta')I_i(q',\theta+\theta')d\theta' \,, \tag{1}$$

where the pixel locations are represented in polar coordinates (q,θ) and N is the number of diffraction patterns. The q-space correlation function $C(q,q',\theta)$ converges to a function that does not depend on the absolute orientation of the sample. For measurements of multiple dilute particles per exposure, the correlation function converges to the single particle correlation function [26]. Therefore, this analysis has advantages when multiple crystals are measured.

For crystals, points (q,θ) can be restricted to the Bragg locations and thus the integral written as a sum over integrated Bragg peaks:

$$C(q,q',\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{m\in\mathcal{P}_i(q)}\sum_{n\in\mathcal{P}_i(q')} I_i(q_m,\theta_m)I_i(q_n,\theta+\theta_m) \,, \tag{2}$$

where m (or n) ranges over a set of Bragg peaks $\mathcal{P}_i(q)$ found in diffraction pattern i with $q_m = q$. The position of each Bragg peak is represented in polar coordinates (q,θ), where θ is an angle around the beam axis. We have defined $\theta \equiv \theta_n - \theta_m$.

The q-space intensity correlation function $C(q,q',\theta)$ is analysed differently depending on the sample. For bulk disordered materials, it is common to look for angular symmetries by performing a Fourier analysis [29]. For single particles, $C(q,q',\theta)$ can be combined with coherent diffractive imaging algorithms to recover an image of the object [38–40].

The intensity correlation function can also be converted into a real-space correlation function known as the pair-angle distribution function (PADF) [37]. The PADF is the natural generalisation of pair-distribution analysis of powder diffraction or SAXS data. It encodes the frequency of various local atomic or molecular arrangements in the sample, which can be highly sensitive to local 3D structure in a bulk material. The PADF has been calculated for simulated amorphous materials and simple face-centred cubic crystals [37]. Experimentally, PADF studies have been performed with disordered porous carbon materials and gold nanocrystals using electrons [30] and bond angle peaks were observed in both of these materials. With X-rays, the PADF technique has been used to study the local 3D structure of self-assembled lipid materials [31].

The PADF can be written in terms of n-body correlation functions as follows,

$$\Theta(r,r',\theta) = \tilde{g}^{(2)}(r,r',\theta) + \tilde{g}^{(3)}(r,r',\theta) + \tilde{g}^{(3)}(r,r',\pi-\theta) + \tilde{g}^{(4)}(r,r',\theta)\ , \tag{3}$$

where r and r' are the distances between two pairs of atoms and θ the relative angle between them. In the 3-body terms the atom pairs share a common atom, and in the 2-body terms a pair is correlated with itself. The geometry is shown in Figure 1. The tilde symbol ($\tilde{g}$) indicates a modification from the standard n-body correlation function from statistical physics by averaging over degrees of freedom that the measurement is insensitive to, which are absolute orientation, absolute position and the relative separation of any two atom pairs. See the Appendix A for the precise definition of these functions. Despite its abstract form, the PADF function can be interpreted relatively easily by identifying angular peaks and interpreting their location. For example, at high resolution r and r' can be set to typical bond distances and the PADF then yields the bond angle distribution, which can contain peaks at precise locations determined by the composition and bonding or the sample. At large distances, we obtain characteristic angular structures related to larger structures, which for protein crystals could include correlations between secondary structures, between proteins or between units cells. The PADF has the potential to access 3D structural information about protein conformation or crystal packing, even when crystal order or data quality prevents high resolution structure determination by conventional crystallography analysis methods.

The real-space PADF is obtained from the experimentally measured q-space correlation function $C(q,q',\theta)$ via a series of transformations based on spherical harmonics and spherical Bessel functions. The precise details of these transformations are described in [37]. The curvature of the Ewald sphere is accounted for and a lack of Friedel symmetry in the 2D diffraction patterns due to Ewald sphere curvature does not impact the results. It is important to note that the whole map from $C(q,q',\theta)$ to the PADF is linear and does not involve phasing. In this sense, the PADF has closer relationship to the Patterson function of a crystal than the electron density. The PADF can also be calculated directly from an atomic model by evaluating the n-body correlation functions, thus facilitating structural interpretation. In the following sections, we present PADF results calculated from experimentally measured structure factor files and structure files taken from the PDB [5].

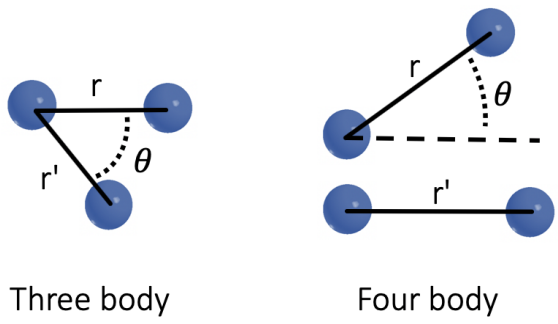

Figure 1. Geometry of three- and four-body correlations. The definition of the atomic coordinates in the pair-angle distribution function (PADF). Note the four-body term is insensitive to the distance between the two pairs.

3. Results

3.1. Correlations from Atomic Models

The most basic method to calculate model correlations is directly from atomic models such as those deposited in the PDB. Calculation of the model PADF proceeds by defining a sphere in which the interatomic distances and angles (r, r' and θ) are calculated for a selected probe atom. Each contact that satisfies the selection criteria (e.g., $r = r'$) is counted and the coordinates of contributing atoms stored for later analysis. These selection criteria can be altered to generate alternative slices of the PADF. This process is repeated for each atom in the asymmetric unit. Finally, the list of contributing contacts is binned to create model PADF plots and a series of corrections are applied to account for features of the experimental PADF plots. These corrections include subtraction of the angular averages and correcting for geometric terms by dividing by $\sin\theta$. In our model calculations, we ignore solvent molecules that are not bound to the structure, i.e., away from the protein surface as they are expected to produce a uniform background to the PADF plots. Tightly bound water molecules, cofactors, coenzymes, etc. will contribute to the model PADF; however, they represent only a small portion of the atoms in any given protein structure (~5% in this study) making the PADF relatively insensitive to their structure. Modelling water molecules near the protein surface that are partially ordered would require additional modelling to interpret [41], which is beyond the scope of this study. Calculating correlations from atomic models helps with interpreting experimental PADFs and also allows us to identify "fingerprint" features related to the structures of biomolecules over extensive length scales.

Currently the PADF of structural motifs in biomolecules is not known. We have used existing experimental data to construct the set of fingerprints features for biomolecules. We take advantage of the huge range of structures deposited in the PDB including data derived from both diffraction and NMR studies. The resulting model PADF contains information on every n-body contact allowing us to extract contributions from particular parts of the substructure, such as a helices, a particular residues, or even atoms.

To explore how increasing degrees of structure effects the PADF plots we have calculated correlations for a series of PDB entries: an intrinsically disordered polypeptide in solution (2M5L), a bundle of three α-helices (1COS [42]), and a >100 kDa transcription protein from the bacterium *Sinorhizobium fredii* (4OMZ) [43]. Each of these structures is described in detail in the following sections.

3.1.1. Disordered Polypeptide

The most fundamental level of structure in any protein is created by the polymeric interatomic bonding giving rise to the primary structure. The contributions to the PADF from this fundamental structure is examined here as it will form the basis for further contributions from higher level structural features.

One intuitive way to visualise features in the PADF is to plot the $\Theta(r = r', \theta)$ surface which can be thought of as encoding angular information (θ) about pairs of atoms equidistant ($r = r'$) from some sample atom. We show the $\Theta(r = r', \theta)$ surface for a disordered polypeptide (PDB entry 2M5L) in Figure 2. On short length scales commensurate with intramolecular bonds (1–2 Å), we observe sharp peaks due to the strong geometric constraints imposed by covalent bonding (Figure 2a). At slightly longer distances (~2.5 Å), the angular dependence becomes weaker with multiple peaks appearing over a range of angles. This reflects the next-nearest neighbour distances, which can adopt a wider range of angles due to the inherent flexibility of the polypeptide chain. At this scale, the distance component is still strongly reflective of intramolecular bonding.

