Rietveld refinement for macromolecular powder diffraction

Maria Spiliopoulou[1*], Dimitris-Panagiotis Triandafillidis[1*], Alexandros Valmas[1], Christos Kosinas[1], Andrew N. Fitch[2], Robert B. Von Dreele[3], Irene Margiolaki[1]

[1] Department of Biology, Section of Genetics, Cell Biology and Development, University of Patras, GR-26500, Patras, Greece

[2] European Synchrotron Radiation Facility, BP 220, Grenoble Cedex 9, F-38043, France

[3] Advanced Photon Source, Argonne, IL 60439, USA

[*] These authors contributed equally

Abstract

X-ray powder diffraction (XRPD) has long been employed extensively for the structural characterization of materials, quantitative analysis of multicomponent mixtures, phase identification, texture and microstructural analysis. At the heart of all those, lies the Rietveld method, which has revolutionized the use of powder diffraction for materials studies. Initially applied to minerals and inorganics, the Rietveld method has progressively been used for more complex materials, for instance zeolites and pharmaceuticals. A major advance of the method came some 20 years ago, when the first protein structure was successfully refined. Due to the sheer complexity of macromolecules, several new approaches and algorithms had to be pioneered or adapted from macromolecular single-crystal diffraction experiments, thus constituting macromolecular XRPD a quite unique field of study. This review aims to provide necessary elements of theory and application of structure solution and refinement via the Rietveld method for macromolecular XRPD data. Practical explanations and highlighted case studies are also presented.

Introduction

A steadily increasing number of studies has underlined the efficiency of powder diffraction in a wide spectrum of fields, including structure determination of zeolites[1], inorganics[2], pharmaceuticals[3], and more recently, biological macromolecules[4–12]. The diffraction of polycrystalline materials is an old enough technique, developed in 1916[13,14], but was initially used only for phase analysis and to test the durability of building materials. The fact that this technique was not exploited for structure solution is mainly attributed to the complexity of analyzing the diffraction data, as well as to the poor quality of collected data, due to the instrumentation available at the time.

Between 1967 and 1969, Hugo Rietveld developed the 'Rietveld method'[15,16] which, in contrast to profile-fitting procedures, does not use integrated intensities of single diffraction peaks, but employs the entire XRPD pattern, partially overcoming the peak overlap problem and allowing for the maximum amount of information to be extracted from the pattern. This method was initially developed for neutron diffraction and later extended for X-rays. Rietveld refinement was originally used for crystal structure refinement and has evolved as a quantitative method for phase analysis of multicomponent mixtures using X-ray powder diffraction data[17], while contributed to a renewed interest in powder diffraction techniques, even to the extent of replacing single-crystal techniques, in certain applications[18].

Nevertheless, the most important problem, to date, arises from the overlapping reflections, since a large number of crystallites diffract simultaneously and only the sum of their signals is observed, thus complicating considerably the analysis of the diffraction data obtained. Structure solution requires

integrated intensities from a three-dimensional indexed reciprocal lattice, however in powder diffraction, the 3D diffraction data collapse to one-dimension powder pattern, thus the resulting peaks are highly overlapping.

In the Rietveld method, each data point is considered as an observation and during the refinement procedure, structural parameters, background coefficients and profile parameters are varied in a least-squares procedure until the calculated powder profile, based on the structural model, best matches the observed pattern. In the original procedure proposed by Rietveld, peaks were separated from the background, but in later developments of the method, the background contribution was also incorporated into the refinements[19].

Rietveld's method is general for all types of materials: it can make use of intensities even for severely overlapping reflections, as well as reflections with no intensity, and thereby made quantitative crystallographic analysis possible for materials which did not form single-crystals, even when complex and with low symmetry[20]. Rietveld refinement is a standardless method, in the sense that it does not use a calibration curve, but it does require a structural model (lattice parameters, space-group symmetry, atomic positions) of each phase being analyzed, from which a powder pattern can be simulated[3]. This article will focus on the structural characterization and the application of the Rietveld method on XRPD data obtained from macromolecular polycrystalline precipitates.

Protein and peptide crystals are sensitive to environmental alterations and extremely fragile, since extensive amounts of solvent are present, creating channels that surround macromolecules with layers of water, which help preserve their structure during crystallization[21,22]. Typically, about half the volume of a macromolecular crystal consists of water, the amount of which is closely related to relative humidity and/or temperature levels around the sample. Even small changes in a sample's environment may cause subsequent adjustments in solvent channels, driving protein molecules not to occupy exactly equivalent positions within or between unit-cells, directly affecting diffraction quality of the crystal, which either leads to collapse of crystal matrix[23] or rarely enhances diffraction capacity[24].

Until late 90s, macromolecular crystal structures were considered too complex for XRPD studies, however an initial experiment by Robert B. Von Dreele with metmyoglobin[11], revealed patterns with unexpectedly sharp peaks (**Figure 1**). Further developments in instrumentation methods and software, have made macromolecular structure solution using XRPD data feasible, ever since. In the original metmyoglobin work, structure refinement was performed using a modified version of the *General Structure Analysis System* (*GSAS*)[25], and later it was incorporated as a standard feature of the *GSAS-II* software[26,27]. Just the following year, another structure was solved and refined, that of human insulin[12]. Although dedicated phasing and structure solution methods are still lacking for macromolecular XRPD, employing equivalent methods and software from single-crystal X-ray diffraction (SCXRD) in combination with stereochemically restrained Rietveld refinement[6,8], has enabled structure refinement of many proteins, such as lysozyme[28], the SH3 domain of ponsin[29], bovine insulin[30] and pharmaceutical peptides[31,32].

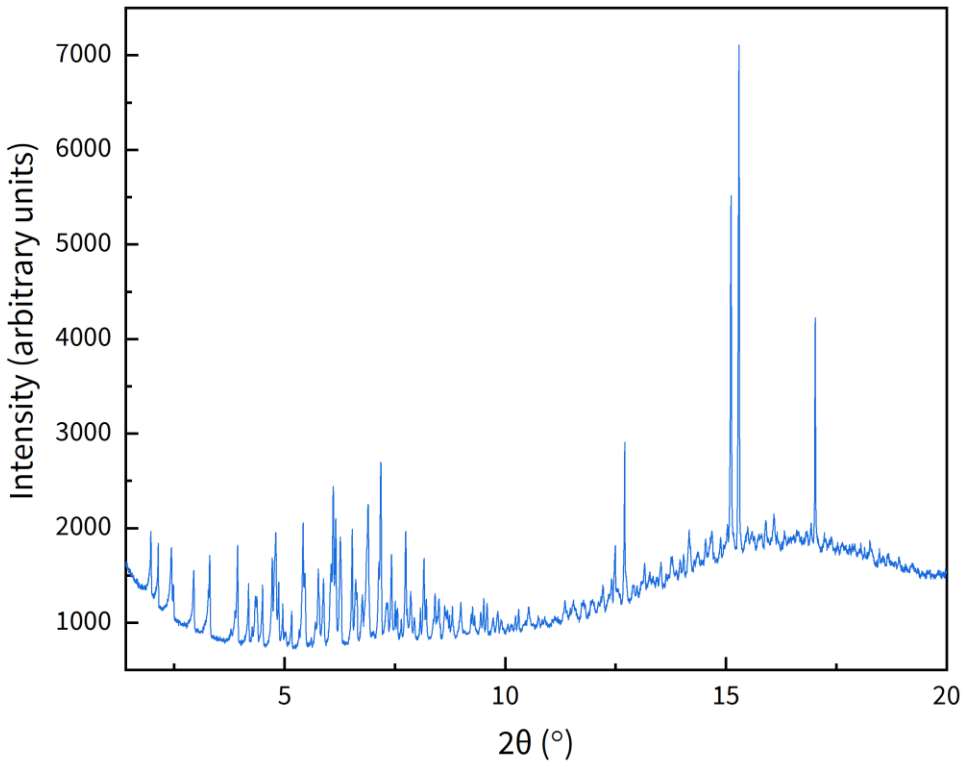

Figure 1. Powder diffraction data of metmyoglobin[11]. Data were collected on beamline X3bq at the National Synchrotron Light Source. The strong peaks observed in the high 2θ region are due to a second crystalline phase from $(NH_4)_2SO_4$.

Macromolecular Rietveld refinement is usually both multipattern and stereochemically restrained. Multiple patterns are employed to take advantage of sample induced anisotropic lattice changes: a process that helps towards partial alleviation of the peak overlap problem[9,29,30,33–37]. On the other hand, stereochemical restraints help reduce the otherwise large number of independent refinable parameters, promoting refinement to converge to physically meaningful structures[5].

Sample preparation

Protein crystallization

Protein crystals are typically acquired by adding a cocktail of crystallization agents (e.g. polymers, salts, buffers) to a highly purified and highly concentrated protein solution. Gentle manipulation of solution properties promotes protein supersaturation which can, potentially, be relieved by the formation of a crystalline solid state. The choice of the initial cocktail is based on a trial-and-error approach using crystallization screens, but once a hit is obtained, it can, rationally and systematically, be optimized. The crystallization process can be visualized using a phase diagram (**Figure 2**), which depicts protein solubility against different physicochemical properties of the solution. In an ideal crystallization experiment, the system will traverse the phase diagram, from the soluble region to the nucleation zone, where critical nuclei are formed, and later to the metastable zone, where nuclei grow to crystals, only to return to the soluble region, where crystal growth is halted. At the same time, the solution will transition from undersaturation to supersaturation and back, hopefully yielding a crystal in the process[38]. Depending on the extent to which the system will 'penetrate' the nucleation zone, a varying number of nuclei will be formed. For SCXRD experiments, one might prefer fewer nuclei, in

order to form a few large crystals, while for XRPD experiments, the available protein must be distributed on thousands of nuclei to form many microcrystals.

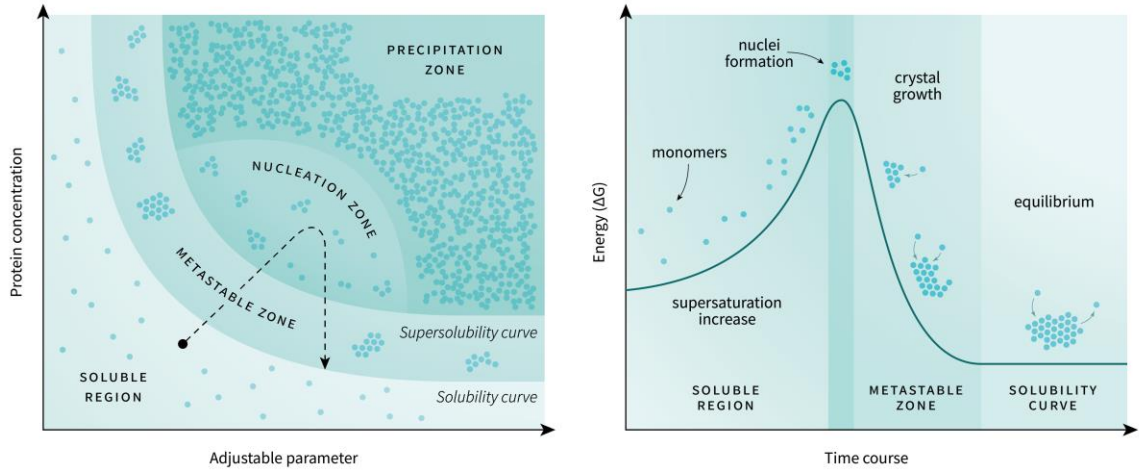

Figure 2. Typical crystallization phase diagram of a protein (**left**). The dashed line shows the ideal course of a crystallization experiment. Thermodynamic description of the course of a successful crystallization experiment (**right**).

In typical protein crystallization practice, a considerable number of experiments are set up and monitored in parallel. Most of these will yield no crystals, but even those that do, may not yield crystals of appropriate quality for SCXRD measurements (**Figure 3**). These crystals might be small in size (less than 10 µm in each direction), exhibit 1D or 2D growth (needles or plates), display micro-architectural lattice defects (incorporated impurities, mosaicity) or macro-architectural defects (e.g. twinned crystals, urchins, crystal clusters). Although a lot of effort has been dedicated to method development for characterizing these crystals with SCXRD[39–41], such samples can be efficiently studied via XRPD[29,42,43]. Of course, one must first be able to scale up the production of crystalline material, since a minimum of approximately 10 µl precipitate is required for a typical experiment, although it is possible to measure even single crystalline constructs, such as a single urchin, using specialized sample holders.

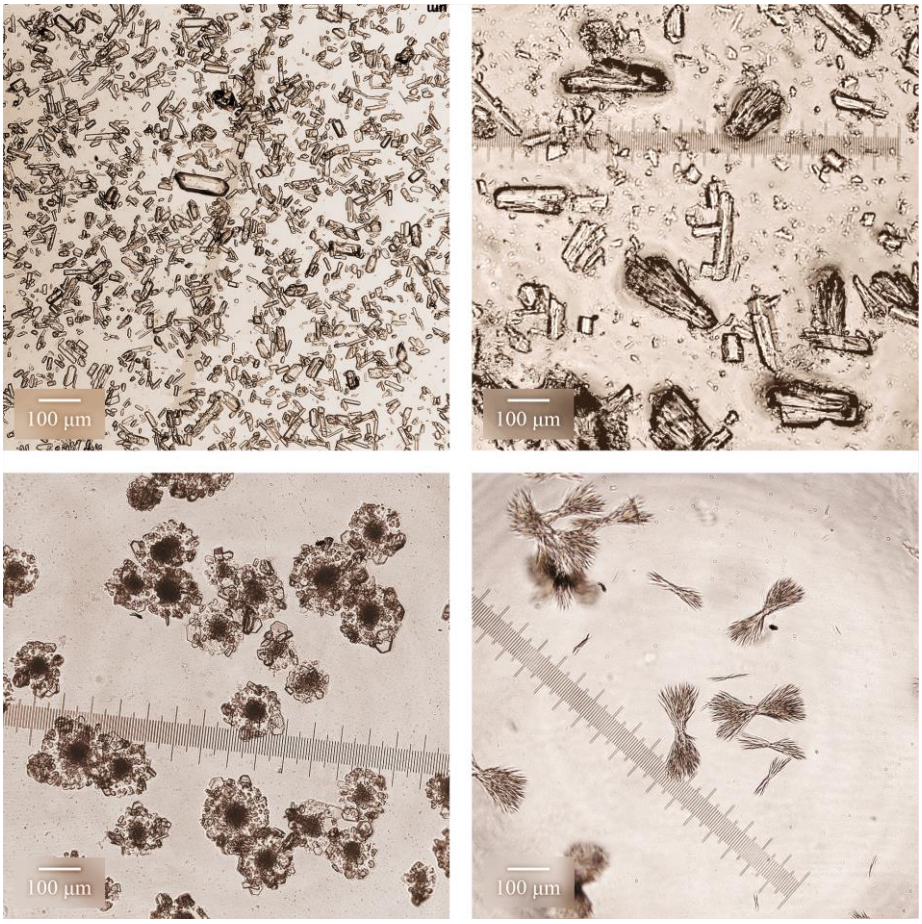

Figure 3. Different crystal morphologies that are typically inaccessible to SCXRD studies but can be effectively studied via XRPD. **Upper left:** very small crystals (octreotide). **Upper right:** crystals with rough edges (insulin). **Lower left:** clusters of crystals (insulin). **Lower left:** needle-shaped formations (polyphenol oxidase).

Scaling up a crystallization experiment might not be as straightforward as increasing the solution volume, since it is usually required to change the crystallization technique used, e.g. from vapor diffusion to batch, which is inadvertently followed by a different equilibration route[44]. Fortunately, due to the advent of serial crystallography at synchrotrons and xFELs, a lot of interest is focused in the conversion of single crystals grown by vapor diffusion, to microcrystals grown in batch[45], which will be of benefit to macromolecular XRPD, as well. It should be noted that scaling up crystallization experiments is usually very expensive in terms of protein consumption. For reference, 200 µl of protein and crystallization mix (typically mixed in 1:1 ratio), could yield as little as 10 µl of microcrystalline precipitate.

The ideal powder sample contains millions of microcrystals, preferably monodisperse in size and originating from a single crystalline phase, but samples with varying crystallite sizes and multiphase mixtures are also routinely studied via XRPD. In contrast to other materials, protein samples contain large amount of 'mother liquor' (i.e. an aqueous mixture of crystallization agents), which should not be confused with the large amount of water included inside the protein crystals. Therefore, protein samples are polycrystalline slurries instead of dry powders.

Sample holders

Sample holders for diffraction experiments are made of different materials, all of which exhibit high mechanical and thermal stability, are insensitive to radiation damage, but most importantly, have a high X-ray transmittance. Depending on the nature and available quantity of the microcrystalline

sample under study, different sample holders are available for use (**Figure 4**). The typical sample holders are borosilicate glass capillary tubes, which are sealed at one end. After transferring an adequate amount of polycrystalline slurry, the capillary is centrifuged to enhance crystal packing density. Excess mother liquor is, then, carefully removed from the capillary via a syringe, in order to prevent sliding of the precipitate, since the capillary is placed horizontally during data collection. Dense crystal packing inside the capillary increases Bragg signal, decreases diffuse scattering from solvent molecules and is necessary for protein samples, which diffract X-rays weakly and are surrounded by large amounts of solvent. Lastly, the capillary is cut to the appropriate length and sealed on the open end with silicone vacuum grease to prevent sample dehydration.

An alternative solution is to use Kapton capillaries (polyimide) with embedded Millipore filters[46]. Kapton capillaries are more robust than glass capillaries, although less chemically inert. Polycrystalline slurry is drawn into the capillary via a syringe and compacted against the embedded filter, thus negating the need for sample centrifugation.

Different sample holders are available for samples with a volume as low as a few µl (e.g. a single crystallization drop), or even single crystalline constructs (e.g. a single urchin). Initially, micro-meshes were used to provide support to these fragile structures, but the need for cryocooling and cryoprotection, due to air exposure, led to development of other sample holders. Such samples can now be loaded between two Silson membranes (silicon nitride) or Kapton foils, which encapsulate the sample and prevent dehydration[47].

XRPD data can also be collected from protein microcrystals using flat stage sample holders. To prevent dehydration, the samples must be enclosed in a constant relative humidity chamber, which also enables humidity and temperature variation experiments[48,49].

