nature reviews methods primers

https://doi.org/10.1038/s43586-025-00433-8

Primer

Check for updates

Macromolecular crystallography

Pavel V. Afonine[1], Armando Albert[2], Kay Diederichs[3], Juan A. Hermoso[2], Eugene Krissinel[4], José Antonio Márquez[5], Santosh Panjikar[6,7], Maria Solà[8], Andrea Thorn[9] & Isabel Usón[8,10]

Abstract

Crystallography provides structural evidence of macromolecules in atomic detail. However, the atomic structure is not the direct outcome of the experiment. Diffraction data need to be processed and the phase problem must be solved to visualize the map from which an atomic model of the macromolecule is interpreted and iteratively improved. Despite this complex process from sample to scientific answer, crystallography is widely accessible in biochemical and biological research. Easy access to the experimental set-ups, free software for academic use and complimentary analytical computing, supported by automation and expert assistance, makes crystallography available to non-crystallographers. This Primer offers a practical and rational introduction to macromolecular crystallography, whether to engage directly or to critically assess results, with a focus on understanding the diffraction data, solving the phase problem, building and refining the atomic model, and interpreting the resulting atomic structure. We provide an overview of what crystallography can achieve, the key decisions and trade-offs involved, and how to evaluate outcomes effectively.

Sections

[1]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [2]Instituto de Química Física Blas Cabrera, Consejo Superior de Investigaciones Científicas, Madrid, Spain. [3]Fachbereich Biologie, Universität Konstanz, Konstanz, Germany. [4]Research Complex at Harwell, Rutherford Appleton Laboratory, Didcot, UK. [5]European Molecular Biology Laboratory, Grenoble, France. [6]Australian Synchrotron, Melbourne, Victoria, Australia. [7]Monash University, Melbourne, Victoria, Australia. [8]Instituto de Biología Molecular de Barcelona, Consejo Superior de Investigaciones Científicas, Barcelona, Spain. [9]Institute for Nanostructure and Solid State Physics, Universität Hamburg, Hamburg, Germany. [10]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. e-mail: uson@ibmb.csic.es

Introduction

Macromolecular crystallography (MX) provides a 3D view of the molecular structures involved in biology, biomedicine and biotechnology. Establishing the atomic structure of nucleic acids, proteins and their complexes mediates our understanding of macromolecular function and informs on how to influence it. The pathway from sample crystal to solved structure in MX (Fig. 1a) starts with placing a crystalline sample between a radiation source and a detector[1]. Crystals amplify the scattered photon signal, impose geometric conditions on the generation of constructive interference and yield a diffraction pattern. In addition, the process of crystallization selects a single conformation from an otherwise dynamic system, trapping molecules in a lattice with constant spatial relation. Haemoglobin, for example, was crystallized decades before it could be stabilized against dehydration to evidence diffraction. Two more decades were needed to solve the structure of this important protein[2].

Diffraction is characterized by the recorded scattered beams from the crystal sample. Following the formulation of diffraction as the phenomenon of reflection by imaginary planes introduced in Bragg's law (Fig. 1b), the scattered beams are universally called Bragg reflections[3]. As the X-ray source and the detector are fixed, the crystal is rotated to bring all accessible reflections into a diffraction condition. The Ewald sphere (Fig. 1c) is a geometric construction to show the directions of the diffracted beams[4]. The distance between the mirroring families of planes in a crystal lattice defines the resolution such that closer spacing corresponds to higher resolution and more level of detail.