Moving even further in r we observe the beginning of the dominant feature of these PADF plots – "arcs" of correlation extending from approximately 3 Å out in r to beyond 20 Å. These features can be explained through the simple geometric models shown in the inset of Figure 2b, considering a set of atoms a, b and c. Atoms b and c are some interatomic distance, r_{bc}, apart, a contribution to the $r = r'$ slice of the PADF will occur when there is some probe atom a positioned perpendicular to

interatomic distance r_{bc}. This results in a simple geometric relation between this perpendicular probe distance (p), characteristic distance $|bc|$, and angle, θ:

$$p(\theta) = \frac{|r_{bc}|}{2\sin(\frac{\theta}{2})} \,. \tag{4}$$

Correlation arcs therefore provide an indication of the most regularly occurring interatomic distances in a structure. For the polypeptide 2M5L, we observe two arcs with distances 1.3 Å and 2.2 Å corresponding to the two strongest peaks in the pair distribution function (PDF) shown in Figure 2c, indicating they are the two most common interatomic distances in the polypeptide. The scheme shown in the inset of Figure 2b demonstrates additional geometric properties of these arcs: the value of $|r_{bc}|$ is found at $\theta = 60^\circ = 120^\circ$ corresponding to an equilateral triangle, on the condition that there is a suitable probe atom a. Figure 2d depicts the atomic structures and the large amount of structural variation in the ensemble, which suppresses angular structure in the PADF aside from the arcs. We have shown that even in the absence of any ordering beyond covalent bonding, the PADF contains clear angular features which can be linked back to molecular geometry and which will occur in any system with covalent bonding.

Our study of the disordered polypeptide is included solely for the theoretical insight it provides about atomic correlations in the primary structure of the peptide chain. We expect the arcs in the PADF to be observed for other polymers or disordered protein systems, as they originate from fixed bond distances and second neighbour distances in the monomer correlated with a third probe atom at an arbitrary distance. PADF analysis may have future applications to study the folding of disordered proteins to complement recent progress in measuring the confirmations of disordered proteins in solution [44]. However, correlation experiments on disordered proteins would be very low signal to noise and, like X-ray imaging of single proteins, would require the brightest XFEL facilities. Protein crystals produce higher signal-to-noise data and are currently feasible to measure. Therefore, for the remainder of our study we restrict our focus to proteins that can form crystals.

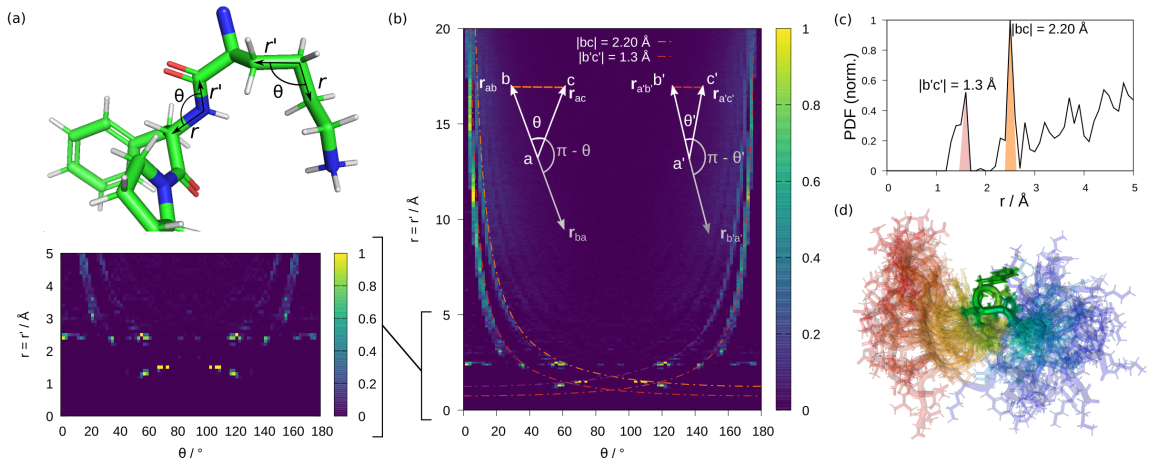

Figure 2. Model PADF plot for a disordered protein in solution. (**a**) At short distances, the PADF plot contains sharp peaks due to the strong geometric constraints imposed by covalent bonding. Examples of contributing three-body-correlations are shown on a section of the polypeptide chain. (**b**) The ensemble average over all stable protein conformations results in a relatively featureless PADF plot at larger $r = r'$ with the exception of the narrow "arcs" tending towards $\theta = 0$ and 180°. (**c**) Atomic pair distribution function for 2M5L with highlights indicating characteristic distances 1.3 Å and 2.2 Å. (**d**) Structures of 2M5L overlapping with the persistent β-turn shown in green.

3.1.2. Alpha Helices

Alpha helices are the most abundant secondary structure element in proteins and their rapid assembly is the basis for folding more complex structures [45–47]. To test the sensitivity of XCCA to the formation and arrangement of alpha helices we have calculated PADF plots using a model system consisting of a bundle of three parallel helices as presented in the PDB entry 1COS [42]. To fully separate the effects of packing helices together we have first calculated the PADF for just one of the helices in this bundle (Figure 3a), then subsequently the bundle of three together (Figure 3b). Finally, we compare the quantity of structural information provided by PADF analysis versus one-dimensional PDFs for each of these structures.

As in the case of the disordered polypeptide, the most immediate features in both plots are the arcs due to the intramolecular bonds with characteristic distances of 1.3 and 2.2 Å. Compared to 2M5L, the single alpha helix (Figure 3a) shows a minor increase in diffuse angular structure which is most prominent in the r range of approximately 5–11 Å which can be tied to the formation of the helix. Figure 3c shows the sums of the PADF across this r range for 2M5L and the single helix, illustrating the additional correlation intensity from $20 < \theta < 60^{\circ}$.

Next, we compare the result of packing three alpha helices together in a bundle, a simple secondary structure feature observed in proteins. Figure 3b shows the model PADF plot for the three-helix bundle. We show lines due to characteristic distances on half of the PADF plot for clarity, and in this case only over the angular range where appreciable correlation intensity can be observed. In addition to the intramolecular bonds at 1.3 Å and 2.2 Å, we observe a strong arc of intensity that we can link to a characteristic distance of 5 Å and limited to angular regions $\theta < 40^{\circ}$, $\theta > 140^{\circ}$. This characteristic distance is the approximate diameter of an alpha helix. The angular range is a result of the fact that the probe atom must itself lie within one of the alpha helices at a distance $r = r'$. The second prominent feature due to alpha-helix packing is a very broad arc appearing at $r = r' > \approx 11$ Å and limited to $\theta < 60^{\circ}$. We can determine approximate upper and lower limits for contributing characteristic distances of approximately 9 to 16 Å (Figure 3b). These distances correspond to the inner and outer distances between neighbouring packed helices, i.e., when "probed" by an atom in the third helix. Although this analysis provides less structural information on the geometry of the secondary structure when compared to crystallography, the advantages become clear in the case of non-crystalline systems where structural information is typically extracted from PDF analysis.

In Figure 3d, we present calculated atomic PDFs for the three systems discussed so far: disordered polypeptide (PDB code: 2M5L – averaged over all 25 configurations), single helix (PDB code: 1COS – blue) and three helix bundle (PDB code: 1COS – red). Each pattern is normalised to the strongest peak at 1.3 Å. Comparing these PDFs, we see that the packing of helices leads to some subtle changes in the PDF at $r > 5$ Å. However, due to the reduction of this data to one dimension, there is little meaningful structural information that can be extracted.