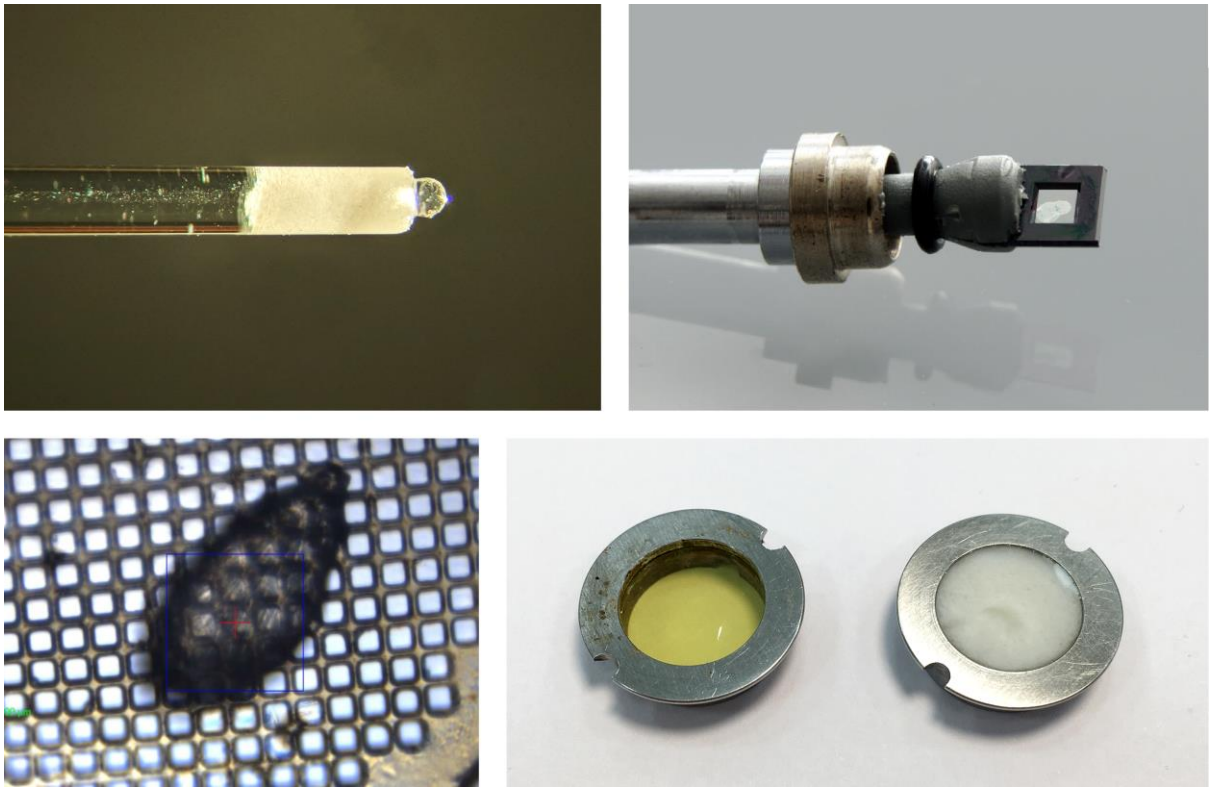

Figure 4. Sample holders for protein powder diffraction experiments. **Upper left:** Borosilicate glass capillary loaded with microcrystalline precipitate, before mother-liquor removal. The capillary is 1 mm in diameter. **Upper right:** Two Silson membrane windows attached to a blade holder[47]. A drop of polycrystalline slurry is added between membranes. The window size is 3 x 3 mm. **Lower left:** A single urchin supported on a micro-mesh with 25 µm openings[47]. **Lower left:** Empty (left) and filled (right) flat stage sample holders, consisting of a Kapton foil attached to a metal ring[48].

Data collection

Cryocooling

During diffraction experiments protein samples are susceptible to radiation damage: a combination of events resulting from sample's absorption of X-ray photons. All of the effects known from SCXRD experiments (e.g loss of *d*-spacing resolution, disulfide bridges rupture, decarboxylation of glutamates and aspartates, dehydroxylation of tyrosines and cleavage of carbon-sulfur bonds in methionines, degradation from free radicals) are also true for XRPD measurements and mainly localized around the area of the crystal illuminated by the X-ray beam[50,51].

In XRPD experiments, radiation damage additionally contributes to peak broadening, due to the gradual loss of crystallinity. However, attempts to suppress radiation damage by cryo-cooling, as is typical in SCXRD, have resulted in phase transitions and formation of ice peaks in turkey egg-white lysozyme studies[52]. The issue arises from the slow cooling rate of the sample due to its substantially larger size compared to single crystals in loops. For reference, a typical glass capillary has a 1 mm diameter, while a single crystal is only 50 μm in length. In SCXRD, the sample is flash-cooled, and only vitrified ice (amorphous) is formed, while in XRPD, the slower dissipation of heat allows formation of crystalline ice, which disrupts the macromolecular lattice and induces phase transitions. Introduction of a cryoprotectant (e.g. glycerol, polyethylene glycol) to the microcrystalline slurry prevents significant peak broadening, as was explored systematically in hen egg-white lysozyme[53], although a poor choice in the type and concentration of cryoprotectant could lead to detrimental results.

Data collection at room temperature is also possible for protein powders, but precautions must be taken. At synchrotron sources, the capillary must be periodically translated, to expose a 'fresh' region to the beam. Complete scans from different regions are manually inspected for signs of radiation damage (e.g. intensity damping, peak shifts and broadening) and then averaged to improve data statistics. At current levels of synchrotron beams' brilliance, the sample has a lifetime of ~2 min, before irreversible damage is observed. The situation is quite different with laboratory diffractometers. The low beam brilliance does not cause significant amounts of radiation damage, thus, the sample can be illuminated by X-rays even for 12 hours. Therefore, laboratory data have far better data statistics than synchrotron data, especially in the high *d*-spacing range, although the greater axial and beam divergence contribute to higher peak overlap[7]. Room temperature data collection is typically preferred, since it eliminates the need for a time-consuming and sample-expensive systematic evaluation of cryoprotectants, while yielding data of comparable quality, when executed carefully.

Instrumentation

One of the most important steps in macromolecular powder diffraction is the choice of instrumental configuration. As demonstrated by early works[11,12,54] protein microcrystals are a few hundred unit-cells in each dimension and, due to their weak intermolecular interactions, are unable to retain lattice defects at room temperature. Therefore, the diffraction peaks are inherently sharp and broadening is mainly due to the instrumentation.

While most studies are carried out at dedicated powder diffraction instruments, data of acceptable quality can also be collected from SCXRD[55] or SAXS setups[56], both at synchrotron beamlines and laboratory instrumentation. In general, it is always tempting to choose the highest angular resolution instrument available, but this comes at the price of poorer *d*-spacing range and data statistics[6,57].

In the case of dedicated powder diffraction instruments, there are two main options available: multi-crystal analyzer detectors and multi-channel position-sensitive detectors. Detectors equipped with crystal analyzers have for long been preferred, due to their superior angular resolution. Data collection with these detectors is inherently slow, due to the requirement for detector scanning, thus, high 2θ regions suffer from low data statistics, owing to prolonged irradiation period. Combined with the loss

of intensity during the secondary diffraction through the analyzer crystals, these detectors can occasionally result in poor *d*-spacing resolution[5].

On the other hand, multi-channel position-sensitive detectors can record data simultaneously over the entire 2θ range, thus minimizing data acquisition times and the effect of radiation damage, resulting typically in data of enhanced *d*-spacing resolution. Position-sensitive detectors are typically 2D (i.e. area detectors), but there are also curved 1D variants, comprised from flat strips arranged in a circle. The angular resolution of position-sensitive detectors is determined by pixel size, sample dimensions and beam size at detector surface. In general, the angular resolution is about an order of magnitude lower than that of multi-crystal analyzer detectors[57]. Peak broadening induced by the sample dimensions can be mostly compensated for, by focusing the beam at the surface of the detector, instead of the sample. This approach has successfully been employed with area detectors[9,58] and recently with 1D detectors[32,59]. In the case of area detectors, counting statistics are increased at higher 2θ angles, because the length of the arc used for intensity integration increases with the diffraction angle, i.e. more individual observations (pixels) are used per ring. Finally, as the pixel size on these detectors continues to fall, further enhancements to the angular resolution of position-sensitive detectors are anticipated.

The choice of instrumentation is ultimately dictated by the experiment requirements. Phase identification or high throughput screening of samples can be accomplished even from data with low counting statistics, while lattice parameters can be extracted accurately enough with less than 30 peaks. Even laboratory patterns with highly overlapping and highly asymmetric peaks are routinely employed for this purpose. On the other hand, structure solution and refinement require high data statistics, high *d*-spacing and angular resolution. Since to date, no single setup satisfies all three conditions, data collection utilizes multiple instrumental configurations[57], while the multipattern Rietveld approach reaps the benefits of individual setups, as will be discussed later.

A comprehensive, but non-exhaustive list of beamlines and detectors that have been recruited for macromolecular XRPD, including their individual strengths and weaknesses, can be found in International Tables: Volume H[6].

Preliminary analysis

Indexing

The first step in any crystallographic data analysis is the determination of the unit-cell, a procedure called ‘indexing’, since, in the process, a set of Miller indices (*hkl*) is assigned to each reflection. In SCXRD, indexing is, in principle, straightfoerward, since successive two-dimensional slices of the three-dimensional reciprocal lattice are recorded. Early indexing programs exploited patterns of distances between reflections that remained consistent through consecutive frames[60], while modern programs can index a lattice even from a single frame by employing Fourier methods[61]. Thus, indexing is highly automated and several checks are in place to ensure the validity of the solution.

In XRPD, the process is much more involved and more prone to errors, owing to the substantial paucity of information: the powder pattern results from all individual reciprocal lattices, instead of the one lattice present in a single crystal. The indexing process and the typical algorithms used are excellently described by Altomare *et al*[62]. XRPD indexing is very sensitive to peak positions, therefore only the low 2θ range is used, where peak positions are less susceptible to small lattice parameters variations. In practice, 20-30 peaks are necessary for successful indexing: having fewer peaks leads to underdetermination of the system of equations to be solved. On the other hand, including more peaks from a higher 2θ range negatively biases the indexing algorithms towards extraction of solutions with smaller unit cells, thus rendering the correct cell harder to determine.

The success of an indexing trial is never guaranteed. It depends highly on the initial choice of peaks, accuracy of peak position, presence of impurity peaks, symmetry and unit-cell size. Higher symmetries and smaller unit-cells are typically easier to index. Even when solutions are yielded, a careful evaluation for each solution must be made, to determine the correct one. Typical evaluation

includes a few cycles of Pawley or Le Bail refinement for the various plausible space group settings. Space group assignment can also be ambiguous, since specific symmetry elements, such as the screw axes, cannot be assigned unequivocally to specific axes.

In macromolecular XRPD, indexing is even more challenging, due to the very large unit-cells present. For reference, a typical macromolecular unit-cell axis falls in the range of 50-100 Å, while extreme cases of up to 300 Å axes have been reported (e.g. for photosystems structures, but also for much smaller proteins[63]). To mitigate the adverse effect of large unit-cells on indexing algorithms, one might assign a smaller wavelength (e.g. λ/10), in order to shift *d* spacings towards smaller values. Of course, the axes for each of the solutions yielded, are also scaled by the same factor.

Macromolecular patterns typically exhibit much lower intensity-to-noise ratios than inorganics, therefore peak assignment is not always easy. The presence of extraneous peaks from foreign phases, further complicates this process. Foreign phases could appear in the sample either due to crystallization of one or more of the crystallization 'cocktail' agents (e.g. salts, ligands), or due to polymorphism of the macromolecule under study. Co-existence of two or more macromolecular phases renders indexing impossible, when all phases are unknown. In such cases, maintaining a local database of known phases of the macromolecule under investigation, could help in identifying one or more of the constituent phases, thus enabling indexing attempts on the remaining peaks of unknown origin.

Pattern fitting

In order to solve a structure using powder data it is necessary to extract as many reflections as possible from the collected profile. Until the 1980s, this was not feasible due to the overlapping nature of the peaks in the powder diffraction profile. However, with the development of high-speed computers with large memories and high-resolution diffractometers, whole powder pattern decomposition (WPPD) became a viable and important part of the analysis of powder data. All the methods are based on pattern decomposition, rather than deconvolution because the exact peak FWHM function is not known *a priori*. There are two methods that are in common usage at the present time, one due to Pawley and one due to Le Bail.

In the Pawley method[64], the unit-cell parameters, background parameters, zero point errors (zero shift), peak shape parameters and all reflection intensities are subjected to non-linear least squares refinement. A unit-cell is used to generate the initial peak list, while peak positions are adjusted by refining the unit-cell lattice parameter, rather than the individual peak positions.

Many of the refinable parameters are identical to those used in a Rietveld refinement program, but the significant difference is that every reflection is assumed to have an arbitrary peak intensity I_{hkl}, which is subjected to refinement, whereas in Rietveld, peak intensity is calculated from the structure factures, F_{hkl}, which, in turn, are calculated from the atomic coordinates of the model structure. The least-squares minimization procedure requires a $(p+n)\times(p+n)$ square matrix, where n is the number of symmetry-independent reflections generated for the 2θ range covered by the data and p the remaining refinable parameters (e.g. peak shape, background, unit-cell).

The correlation between peak intensities increases with increasing overlap, and reaches 100% for peaks with identical *d*-spacing within the limits of data resolution[65]. A major complication in overlapping peak extraction is that a Pawley refinement tends to converge to an arbitrarily large positive intensity for one constituent peak and negative intensity for another peak, while their sum fits the observed composite peak. To overcome the extraction of negative intensities, which lack physical meaning, restraints can be imposed on the constituent peak intensities. The restraints can either 'force' the intensities to have positive values (e.g. by extracting $|F|$ instead of I[66]) or equipartition the intensities of composite overlapping reflections to their constituent peaks. However, such approaches are much more computationally expensive, and thus slower.

In the case of the Le Bail method[67], the same parameters are included in the fitting procedure. The algorithm involves refinement of unit-cell, profile parameters and peak intensities in order to fit the

measured powder diffraction pattern. All parameters except for structural constants, namely atomic positions, occupancies, and displacement factors, are subjected to least-squares refinement. The Le Bail method iterates the same Rietveld formula, with only a slight modification of the code. Since $|F_c|^2$ cannot be calculated from atomic coordinates, all calculated peak intensities are initially set to an arbitrary value, e.g. $|F_c|^2 = 1$. These structure factors are iteratively adjusted for each peak until they best match the observed peak intensities, in a separate process[65]. These are then entered back in the Rietveld intensity decomposition formula as if they had been derived from a structural model. Hence, the intensities of the individual peaks are not treated as least-squares parameters and are not directly refined. Since identical peak intensities are used as starting values, the intensities of (almost) fully overlapping reflections tend to be equipartitioned after the refinement converges. Negative intensities in Le Bail refinement are less common than in Pawley, since all constituent peaks are apportioned the same arbitrary starting intensity value, but they can still occur[68].

The main difference of Pawley and Le Bail methods is the intensity extraction approach. Pawley refinement partitions overlapping intensities more effectively, depending on the restraints imposed, but there is a cost on computation time. On the other hand, Le Bail refinement is much faster and, since it uses a modified Rietveld code, it is preferred as a stepping-stone to calculate starting values for peak shape and lattice parameters for subsequent Rietveld refinements.

In most scenarios, the Pawley and Le Bail methods are used for refining lattice and peak shape parameters of known phases. If the structure is also known, it can later be refined via Rietveld, however it is important to remember that the extracted intensities from Pawley and Le Bail refinements can be used for structure determination using direct methods[67,69] or charge flipping, even though this approach has not been applied to macromolecular XRPD, due to the sheer complexity of the structures under study.

Observed intensities extraction

In contrast to SCXRD, extraction of observed intensities in XRPD is non-trivial, due to peak overlap. Since only the intensity of the composite peak is measured, it is difficult to estimate the constituent peak intensities, especially for peaks under exact overlap. The solution proposed by Rietveld[15] was to estimate the individual observed peak intensities by partitioning the total observed intensity (obtained after background subtraction and integration) based on the contributions of the constituent peaks to the total calculated intensity[20] (**Figure 5**).

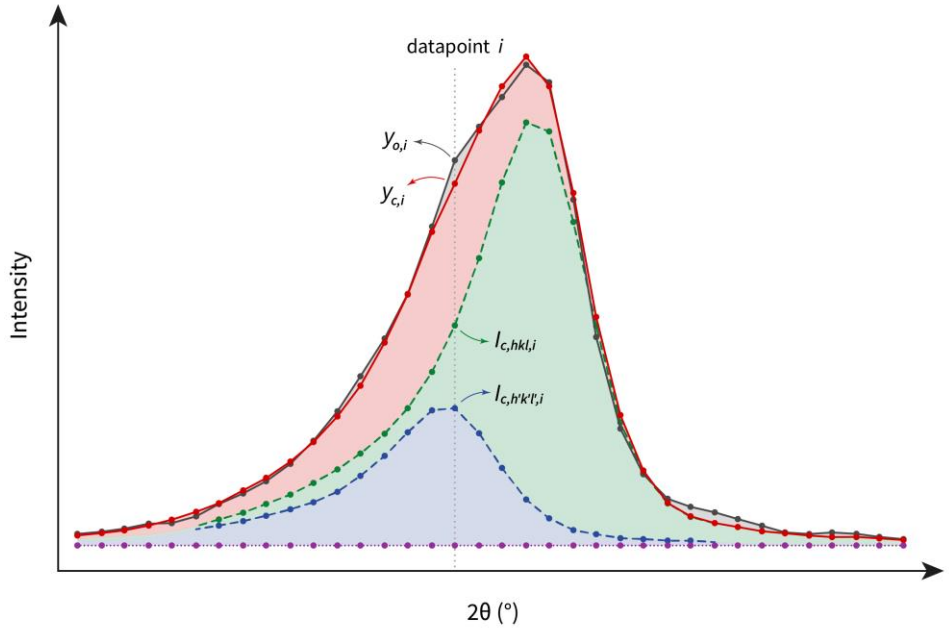

Figure 5. Extraction of observed intensities from overlapping reflections in the Rietveld method is based on the contributions of individual peaks to the total calculated intensity. Gray and red profiles correspond to observed and calculated profiles of the composite peak. Green and blue profiles are the calculated profiles of the overlapping, constituent peaks. The shaded regions under the profile lines are the corresponding total peak intensity. The purple line is the refined background, which is subtracted before area integration of the peak profiles.

The intensity value for each datapoint, i, in the observed profile ($y_{o,i}$) is apportioned to each contributing reflection, hkl, by applying the ratio of constituent calculated peak intensity ($I_{c,hkl,i}$) to total calculated peak intensity for the same datapoint ($y_{c,i}$), or:

$$I_{o,hkl,i} = y_{o,i} \frac{I_{c,hkl,i}}{y_{c,i}} \quad \textbf{(1)}$$

The total calculated peak intensity per datapoint ($y_{c,i}$) is the sum of the individual calculated peak intensities at the same datapoint, i:

$$y_{c,i} = \sum_{hkl}^{reflections} I_{c,hkl,i} \quad \textbf{(2)}$$

The overall observed intensity for each reflection, is then acquired by summation of the individual contributions, $I_{o,hkl,i}$, over all datapoints in the composite peak, or:

$$I_{o,hkl} = \sum_{i}^{datapoints} I_{o,hkl,i} \quad \textbf{(3)}$$

The intensity values, either observed or calculated, are then converted to equivalent structure factors, by applying various correction terms (c) for Lorentz polarization, multiplicity, absorption and preferred orientation, using the following relation:

$$I_{hkl} = c_{hkl} \cdot F_{hkl}^2 \quad \textbf{(4)}$$

Since the individual observed intensities and, by extension, the observed structure factors are determined using the contributions of calculated intensities, derived from the structural model, the estimates are only as good as the quality of the model. Due to this peculiarity, the observed structure factors are inherently biased towards the model, but this is often the best estimate that can be made[20]. As the model is progressively improved, so do the observed structure factors estimates.

Structure solution

The crystallographic phase problem

The Rietveld method cannot be used for *ab initio* structure determination. Rietveld refinement requires introduction of known structural information for the various crystalline phases contained in the pattern under study. Therefore, a crucial point that determines the outcome of the analysis is the choice of the most appropriate starting model.

From the experimentally extracted and refined intensities and as well as the computationally determined phases, we end up, through the Fourier transform, to a function of electron density. Based on the electron density map, i.e. a map of the distribution of electrons in three-dimensional space, atom positions can be distinguished and, finally, the overall structural model can be obtained. Protein crystals do not favor the creation of high-quality maps, where it is unequivocally possible to distinguish all individual atoms, especially in the early steps of refinement, mainly due to their reduced organization and increased mobility, owing both to high solvent percentages and the fact that protein crystals are fragile. In order to identify all the atoms that make up the protein, all the available information is utilized, i.e. the amino acid sequence, the number of molecules and the percentage of solvent per unit-cell.

Structure factors are proportional to the intensity of reflections and can be expressed as a wave or vector of the form:

$$\mathbf{F}_{hkl} = |F_{hkl}| \cdot e^{i\varphi_{hkl}}, \quad I_{hkl} \propto |F_{hkl}|^2 \tag{5}$$

Therefore, their calculation requires not only their absolute values but also the diffractive wave direction when it was recorded. Knowing both the amplitude and direction of the waves, permits direct calculation of the electron density in the unit-cell and consequently determination of the atomic positions. However, the measurement of reflection intensities results only in the amplitudes of the structure factors ($|F_{hkl}|$), while their phases (φ_{hkl}) remain unknown. The inability to measure phases directly is the crystallographic 'phase problem' and can be calculated by both direct and indirect methods[70]. Indirect methods lead to structure solution based on some known phases, whereas direct methods exploit the fact that phase relations exist between certain sets of structure factors and recover all the phases[71]. Several approaches for solving the phase problem exist, and the choice of phasing strategy depends on our state of prior knowledge and technical or biochemical resources.

Molecular replacement

A general indirect approach to determining initial phases is molecular replacement (MR)[72]. An initial search model for a protein molecule is selected amongst other known structures of proteins that have similar amino-acid sequences, with a homology greater than 30%[73]. This model is then oriented and positioned in the unit-cell until the solution with the best fit between calculated diffraction data from the replaced model and the experimental data from the unknown structure is obtained (**Figure 6**). There are no more than six degrees of freedom to be determined per molecule when solving a structure by molecular replacement. The first three give the orientation, while the latter three give the position of the molecule with respect to the symmetry elements of the space group. The applicability of MR in macromolecular XRPD was first demonstrated by Von Dreele *et al*.[12] with the case of $T_3R_3^f$ insulin. In this work, the correct placement of the molecule in the asymmetric unit (asu) was performed by defining the entire molecule as a rigid body prior to Rietveld refinement. However, later

works[29,43,52,74] employed standard SCXRD software for MR, such as *MOLREP*, using integrated intensities from restrained Pawley or Le Bail refinement as input. *MOLREP*[75] can also be used for automated MR in XRPD, by proposing the number of molecule copies present in the asu, as well as their position and orientation[5]. After determining the position and orientation of the search model in the unit-cell, and replacing it with the new model, it is of critical importance to go through the structure and identify the regions that are different in the new protein compared to the initial model, and evaluate the electron density by generating Patterson maps[76].

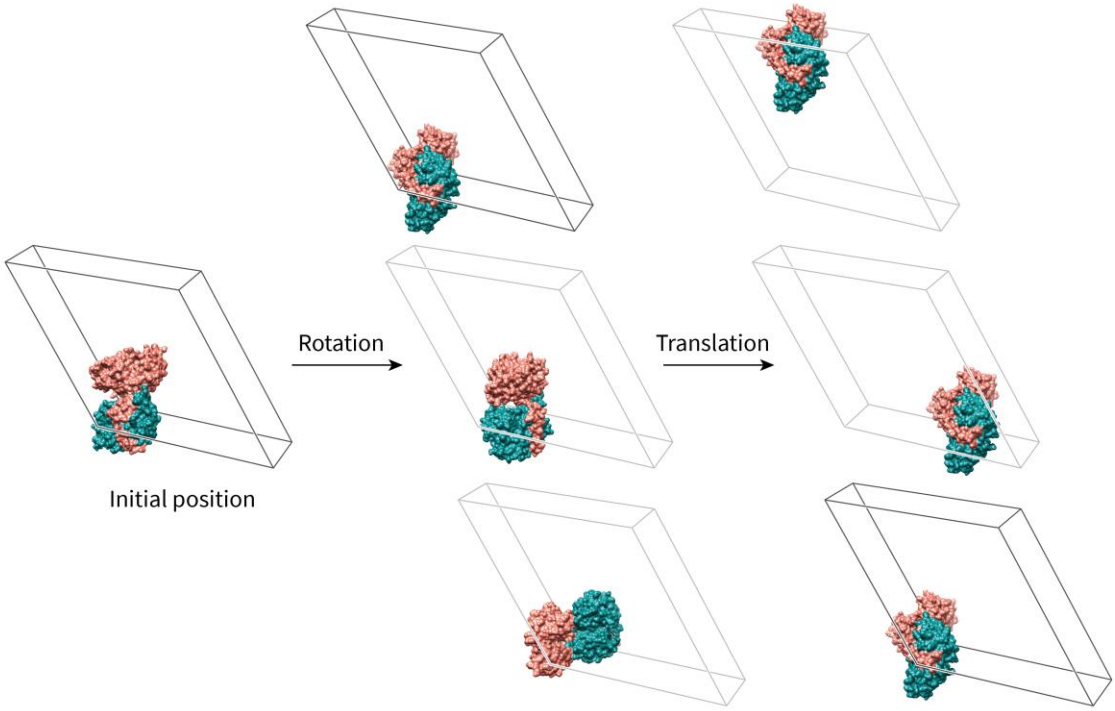

Figure 6. The molecular replacement method evaluates all possible orientations and positions of the protein until the best agreement between the calculated and observed data is achieved. The model is initially placed in the experimental unit cell and the best position resulting from rotation is chosen and constitutes the orientation, which is then translated along x, y, z axis. Dark-colored unit cells correspond to the best candidates after each process.

Isomorphous replacement

In cases where a starting model is not available, a common method used, is isomorphous replacement (IR)[77], which requires the attachment of heavy atoms to the protein molecule in order to determine the phases of X-ray reflections. Isomorphous replacement methods were the first successfully applied phasing methods in protein crystallography. IR was theoretically developed from the mid-1930s and first applied to protein crystals for the determination of centric reflection in hemoglobin in the mid-1950s[78]. The first step to apply this method, is to co-crystallize the molecule of interest with heavy atoms or soak the existing crystalline sample in a heavy atom solution, providing, thus, a derivative crystal. Subsequently, diffraction data need to be collected from the protein with and without the added heavy atom (H), called derivative (PH) and native (P) forms, respectively. Provided that the two patterns are isomorphous, i.e. no significant changes are observed in lattice dimensions and symmetry, intensity differences between a native and a derivative pattern can be exploited for substructure solution. The methods are generally called isomorphous replacement methods, and, depending on how many derivatives are used, there are single isomorphous replacement (SIR)[79] or multiple isomorphous replacement (MIR) methods[73]. The obtained diffraction patterns have strong intensity differences due to the heavy atom's contribution, thus coordinates of these atoms can be determined using direct methods. The amplitudes of a reflection are measured for the native crystal,

$|F_P|$, and for the derivative crystal, $|F_{PH}|$. The isomorphous difference, $|F_H| \cong |F_{PH}| - |F_P|$, can be used as an estimate of the heavy-atom structure-factor amplitude to determine the heavy atom's positions using Patterson or direct methods[76]. Knowledge of these coordinates allows the calculation of the heavy atom's contribution to each structure factor. Consequently, amplitudes of the native and derivative protein can be determined and reasonable assumptions about the possible phases can be made (**Figure 7**). IR has been applied to XRPD as well, although to date it has led only to medium resolution structures[33].

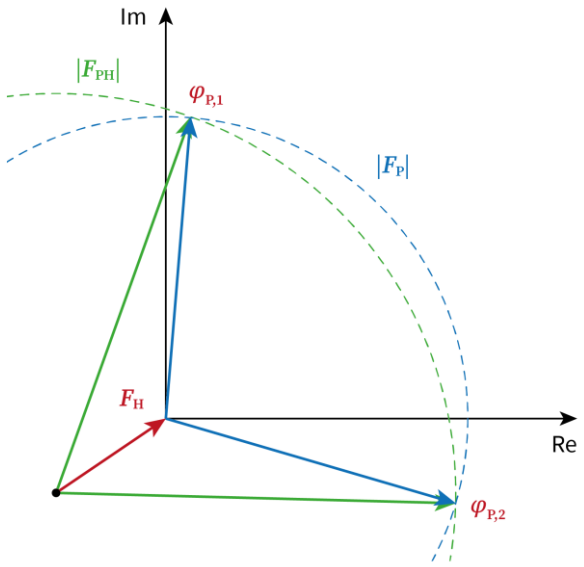

Figure 7. In isomorphous replacement, differences in scattered intensities of native (P) and derivative (PH) crystals will largely reflect the contribution of heavy atoms to the diffraction profile. Intensity measurements reveal the magnitude of $|F_{PH}|$ and $|F_P|$ structure factors, each of which can form a circle covering all the possible vectors for each of them, colored in green and blue, respectively. From the difference data, the position of heavy atom can be determined and, therefore, the entire complex structure factor of $\mathbf{F}_H$ is derived (red vector). The two possible solutions for the phase angle ($\varphi_{P,1}$ and $\varphi_{P,2}$) can be determined by the intersection of the two circles. In practice, only one solution is correct.

Anomalous dispersion

Dispersion is generally a change of a property with frequency (or energy), thus, X-ray dispersion is the energy-dependent change of the scattering factor of X-rays when absorption occurs near the absorption edge of an element[73]. The atomic scattering factor contains three components[80]:

$$f = f_0 + \Delta f' + i\Delta f'' \quad \textbf{(6)}$$

a normal scattering term f_0, that is dependent on the scattering angle ($\sin\theta/\lambda$), and two terms, $\Delta f'$ and $\Delta f''$, that are dependent on wavelength and represent the anomalous scattering that occurs at the absorption edge when the X-ray photon energy is sufficient to extract an electron from an inner shell (**Figure 8**). The method of anomalous dispersion is based on the phenomenon of abnormal scattering observed when an incident wavelength is selected near the absorption edge of a chemical element in the sample[70]. Anomalous dispersion causes significant changes in specific reflections, Friedel pairs ($|F_{hkl}| = |F_{\bar{h}\bar{k}\bar{l}}|$) related to these elements, revealing the phases of these reflections[6]. When a single wavelength is used for an anomalous scattering experiment, the method is called single-wavelength anomalous diffraction (SAD), whereas when multiple wavelengths are used, the method is called multi-wavelength anomalous diffraction (MAD). MAD method breaks the two-fold phase ambiguity by including a second, dispersive data set at a different wavelength in the phasing equations. Although MAD methods require exactly tunable X-ray wavelengths and, thus, a synchrotron source, they have revolutionized the process of solving novel macromolecular structures using SCXRD[81].

Unfortunately, in XRPD the Friedel pairs are always under exact overlap, thus differences in their amplitudes cannot be inferred, however it has been shown that dispersive differences can be measured[82,83], even though this method has not been used for protein structure solution, yet.

According to the Patterson method, calculating the homonymous function for each point of the unit-cell, leads to the construction of a three-dimensional vector map. This map consists of peaks that correspond to the final ends of the vectors which unite, in pairs, all the individual atoms of the structure. However, due to the complexity of the Patterson function and the corresponding map for macromolecular structures, in practice the application of the method is usually applied in combination with other methods, to infer the position of heavy atoms or groups[70].

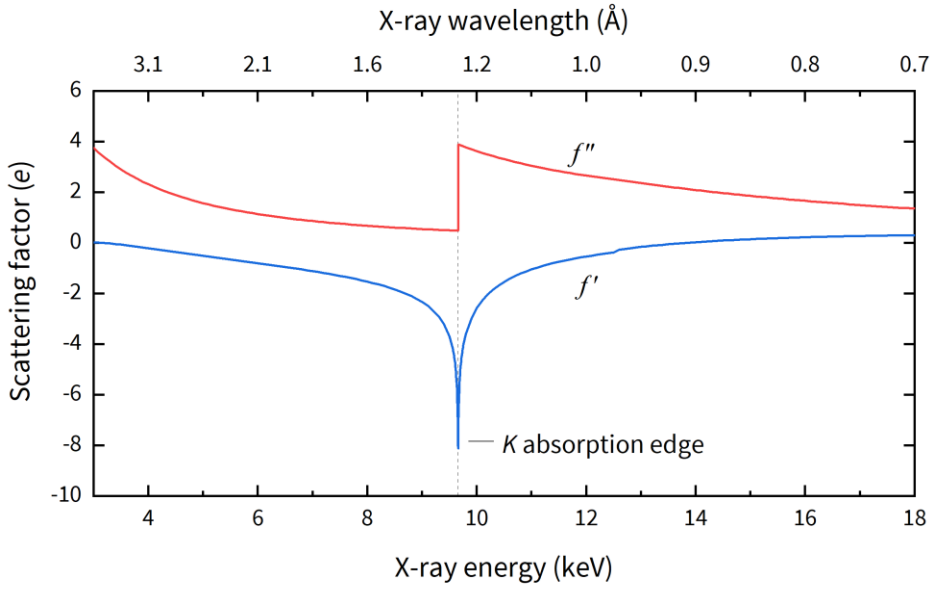

Figure 8. Anomalous scattering factor terms, *f′* and *f″*, for zinc. A sharp discontinuity is observed for both terms around the absorption edge of the element (Zn-*K*: 9.66 keV or 1.28 Å).

Structure refinement

Peak overlap

A powder histogram is a one-dimensional projection of the three-dimensional diffraction data. Since the number of reflections increases by $(\sin\theta/\lambda)^3$, the peak overlap problem is very common at higher angles[84]. Peak overlap falls in two categories: exact or systematic and partial or accidental overlap. Exact peak overlap is dictated by crystal symmetry facts, as Miller planes corresponding to the same (*hkl*) family are characterized by equal *d*-spacings, and therefore contribute to a single diffraction peak, upon the collapse of the 3D reciprocal lattice into one dimension. On the other hand, partial peak overlap is for non-symmetry-related peaks, with nearly equal *d*-spacings, and is ultimately dependent on the instrumental angular resolution and sample quality.

The loss of information due to peak overlap in a powder pattern is severe and can be quantified. The degree of overlap of individual reflections depends on their separation in $\Delta d/d$ and their peak widths[65]. Ultimately, the intensity information content in the powder pattern will be determined by these factors. The average number of reflections within a resolution shell of width Δd^* at a radius $d^* = 1/d = \sin\theta/\lambda$ in reciprocal space can be approximated, for a triclinic material with unit cell volume *V* by[84]:

$$\Delta N(d^*) \cong 2\pi V d^{*2} \Delta d^* \quad \textbf{(7)}$$

The value steadily increases with a constant Δd^*. If the expression is converted to equidistant intervals in 2θ using the Bragg equation (since $d^* = \sin\theta/\lambda$), peak overlap is more severe at higher angles. For protein samples, the large unit-cells further aggravate this problem and significant loss of information can be observed at smaller 2θ angles. Peak overlap, whether exact or partial, is exacerbated at higher angles (**Figure 9**), however the former is also prone to lattice symmetry, with higher order symmetry (e.g. cubic) resulting in more exact overlap than lower order symmetries (e.g. monoclinic). Partial overlap is significantly alleviated by the use of high angular resolution synchrotron instrumentation[65].