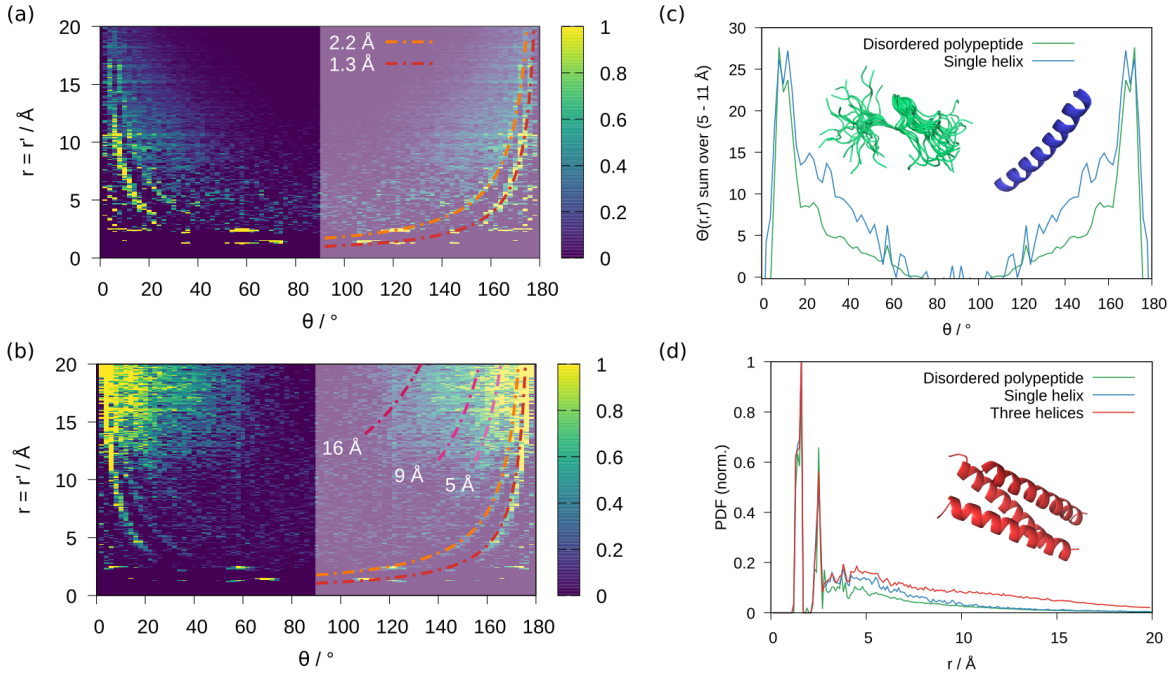

Figure 3. Model PADF plot for α helices. Plots are shown for (**a**) a single helix 40 Å long and (**b**) a bundle of three such helices. (**c**) Comparative sections through the PADF $\Theta(r, r', \theta)$ at $r = r' = 5$–11 Å for disordered polypeptide (green) and single alpha helix (blue) (**d**) Normalised atomic pair distribution functions for 2M5L (green), helix A in 1COS (blue) and all three helices in 1COS (red) as shown below. Cartoon depictions of each of the three models are shown as insets.

3.2. q-Space Correlations from the Structure Factors

As described in Section 2, PADF volumes of molecular systems can be calculated from X-ray diffraction data. This is how the PADF could be measured experimentally when an atomic model is not available. In this approach, the q-space correlation volumes of the system is calculated (see Equation (2)), after which we perform a series of transformations based on spherical harmonics and spherical Bessel functions to obtain the real space PADF volume. This process is outlined in previous work [37].

Experimentally, the q-space correlation function can be calculated using Bragg peaks detected in 2D diffraction data. Statistical convergence requires the correlations to be averaged over an ensemble of diffraction patterns measured from randomly oriented crystals, such as those measured in serial crystallography experiments. However, here we take a more direct route by using 3D structure factor data from previous crystallography experiments, i.e., after the data has been indexed and merged. To calculate the q-space correlation, $C(q, q', \theta)$, the structure factor CIFs for a single alpha helix protein (PDB entry 1AL1) and bundle of three alpha helices proteins (PDB entry 1COS) were downloaded from the PDB. The files were parsed for crystal cell dimensions a, b, c, crystal cell angles α, β, γ, Bragg reflections h, k, l, scattering factors F_{hkl} and space group symmetry. The scattering factors were squared to produce the scattering intensities I_{hkl}, and the space group symmetry of the crystal was used to calculate the symmetric Bragg reflections. For each Bragg reflection, the corresponding scattering vector $\mathbf{q}$ was found from calculating reciprocal lattice vector lengths $\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*$:

$$\mathbf{q} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* \,. \tag{5}$$

Once a list of scattering vectors had been calculated, a q-space correlation volume was created per Equation (2). The correlation volume consisted of a 3D matrix where we store the correlation intensity, and has three axes for each input of the correlation function $C(q, q', \theta)$. For each pair of scattering

vectors **q** and **q**′, we calculated the angle θ between them and their magnitudes q, q' to determine their position in the correlation volume. After determining the position within the correlation volume, the value in that position was added by the multiplication of the **q** and **q**′ scattering intensities. To save computation time, a maximum **q** magnitude (q_{max}) was implemented such that any scattering vector **q** with magnitude greater than q_{max} would not be correlated. For the structure factors of the single helix protein (1AL1) and the three helix bundle (1COS), we selected a q_{max} of 0.37 Å^{-1}, which was the magnitude of the largest scattering vector of the 1AL1 vectors. Although this reduced the number of correlated vectors in the 1COS case, this meant that the correlation volumes of each protein can be compared at the same resolution. The total number of correlated scattering vectors for 1COS and 1AL1 was 13,984 and 20,880, respectively. In terms of computational complexity, calculation time is not drastically dependant on the complexity of the protein, but on the number of the scattering vectors required to accurately reconstruct the protein unit cell. The correlation calculation goes as $O(n^2)$, where n is the number of scattering vectors. This is illustrated in the case of selected proteins, where although protein 1AL1 is a single turn alpha helix and it is more computationally intensive then 1COS, a multiple turn alpha helix, because it has more scattering vectors within the q_{max} range. The correlation volumes were 256 bins in q axis directions, and 360 bins in the θ axis directions, providing 0.5° angular resolution and 0.144 $\times 10^{-3}$Å^{-1} q-space resolution.

In Figure 4a,b, we see two maps pertaining to the q-space correlation volumes of proteins 1AL1 and 1COS, respectively. The maps are taken from the $q = q'$ plane of the correlation volume, which represents correlations between vectors with the same magnitude (e.g., within the same scattering ring). Above and to the right of these maps are line plots illustrating the average intensity through each axis in the map. These plots demonstrate distinct banding though the q axis, with notably low correlation between $q = 0.21$ Å^{-1} and $q = 0.23$ Å^{-1} in plot (a) and between $q = 0.15$ Å^{-1} and $q = 0.19$ Å^{-1} in (b). We also observe symmetry about the 90° correlation angle axis, due to Friedel symmetry. Note that for every Bragg reflection (h,k,l), there is an equivalent Friedel pair ($-h$,$-k$,$-l$) with approximately the same intensity. If one scattering vector q correlated with another vector q' subtends an angle θ, then the Friedel pair of the correlating vector q would subtend an angle of $180° - \theta$ with the vector q'.

Banding in the correlation length axis is due to the discrete number of allowed correlation magnitudes. Bragg reflections can only fall on integer indices, and are hence discrete. The scattering vectors **q** are calculated from the Bragg reflection indices and, hence, are also discrete. With higher indexed reflections, more allowed scattering vectors would be observed and show less banding. This banding is observed in the volatility in mean intensity though the q axis.