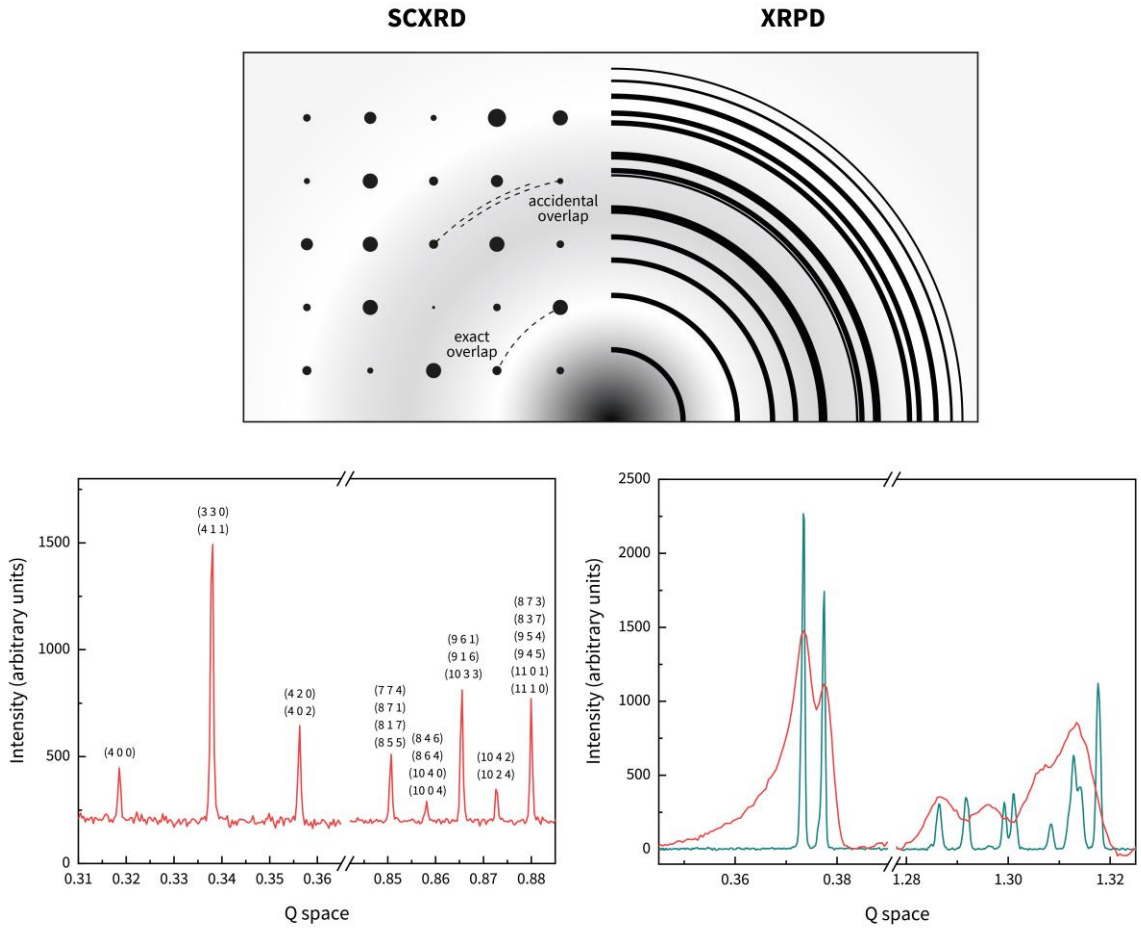

Figure 9. Upper: SCXRD and XRPD diffraction patterns illustrating the cause of the two forms of overlapping reflections (exact and accidental). **Bottom left:** Partial overlap is exacerbated by instrumental angular resolution. Data from octreotide microcrystals of orthorhombic symmetry ($P2_12_12_1$) collected from a synchrotron (teal) and laboratory (red) setup. **Bottom right:** The frequency of exact overlap increases with diffraction angle and is more prevalent with higher symmetry. Data from insulin microcrystals of cubic symmetry ($I2_13$). $Q = 4\pi \sin\theta/\lambda$ is momentum transfer.

Multipattern refinement

Apart from instrumentation improvements, partial peak overlap can be relieved experimentally by inducing anisotropic changes to lattice dimensions and collecting multiple datasets[7,9]. Such changes can lead to slightly different positions of partially overlapped neighboring peaks along the 2θ axis and, hopefully, improve their separation. Subtle changes in the crystal lattice can be induced by varying physical properties of the sample, but care must be taken not to disturb or disrupt the lattice

completely. Using a set of patterns, across which a physical parameter is varied and forcing the refinement algorithm to use similar intensity values for the same peak throughout all different datasets, leads to a significantly more accurate intensity extraction[85]. This method has been successfully applied in diffraction patterns collected from polycrystalline samples produced under altered crystallization conditions in terms of pH[7,34,55,63,86] or by anisotropic thermal expansion[87,88]. Another interesting approach is radiation-induced lattice changes, where the sample is slightly overexposed to the X-ray beam in order to provide datasets with and without radiation damage[30,35]. However, since radiation damage is a destructive process, care must be taken to retain a moderate dose and prevent non-isomorphism of different patterns.

Multipattern refinement can be incorporated both in Pawley and Rietveld refinement workflows. In Pawley refinement using *PRODD*[85,89], a single set of intensities is extracted from multiple datasets, with values more accurately determined than those extracted from any single dataset alone. In Rietveld refinement using *GSAS*[25], datasets from different sources (e.g. synchrotron beamlines and lab diffractometers) are simultaneously refined and the higher quality regions of each dataset are leveraged. Combining multiple datasets together, where either the unit-cell parameters or the preferred orientation is varied (although the latter is typically not observed in protein samples), allows the contributing reflections within a cluster of overlapped peaks to be more easily distinguished and has been employed at great extend in macromolecular XRPD[8].

Refinable parameters

During Rietveld analysis structure, lattice parameters, peak shape function and preferred orientation of the microcrystals are determined. Pawley and Le Bail methods are valuable for obtaining starting values for non-structural parameters in a crystallographic fit or, occasionally, for sample characterization, where perhaps only peak-broadening terms are desired. In our workflow, Pawley refinement is performed using *PRODD*[85,89] or *HighScore Plus*[90], to extract accurate unit-cell, while peak shape and background parameters are typically optimized via Le Bail refinement, using *GSAS*[25], which is also employed for Rietveld refinement.

Although theoretical diffraction peaks from an ideal lattice follow the Dirac delta function, the observed peak shape in XRPD is the consequence of a variety of effects arising from the non-ideality of the source, and sample, as well as aberrations introduced by optics and detector geometry. In the case of X-rays, peak shapes deviate from the Gaussian representation used for treating early, low-resolution neutron diffraction data[91].

The intensity for a Gaussian peak, $G(2\theta)$, centered at $2\theta_{hkl}$, is obtained by:

$$G(2\theta) = a_{\mathrm{G}} \cdot e^{-b_{\mathrm{G}}(2\theta - 2\theta_{hkl})^2}, \text{ with } a_{\mathrm{G}} = \frac{2}{H_{\mathrm{G}}}\sqrt{\frac{\ln 2}{\pi}} \text{ and } b_{\mathrm{G}} = \frac{4\ln 2}{H_{\mathrm{G}}^2} \quad \textbf{(8)}$$

The full-width at half-maximum (FWHM) for a Gaussian peak, H_{G}, is given by[92,93]:

$$H_{\mathrm{G}} = \sigma_{\mathrm{G}}\sqrt{8\ln 2} = \sqrt{U\tan^2\theta_{hkl} + V\tan\theta_{hkl} + W} \quad \textbf{(9)}$$

where σ_{G} is the variance of a Gaussian peak. *U*, *V* and *W* are refinable parameters related, collectively, to source size, collimators, monochromators and crystal mosaicity.

The intensity for a Lorentzian peak, $L(2\theta)$, centered at $2\theta_{hkl}$, is obtained by:

$$L(2\theta) = \frac{a_{\mathrm{L}}}{1 + b_{\mathrm{L}}(2\theta - 2\theta_{hkl})^2}, \text{ with } a_{\mathrm{L}} = \frac{2}{\pi H_{\mathrm{L}}} \text{ and } b_{\mathrm{L}} = \frac{4}{H_{\mathrm{L}}^2} \quad \textbf{(10)}$$

The Lorentzian FWHM, H_{L}, is modeled after[94]:

$$H_{\mathrm{L}} = 2\gamma_{\mathrm{L}} = \frac{X}{\cos\theta_{hkl}} + Y\tan\theta_{hkl} + Z \quad \textbf{(11)}$$

where γ_{L} is the scale parameter of a Lorentzian peak. *X*, *Y* and *Z* are refinable parameters related crystallite size and microstrain broadening. More complex descriptions for *X* and *Y* are typically used in refinement routines[25]. Each of the six refinable parameters is directly related to physical properties of the sample. Most notably, *X* is related to crystallite size broadening, while *U* and *Y* are related to microstrain broadening[95].

The peak shape of XRPD profiles is the result of a convolution of the instrument and sample contributions, therefore it is best described by a Voigt function (i.e. convolution of Gaussian and Lorentzian functions). However, calculations for Voigt functions are computationally expensive, therefore a pseudo-Voigt function[96] is often used for the description of peak shape profiles. A pseudo-Voigt peak, $pV(2\theta)$, centered at $2\theta_{hkl}$, is the normalized sum of a Gaussian and Lorentzian peak:

$$pV(2\theta) = \eta L(2\theta) + (1-\eta)G(2\theta), \text{ with } 0 \leq \eta \leq 1 \quad \textbf{(12)}$$

where η is the mixing coefficient. Initially, η was an arbitrary parameter, but since it could not be related to any physical properties of the sample, its value was fixed, in order for the pseudo-Voigt function to closely approximate a true Voigt function[93]:

$$\eta = 1.36603\frac{H_{\mathrm{L}}}{H_{\mathrm{pV}}} - 0.47719\left(\frac{H_{\mathrm{L}}}{H_{\mathrm{pV}}}\right)^2 + 0.11116\left(\frac{H_{\mathrm{L}}}{H_{\mathrm{pV}}}\right)^3 \quad \textbf{(13)}$$

Following, the pseudo-Voigt FWHM, H_{pV}, also has a fixed value, arising from the constituent Gaussian and Lorentzian FWHMs[93]:

$$H_{\mathrm{pV}} = \sqrt[5]{H_G^5 + 2.69269 \cdot H_G^4 H_L + 2.42843 \cdot H_G^3 H_L^2 + 4.47163 \cdot H_G^2 H_L^3 + 0.07842 \cdot H_G H_L^4 + H_L^5} \quad \textbf{(14)}$$

Further enhancements to the basic pseudo-Voigt function are usually required as the shape of the peaks is not symmetric, but at low scattering angles peaks display some asymmetry due to axial divergence. The source of the asymmetry is the finite length of the detector slit, which records an arc of the Debye-Scherrer diffraction cone. The curvature of the cone is proportional to $|2\theta - 90°|$, therefore decreases towards 90°. A description of the detector slit intercept function, $D(2\theta)$, was first formulated for neutron diffraction as[97]:

$$D(2\theta) = \frac{L^2 \sin 2\theta_{hkl}}{4HS \cdot h(2\theta)\cos(2\theta_{hkl} - 2\theta)} w(2\theta) \quad \textbf{(15)}$$

where $2H$ is the detector slit length, $2S$ is the sample illumination length, L is the diffractometer radius, $h(2\theta)$ is a projection function of the diffraction cone to the detector slit plane (**Figure 10**), defined as:

$$h(2\theta) = L\sqrt{\frac{\cos^2 2\theta}{\cos^2 2\theta_{hkl}} - 1} \quad \textbf{(16)}$$

$w(2\theta)$ is a weighting function, which determines the extent of asymmetry (**Figure 10**) at different distances from the theoretical peak center, $2\theta_{hkl}$, and is defined as:

$$w(2\theta) = \begin{cases} 0 & 2\theta \leq 2\theta_{\min} \text{ or } 2\theta > 2\theta_{hkl} \\ H + S - h(2\theta) & 2\theta_{\min} \leq 2\theta \leq 2\theta_{\mathrm{infl}} \\ 2 \cdot \min(H, S) & 2\theta_{\mathrm{infl}} \leq 2\theta \leq 2\theta_{hkl} \end{cases} \quad \textbf{(17)}$$

where the two angles $2\theta_{\min}$ and $2\theta_{\mathrm{infl}}$ are:

$$2\theta_{\min} = 2\theta_{hkl} - \cos^{-1}\left[\cos 2\theta_{hkl} \cdot \sqrt{\left(\frac{H+S}{L}\right)^2 + 1}\right] \quad \textbf{(18)}$$

$$2\theta_{\mathrm{infl}} = 2\theta_{hkl} - \cos^{-1}\left[\cos 2\theta_{hkl} \cdot \sqrt{\left(\frac{H-S}{L}\right)^2 + 1}\right] \quad \textbf{(19)}$$

The final asymmetric pseudo-Voigt function is acquired by convolution of the pseudo-Voigt function with the slit intercept function:

$$pV_{\mathrm{FCJ}}(2\theta) = \int_0^{2\theta_{min}} pV(2\theta - \tau)D(\tau)d\tau = \frac{L^2 \sin 2\theta_{hkl}}{4HS} \int_0^{2\theta_{min}} \frac{w(\tau)pV(2\theta - \tau)}{h(\tau)\cos(2\theta_{hkl} - \tau)}d\tau \quad \textbf{(20)}$$

Analytical solution of this convolution integral is not possible, thus a numerical integration was implemented for X-rays by Finger *et al.*[98] Although this formula is complicated at first sight, only two refinable parameters are required to successfully describe the peak asymmetry, S/L and H/L, which are usually equal in well-designed instruments. A notable observation is the fact that the weighting function, $w(2\theta)$, shifts the peak center from $2\theta_{hkl}$ to a slightly lower angle (**Figure 10**).

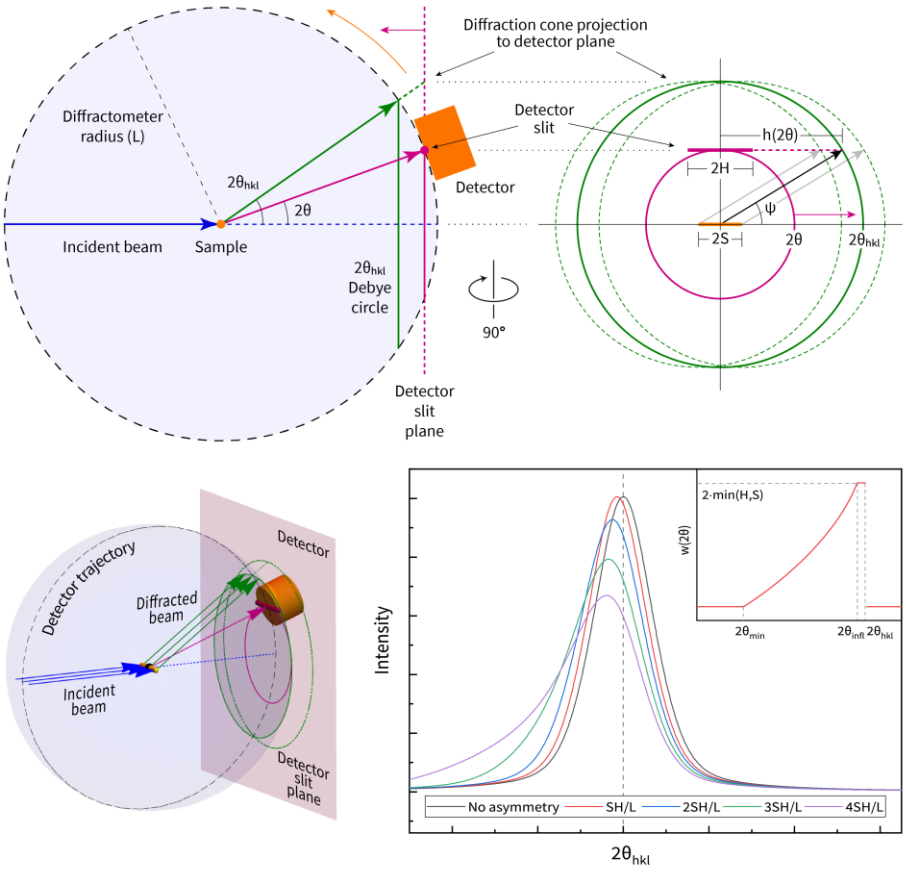

Figure 10. Peak asymmetry of diffraction lines is caused by the finite detector slit length and sample illumination length. The overall diffractometer geometry, as described by van Laar and Yelon[97], is shown on the left. The projection of the diffraction cone to the detector slit plane is shown on upper right. Increasing amounts of asymmetry, defined as increasing values for *S/L* and *H/L*, result in progressively larger angular dislocations of the peak center (bottom right). The weighting function, *w*(2*θ*), is shown as an inset.

In Rietveld analysis, background parameters of the diffraction patterns are also determined. Background originates mainly from an amorphous phase in the sample and the diffuse scattering of the soluble substance surrounding the crystals, and secondarily from the diffusion of radiation caused by air, as well as from the sample holder (glass or kapton capillary). To calculate the background contribution there are many relationships used, the most common of which being a polynomial function of the form:

$$B(2\theta) = \sum_{n} b_n (2\theta)^n \tag{21}$$

where b_n are refinable background parameters. Different number of terms are used depending on the complexity of the background.

The exact determination of background is a significant step which has to be performed accurately in a Rietveld refinement, since the atomic thermal parameters (B_{iso} or U_{iso}) are strongly correlated with it. Especially at high 2θ angles, when the angular resolution is lower and the peak overlap is higher, the background extraction can be more complicated. Therefore, the background is initially estimated by the user and subsequently subjected to refinement along with the rest of the parameter set.

Lastly, but most importantly, peak intensities are determined by the crystal structure itself. Each atom in the molecular structure is described by three positional parameters, *x*, *y*, *z*, and an atomic displacement parameter (ADP), which determines the thermal motion of each atom, or how localized the electron density is around the *x*, *y*, *z* position. Once all atomic positions have been refined, one then usually adds isotropic atomic displacement parameters ($B_{\mathrm{iso}} = 8\pi^2 U_{\mathrm{iso}}$) for individual atoms to the refinement, starting with atoms that have the greatest scattering power. In more complex structures it may be necessary to assign a single ADP value to a group of atoms with similar mass and coordination[20]. The four positional and ADP parameters, combined with the unit-cell parameters and symmetry, determine the structure factors and, therefore, the total peak intensity.

Further modifications to peak intensity can arise from sample's preferred orientation, which is typically not observed for protein sample, since the microcrystals are in a slurry, but also from the solvent content within the crystals. Macromolecular crystals contain 20-80% solvent between the macromolecular lattice, mainly water, but also other components of the crystallization 'cocktail', typically called 'mother-liquor'. Although this solvent is highly mobile and lacks any crystalline structure, pockets of solvent are periodically present in the crystal lattice[99]. Thus, there is an average electron density ($\sim 0.3 - 0.4\ e/\text{Å}^3$) resulting from all solvent molecules, which follows the same order as the macromolecular lattice itself and modifies the structure factor values. The solvent content density is accounted for using the exponential scaling model, which is a direct application of the Babinet's principle to the calculated structure factors:

$$|F_{\mathrm{C}}| = |F_{\mathrm{P}}| - A_{\mathrm{S}}|F_{\mathrm{P}}| e^{-8\pi^2 B_S \frac{\sin^2\theta}{\lambda}} \tag{22}$$

where $|F_{\mathrm{P}}|$ are the amplitudes of the protein structure factors. The solvent structure factors are assumed to be proportional to the protein structure factors, while opposite in phase. A_S reflects the ratio of the solvent electron density to the protein electron density, whereas B_S is proportional to the data resolution. The exponential scaling model can be extended to modify the atomic structure factors, f_0, using:

$$f = f_0 - A_S e^{-8\pi^2 B_S \frac{\sin^2\theta}{\lambda}} \tag{23}$$

where A_S and B_S are considered refinable parameters and are varied independently for each pattern. The assumption that solvent structure factors are corresponding to protein structure factors, ultimately

depends on data resolution. The approximation holds true up to medium resolution (~ 5 Å). It was originally applied to SCXRD by Moews and Kretsinger[100] and later applied to XRPD[11,28,52]. Although more sophisticated masking methods for solvent contribution estimations are now available for SCXRD[99], this approach is accurate enough for XRPD, since the resolution is typically lower.

All refinable parameters are varied either simultaneously or individually in a least square minimization algorithm, where a simulated profile, generated based on the values of these parameters, is compared against the experimental profile, and systematically altered until convergence is reached.

Stereochemically-restrained Rietveld refinement

Data-to-parameters ratio

In order to achieve a stable structure refinement, the data-to-parameters ratio (*r*) must be high:

$$r = \frac{n}{p} \tag{24}$$

where *n* is the total number of reflections in a given *d*-spacing range and *p* is the number of refinable parameters. The number of unique reflections up to a resolution limit of $d_{\min}$ can be approximated by[73]:

$$N_{\text{unique}} \cong \frac{2\pi}{3z} \frac{V_{\text{cell}}}{d_{\min}^3} \tag{25}$$

where *z* is the general position multiplicity for a given point group and V_{cell} the unit-cell volume. The number of unique reflections is strongly affected by *z*, which ranges from 1, in the triclinic *P*1 case, up to 96, for example in the cubic *F*432. Of course, the total number of observed reflections (*n*) is further restricted by the sample and instrumentation properties. Proteins crystals have weak scattering power and contain large amounts of amorphous solvent, thus high 2θ reflections have low intensity and are commonly lost in background noise.

For each atom in the structure, a total of (at least) four parameters are refined: three positional (*x*, *y*, *z*) and one atomic displacement parameter (U_{iso} or B_{iso}). Proteins contain tens to hundreds of thousands of atoms in the asymmetric unit, thus the number of refinable parameters is quite large. In practice, roughly 100 additional parameters are required for bulk solvent correction, background refinement, peak shape and scaling.

Even in the ideal scenario of a small protein (few atoms) and low symmetry (many unique reflections), the data-to-parameters ratio is low. For example, PDB entry 1lzt corresponds to triclinic lysozyme, with 1221 atoms in the asu. To achieve the minimum required ratio of $r = 1$, we need four times the number of atoms, in reflections, i.e. 4884 unique reflections, which translates to a required diffraction limit of 2.24 Å (**Figure 11**). If the crystal diffracts below 2.24 Å, the system is undetermined, and refinement is impossible. However, even when $r > 1$, the refinement is quite unstable and commonly converges to false minima [101], due to the low degree of overdetermination and the non-linear nature of the equations (e.g. trigonometric functions in structure factors). The refinement approaches numerical stability[73], when $r > 10$, which translates to a required diffraction limit of 1.04 Å. The required atomic or near atomic resolution in proteins is rarely feasible, while the situation is worsened at higher symmetries or larger proteins.

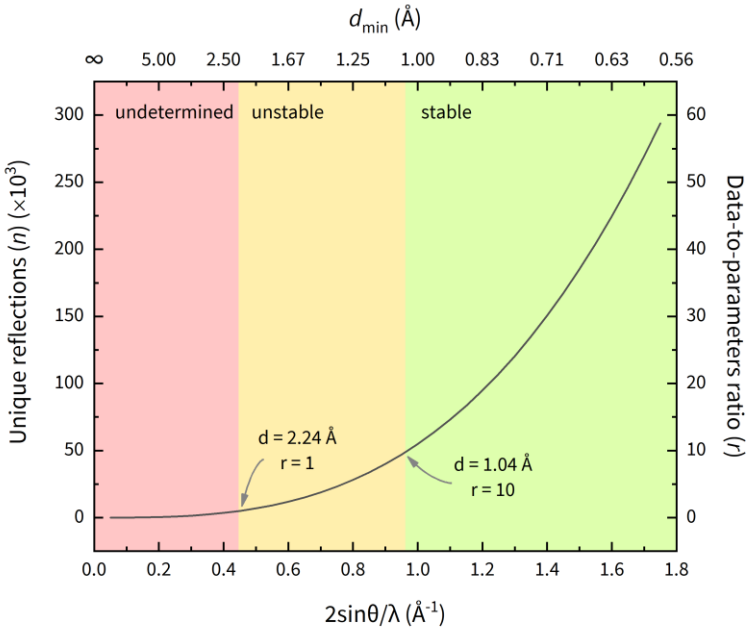

Figure 11. The number of unique reflections and data-to-parameters ratio (r) with regards to d-spacing resolution for triclinic lysozyme (PDB entry: 1lzt). When $r < 1$, the system is underdetermined. At $r > 1$, refinement is feasible, but only if $r > 10$ the refinement is stable.

Stereochemical restraints

The low data-to-parameters ratio in macromolecular crystallography is present both in SCXRD and XRPD and should not be confused with peak overlap. Peak overlap results in ambiguity in peak intensities, but, given an equal d-spacing resolution, the same number of reflections will be extracted from a SCXRD and XRPD pattern, although the intensity of the latter could be highly inaccurate. To overcome this innate limitation of protein diffraction, additional data must be provided. It is fortuitous that atoms in proteins are organized in amino acids: small molecules of approximately 10-30 atoms, with well-studied stereochemistry. Thus, although the overall structure of a protein might be unknown, properties about the local structure of each amino acid are known. Proteins are treated as a collection of amino acid rigid bodies, while the movement of atoms therein is heavily restrained.

The prior knowledge of amino acid stereogeometry is vast and ranges from bond lengths and angles, to torsion angles and coplanarity of atoms. In a stereochemically restrained refinement, atoms are allowed to move, but only in a subset of real-space, as defined by the restraints. The allowed (or ideal) values for the restraints follow sharp distributions (i.e. low standard deviations) and were first tabulated by Engh and Huber[102], based on high-quality small molecule structures from the Cambridge Structural Database.

The stereochemical restraints are introduced to the least-square minimization function, M, as added terms, therefore they are treated, mathematically, as data. The immediate consequence is that the data-to-parameters ratio is significantly improved and structure refinement is rendered feasible even for medium resolution (~3 Å) diffraction data. Restraints are introduced for each substructure, j, of each amino acid, whether that is a pair or a group of atoms, depending on the restraint type. The final minimization function used in refinement is the following[11]:

$$\begin{aligned} M = \sum w_{yi}(y_{oi} - y_{ci})^2 &+ f_d \sum w_{dj}\left(d_{oj} - d_{cj}\right)^2 + f_a \sum w_{aj}\left(a_{oj} - a_{cj}\right)^2 \\ &+ f_\tau \sum w_{\tau j}\left(\tau_{oj} - \tau_{cj}\right)^2 + f_p \sum w_{pj}\left(-p_{cj}\right)^2 + f_x \sum w_{xj}\left(x_{oj} - x_{cj}\right)^2 \\ &+ f_v \sum w_{vj}\left(v_{oj} - v_{cj}\right)^4 + \cdots \end{aligned} \qquad \textbf{(26)}$$

where y_i is the pattern intensity at position i, d_j is a bond distance, a_j is a bond angle, τ_j is a torsion angle, p_j is the deviation from the best plane, x_j is a chiral volume, v_j is a non-bonded atoms 'bump'. The subscripts o and c correspond to observed and calculated values of the respective quantity. Calculated values for stereoconfiguration are provided by the restraints. Each squared difference term is weighted by a w term, which for intensities y_i equals to $\sigma_{y_{oi}}^{-2}$, and for a restraint R_j equals to $\sigma_{R_{ci}}^{-2}$ (i.e. the standard deviation of the calculated or ideal value). Lastly, each summation term, which corresponds to a suite of terms of the same nature (e.g. bond angles), is further weighted by a f term. The values of the f terms must be carefully chosen to prevent the dominance of a suite of terms in the refinement and are progressively tweaked throughout the refinement[103].

Depending on the system under investigation, additional restraints may be added to the minimization function, with regards to non-crystallographic symmetry, experimental phases or atomic displacement parameters, to name a few[104].

Each of the restraints refers to a different aspect of the overall stereoconfiguration, and, therefore, is calculated differently, and in some cases, counterintuitively. A few of the most important restraints are shown in **Figure 12** and will now be discussed.

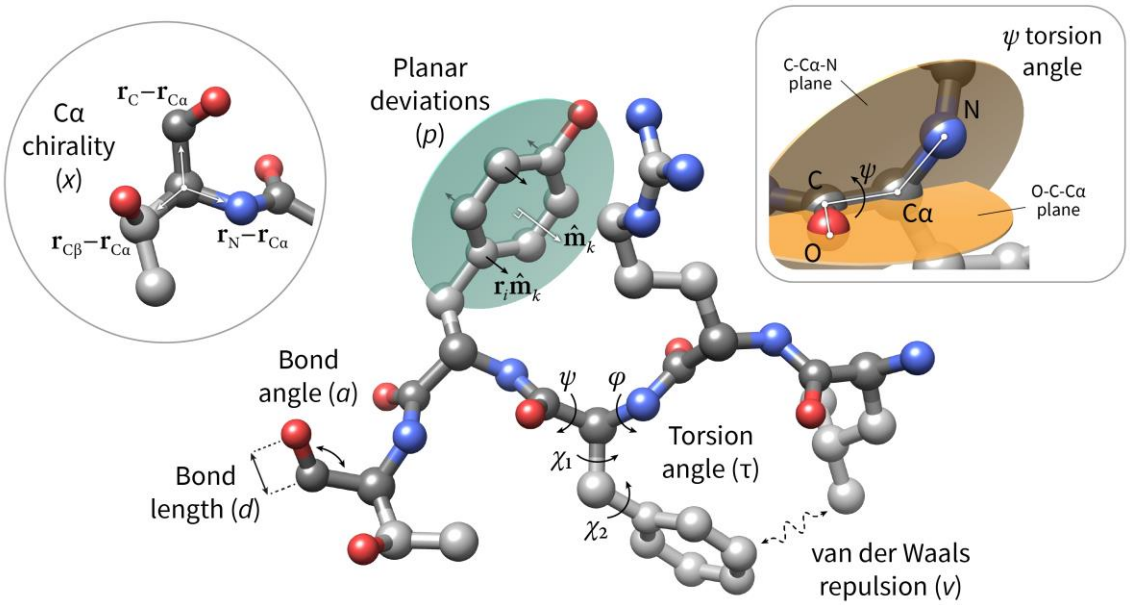

Figure 12. Overview of the main restraints imposed to amino acids atoms movement during refinement. Carbon, nitrogen and oxygen atoms are colored gray, blue and red, respectively. Backbone carbon atoms are colored darker.

Distance, d_{12}, between two bonded atoms 1, 2, is defined as the magnitude of the vector that connects atom 1 to 2:

$$\mathbf{d}_{12} = \mathbf{r}_2 - \mathbf{r}_1, \qquad d_{12} = |\mathbf{r}_2 - \mathbf{r}_1| \tag{27}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the atomic position vectors for atoms 1 and 2, respectively. Bond angles (a) and torsion angles (or dihedral angles, τ) restraints can be calculated as distances between neighboring atoms 1, 3 and 1, 4, respectively:

$$a_{13} = |\mathbf{r}_3 - \mathbf{r}_1| \tag{28}$$

$$\tau_{14} = |\mathbf{r}_4 - \mathbf{r}_1| \tag{29}$$

Alternatively, a more intuitive definition is using the actual angles, which can be calculated from following formulas[25]:

$$a_{13} = \cos^{-1}\left[\frac{\mathbf{d}_{21} \cdot \mathbf{d}_{23}}{|\mathbf{d}_{21}| \cdot |\mathbf{d}_{23}|}\right] \tag{30}$$

$$\tau_{14} = \cos^{-1}\left[\frac{(\mathbf{d}_{21} \cdot \mathbf{d}_{32}) \times (\mathbf{d}_{32} \cdot \mathbf{d}_{43})}{|\mathbf{d}_{21}||\mathbf{d}_{32}|^2|\mathbf{d}_{43}| \cdot \sin a_{13} \cdot \sin a_{24}}\right] \tag{31}$$

Torsion angles are of special importance in protein structures. Each torsion angle around a rotatable bond (2-3) is defined for a set of four bonded atoms (1-2-3-4). A unique plane can be defined for each of the two subsets of three connected atoms (1-2-3 and 2-3-4). The angle between the two planes is called torsion angle and, physically, corresponds to rotation around the bond between atoms 2, 3[105]. There are two special torsion angles, called φ and ψ, that are formed by backbone atoms. Specifically, atoms 2 and 3 in a φ angle are the amino group nitrogen atom and the main Cα carbon, while in ψ angle they are the main Cα carbon and the carboxyl group carbon atom. Different pairs of values result in different secondary structure elements of the protein. All along the refinement process, the quality of φ, ψ angle pairs is evaluated using Ramachandran plots, which describe the energetically allowed combinations[106].

The regions on the plot (**Figure 13**) with the highest density are the so-called 'allowed' or 'low-energy' regions. Some values of φ and ψ are forbidden, since they will bring the atoms too close to each other, in a steric clash. For a high-quality and resolution experimental structure these regions are usually empty or almost empty: very few amino acid residues in proteins have their torsion angles within these 'disallowed' regions. Exceptions to this rule result in some strain in the polypeptide chain. In such cases additional interactions will be present to stabilize the structure. Normally such conformations have functional significance and may be conserved within a protein family[107].

In *GSAS*, coupled torsion angles pairs (φ, ψ or χ_1, χ_2) are restrained by pseudopotentials that are generated by summation of 3-7, two-dimensional Gaussian distributions[28]. These pseudopotentials reflect the energetically favored and forbidden regions for the two torsion angles. Each Gaussian distribution requires six coefficients, which were originally determined by fitting to distributions of paired torsion angles extracted from high-resolution protein structures.

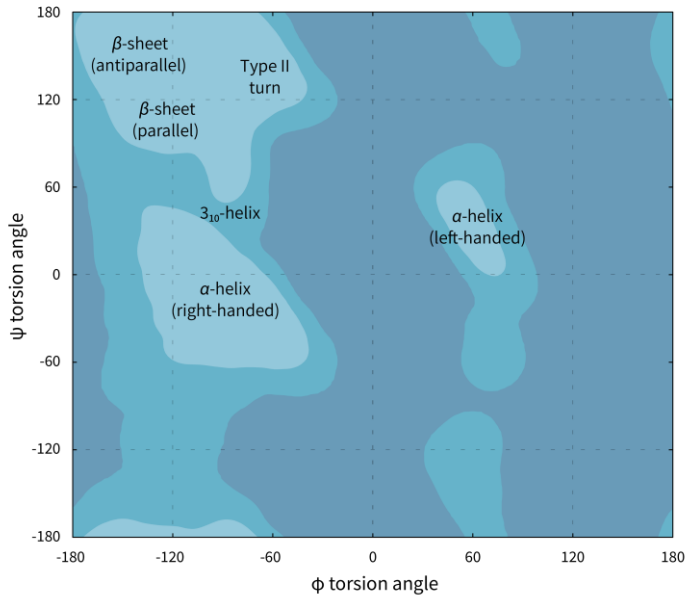

Figure 13. Ramachandran plot of the energetically favored φ, ψ torsion angle pairs. Brightly colored regions are highly favored, while darker colored regions are not permissible. Different combinations of φ, ψ angles result in distinct secondary structure elements.

Another important aspect of certain amino acids is the coplanarity of specific groups, e.g. phenyl, indole, imidazole or guanidyl rings, carboxyl and amide side chains, as well as the peptide bonds. Planar restraints are relatively strong and are implemented as deviation of the individual atoms from their common least squares plane. For each atom, *i*, in a planar group, the deviation, p_i, is calculated as follows[73]:

$$p_i = \hat{\mathbf{m}}_k \mathbf{r}_i - d_{i,k} \tag{32}$$

where $\hat{\mathbf{m}}_k$ is the unit vector normal to the least-square plane *k*, $\mathbf{r}_i$ is the atomic position vector of atom *i* and $d_{i,k}$ the scalar distance of atom i from the least-square plane *k*. Alternatively, the dot product $\hat{\mathbf{m}}_k \mathbf{r}_i$ corresponds to the projection of vector $\mathbf{r}_i$ to $\hat{\mathbf{m}}_k$.

Another crucial aspect of amino acid stereoconfiguration is their chirality. The central α-carbon atom of an amino acid is asymmetric or chiral, because it is tetrahedrally connected to four distinct chemical groups: a hydrogen atom, a carboxyl group, an amino group and a variable side chain (-R), which is characteristic of each amino acid. The distance restraints discussed earlier (bond angles, torsion angles) are insensitive to handedness, thus additional chirality restraints must be imposed to preserve the correct handedness during refinement. For each Cα atom, the chiral volume is defined as[101]:

$$x_{C\alpha} = (\mathbf{r}_N - \mathbf{r}_{C\alpha}) \cdot \left[(\mathbf{r}_C - \mathbf{r}_{C\alpha}) \times (\mathbf{r}_{C\beta} - \mathbf{r}_{C\alpha})\right] \tag{33}$$

where $\mathbf{r}_i - \mathbf{r}_{C\alpha}$ is a vector starting from Cα and pointing towards atom *i*. The sign of the chiral volume depends on the handedness of Cα (positive for L-amino acids and negative for D-amino acids), while the magnitude equals to the volume of the parallelepiped formed by the three vectors.

Lastly, apart from stereochemical restraints between bonded atoms, non-bonded atoms contacts should also be taken into consideration. The interactions are characterized by Lennard-Jones or Buckingham potential energy functions $U(d)$, which are strongly repulsive when the distance, *d*, between two atoms is small, while slightly attractive when *d* is large (**Figure 14**). A Lennard-Jones (LJ) potential function follows the expression[108]:

$$U_{LJ}(d) = \varepsilon \left[\left(\frac{d_{\min}}{d} \right)^{12} - 2 \left(\frac{d_{\min}}{d} \right)^{6} \right] \tag{34}$$

where ε is the potential well depth and $d_{\min}$ is the distance where the potential has the minimum value, i.e. $U(d_{\min}) = \varepsilon$. The d^{-12} term is responsible for repulsion and the d^{-6} term for attraction. The LJ function is quite computationally expensive to calculate over each potential contact, and since only the repulsive term is needed to prevent atoms from bumping into each other, it is approximated by a quadratic equation[101]:

$$U(d_{ij}) - U(d_{ij(\min)}) \cong \begin{cases} \frac{1}{\sigma_{ij}^{2n}} \left(d_{ij} - d_{ij(\min)}\right)^{2n} & \text{for } d_{ij} \le d_{ij(\min)} \\ 0 & \text{for } d_{ij} > d_{ij(\min)} \end{cases} \tag{35}$$

where d_{ij} is the distance between atoms *i*, *j*, $d_{ij(\min)}$ is the distance where the potential is minimum, depending on *i*, *j* nature and type of contact and σ_{ij} is a weighting factor, also dependent on the nature of *i*, *j*. The exponent *n* typically equals to 2, when *d* is the range of $d_{\min} - 2$Å and $d_{\min}$. Values for σ and $d_{\min}$ are tabulated for different pairs of atoms, to further reduce calculation efforts. Therefore, only when the atoms come closer than a predefined distance $d_{\min}$, will this 'anti-bumping' restraint come to play.