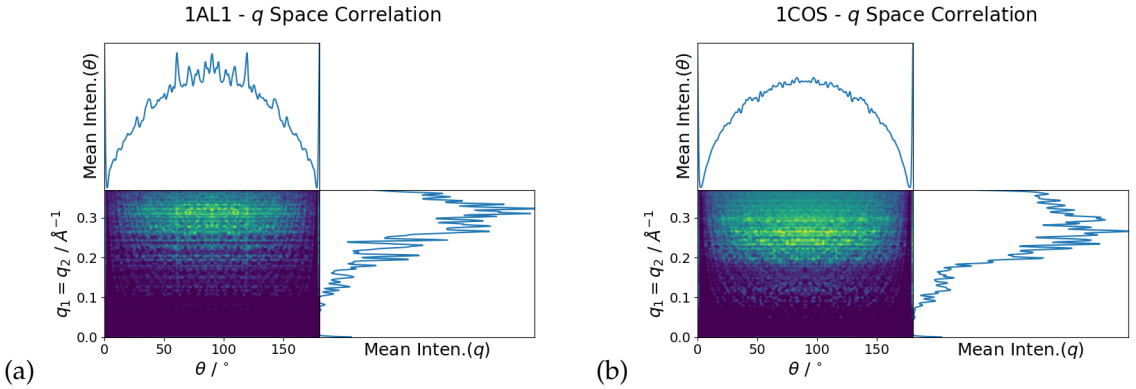

Figure 4. q-space correlations of (**a**) single helix protein (1AL1) and (**b**) three-helix bundle (1COS), generated from structure factors. The plot intensity has been scaled to the power of 0.25, and convolved with a Gaussian blur, with full width at half maximum (FWHM) of 6.8 × 10^{-3} Å^{-1} by 2.3°. Scaling allows highlighting of the lower intensity peaks, and the blur allows the small peak areas to become more visible. All of the intensity values are then scaled between 1 and 0 for the maximum and minimum values. Line plots above and right of the images show average intensity through each axis.

Similar to the atomic model PADF correlations in the previous section, we see arcs in the q-space correlations in Figure 4 that are steep at angles in the range of 0° to 40°, and 140° to 180°, and flatten towards the centre of the plots, outside of these ranges. Each arc occurs due to the correlations of pairs of scattering vectors that are equidistant apart on concentric diffraction rings. This is illustrated in Figure 5. Here, we show two concentric diffraction rings, and for each ring, two scattering vectors $\mathbf{q}$ and $\mathbf{q}'$ that are to be correlated together. The chord distance, dashed line, between $\mathbf{q}$ and $\mathbf{q}'$ is the same for the two scattering rings. By increasing the radius of the diffraction ring (increasing q), to maintain the same chord distance, the angle subtended by the scattering vectors must decrease. Plotting the radius q as a function of subtended angle produces the arc features observed in the q-space correlation plots, where each peak along the arc corresponds to a diffraction ring that has the chord distance associated with the arc. Geometrically, the arc follows a fit expression of

$$q(\theta) = \frac{c}{2\sin(\frac{\theta}{2})} \tag{6}$$

where c is the characteristic chord length. Note that this geometry also explains arcs found in the PADF plots generated from atomic models in the previous section.

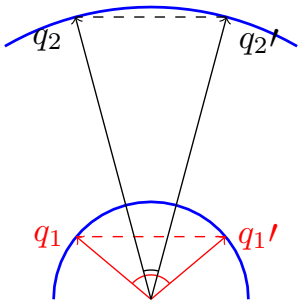

Figure 5. Equidistant chords on concentric scattering rings form the arcs in the q-space correlation plot.

3.3. PADF Calculations Converted from the Q-Space Correlations Functions

Although Figure 4 demonstrates that the q-space intensity plots contain a significant amount of information, it is difficult to relate this information back to the molecular structure. However, as previously stated, we can produce PADF plots from q-space correlation volumes [37]. Using this method, we will compare the q-space correlation volume generated PADFs to PADFs predicted from atomic models. Once the PADF was created from the q-space volume, we extracted the $r = r'$ plane to generate the plots shown in Figure 6a,b. We also extracted line plots though θ at various r distances, to further illustrate the differences between the PADF intensities. The line plots for 1AL1 and 1COS are shown in Figure 6c,d respectively. From the line plots, we see that the intensity profile is comparable for $r = 3$ Å, but diverge with increasing r.

Within these plots, we see arcs that are symmetric about the $\theta = 90°$ axis, as observed in the previous section. For a range between 0 and 10 Å, we see that the arcs closely follow the expected fits for characteristic chord lengths of 2.2 Å and 6 Å distances. This corresponds well to the characteristic distances found in the previous section of 2.2 Å and 5 Å.

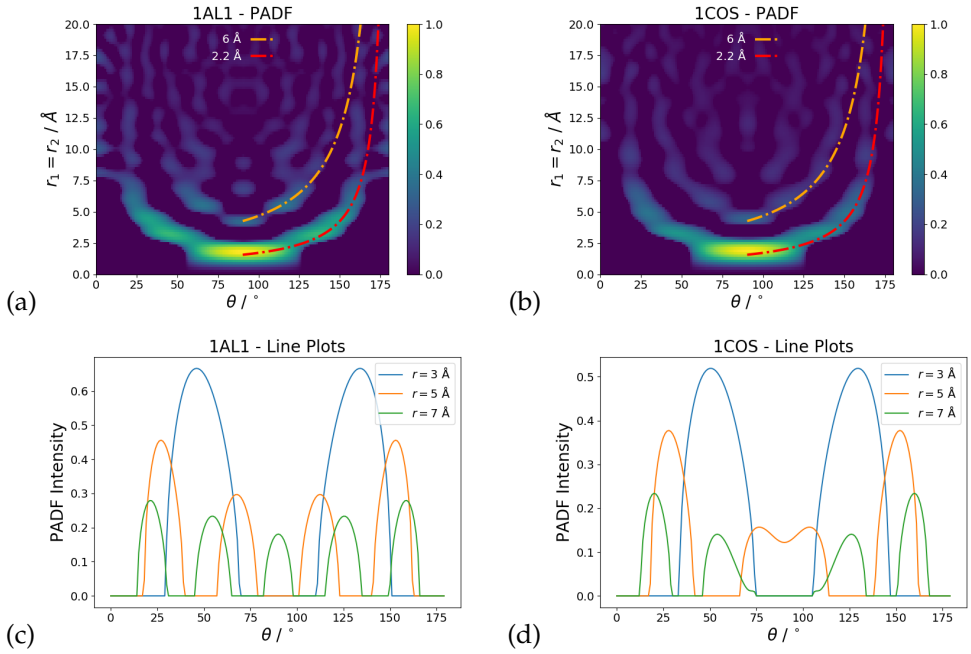

Figure 6. PADF correlation of (**a**) a single helix protein (1AL1) and (**b**) a three-helix bundle (1COS), generated from the q-space correlation function $C(q, q', \theta)$ (see Equation (2)). The intensity within these plots is clipped to show only positive correlation, scaled to the power of 0.25 to flatten peak intensity. This qualitatively shows regions of correlation, where we might expect to see particular arrangements of atoms, but removes the relative intensity of finding such arrangements. Plots (**c**,**d**) show PADF intensity line plots for selected distances as a function of θ for 1AL1 and 1COS respectively.

Comparison between the atomic model PADF and q-space correlation PADF is illustrated in Figure 7. Figure 7a outlines selected arcs (dashed lines) and peaks (circle markers) characteristic of alpha helices found in the atomic model PADF. These peaks are more visible in Figure 7b, which utilises a Gaussian blur and an intensity threshold to highlight peaks in the 40°–90° range. The Gaussian blur had a full width of half maximum of 12° and 2.2 Å. These selected peaks do not occur in the disordered peptide chain, a single helix or a isolated three-helix bundle. They are due to the symmetric packing of the asymmetric unit into the unit cell, which for the single alpha helix protein 1AL1 has 48 symmetric partners. The selected peaks that occur in the atomic model PADF correspond to peaks in Figure 7c, which is the PADF generated from the q-space correlation. Hence, the PADF generated from experimental data is sensitive to the packing of the molecules into the unit cell.

This is further illustrated in Figure 7d, where three symmetric proteins form an equilateral triangle between equivalent residues on the proteins. We can observe correlations within this arrangement in the $r = r'$ plane, by viewing the intensity on a vertical line through 60°. Correlation between the alpha carbons of the residue LEU3 yields lengths of 10.1 Å and 60°, which appears within range of the red marker in Figure 7a–c. Furthermore, correlation with the beta and gamma carbons on the LEU3 residue (not shown in the cartoon representation) gives distances of 5.8 Å and 8 Å, respectively, which are also close to peaks in the PADF plots. Note that the PADF volume considers the correlations of all atoms with all other atoms, and thus peaks can also be found when correlating between non-equivalent residues. These peaks, however, would not necessarily occur in the $r = r'$ plane and, hence, require some other representation of the volume.