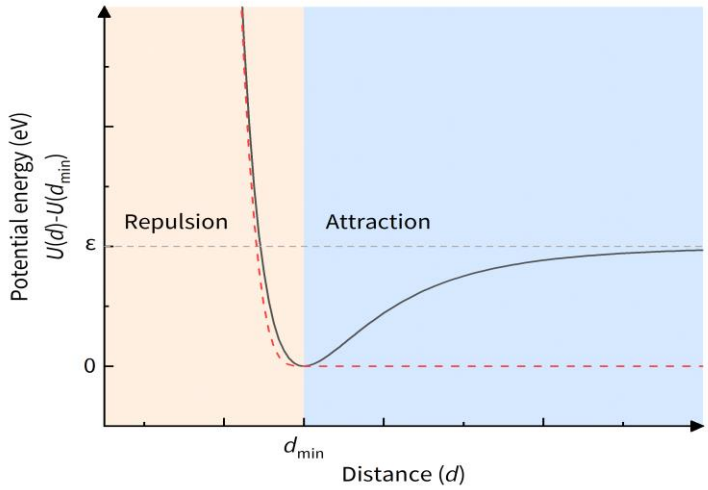

Figure 14. A Lennard-Jones potential function (gray), shifted upwards by ε (i.e. the well depth) and its quadratic approximation (red) that is used to prevent bumping of non-bonded atoms. The 'anti-bumping' function is slightly offset to the left, for clarity.

Overall, stereochemical restraints are required both in macromolecular SCXRD and XRPD, however they are more important to the latter, since the quality of the extracted structure factors is lower. Restraints not only act as pseudo-observations, but additionally prevent refinement from converging to non-logical solutions, i.e. conformations that are not physically possible. As mentioned earlier, the restraint weighting factors, *f*, are initially set to high values, to downweight the reflections contribution, and are progressively softened to improve the agreement factors[103].

Flexible rigid-body approach

More recently, a new approach was introduced for macromolecular XRPD data: the flexible rigid-body refinement (FRB)[27,30]. Instead of restraining the movement of each atom in a protein, each amino acid is considered as a rigid body, which can only pivot around the Cα carbon atom. Only six parameters are required for each amino acid: three positional, for the location of the Cα carbon, and three rotational parameters (described in terms of a quaternion), for the orientation of the amino acid. Additional parameters are introduced for the torsion angles (thus the name 'flexible') both for backbone and side chains, while bond lengths and angles are kept fixed at ideal values. The quaternion description of the residue orientation is preferred compared to a triplet of Eulerian angles, since it does not suffer from 'gimbal lock', i.e. the loss of one degree of freedom.

FRB refinement is especially useful for medium resolution data or unit-cells with very high symmetry, where the data-to-parameters ratio is so low, that not even stereochemically-restrained refinement is feasible, since it reduces the number of refinable parameters to a third of those required for free-atom refinement. In addition, this approach is much faster, since fewer parameters are optimized concurrently, and eases the interpretation of poor-quality data.

Model evaluation

Agreement factors

Throughout during Rietveld refinement, it is crucial to quantify the quality of the fit between the model and the data. Several reliability (*R*) indices exist for this purpose, each for checking a slightly different aspect of the fit. The profile R-factor (R_p) is calculated as follows[20]:

$$R_{\mathrm{p}} = \frac{\sum |y_o(2\theta) - y_c(2\theta)|}{\sum |y_o(2\theta)|} \quad \textbf{(36)}$$

where y_o and y_c are the observed and calculated intensities at position 2θ. R_{p} basically compares the differences between the observed and calculated profiles. Of course, differences in peak intensities have more importance than in non-peak regions, which gives rise to the weighted R-factor[20]:

$$R_{\mathrm{wp}} = \sqrt{\frac{\sum w(2\theta)[y_o(2\theta) - y_c(2\theta)]^2}{\sum w(2\theta)[y_o(2\theta)]^2}} \quad \textbf{(37)}$$

where the weight, *w*, is defined as:

$$w(2\theta) = \frac{1}{\sigma^2_{y_o}(2\theta)} \quad \textbf{(38)}$$

R_{wp} factor allows for a more meaningful comparison of the profiles. It should be noted that R_{wp} is sensitive to the background contribution. Patterns where a large portion of the total intensity arises from the background alone, tend to give small values of R_{wp}, even if the model is quite incorrect[109,110].

Ideally, the final R_{wp} should approach a target value, which is called expected R-factor (R_{exp}), and depends on data quality terms, such as the counting statistics[109]:

$$R_{\mathrm{exp}} = \sqrt{\frac{n - p + r}{\sum w(2\theta)[y_o(2\theta)]^2}} \quad \textbf{(39)}$$

where n is the number of observed data points, p the number of refinable parameters and r the number of restraints imposed on the data.

During refinement, the least squares procedure minimizes the following quantity:

$$M = \sum w(2\theta)[y_o(2\theta) - y_c(2\theta)]^2 \quad \textbf{(40)}$$

or its extended counterpart **(26)**. This quantity is also called χ^2 and should progressively decrease as the model produces a better agreement with the experimental data, although small increases can occur when refining correlated parameters[110]. χ^2 estimates how closely the model explains the observations. Because the value of χ^2 has no meaning on an absolute scale, it is often scaled by the degrees of freedom of the system, which equal to the number of observed datapoints (n), minus the refinable parameters (p), plus the number of restraints (r), since they are considered as additional datapoints. This gives rise to the reduced χ^2 factor (χ^2_r), which can also be calculated from R_{wp} and R_{exp}[20]:

$$\chi^2_r = \frac{\chi^2}{n - p + r} = \left(\frac{R_{\mathrm{wp}}}{R_{\mathrm{exp}}}\right)^2 \quad \textbf{(41)}$$

In an ideal scenario, χ^2_r should be equal to 1, however in practice, it is always larger than 1. χ^2_r should never drop below 1. If a refinement ends with $\chi^2_r < 1$, then either the standard uncertainties (*w*) in the observed data are overestimated, or the model is fitting noise[110].

In SCXRD a similar term is used, called goodness of fit, (GOF):

$$\mathrm{GOF} = \sqrt{\chi^2_r} \quad \textbf{(42)}$$

Finally, a non-statistical factor, R_{F^2}, is particularly useful for evaluating the agreement between the structure factors derived by the structural model and experimental data and is defined as[20]:

$$R_{F^2} = \frac{\sum\left(F_{oj}^2 - F_{cj}^2\right)}{\sum F_{oj}^2} \quad \textbf{(43)}$$

For the calculation of this factor, the values of the calculated structure factors (F_c) are derived directly from the crystallographic model, while the values of the experimental structure factors (F_o) are estimated via Rietveld refinement. The accuracy of the estimation heavily depends on peak overlap. In general, as the structural model improves and approaches the experimental data, experimental structure factors estimation becomes more accurate, and R_{F^2} decreases towards 0% in cases of absolute match with the calculated structure factors.

In SCXRD experiments, during the evaluation of agreement between the experimental data and the structural model, R and R_{free} factors are used[111]:

$$R = \frac{\sum\left|\left|\mathbf{F}_{oj}\right| - \left|\mathbf{F}_{cj}\right|\right|}{\sum\left|\mathbf{F}_{oj}\right|} \quad \textbf{(44)}$$

where R_{free} is calculated from a subset of randomly selected reflections, *j*, that are not used in refinement, while the rest, which are used in refinement are employed for calculating *R*. The goal here is for the two factors to be in agreement at the end of the refinement.

R and R_{free} are valuable factors in SCXRD, however, they cannot be applied to XRPD due to peak overlap. The values of observed structure factors cannot be derived directly from the pattern itself since it requires the decomposition of overlapping reflections. In turn, this decomposition requires a structural model, thus, making the derived set of structure factors inherently biased. Therefore, if one would attempt to calculate R_{free} from a set of random reflections, it would always be in agreement with R, at any stage of Rietveld refinement. In addition, if the total random reference reflections include two reflections that overlap precisely, then, statistically, the R_{free} indicator loses its significance and is no longer reliable[85].

Structure evaluation

Although low agreement factors are a great indication the refinement is going the right way, additional considerations must be considered with regards to the overall stereochemical quality of the resulting model. In general, a balance between good agreement factors and good stereochemistry is required.

In XRPD model bias is manifested both in calculated phases and observed structure factors, rendering any type of electron density map biased, to some degree. To reduce the phase bias, total omit maps are generated using *SFCHECK*[112], in order to evaluate the fit of the structure to the electron density. Total omit maps[113,114] are generated by combining omit maps for subregions of the asymmetric unit, which are calculated by flattening the electron density inside specific regions ($\rho_{region} = \langle\rho_{cell}\rangle$) and calculating the reverse Fourier transform of the remaining electron density. Phases for each subregion map are derived by the contributions of all atoms in the asu, except those in the subregion, therefore they are not biased by the atoms present within the subregion. The process is repeated until the entire asu has been sampled. This approach tends to reduce model bias in the final electron density map and enables investigation of incorrect placement of water molecules or amino acid configurations.

Unexplained ‘blobs’ of electron density typically indicate missing atoms in the structure, such as water molecules or other small molecules, such as ions or ligands, or even incorrect residue side chains placement. The choice of the appropriate molecule depends on the ‘blob’ shape and size, while the available options are derived by a list of substances used in the crystallization ‘cocktail’. Addition

of such molecules is typically performed in *Coot*[115], and new maps are calculated to evaluate whether they can explain the 'blobs'.

Once all amino acids exhibit a reasonable fit to the electron density and an initial structural model of the protein has been determined, including water or other small molecules, evaluations of the overall stereochemistry are performed, typically in *PROCHECK*[116]. Depending on how good the electron density resolution is, harsher restrictions are applied during these evaluations. Although stereochemical restraints are imposed during the refinement, the lowered weight factors used at the late stages of refinement can potentially allow disfavored configurations of amino acids. At that stage, energy minimizations, typically with *Swiss-PdbViewer*[117], enable more automated improvements to the overall structure, instead of manual tweaking of individual amino acid side chains, although these are inevitably required for some regions of the structure.

Lastly, the model stereochemistry is evaluated with *MolProbity*[118,119], which analyzes a variety of structural features, for example Ramachandran outliers, Cβ deviation, energetically unfavorable rotamers, atom contacts and clashes. A detailed report is outputted, including an overall 'MolProbity score', which is also provided as a percentile score, relative to the cohort of deposited PDB structures, within ± 0.25 Å of the model's resolution.

Another interesting type of stereochemical evaluation is performed with *Errat*[120]. *Errat* is based on the observation that different atom types are non-randomly distributed, with respect to each other, in protein structures, therefore incorrect regions of the structure could be detected on this basis. It analyzes clusters of non-bonded atoms, looking for deviations from known interaction patterns, as derived by high-quality protein structures, and outputs an 'error score' per residue. Finally, if the model meets the required stereochemical criteria, it is deposited to the Protein Data Bank[121] (PDB), accompanied by the extracted observed structure factors.

Case studies

The first protein structure refined via Rietveld was metmyoglobin, at 3.3 Å[11]. In this paper, Von Dreele introduced the concept of stereochemically-restrained Rietveld refinement to protein powder data, coupled with solvent correction and Ramachandran pseudopotentials restraints. Later, the multipattern component was introduced in hen egg-white lysozyme studies, which, combined with area detectors (**Figure 15**) instead of crystal analyzer detectors, led to a significant improvement in data quality and allowed the placement of water molecules in the model[9]. More recently, the flexible rigid-body approach was employed for successful refinement of bovine insulin[30].

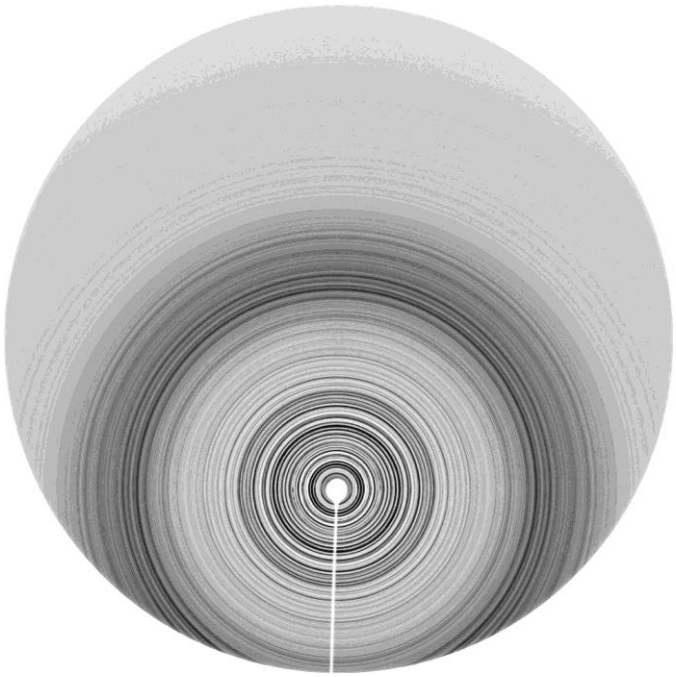

Figure 15. XRPD pattern of hen-egg white lysozyme microcrystals collected on beamline 1-BM at the Advanced Photon Source [9]. Approximately 9000 unique reflections are present in the detected range (55.66 Å - 1.85 Å).

As it was shown early on, protein powder diffraction could be used for identification of ligand binding to protein molecules[28,54]. Protein-ligand complexes are the first essential step for structure-based drug design. Detection of ligand binding typically requires high-resolution diffraction data, thus these early studies demonstrated the applicability of powder diffraction for this purpose. In cases where the data resolution does not allow for structure solution and refinement, ligand binding can be deduced by the presence of unique crystalline polymorphs, thus indexing followed by Pawley or Le Bail refinement allow for accurate characterization of the lattice dimensions, until collection of higher resolution data. This strategy has been employed successfully for human insulin, where 10 crystalline polymorphs have been identified with 9 ligands, resulting in ~30 unique combinations, as reviewed extensively elsewhere[8].

Phasing for protein powder data was initially focused on molecular replacement. A crucial step in this process is the accurate extraction of intensities. Since intensity extraction is inherently biased in XRPD, due to peak overlap, molecular replacement does not always propose the correct solution. In the cases of hen and turkey egg-white lysozyme[34,52], the correct solution was the second-best solution proposed by *MOLREP*, even though the sequence identity was almost 100%. However, structure solution has been performed successfully using starting models with lower sequence identity from different species[74]. Perhaps the most notable case is that of the SH3 domain of ponsin, which employed a starting model with 38% sequence identity[29]. This study demonstrated that powder diffraction could be used not only for phase identification and derivative structure refinement, but also for novel structure solution, since no single-crystal structure was available at the time.

In cases where no starting models are available, isomorphous replacement has also been employed successfully for structure solution using powder data. The feasibility of the approach was initially demonstrated by detection of heavy atoms in difference maps using xenon derivatives of HEWL[58] and later gadolinium derivatives for HEWL and uranium derivatives for porcine pancreatic elastase[37]. Eventually, the first medium resolution structure solution was achieved via MIR using a gadolinium and holmium HEWL derivative[33]. MIR allowed resolution of the two-fold phase ambiguity and identify the correct chirality of the protein molecules. However, due to the medium resolution of the data (5.3 Å), only selected features of the protein secondary structure (α-helices) could be determined.

Although structure solution and refinement of protein structures using powder data has proven to be feasible, the successful applications in the literature are quite limited, due to the complexity of data analysis and the severe lack of information in the observed pattern resulting from peak overlap. To date, only 19 macromolecular structures have been solved to adequate resolution and stereochemical quality to meet the strict deposition standards of the Protein Data Bank. However, powder diffraction has been employed extensively for preliminary structural information, for instance phase identification in a variety of microcrystalline protein precipitates[42,43,122–128], apart from the systematic studies of human insulin polymorphism[8]. On the other hand, the situation is quite more promising for oligopeptides. Due to their significantly reduced size, compared to macromolecules, peptide structures are more commonly solved and refined from powder diffraction data, when single crystals are not available[31,32,129–131].

Overall, Rietveld refinement of macromolecules is a complex process that requires a strong background on both powder diffraction and structural biology. Since its first application to metmyoglobin, the Rietveld method has progressively been adapted to tackle efficiently the sheer complexity of macromolecular structures. Concurrently, development of advanced detectors and X-ray optics, in combination with novel data collection strategies, have dramatically improved the quality of powder diffraction data, mitigating notably the peak overlap and radiation damage observed in the patterns. However, there is still an unmet instrumental challenge on how to fully exploit the inherently sharp protein peaks, while mitigating radiation-related sample degradation. Further improvements are essential to enhance and automate data analysis, which currently is quite complex and time-consuming. Development of dedicated phasing software for powder diffraction is especially needed. However, powder diffraction has proven to be an important, yet complimentary, tool for structural biology.

Acknowledgement

This research is funded by …. Argonne National Laboratory's work was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Science, under Contract No. DE-AC02- 06CH11357.

References

(1) McCusker, L. B.; Baerlocher, Ch. Zeolites. In *International Tables for Crystallography Volume H: Powder diffraction*; 2019; pp 452–464. https://doi.org/10.1107/97809553602060000961.

(2) Černý, R.; Favre-Nicolin, V. Solving and Refining Inorganic Structures. In *International Tables for Crystallography Volume H: Powder diffraction*; 2019; pp 442–451. https://doi.org/10.1107/97809553602060000960.

(3) Bernstein, J.; Reutzel-Edens, S. M. Powder Diffraction and Pharmaceuticals. In *International Tables for Crystallography*, *Volume H: Powder Diffraction*; 2019; pp 767–781. https://doi.org/10.1107/97809553602060000979.

(4) Karavassili, F.; Valmas, A.; Fili, S.; Georgiou, C.; Margiolaki, I. In Quest for Improved Drugs against Diabetes: The Added Value of X-Ray Powder Diffraction Methods. *Biomolecules* **2017**, *7* (4), 63. https://doi.org/10.3390/biom7030063.

(5) Karavassili, F.; Margiolaki, I. Macromolecular Powder Diffraction: Ready for Genuine Biological Problems. *Protein Pept. Lett.* **2016**, *23* (3), 232–241. https://doi.org/10.2174/0929866523666160120152839.

(6) Margiolaki, I. Macromolecular Powder Diffraction. In *International Tables for Crystallography Volume H: Powder diffraction*; 2019; pp 718–736. https://doi.org/10.1107/97809553602060000975.