Conversely, not all equivalently correlated residues will form peaks at 60°. Regions of anti-correlation, where it would be considered unlikely to find a particular arrangement, can change the relative peak intensity and remove a peak altogether.

For distances $r > 12$ Å, the PADF plots calculated from experimental structure factors deviate from expected distributions from the atomic models. The experimental PADF plots show an angular

modulation, where the atomic models show sharp arcs and a diffuse structure. This is likely due to the finite angular bandwidth used in the experimental PADF, producing aliasing artefacts. Although the information is still related to the sample structure, it is no longer possible to manually identify molecular structure with specific peaks in this region. Instead, computational modelling will be required to relate molecular structure to aliased PADF data.

The aliasing is a consequence of the angular resolution of the reconstruction, which is determined by a number of factors. The number of spherical harmonic terms used to reconstruct the PADF influences the angular sampling, for example, by calculating to the 58th spherical harmonic, we expect 58 oscillations in intensity at high correlation length over 360°. This comes to approximately 6.2° angular resolution. In the plots shown in Figure 6, the correlation is run to 60 Å, but cropped to 13 Å. Other factors that influence the aliasing include the maximum real space distance computed and the binning of the distance measurement, as well as the number and maximum magnitude of scattering vectors in the q-space correlation volume input, and the binning of the q-space correlation. This could potentially limit the length scales that can be investigated by this method, however, further analysis is required to understand the interplay of these variables, and the accuracy of the PADFs at variable limits.

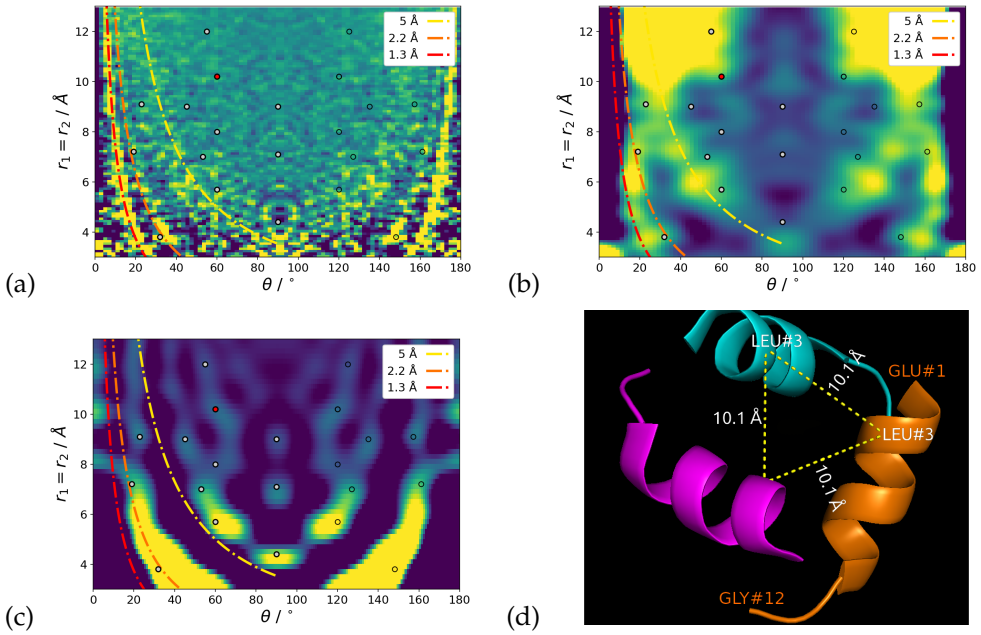

Figure 7. Comparison between PADFs of a single-helix protein (1AL1) generated via different methods. (**a**) PADF generated from atomic model, no intensity scaling. (**b**) PADF generated from atomic model, enhancing peaks between 40° and 90°. (**c**) PADF generated from q-space correlation, showing corresponding peak positions. (**d**) Atomic models of three symmetric 1AL1 proteins in the crystal, showing one of the possible contributions on the peak intensity at approximately 60°, 10 Å.

3.4. A Test Case: NoIR from Sinorhizobium fredii

Finally, we discuss a test case of the protein NoIR from *Sinorhizobium fredii* (PDB entry 4OMZ). This protein was chosen on the basis of molecular weight (>100 kDa) with a very high alpha helix propensity and few other secondary structural features (59% alpha helix, 7% beta sheet). Furthermore, structure factor files are available for direct comparison between model- and structure-factor-derived PADFs. In Figure 8a, we see the characteristic band at 5 Å indicative of folded alpha helices as well as the broad arc feature beginning at $r = r' = 14$ Å, first observed in the three-helix bundle (1COS) and likely due to inter-helix correlations (see Figure 8b). The PADF was blurred with a Gaussian kernel to lower the angular resolution and facilitate comparison to the PADF generated from the structure factors. At lower resolution, the diffuse structure is maintained and not all arcs are clearly separated.

To generate the structure factor PADF of NoIR protein, we created the q-space correlation volume as previously described, with a q_{max} of 0.1849 Å^{-1}. We reduced the q_{max} value by half compared to correlations of single helix protein and the three helix bundle, due to computation time constraints. The total number of correlated scattering vectors was 13,200.

The PADF generated from the structure factors of 4OMZ is shown in Figure 8d. Here, we reconstructed the PADF to a maximum distance of 120 Å, but cropped to 20 Å as in the previous cases for single helix and the three helix bundle. We doubled the maximum reconstruction distance because we halved the q_{max} value of the q-space correlation. This maintains the resolution and pixel size for comparison to the reconstructions of the single helix and the three helix bundle, due to the inverse relationship between scattering vector magnitude and resolution. Correlation arcs are visible near the characteristic distances of 5 Å and 9.5 Å. The angular PADF structure at angles between 60° and 120° is less textured than for the three helix bundle (1COS), qualitatively consistent with the diffuse angular structure of Figures 8a,c. However, the peaks at approximately 90° do not correspond to peaks predicted by the atomic model, suggesting the influence of finite angular and radial resolution. Similar to the 1AL1 and 1COS reconstructions, we calculated the PADF to the 58th spherical harmonic, which corresponds to approximately a 6.8° angular resolution.

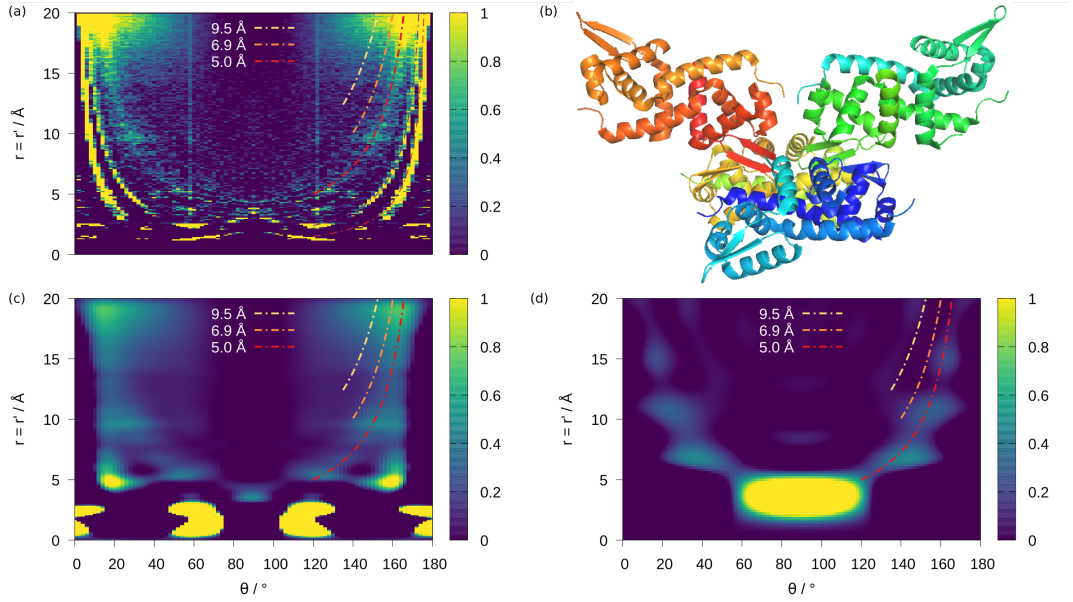

Figure 8. Comparison between PADFs of protein NolR from *Sinorhizobium fredii* (PDB code: 4OMZ) generated via different methods. (**a**) PADF generated from atomic model, no intensity scaling; (**b**) the ribbon diagram of the protein structure; (**c**) PADF generated from atomic model with lower angular resolution for comparison to PADF generated from the structure factors; and (**d**) PADF generated from the q-space correlation of the structure factors.

4. Discussion

From the PADF plots of model structures, we have identified several angular correlation "fingerprints" that can be uniquely associated to structural motifs in proteins. Calculations from atomic models provide the ideal form of these fingerprints, unaffected by the resolution limitations and errors of a real experiment. From the atomic models, we have shown that the PADF plots of disordered polypeptide chains show clear bonding peaks at high resolution and two clear angular arcs. We note that bond-angle peaks are an average over all bonds in the structure and are unlikely to be sensitive to chemical changes in small localised regions of the chain. The arcs are well explained by the three-body correlations where one pair distance is fixed to a bond distance (≈1.3 Å) or a second neighbour distance (≈2.2 Å). Varying the distance to a third atom on the chain, generates the almost continuous arcs. Therefore, the PADF "fingerprint" of a completely disordered chain is to see these arcs

in the absence of other angular structure. We then saw that a single alpha helix structure displays extra diffuse arcs associated with the inner and outer diameter of the helix. This diffuse angular structure becomes richer when three helices are formed into bundle. The packing of many alpha helices into a unit cell of high symmetry generates complex texture of peaks in the diffuse angular regions of the PADF. It is clear from these correlation "fingerprints" that PADF plots can contain a significant amount of structural information that is not available in powder diffraction or SAXS plots, which lack angular information.

Although there are a number of secondary structural motifs, we have focused on alpha helices in our study. Alpha helices are the most prevalent and easily predictable, and can be found in a wide range of protein types. Unlike more flexible motifs (e.g., beta sheets), they display a highly constrained geometry which produces strong angular correlations. This is most clearly demonstrated by the far smaller dihedral-angle domains for alpha helices versus beta sheets in Ramachandran plots. Alpha helices also frequently assemble into regular bundles creating well-defined correlations as explored above. The approximately flat structure of beta-sheets will result in correlations with a weak or non-existent angular dependency, making their detection by "fingerprints" of angular peak positions challenging.

Calculation of the model PADF can be computationally expensive. In the current implementation, the time taken to construct the model PADF is dependent upon the number of atoms in the asymmetric unit and the number of atoms within the probe sphere. This second term introduces a r^3 dependence on probe radius as the number density of non-H atoms in the proteins we have studied is approximately constant at 0.03 Å^{-3}. Currently, the calculation of a model PADF to $r_{max} = 20$ Å for 1AL1 consisting of 100 atoms in the asymmetric unit takes approximately 900 s, extending up to 27,000 s for an asymmetric unit of 709 atoms on a standard desktop computer.

Improving the calculation speed is an important step for improving the analytical use of this technique in a similar fashion to the importance of computational power in the rise of crystallography. Most importantly, the model PADF is a statistical function and can be sampled by Monte Carlo methods for systems, such as proteins with hundreds to thousands of atoms in the asymmetric unit. Currently, we perform a simple convergence test after each cycle, comparing the model PADF for cycle n and $n-1$ and calculating the cosine similarity. Depending on the aim of the calculation the function can converge within tens of cycles. This is especially true in cases where the PADF will have experimental resolution limits applied, see, e.g., Figure 8. In a more practical point underlying routines could be implemented in a faster compiled language such as c/c++, instead of python, and take advantage of additional parallelisation.

The calculation of model PADF plots is not inherently limited to atomistic models. While this approach takes an individual atom in the asymmetric unit as the probe, the scattering density can instead be represented by an arbitrary electron density. This density can then be probed at random and correlations within a probe radius calculated, again most likely through Monte Carlo methods. This approach could be especially useful for modelling structures containing intrinsically disordered domains or materials such as liquid crystals.

Beyond theory, it is important to understand how much of the PADF correlation information is accessible in the experiment. Therefore, we generated PADF plots from the structure factors of crystals of our model structures. We saw clearly that the bond angle PADF peaks were not resolved at typical resolutions of protein crystallography >2 Å. The characteristic angular arcs of the polypeptide chain and individual helices are clearly identifiable up to around ≈12 Å. At these small distance scales, we also observe significant complex angular peak structure from the geometric packing of the α-helices into the unit cell. It is highly interesting that this angular structure was a small modulation of the diffuse correlations from the atomic models, but appears clearly in the analysis of experimental data. This is due to the radial and angular scaling of the three-body correlations that defined in the PADF, producing an altered contrast from the atomic calculations. In this case, the PADF contrast is beneficial for visual identification of the angular "fingerprints".

Beyond 12 Å, the finite angular resolution used in the analysis of experimental data causes a significant modulation of the angular structure. This modulation changes the positions of angular peaks and prevents the identification of atomic structures that cause each angular peak. Consequently, the angular peak positions at larger distances do not correspond as well to specific three-body combinations in the atomic model. However, as this effect is not due to errors, but to finite angular resolution, the PADF at larger distances is useful for structural modelling or fitting.

The finite resolution of experimental data places limits on the visual identification of more protein complex structures using angular peak positions. Our test case, NoIR from *Sinorhizobium fredii*, highlights the challenge of relating PADF peak positions to local atomic arrangements in complex protein structures. Qualitatively, the PADF plots are too diffuse to visually identify "fingerprints" for complex protein structures. We expect this to hold true for proteins that have fewer alpha helices and more diverse composition of secondary structure elements than NoIR. Computational structure refinement algorithms would need to be developed to quantitatively match protein structure to the continuous PADF distribution. Based on our results, we expect the potential sensitivity of computational refinement methods from PADF plots to be greater than powder diffraction or SAXS, but may be less than conventional crystallography.

To compute the PADF from 3D datasets of experimental structure factors, as we have studied here, the experimental requirements are exactly the same as standard crystallography or serial crystallography experiments. The most data-intensive experiment is serial crystallography experiments with XFELs, where at least 10^4 indexable patterns are typically required for a complete 3D dataset at sufficient resolution for structure determination [48]. Due to inefficiency in data collection and processing, this may require of the order of 10^6 X-ray pulses, which takes a few hours of data collection with a pulse repetition rate ~100 Hz. PADF analysis could be applied to such datasets with similar results to those presented in our study.

PADF analysis can be applied directly to Bragg peaks selected from the 2D diffraction patterns without determining crystal orientations. This may introduce new sources of noise that are suppressed when the 2D data is merged into a 3D dataset first. Intensity fluctuations due to peak partiality, incident beam intensity and crystal size may be detected when PADF analysis is applied directly to Bragg peaks selected from the 2D diffraction patterns. This in turn may increase the number of diffraction patters required. However, analysing 2D patterns also permits more crystals to be measured per exposure. Further study is required to quantify the data requirements for analysing 2D diffraction patterns.

The data used in this study was derived from single-crystal protein crystallography experiments sourced from the PDB. One can envision applying the PADF technique to serial diffraction data from ensembles of (nano)crystals. The upper limit on the number of crystals PADF analysis can tolerate is reached when isotropic 2D powder rings form on the detector. Textured, anisotropic powder rings still contain the angular information required to recover the PADF, even if individual Bragg reflections can no longer be indexed. If individual Bragg reflections cannot be isolated, then the entire detector could be analysed as is performed in XCCA measurements of single particles and bulk materials. This could then lead to innovations in experimental design in, for example, time-resolved measurements or crystal growth studies. However, further study is required to understand how crystal size, the number of crystals per exposure and background scattering impact the quality of the PADF correlations, before experimental innovations can be attempted.