(7) Margiolaki, I.; Wright, J. P. Powder Crystallography on Macromolecules. *Acta Crystallogr. A* **2008**, *64* (1), 169–180. https://doi.org/10.1107/S0108767307043735.

(8) Spiliopoulou, M.; Valmas, A.; Triandafillidis, D.-P.; Kosinas, C.; Fitch, A.; Karavassili, F.; Margiolaki, I. Applications of X-Ray Powder Diffraction in Protein Crystallography and Drug Screening. *Crystals* **2020**, *10* (2), 54. https://doi.org/10.3390/cryst10020054.

(9) Von Dreele, R. B. Multipattern Rietveld Refinement of Protein Powder Data: An Approach to Higher Resolution. *J. Appl. Crystallogr.* **2007**, *40* (1), 133–143. https://doi.org/10.1107/S0021889806045493.

(10) Von Dreele, R. B. Protein Crystal Structure Analysis from High-Resolution X-Ray Powder-Diffraction Data. In *Methods in Enzymology*; Elsevier, 2003; Vol. 368, pp 254–267. https://doi.org/10.1016/S0076-6879(03)68014-6.

(11) Von Dreele, R. B. Combined Rietveld and Stereochemical Restraint Refinement of a Protein Crystal Structure. *J. Appl. Crystallogr.* **1999**, *32* (6), 1084–1089. https://doi.org/10.1107/S002188989901064X.

(12) Von Dreele, R. B.; Stephens, P. W.; Smith, G. D.; Blessing, R. H. The First Protein Crystal Structure Determined from High-Resolution X-Ray Powder Diffraction Data: A Variant of T3R3 Human Insulin–Zinc Complex Produced by Grinding. *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56* (12), 1549–1553. https://doi.org/10.1107/S0907444900013901.

(13) Debye, P.; Scherrer, P. Interferenzen an Regellos Orientierten Teilchen Im Röntgenlicht. I. In *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen*; Mathematisch-Physikalische Klasse; Weidmannsche Buchhandlung: Berlin, 1916; pp 1–15.

(14) Debye, P.; Scherrer, P. Interferenzen an Regellos Orientierten Teilchen Im Röntgenlicht. II. In *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen*; Mathematisch-Physikalische Klasse; Weidmannsche Buchhandlung: Berlin, 1916; pp 16–26.

(15) Rietveld, H. M. A Profile Refinement Method for Nuclear and Magnetic Structures. *J. Appl. Crystallogr.* **1969**, *2* (2), 65–71. https://doi.org/10.1107/S0021889869006558.

(16) Rietveld, H. M. Line Profiles of Neutron Powder-Diffraction Peaks for Structure Refinement. *Acta Crystallogr.* **1967**, *22* (1), 151–152. https://doi.org/10.1107/S0365110X67000234.

(17) Wiles, D. B.; Young, R. A. A New Computer Program for Rietveld Analysis of X-Ray Powder Diffraction Patterns. *J. Appl. Crystallogr.* **1981**, *14* (2), 149–151. https://doi.org/10.1107/S0021889881008996.

(18) Rietveld, H. M. The Rietveld Method. *Phys. Scr.* **2014**, *89* (9), 098002. https://doi.org/10.1088/0031-8949/89/9/098002.

(19) Klug, H. P.; Alexander, L. E. *X-Ray Diffraction Procedures for Polycrystalline and Amorphous Materials*, 2nd ed.; Wiley, 1974.

(20) Toby, B. H. Rietveld Refinement. In *International Tables for Crystallography Volume H: Powder diffraction*; 2019; pp 465–472. https://doi.org/10.1107/97809553602060000962.

(21) Khatchaturyan, A. G. Phase Transformations in Crystals Composed of Macromolecules. *Sov. Phys. - J. Exp. Theor. Phys.* **1977**, *45* (3), 601–604.

(22) Morozov, V. N.; Morozova, T. Y. Viscoelastic Properties of Protein Crystals: Triclinic Crystals of Hen Egg White Lysozyme in Different Conditions. *Biopolymers* **1981**, *20* (3), 451–467. https://doi.org/10.1002/bip.1981.360200304.

(23) McPherson, A. Protein Crystallization in the Structural Genomics Era. *J. Struct. Funct. Genomics* **2004**, *5* (1/2), 3–12. https://doi.org/10.1023/B:JSFG.0000029199.43875.92.

(24) Bowler, M. W.; Mueller, U.; Weiss, M. S.; Sanchez-Weatherby, J.; Sorensen, T. L.-M.; Thunnissen, M. M. G. M.; Ursby, T.; Gobbo, A.; Russi, S.; Bowler, M. G.; Brockhauser, S.; Svensson, O.; Cipriani, F. Automation and Experience of Controlled Crystal Dehydration: Results from the European Synchrotron HC1 Collaboration. *Cryst. Growth Des.* **2015**, *15* (3), 1043–1054. https://doi.org/10.1021/cg500890r.

(25) Larson, A. C.; Von Dreele, R. B. General Structure Analysis System (GSAS). In *Los Alamos National Laboratory Report LAUR 86-748*; UC Regents, California, USA, 2004.

(26) Toby, B. H.; Von Dreele, R. B. GSAS-II: The Genesis of a Modern Open-Source All Purpose Crystallography Software Package. *J. Appl. Crystallogr.* **2013**, *46* (2), 544–549. https://doi.org/10.1107/S0021889813003531.

(27) Von Dreele, R. B. Protein Refinement with GSAS-II. *Powder Diffr.* **2019**, *34* (S1), S32–S35. https://doi.org/10.1017/S0885715619000204.

(28) Von Dreele, R. B. Binding of N-Acetylglucosamine Oligosaccharides to Hen Egg-White Lysozyme: A Powder Diffraction Study. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61* (1), 22–32. https://doi.org/10.1107/S0907444904025715.

(29) Margiolaki, I.; Wright, J. P.; Wilmanns, M.; Fitch, A. N.; Pinotsis, N. Second SH3 Domain of Ponsin Solved from Powder Diffraction. *J. Am. Chem. Soc.* **2007**, *129* (38), 11865–11871. https://doi.org/10.1021/ja073846c.

(30) Margiolaki, I.; Giannopoulou, A. E.; Wright, J. P.; Knight, L.; Norrman, M.; Schluckebier, G.; Fitch, A. N.; Von Dreele, R. B. High-Resolution Powder X-Ray Data Reveal the T6 Hexameric Form of Bovine Insulin. *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69* (6), 978–990. https://doi.org/10.1107/S0907444913003867.

(31) Fili, S.; Valmas, A.; Spiliopoulou, M.; Kontou, P.; Fitch, A.; Beckers, D.; Degen, T.; Barlos, K.; Barlos, K. K.; Karavassili, F.; Margiolaki, I. Revisiting the Structure of a Synthetic Somatostatin Analogue for Peptide Drug Design. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2019**, *75* (4), 611–620. https://doi.org/10.1107/S2052520619006012.

(32) Spiliopoulou, M.; Karavassili, F.; Triandafillidis, D.-P.; Valmas, A.; Kosinas, C.; Barlos, K.; Barlos, K. K.; Reinle-Schmitt, M.; Gozzo, F.; Margiolaki, I. Macromolecular Powder Diffraction Using the Mythen II Detector: Case Study of a Pharmaceutical Peptide, Octreotide. **2020**, *(to be published)*.

(33) Basso, S.; Besnard, C.; Wright, J. P.; Margiolaki, I.; Fitch, A.; Pattison, P.; Schiltz, M. Features of the Secondary Structure of a Protein Molecule from Powder Diffraction Data. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (7), 756–761. https://doi.org/10.1107/S0907444910010723.

(34) Basso, S.; Fitch, A. N.; Fox, G. C.; Margiolaki, I.; Wright, J. P. High-Throughput Phase-Diagram Mapping via Powder Diffraction: A Case Study of HEWL versus PH. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61* (12), 1612–1625. https://doi.org/10.1107/S0907444905031963.

(35) Besnard, C.; Camus, F.; Fleurant, M.; Dahlström, A.; Wright, J. P.; Margiolaki, I.; Pattison, P.; Schiltz, M. Exploiting X-Ray Induced Anisotropic Lattice Changes to Improve Intensity Extraction in Protein Powder Diffraction: Application to Heavy Atom Detection. *Z. Für Krist. Suppl.* **2007**, *26*, 39–44.

(36) Wright, J. P.; Markvardsen, A. J.; Margiolaki, I. Likelihood Methods with Protein Powder Diffraction Data. *Z. Für Krist. Suppl.* **2007**, *26*, 27–32.

(37) Wright, J. P.; Besnard, C.; Margiolaki, I.; Basso, S.; Camus, F.; Fitch, A. N.; Fox, G. C.; Pattison, P.; Schiltz, M. Molecular Envelopes Derived from Protein Powder Diffraction Data. *J. Appl. Crystallogr.* **2008**, *41* (2), 329–339. https://doi.org/10.1107/S0021889808002732.

(38) McPherson, A.; Gavira, J. A. Introduction to Protein Crystallization. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2013**, *70* (Pt 1), 2–20. https://doi.org/10.1107/S2053230X13033141.

(39) Campeotto, I.; Lebedev, A.; Schreurs, A. M. M.; Kroon-Batenburg, L. M. J.; Lowe, E.; Philips, S. E. V.; Murshudov, G. N.; Pearson, A. R. Pathological Macromolecular Crystallographic Data Affected by Twinning, Partial-Disorder and Exhibiting Multiple Lattices for Testing of Data Processing and Refinement Tools. *Sci. Rep.* **2018**, *8* (14876). https://doi.org/10.1038/s41598-018-32962-6.

(40) Poulsen, H. F.; Vaughan, G. B. M. Multigrain Crystallography and Three-Dimensional Grain Mapping. In *International Tables for Crystallography Volume H: Powder diffraction*; 2019; pp 601–616. https://doi.org/10.1107/97809553602060000970.

(41) Sørensen, H. O.; Schmidt, S.; Wright, J. P.; Vaughan, G. B. M.; Techert, S.; Garman, E. F.; Oddershede, J.; Davaasambuu, J.; Paithankar, K. S.; Gundlach, C.; Poulsen, H. F. Multigrain Crystallography. *Z. Für Krist.* **2012**, *227* (1), 63–78. https://doi.org/10.1524/zkri.2012.1438.

(42) Collings, I.; Watier, Y.; Giffard, M.; Dagogo, S.; Kahn, R.; Bonneté, F.; Wright, J. P.; Fitch, A. N.; Margiolaki, I. Polymorphism of Microcrystalline Urate Oxidase from Aspergillus Flavus. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (5), 539–548. https://doi.org/10.1107/S0907444910005354.

(43) Papageorgiou, N.; Watier, Y.; Saunders, L.; Coutard, B.; Lantez, V.; Gould, E. A.; Fitch, A. N.; Wright, J. P.; Canard, B.; Margiolaki, I. Preliminary Insights into the Non Structural

Protein 3 Macro Domain of the Mayaro Virus by Powder Diffraction. *Z. Für Krist.* **2010**, *225* (12). https://doi.org/10.1524/zkri.2010.1348.

(44) Chayen, N. E. Turning Protein Crystallisation from an Art into a Science. *Curr. Opin. Struct. Biol.* **2004**, *14* (5), 577–583. https://doi.org/10.1016/j.sbi.2004.08.002.

(45) Beale, J. H.; Bolton, R.; Marshall, S. A.; Beale, E. V.; Carr, S. B.; Ebrahim, A.; Moreno-Chicano, T.; Hough, M. A.; Worrall, J. A. R.; Tews, I.; Owen, R. L. Successful Sample Preparation for Serial Crystallography Experiments. *J. Appl. Crystallogr.* **2019**, *52* (6), 1385–1396. https://doi.org/10.1107/S1600576719013517.

(46) Von Dreele, R. B. A Rapidly Filled Capillary Mount for Both Dry Powder and Polycrystalline Slurry Samples. *J. Appl. Crystallogr.* **2006**, *39* (1), 124–126. https://doi.org/10.1107/S002188980503284X.

(47) Watier, Y. Powder Diffraction Studies of Proteins. PhD Thesis, Université de Grenoble, Grenoble, France, 2011.

(48) Logotheti, S.; Valmas, A.; Trampari, S.; Saslis, S.; Spiliopoulou, M.; Beckers, D.; Degen, T.; Nenert, G.; Fitch, A. N.; Margiolaki, I. Effects of In-Situ Humidity Variation at Tetragonal HEWL Polycrystalline Powder. *J. Appl. Crystallogr.* **2019**, *(in press)*.

(49) Trampari, S.; Valmas, A.; Logotheti, S.; Saslis, S.; Fili, S.; Spiliopoulou, M.; Beckers, D.; Degen, T.; Nénert, G.; Fitch, A. N.; Calamiotou, M.; Karavassili, F.; Margiolaki, I. In Situ Detection of a Novel Lysozyme Monoclinic Crystal Form upon Controlled Relative Humidity Variation. *J. Appl. Crystallogr.* **2018**, *51* (6), 1671–1683. https://doi.org/10.1107/S1600576718013936.

(50) Garman, E. F. Radiation Damage in Macromolecular Crystallography: What Is It and Why Should We Care? *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (4), 339–351. https://doi.org/10.1107/S0907444910008656.

(51) Wright, J. P.; Pechkova, E.; Nicolini, C. Synchrotron Powder Diffraction Study of Radiation Damage in Langmuir Blodgett Nanotemplate Crystallised Protein. *Am. J. Biochem. Biotechnol.* **2014**, *10* (3), 162–168. https://doi.org/10.3844/ajbbsp.2014.162.168.

(52) Margiolaki, I.; Wright, J. P.; Fitch, A. N.; Fox, G. C.; Von Dreele, R. B. Synchrotron X-Ray Powder Diffraction Study of Hexagonal Turkey Egg-White Lysozyme. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61* (4), 423–432. https://doi.org/10.1107/S0907444905001393.

(53) Jenner, M. J.; Wright, J. P.; Margiolaki, I.; Fitch, A. N. Successful Protein Cryocooling for Powder Diffraction. *J. Appl. Crystallogr.* **2007**, *40* (1), 121–124. https://doi.org/10.1107/S0021889806044943.

(54) Von Dreele, R. B. Binding of N-Acetylglucosamine to Chicken Egg Lysozyme: A Powder Diffraction Study. *Acta Crystallogr. Sect. D* **2001**, *57* (12), 1836–1842. https://doi.org/10.1107/S0907444901015748.

(55) Karavassili, F.; Giannopoulou, A. E.; Kotsiliti, E.; Knight, L.; Norrman, M.; Schluckebier, G.; Drube, L.; Fitch, A. N.; Wright, J. P.; Margiolaki, I. Structural Studies of Human Insulin Cocrystallized with Phenol or Resorcinol via Powder Diffraction. *Acta Crystallogr. D Biol. Crystallogr.* **2012**, *68* (12), 1632–1641. https://doi.org/10.1107/S0907444912039339.

(56) Lahey-Rudolph, J. M.; Schönherr, R.; Jeffries, C. M.; Blanchet, C. E.; Boger, J.; Ferreira Ramos, A. S.; Riekehr, W. M.; Triandafillidis, D.-P.; Valmas, A.; Margiolaki, I.; Svergun, D.; Redecke, L. Rapid Screening of in Cellulo Grown Protein Crystals via a SAXS-XRPD Synergistic Approach. *J. Appl. Crystallogr.* **2020**, *(submission code: EI5059)*.

(57) Margiolaki, I.; Wright, J. P.; Fitch, A. N.; Fox, G. C.; Dreele, R. B. V.; Miura, K.; Gozzo, F.; Besnard, C.; Camus, F.; Pattison, P.; Degen, T. Powder Diffraction Studies on Proteins: An Overview of Data Collection Approaches. *Z. Für Krist. Suppl.* **2007**, *26*, 1–14. https://doi.org/10.1524/zksu.2007.2007.suppl_26.1.

(58) Von Dreele, R. B.; Lee, P. L.; Zhang, Y. Protein Polycrystallography. In *Ninth European Powder Diffraction Conference*; Zeitschrift für Kristallographie / Supplemente: München, 2006; Vol. 23, pp 3–8. https://doi.org/10.1524/9783486992526-004.

(59) Triandafillidis, D. P.; Parthenios, N.; Spiliopoulou, M.; Valmas, A.; Gozzo, F.; Reinle-Schmitt, M.; Weiss, M. S.; Wollenhaupt, J.; Beckers, D.; Degen, T.; Pop, M.; Fitch, A.; Karavassili, F.; Margiolaki, I. Insulin Polymorphism Induced by Two Distinct Polyphenols:

New Crystal Forms and Advances in Macromolecular Powder Diffraction. *Acta Crystallogr. Sect. Struct. Biol.* **2020**, *(submission code: GM5074)*.

(60) Kabsch, W. Evaluation of Single-Crystal X-Ray Diffraction Data from a Position-Sensitive Detector. *J. Appl. Crystallogr.* **1988**, *21* (6), 916–924. https://doi.org/10.1107/S0021889888007903.

(61) Rossmann, M. G. Automatic Indexing of Oscillation Images. In *International Tables for Crystallography Volume F: Crystallography of Biological Macromolecules*; Rossmann, M. G., Arnold, E., Eds.; International Tables for Crystallography; Springer Netherlands: Dordrecht, 2001; pp 209–211. https://doi.org/10.1107/97809553602060000674.

(62) Altomare, A.; Cuocci, C.; Moliterni, A.; Rizzi, R. Indexing a Powder Diffraction Pattern; 2019; pp 270–281. https://doi.org/10.1107/97809553602060000949.

(63) Fili, S.; Valmas, A.; Norrman, M.; Schluckebier, G.; Beckers, D.; Degen, T.; Wright, J.; Fitch, A.; Gozzo, F.; Giannopoulou, A. E.; Karavassili, F.; Margiolaki, I. Human Insulin Polymorphism upon Ligand Binding and PH Variation: The Case of 4-Ethylresorcinol. *IUCrJ* **2015**, *2* (5), 534–544. https://doi.org/10.1107/S2052252515013159.

(64) Pawley, G. S. Unit-Cell Refinement from Powder Diffraction Scans. *J. Appl. Crystallogr.* **1981**, *14* (6), 357–361. https://doi.org/10.1107/S0021889881009618.

(65) Dinnebier, R. E.; Leineweber, A.; Evans, J. S. O. *Rietveld Refinement: Practical Powder Diffraction Pattern Analysis Using TOPAS*; De Gruyter: Berlin ; Boston, 2019.