5. Conclusions

We have applied PADF analysis to crystallography data to determine how protein structure translates into multi-atom angular correlations. We have found that atomic bonding, polypeptide chains, alpha helices and the bundling of alpha helices all have distinctive angular correlations (or "fingerprints"). From experimental data, the angular arcs generated by the polypeptide chains and alpha helices were clearly identifiable, as were correlations from the packing of symmetry equivalent molecules in the unit cell. There is clearly more structural information in the angular correlations

to identify these structures than in the pair-distribution function. However, the finite angular and radial resolution causes significant impact on the PADF distribution calculated from diffraction data. This suggests that the full potential of protein PADF analysis will be realised in combination with structural modelling.

Author Contributions: Conceptualisation, A.V.M., P.A., J.B. and T.L.G.; methodology, P.A., J.B. and A.V.M.; software, P.A., J.B. and A.V.M.; investigation, P.A. and J.B.; writing—original draft preparation, P.A., J.B. and A.V.M.; writing—review and editing, A.V.M., P.A., J.B. and T.L.G.; supervision, A.V.M., J.B. and T.L.G.; project administration, A.V.M. and T.L.G.; funding acquisition, A.V.M. and T.L.G. All authors have read and agreed to the published version of the manuscript.

Funding: A.V.M. and T.L.G. acknowledge the funding support from the Australian Research Council Discovery Project grant (DP190103027).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results'.

Abbreviations

The following abbreviations are used in this manuscript.

PADF	Pair-angle distribution function
PDF	Pair-distribution function
PDB	Protein Data Bank
XCCA	X-ray cross-correlation analysis
SAXS	Small-angle X-ray scattering
XFEL	X-ray free-electron laser
cryo-EM	Cryo-electron microscopy
NMR	Nuclear magnetic resonance

Appendix A. Definitions of the Modified N-Body Correlation Functions

The modified correlation functions are given by

$$\tilde{g}^{(2)}(r,r',\theta) = \int \tilde{g}^{(2)}(\mathbf{r},\mathbf{r}')\delta\left(\cos\theta - \frac{\mathbf{r}\cdot\mathbf{r}'}{|\mathbf{r}||\mathbf{r}'|}\right) d\Omega d\Omega' \,, \tag{A1}$$

$$\tilde{g}^{(3)}(r,r',\theta) = \int g^{(3)}(\mathbf{r}_1,\mathbf{r},\mathbf{r}')d\Omega d\phi' d\mathbf{r}_1 \tag{A2}$$

and

$$\tilde{g}^{(4)}(r,r',\theta) = \int g^{(4)}(\mathbf{r}_1,\mathbf{r},\mathbf{r}_3,\mathbf{r}')d\Omega d\phi' d\mathbf{r}_1 d\mathbf{r}_3 \,, \tag{A3}$$

where $\mathbf{r}_1$ and $\mathbf{r}_3$ are coordinates of a reference atom in each pair that is integrate out, and Ω and ϕ' are angular coordinates that specify the absolute orientation of the three- or four-atom group, respectively. We have also defined:

$$\tilde{g}^{(2)}(\mathbf{r},\mathbf{r}') = \frac{1}{\rho^3 V}\langle \sum_{i_1=1}^{N}\sum_{i_2\neq i_1}\Big[\delta(\mathbf{r}-\mathbf{r}_{i_{12}})\delta(\mathbf{r}'-\mathbf{r}_{i_{12}}) + \delta(\mathbf{r}-\mathbf{r}_{i_{12}})\delta(\mathbf{r}'+\mathbf{r}_{i_{12}})\Big]\rangle \,. \tag{A4}$$

References

1. Lappano, R.; Maggiolini, M. G protein-coupled receptors: Novel targets for drug discovery in cancer. *Nat. Rev. Drug Discov.* **2011**, *10*, 47–60. [CrossRef] [PubMed]
2. Cournia, Z.; Allen, T.W.; Andricioaei, I.; Antonny, B.; Baum, D.; Brannigan, G.; Buchete, N.V.; Deckman, J.T.; Delemotte, L.; del Val, C.; et al. Membrane protein structure, function, and dynamics: A perspective from experiments and theory. *J. Membr. Biol.* **2015**, *248*, 611–640. [CrossRef] [PubMed]
3. Nikaido, H. Molecular basis of bacterial outer membrane permeability revisited. *Microbiol. Mol. Biol. Rev.* **2003**, *67*, 593–656. [CrossRef] [PubMed]
4. Sharom, F.J. ABC multidrug transporters: Structure, function and role in chemoresistance. *Pharmacogenomics* **2008**, *9*, 105–127. [CrossRef]
5. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; et al. The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58*, 899–907. [CrossRef]
6. Kirkwood, J.; Hargreaves, D.; O'Keefe, S.; Wilson, J. Analysis of crystallization data in the Protein Data Bank. *Acta Crystallogr. Sect. F* **2015**, *71*, 1228–1234. [CrossRef]
7. Beale, J.H. Macromolecular X-ray crystallography: Soon to be a road less travelled? *Acta Crystallogr. Sect. D Struct. Biol.* **2020**, *76*, 400–405. [CrossRef]
8. Duarte, J.M.; Srebniak, A.; Schärer, M.A.; Capitani, G. Protein interface classification by evolutionary analysis. *BMC Bioinform.* **2012**, *13*, 334. [CrossRef]
9. Baskaran, K.; Duarte, J.M.; Biyani, N.; Bliven, S.; Capitani, G. A PDB-wide, evolution-based assessment of protein-protein interfaces. *BMC Struct. Biol.* **2014**, *14*, 22. [CrossRef]
10. Wüthrich, K. The way to NMR structures of proteins. *Nat. Struct. Biol.* **2001**, *8*, 923–925. [CrossRef]
11. Ishima, R.; Torchia, D.A. Protein dynamics from NMR. *Nat. Struct. Biol.* **2000**, *7*, 740–743. [CrossRef] [PubMed]
12. Caffrey, M. A lipid's eye view of membrane protein crystallization in mesophases. *Curr. Opin. Struct. Biol.* **2000**, *10*, 486–497. [CrossRef]
13. Johansson, L.C.; Arnlund, D.; White, T.A.; Katona, G.; DePonte, D.P.; Weierstall, U.; Doak, R.B.; Shoeman, R.L.; Lomb, L.; Malmerberg, E.; et al. Lipidic phase membrane protein serial femtosecond crystallography. *Nat. Methods* **2012**, *9*, 263–265. [CrossRef] [PubMed]
14. Lundstrom, K. Structural genomics for membrane proteins. *Cell. Mol. Life Sci. CMLS* **2006**, *63*, 2597–2607. [CrossRef]
15. Carpenter, E.P.; Beis, K.; Cameron, A.D.; Iwata, S. Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.* **2008**, *18*, 581–586. [CrossRef]
16. Holton, J.M. A beginner's guide to radiation damage. *J. Synchrotron Radiat.* **2009**, *16*, 133–142. [CrossRef]
17. Svergun, D.I.; Koch, M.H.J. Small-angle scattering studies of biological macromolecules in solution. *Rep. Prog. Phys.* **2003**, *66*, 1735–1782. [CrossRef]
18. Levantino, M.; Yorke, B.A.; Monteiro, D.C.F.; Cammarata, M.; Pearson, A.R. Using synchrotrons and XFELs for time-resolved X-ray crystallography and solution scattering experiments on biomolecules. *Curr. Opin. Struct. Biol.* **2015**, *35*, 41–48. [CrossRef]
19. Chapman, H.N.; Caleman, C.; Timneanu, N. Diffraction before destruction. *Philos. Trans. R. Soc. B Biol. Sci.* **2014**, *369*, 20130313. [CrossRef]
20. Nogly, P.; James, D.; Wang, D.; White, T.A.; Zatsepin, N.; Shilova, A.; Nelson, G.; Liu, H.; Johansson, L.; Heymann, M.; et al. Lipidic cubic phase serial millisecond crystallography using synchrotron radiation. *IUCrJ* **2015**, *2*, 168–176. [CrossRef]
21. Botha, S.; Nass, K.; Barends, T.R.M.; Kabsch, W.; Latz, B.; Dworkowski, F.; Foucar, L.; Panepucci, E.; Wang, M.; Shoeman, R.L.; et al. Room-temperature serial crystallography at synchrotron X-ray sources using slowly flowing free-standing high-viscosity microstreams. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2015**, *71*, 387–397. [CrossRef] [PubMed]
22. Nogly, P.; Panneels, V.; Nelson, G.; Gati, C.; Kimura, T.; Milne, C.; Milathianaki, D.; Kubo, M.; Wu, W.; Conrad, C.; et al. Lipidic cubic phase injector is a viable crystal delivery system for time-resolved serial crystallography. *Nat. Commun.* **2016**, *7*, 12314. [CrossRef] [PubMed]