(66) Engel, G. E.; Wilke, S.; König, O.; Harris, K. D. M.; Leusen, F. J. J. PowderSolve– a Complete Package for Crystal Structure Solution from Powder Diffraction Patterns. *J. Appl. Crystallogr.* **1999**, *32* (6), 1169–1179. https://doi.org/10.1107/S0021889899009930.

(67) Le Bail, A.; Duroy, H.; Fourquet, J. L. Ab-Initio Structure Determination of LiSbWO6 by X-Ray Powder Diffraction. *Mater. Res. Bull.* **1988**, *23* (3), 447–452. https://doi.org/10.1016/0025-5408(88)90019-0.

(68) Le Bail, A. Whole Powder Pattern Decomposition Methods and Applications: A Retrospection. *Powder Diffr.* **2005**, *20* (4), 316–326. https://doi.org/10.1154/1.2135315.

(69) Lehmann, M. S.; Christensen, A. N.; Fjellvåg, H.; Feidenhans'l, R.; Nielsen, M. Structure Determination by Use of Pattern Decomposition and the Rietveld Method on Synchrotron X-Ray and Neutron Powder Data; the Structures of Al2Y4O9 and T2O4. *J. Appl. Crystallogr.* **1987**, *20* (2), 123–129. https://doi.org/10.1107/S0021889887087016.

(70) Taylor, G. The Phase Problem. *Acta Crystallogr. D Biol. Crystallogr.* **2003**, *59* (11), 1881–1890. https://doi.org/10.1107/S0907444903017815.

(71) Sheldrick, G. M.; Hauptman, H. A.; Weeks, C. M.; Miller, R.; Usón, I. Ab Initio Phasing. In *International Tables for Crystallography Volume F: Crystallography of Biological Macromolecules*; International Tables for Crystallography; 2012; pp 413–432. https://doi.org/10.1107/97809553602060000850.

(72) Rossmann, M. G. The Molecular Replacement Method. *Acta Crystallogr. Sect. A* **1990**, *46* (2), 73–82. https://doi.org/10.1107/S0108767389009815.

(73) Rupp, B. *Biomolecular Crystallography*; Garland Science, 2010.

(74) Doebbler, J. A.; von Dreele, R. B. Macromolecular Powder Diffraction: Structure Solution via Molecular Replacement. *Z. Für Krist. Suppl.* **2009**, *2009* (30), 33–37. https://doi.org/10.1524/zksu.2009.0005.

(75) Vagin, A.; Teplyakov, A. Molecular Replacement with MOLREP. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (1), 22–25. https://doi.org/10.1107/S0907444909042589.

(76) Patterson, A. L. A Fourier Series Method for the Determination of the Components of Interatomic Distances in Crystals. *Phys. Rev.* **1934**, *46* (5), 372–376. https://doi.org/10.1103/PhysRev.46.372.

(77) Perutz, M. F. Isomorphous Replacement and Phase Determination in Non-Centrosymmetric Space Groups. *Acta Crystallogr.* **1956**, *9* (11), 867–873. https://doi.org/10.1107/S0365110X56002485.

(78) Green, D. W.; Ingram, V. M.; Perutz, M. F.; Bragg, W. L. The Structure of Haemoglobin - IV. Sign Determination by the Isomorphous Replacement Method. *Proc. R. Soc. Lond. Ser. Math. Phys. Sci.* **1954**, *225* (1162), 287–307. https://doi.org/10.1098/rspa.1954.0203.

(79) Blow, D. M.; Rossmann, M. G. The Single Isomorphous Replacement Method. *Acta Crystallogr.* **1961**, *14* (11), 1195–1202. https://doi.org/10.1107/S0365110X61003454.
(80) Cromer, D. T. Anomalous Dispersion Corrections Computed from Self-Consistent Field Relativistic Dirac–Slater Wave Functions. *Acta Crystallogr.* **1965**, *18* (1), 17–23. https://doi.org/10.1107/S0365110X6500004X.
(81) Peterson, M. R.; Harrop, S. J.; McSweeney, S. M.; Leonard, G. A.; Thompson, A. W.; Hunter, W. N.; Helliwell, J. R. MAD Phasing Strategies Explored with a Brominated Oligonucleotide Crystal at 1.65Å Resolution. *J. Synchrotron Radiat.* **1996**, *3* (1), 24–34. https://doi.org/10.1107/S0909049595013288.
(82) Helliwell, J. R. *Macromolecular Crystallography with Synchrotron Radiation*; Cambridge University Press, 2005.
(83) Helliwell, J. R.; Bell, A. M. T.; Bryant, P.; Fisher, S.; Habash, J.; Helliwell, M.; Margiolaki, I.; Kaenket, S.; Watier, Y.; Wright, J. P.; Yalamanchilli, S. Time-Dependent Analysis of K2PtBr6 Binding to Lysozyme Studied by Protein Powder and Single Crystal X-Ray Analysis. *Z. Für Krist. - Cryst. Mater.* **2010**, *225* (12), 570–575. https://doi.org/10.1524/zkri.2010.1349.
(84) *Structure Determination from Powder Diffraction Data*; David, W. I. F., Shankland, K., McCusker, L. B., Baerlocher, Ch., Eds.; IUCr monographs on crystallography; Oxford University Press: Oxford ; New York, 2002.
(85) Wright, J. P. Extraction and Use of Correlated Integrated Intensities with Powder Diffraction Data. *Z. Für Krist. - Cryst. Mater.* **2004**, *219* (12). https://doi.org/10.1524/zkri.219.12.791.55857.
(86) Valmas, A.; Magiouf, K.; Fili, S.; Norrman, M.; Schluckebier, G.; Beckers, D.; Degen, T.; Wright, J.; Fitch, A.; Gozzo, F.; Giannopoulou, A. E.; Karavassili, F.; Margiolaki, I. Novel Crystalline Phase and First-Order Phase Transitions of Human Insulin Complexed with Two Distinct Phenol Derivatives. *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71* (4), 819–828. https://doi.org/10.1107/S1399004715001376.
(87) Brunelli, M.; Wright, J. P.; Vaughan, G. B. M.; Mora, A. J.; Fitch, A. N. Solving Larger Molecular Crystal Structures from Powder Diffraction Data by Exploiting Anisotropic Thermal Expansion. *Angew. Chem. Int. Ed.* **2003**, *42* (18), 2029–2032. https://doi.org/10.1002/anie.200250607.
(88) Shankland, K.; David, W. I. F.; Sivia, D. S. Routine Ab Initio Structure Determination of Chlorothiazide by X-Ray Powder Diffraction Using Optimised Data Collection and Analysis Strategies. *J. Mater. Chem.* **1997**, *7* (3), 569–572. https://doi.org/10.1039/a606998c.
(89) Wright, J. P.; Forsyth, J. B. PRODD: Profile Refinement of Diffraction Data Using the Cambridge Crystallographic Subroutine Library. **2000**.
(90) Degen, T.; Sadki, M.; Bron, E.; König, U.; Nénert, G. The HighScore Suite. *Powder Diffr.* **2014**, *29* (S2), S13–S18. https://doi.org/10.1017/S0885715614000840.
(91) Von Dreele, R. B. Powder Diffraction Peak Profiles. In *International Tables for Crystallography Volume H: Powder diffraction*; 2019; pp 263–269. https://doi.org/10.1107/97809553602060000948.
(92) Caglioti, G.; Paoletti, A.; Ricci, F. P. Choice of Collimators for a Crystal Spectrometer for Neutron Diffraction. *Nucl. Instrum.* **1958**, *3* (4), 223–228. https://doi.org/10.1016/0369-643X(58)90029-X.
(93) Thompson, P.; Cox, D. E.; Hastings, J. B. Rietveld Refinement of Debye–Scherrer Synchrotron X-Ray Data from Al2O3. *J. Appl. Crystallogr.* **1987**, *20* (2), 79–83. https://doi.org/10.1107/S0021889887087090.
(94) Stokes, A. R.; Wilson, A. J. C. The Diffraction of X Rays by Distorted Crystal Aggregates - I. *Proc. Phys. Soc.* **1944**, *56* (3), 174–181. https://doi.org/10.1088/0959-5309/56/3/303.
(95) Stephens, P. W. The Physics of Diffraction from Powders. In *International Tables for Crystallography Volume H: Powder diffraction*; 2019; pp 252–262. https://doi.org/10.1107/97809553602060000947.
(96) Wertheim, G. K.; Butler, M. A.; West, K. W.; Buchanan, D. N. E. Determination of the Gaussian and Lorentzian Content of Experimental Line Shapes. *Rev. Sci. Instrum.* **1974**, *45* (11), 1369–1371. https://doi.org/10.1063/1.1686503.

(97) van Laar, B.; Yelon, W. B. The Peak in Neutron Powder Diffraction. *J. Appl. Crystallogr.* **1984**, *17* (2), 47–54. https://doi.org/10.1107/S0021889884011006.
(98) Finger, L. W.; Cox, D. E.; Jephcoat, A. P. A Correction for Powder Diffraction Peak Asymmetry Due to Axial Divergence. *J. Appl. Crystallogr.* **1994**, *27* (6), 892–900. https://doi.org/10.1107/S0021889894004218.
(99) Weichenberger, C. X.; Afonine, P. V.; Kantardjieff, K.; Rupp, B. The Solvent Component of Macromolecular Crystals. *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71* (5), 1023–1038. https://doi.org/10.1107/S1399004715006045.
(100) Moews, P. C.; Kretsinger, R. H. Refinement of the Structure of Carp Muscle Calcium-Binding Parvalbumin by Model Building and Difference Fourier Analysis. *J Mol Biol* **1975**, *91*, 201–228. https://doi.org/10.1016/0022-2836(75)90160-6.
(101) Hendrickson, W. A. Stereochemically Restrained Refinement of Macromolecular Structures. In *Methods in Enzymology*; Elsevier, 1985; Vol. 115, pp 252–270. https://doi.org/10.1016/0076-6879(85)15021-4.
(102) Engh, R. A.; Huber, R. Accurate Bond and Angle Parameters for X-Ray Protein Structure Refinement. *Acta Crystallogr. A* **1991**, *47* (4), 392–400. https://doi.org/10.1107/S0108767391001071.
(103) Sheldrick, G. M. Crystal Structure Refinement Incorporating Chemical Information. In *Direct Methods for Solving Macromolecular Structures*; Fortier, S., Ed.; NATO ASI Series; Springer Netherlands: Dordrecht, 1998; pp 119–130. https://doi.org/10.1007/978-94-015-9093-8_12.
(104) Steiner, R. A.; Tucker, J. A. Keep It Together: Restraints in Crystallographic Refinement of Macromolecule–Ligand Complexes. *Acta Crystallogr. Sect. Struct. Biol.* **2017**, *73* (Pt 2), 93–102. https://doi.org/10.1107/S2059798316017964.
(105) Evans, P. R. An Introduction to Stereochemical Restraints. *Acta Crystallogr. D Biol. Crystallogr.* **2007**, *63* (Pt 1), 58–61. https://doi.org/10.1107/S090744490604604X.
(106) Ramakrishnan, C.; Ramachandran, G. N. Stereochemical Criteria for Polypeptide and Protein Chain Conformations: II. Allowed Conformations for a Pair of Peptide Units. *Biophys. J.* **1965**, *5* (6), 909–933. https://doi.org/10.1016/S0006-3495(65)86759-5.
(107) Pal, D.; Chakrabarti, P. On Residues in the Disallowed Region of the Ramachandran Map. *Biopolymers* **2002**, *63* (3), 195–206. https://doi.org/10.1002/bip.10051.
(108) Jones, J. E.; Chapman, S. On the Determination of Molecular Fields. —II. From the Equation of State of a Gas. *Proc. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character* **1924**, *106* (738), 463–477. https://doi.org/10.1098/rspa.1924.0082.
(109) McCusker, L. B.; Von Dreele, R. B.; Cox, D. E.; Louër, D.; Scardi, P. Rietveld Refinement Guidelines. *J. Appl. Crystallogr.* **1999**, *32*, 36–50. https://dx.doi.org/10.1107/S0021889898009856.
(110) Toby, B. H. R Factors in Rietveld Analysis: How Good Is Good Enough? *Powder Diffr.* **2006**, *21* (01), 67–70. https://doi.org/10.1154/1.2179804.
(111) Rhodes, G. *Crystallography Made Crystal Clear*, 3rd ed.; Academic Press, 2006.
(112) Vaguine, A. A.; Richelle, J.; Wodak, S. J. SFCHECK: A Unified Set of Procedures for Evaluating the Quality of Macromolecular Structure-Factor Data and Their Agreement with the Atomic Model. *Acta Crystallogr. D Biol. Crystallogr.* **1999**, *55* (1), 191–205. https://doi.org/10.1107/S0907444998006684.
(113) Bhat, T. N. Calculation of an OMIT Map. *J. Appl. Crystallogr.* **1988**, *21* (3), 279–281. https://doi.org/10.1107/S0021889887012755.
(114) Bhat, T. N.; Cohen, G. H. OMITMAP: An Electron Density Map Suitable for the Examination of Errors in a Macromolecular Model. *J. Appl. Crystallogr.* **1984**, *17* (4), 244–248. https://doi.org/10.1107/S0021889884011456.
(115) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (4), 486–501. https://doi.org/10.1107/S0907444910007493.
(116) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: A Program to Check the Stereochemical Quality of Protein Structures. *J. Appl. Crystallogr.* **1993**, *26* (2), 283–291. https://doi.org/10.1107/S0021889892009944.

(117) Guex, N.; Peitsch, M. C. SWISS-MODEL and the Swiss-Pdb Viewer: An Environment for Comparative Protein Modeling. *Electrophoresis* **1997**, *18* (15), 2714–2723. https://doi.org/10.1002/elps.1150181505.

(118) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: All-Atom Structure Validation for Macromolecular Crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (1), 12–21. https://doi.org/10.1107/S0907444909042073.

(119) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall, W. B.; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. MolProbity: More and Better Reference Data for Improved All-atom Structure Validation. *Protein Sci. Publ. Protein Soc.* **2018**, *27* (1), 293–315. https://doi.org/10.1002/pro.3330.

(120) Colovos, C.; Yeates, T. O. Verification of Protein Structures: Patterns of Nonbonded Atomic Interactions. *Protein Sci.* **1993**, *2* (9), 1511–1519. https://doi.org/10.1002/pro.5560020916.

(121) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(122) Dilanian, R. A.; Darmanin, C.; Varghese, J. N.; Wilkins, S. W.; Oka, T.; Yagi, N.; Quiney, H. M.; Nugent, K. A. A New Approach for Structure Analysis of Two-Dimensional Membrane Protein Crystals Using X-Ray Powder Diffraction Data. *Protein Sci.* **2011**, *20* (2), 457–464. https://doi.org/10.1002/pro.572.

(123) Fili, S.; Valmas, A.; Christopoulou, M.; Spiliopoulou, M.; Nikolopoulos, N.; Lichière, J.; Logotheti, S.; Karavassili, F.; Rosmaraki, E.; Fitch, A.; Wright, J.; Beckers, D.; Degen, T.; Nénert, G.; Hilgenfeld, R.; Papageorgiou, N.; Canard, B.; Coutard, B.; Margiolaki, I. Coxsackievirus B3 Protease 3C: Expression, Purification, Crystallization and Preliminary Structural Insights. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2016**, *72* (12), 877–884. https://doi.org/10.1107/S2053230X16018513.

(124) Frankaer, C. G.; Moroz, O. V.; Turkenburg, J. P.; Aspmo, S. I.; Thymark, M.; Friis, E. P.; Stahl, K.; Nielsen, J. E.; Wilson, K. S.; Harris, P. Analysis of an Industrial Production Suspension of Bacillus Lentus Subtilisin Crystals by Powder Diffraction: A Powerful Quality-Control Tool. *Acta Crystallogr. D Biol. Crystallogr.* **2014**, *70* (4), 1115–1123. https://doi.org/10.1107/S1399004714001497.

(125) Hunter, M. S.; DePonte, D. P.; Shapiro, D. A.; Kirian, R. A.; Wang, X.; Starodub, D.; Marchesini, S.; Weierstall, U.; Doak, R. B.; Spence, J. C. H.; Fromme, P. X-Ray Diffraction from Membrane Protein Nanocrystals. *Biophys. J.* **2011**, *100* (1), 198–206. https://doi.org/10.1016/j.bpj.2010.10.049.

(126) Inouye, H.; Kirschner, D. A. X-Ray Fiber and Powder Diffraction of PrP Prion Peptides. In *Advances in Protein Chemistry*; Fibrous Proteins: Amyloids, Prions and Beta Proteins; Academic Press, 2006; Vol. 73, pp 181–215. https://doi.org/10.1016/S0065-3233(06)73006-6.

(127) Norrman, M.; Ståhl, K.; Schluckebier, G.; Al-Karadaghi, S. Characterization of Insulin Microcrystals Using Powder Diffraction and Multivariate Data Analysis. *J. Appl. Crystallogr.* **2006**, *39* (3), 391–400. https://doi.org/10.1107/S0021889806011058.

(128) Valmas, A.; Fili, S.; Nikolopoulos, N.; Spiliopoulou, M.; Christopoulou, M.; Karavassili, F.; Kosinas, C.; Bastalias, K.; Rosmaraki, E.; Lichiére, J.; Fitch, A.; Beckers, D.; Degen, T.; Papageorgiou, N.; Canard, B.; Coutard, B.; Margiolaki, I. Dengue Virus 3 NS5 Methyltransferase Domain: Expression, Purification, Crystallization and First Structural Data from Microcrystalline Specimens. *Z. Für Krist. - Cryst. Mater.* **2017**, *233* (5), 309–316. https://doi.org/10.1515/zkri-2017-2091.

(129) Das, U.; Naskar, J.; Mukherjee, A. K. Conformational Analysis of an Acyclic Tetrapeptide: Ab-Initio Structure Determination from X-Ray Powder Diffraction, Hirshfeld Surface Analysis and Electronic Structure: Conformational Study of an Acyclic Tetrapeptide. *J. Pept. Sci.* **2015**, *21* (12), 845–852. https://doi.org/10.1002/psc.2825.

(130) Fujii, K.; Young, M. T.; Harris, K. D. M. Exploiting Powder X-Ray Diffraction for Direct Structure Determination in Structural Biology: The P2X4 Receptor Trafficking Motif YEQGL. *J. Struct. Biol.* **2011**, *174* (3), 461–467. https://doi.org/10.1016/j.jsb.2011.03.001.
(131) Tedesco, E.; Harris, K. D. M.; Johnston, R. L.; Turner, G. W.; Raja, K. M. P.; Balaram, P. Ab Initio Structure Determination of a Peptide β-Turn from Powder X-Ray Diffraction Data. *Chem. Commun.* **2001**, No. 16, 1460–1461. https://doi.org/10.1039/b103876c.