23. Schulz, E.C.; Mehrabi, P.; Müller-Werkmeister, H.M.; Tellkamp, F.; Jha, A.; Stuart, W.; Persch, E.; De Gasparo, R.; Diederich, F.; Pai, E.F.; et al. The hit-and-return system enables efficient time-resolved serial synchrotron crystallography. *Nat. Methods* **2018**, *15*, 901–904. [CrossRef] [PubMed]
24. Tenboer, J.; Basu, S.; Zatsepin, N.; Pande, K.; Milathianaki, D.; Frank, M.; Hunter, M.; Boutet, S.; Williams, G.J.; Koglin, J.E.; et al. Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science* **2014**, *346*, 1242–1246. [CrossRef]
25. Young, I.D.; Ibrahim, M.; Chatterjee, R.; Gul, S.; Fuller, F.D.; Koroidov, S.; Brewster, A.S.; Tran, R.; Alonso-Mori, R.; Kroll, T.; et al. Structure of photosystem II and substrate binding at room temperature. *Nature* **2016**, *540*, 453–457. [CrossRef]
26. Kam, Z. Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. *Macromolecules* **1977**, *10*, 927–934. [CrossRef]
27. Saldin, D.K.; Poon, H.C.; Shneerson, V.L.; Howells, M.; Chapman, H.N.; Kirian, R.A.; Schmidt, K.E.; Spence, J.C.H. Beyond small-angle X-ray scattering: Exploiting angular correlations. *Phys. Rev. B* **2010**, *81*, 174105. [CrossRef]
28. Kirian, R.A. Structure determination through correlated fluctuations in X-ray scattering. *J. Phys. B At. Mol. Opt. Phys.* **2012**, *45*, 223001. [CrossRef]
29. Wochner, P.; Gutt, C.; Autenrieth, T.; Demmer, T.; Bugaev, V.; Ortiz, A.D.; Duri, A.; Zontone, F.; Grübel, G.; Dosch, H. X-ray cross correlation analysis uncovers hidden local symmetries in disordered matter. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 11511–11514. [CrossRef]
30. Martin, A.V.; Bøjesen, E.D.; Petersen, T.C.; Hu, C.; Biggs, M.J.; Weyland, M.; Liu, A.C.Y. Detection of ring and adatom defects in activated disordered carbon via fluctuation nanobeam electron diffraction. *Small* **2020**, *16*, 2000828. [CrossRef]
31. Martin, A.V.; Kozlov, A.; Berntsen, P.; Roque, F.G.; Flueckiger, L.; Saha, S.; Greaves, T.L.; Conn, C.E.; Hawley, A.M.; Ryan, T.M.; et al. Fluctuation X-ray diffraction reveals three-dimensional nanostructure and disorder in self-assembled lipid phases. *Commun. Mater.* **2020**, *1*, 40. [CrossRef]
32. Kurta, R.P.; Donatelli, J.J.; Yoon, C.H.; Berntsen, P.; Bielecki, J.; Daurer, B.J.; Demirci, H.; Fromme, P.; Hantke, M.F.; Maia, F.R.; et al. Correlations in scattered X-ray laser pulses reveal nanoscale structural features of viruses. *Phys. Rev. Lett.* **2017**, *119*, 1–7. [CrossRef]
33. Lokteva, I.; Koof, M.; Walther, M.; Grübel, G.; Lehmkühler, F. Monitoring nanocrystal self-assembly in real time using in situ small-angle X-ray scattering. *Small* **2019**, *15*, 1–8. [CrossRef] [PubMed]
34. Zaluzhnyy, I.A.; Kurta, R.P.; Menushenkov, A.P.; Ostrovskii, B.I.; Vartanyants, I.A. Direct reconstruction of the two-dimensional pair distribution function in partially ordered systems with angular correlations. *Phys. Rev. E* **2016**, *94*, 1–5. [CrossRef] [PubMed]
35. Kurta, R.P.; Altarelli, M.; Vartanyants, I.A. Structural Analysis by X-ray Intensity Angular Cross Correlations. In *Advances in Chemical Physics*; John Wiley & Sons, Ltd: Hoboken, NJ, USA, 2016; Volume 161, 1–39.
36. Zaluzhnyy, I.A.; Kurta, R.P.; Scheele, M.; Schreiber, F.; Ostrovskii, B.I.; Vartanyants, I.A. Angular X-ray Cross-Correlation Analysis (AXCCA): Basic concepts and recent applications to soft matter and nanomaterials. *Materials* **2019**, *12*, 3464.
37. Martin, A.V. Orientational order of liquids and glasses via fluctuation diffraction. *IUCrJ* **2017**, *4*, 24–36. [CrossRef]
38. Donatelli, J.J.; Zwart, P.H.; Sethian, J.A. Iterative phasing for fluctuation X-ray scattering. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 10286–10291. [CrossRef]
39. Saldin, D.K.; Shneerson, V.L.; Fung, R.; Ourmazd, A. Structure of isolated biomolecules obtained from ultrashort X-ray pulses: Exploiting the symmetry of random orientations. *J. Phys. Condens. Matter* **2009**, *21*, 134014. [CrossRef]
40. Starodub, D.; Aquila, A.; Bajt, S.; Barthelmess, M.; Barty, A.; Bostedt, C.; Bozek, J.D.; Coppola, N.; Doak, R.B.; Epp, S.W.; et al. Single-particle structure determination by correlations of snapshot X-ray diffraction patterns. *Nat. Commun.* **2012**, *3*, 1–7. [CrossRef]
41. Virtanen, J.J.; Makowski, L.; Sosnick, T.R.; Freed, K.F. Modeling the hydration layer around proteins: HyPred. *Biophys. J.* **2010**, *99*, 1611–1619. [CrossRef] [PubMed]
42. Lovejoy, B.; Choe, S.; Cascio, D.; McRorie, D.K.; DeGrado, W.F.; Eisenberg, D. Crystal structure of a synthetic triple-stranded alpha-helical bundle. *Science* **1993**, *259*, 1288–1293. [CrossRef] [PubMed]

43. Lee, S.G.; Krishnan, H.B.; Jez, J.M. Structural basis for regulation of rhizobial nodulation and symbiosis gene expression by the regulatory protein NolR. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 6509–6514. [CrossRef] [PubMed]
44. Riback, J.A.; Bowman, M.A.; Zmyslowski, A.M.; Knoverek, C.R.; Jumper, J.M.; Hinshaw, J.R.; Kaye, E.B.; Freed, K.F.; Clark, P.L.; Sosnick, T.R. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **2017**, *358*, 238–241. [CrossRef]
45. Kubelka, J.; Hofrichter, J.; Eaton, W.A. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76–88. [CrossRef]
46. Shaw, D.E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R.O.; Eastwood, M.P.; Bank, J.A.; Jumper, J.M.; Salmon, J.K.; Shan, Y.; et al. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341–346. [CrossRef]
47. Lin, M.M.; Mohammed, O.F.; Jas, G.S.; Zewail, A.H. Speed limit of protein folding evidenced in secondary structure dynamics. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 16622–16627. [CrossRef]
48. Boutet, S.; Lomb, L.; Williams, G.J.; Barends, T.R.M.; Aquila, A.; Doak, R.B.; Weierstall, U.; DePonte, D.P.; Steinbrener, J.; Shoeman, R.L.; et al. High-resolution protein structure determination by serial femtosecond crystallography. *Science* **2012**, *337*, 362–364. [CrossRef]