X-rays interact with the electrons present in macromolecules. The relationship between electron density in real space and structure factors in reciprocal space is described by the Fourier transform (Fig. 1d). For a set of known atomic positions, given lattice metrics and symmetry properties, the structure factors can be calculated (Fig. 1e). The electron density distribution, $\rho(r)$, in real space can be reconstructed from the complex structure factors, $F(hkl)$, with a known amplitude and phase in reciprocal space using the inverse Fourier transform (Fig. 1d). Expressing diffraction in real or reciprocal space is equivalent and both representations are found at every step of a crystallographic determination. The phase problem, central to crystallography, arises because diffraction experiments measure only the intensity of scattered waves and not the phase of the waves, which is crucial for reconstructing the electron density map. Once phase information is obtained, a Fourier synthesis is used to reconstruct the electron density map, whose accuracy depends on the correctness of the phase estimates as well as on the completeness and diffraction angle limit reached in the experiment. X-rays have wavelength in the range of interatomic distances in molecules and crystals, so the map can reach atomic-level detail for well-ordered crystals.

The importance of crystallography as a model for experimental sciences has been endorsed by efforts that support universal access to crystallographic experiments, exemplified in the construction and access to synchrotrons. Common standards have been introduced and upheld by the Protein Data Bank (PDB)[5], providing open access to crystallographic data[6]. Notably, even heterogeneous experiments and parameterizations are brought into comparable description sharing formats and standards. The PDB repository holds results from crystallography (83% as of March 2025), nuclear magnetic resonance (NMR; 6%), and cryo-electron microscopy (11%) experiments. Increasingly present are structures that were solved by combining data from these major techniques with others such as small-angle X-ray scattering (SAXS), crosslinking mass spectrometry, atomic force microscopy, Förster resonance energy transfer and more. Supplementary Fig. 1 depicts the sample types used and example data produced from these complementary techniques.

For a method that crucially depends on calculations, the development of fast, accessible computers and the availability of web services for data analysis has been instrumental. Data can be processed in a personal computer and extensive analyses can be conducted with free, expert web or synchrotron services such as CCP4 Cloud[7] or Auto-rickshaw[8]. The workflow illustrated in Fig. 1a is sequential, but recent years have seen an integration of all computational methods following data processing. Methods to balance prior knowledge with experimental data fuse the structure solution, model construction, refinement and validation stages into a simultaneous, rather than sequential, process[9]. Several automation strategies, including crystal-mounting robots and methods for high-throughput protein expression, purification and crystallization have been made possible by developments in hardware. Software developments (see the user community crystallographic wiki) include advanced data processing, phasing, refinement and validation pipelines; for example, SHELX, a suite of crystallographic programs extending from mineralogy and material science to chemical and biological applications[10]. Integration of MX methods into software suites such as CCP4 (ref. 11) and Phenix[12] has nucleated academic development, whereas GlobalPhasing[13,14] has led industrial development. Better accounting for errors and unknown parameters with Bayesian statistics leads to improved extraction of weak signals[15]. The recent steep rise of artificial intelligence has been swiftly integrated into crystallographic methods for planning the experiment, interpreting results and determining the structure[16,17]. However, although a purely computational model of a macromolecule or its complexes provides a valuable structural hypothesis, only experiments are conclusive and may reveal unexpected results. Furthermore, predictive models do not establish ligands, covalent modifications or other environmental factors, as recently reviewed[18].

This Primer aims to bring students or experts from a different field to MX, providing guidance for experimental design and critical appraisal of results. A basic understanding of diffraction and crystal structure is covered in the Supplementary Information, expanded with online resources such as the Crystallography pages of the Madrid Crystallography and CryoEM School (CSIC Crystallography) and the International Union of Crystallography (IUCr) dictionary, to complement the detailed applications used in the Primer. We explain why certain experimental choices are made while acknowledging limitations of the technique, such as radiation damage or resolution. In analysing complicated data, we discuss data processing and refinement and provide a thorough background of the Fourier transform required for solving the phase problem in crystallography and implications on model bias. Last, common errors in crystal structures and improving the future of MX experimentation are covered.

Experimentation

The typical experimental workflow for MX progresses through sample preparation, crystallization, crystal delivery and diffraction data collection. Large-scale genomics studies, based on the premise that structure drives function[19,20], advanced automation efforts to achieve high throughput. Today, routine crystallographic experiments can be carried out remotely or supported by dedicated facilities allowing access to non-experts. See Supplementary Table 1 for examples of different user facilities.

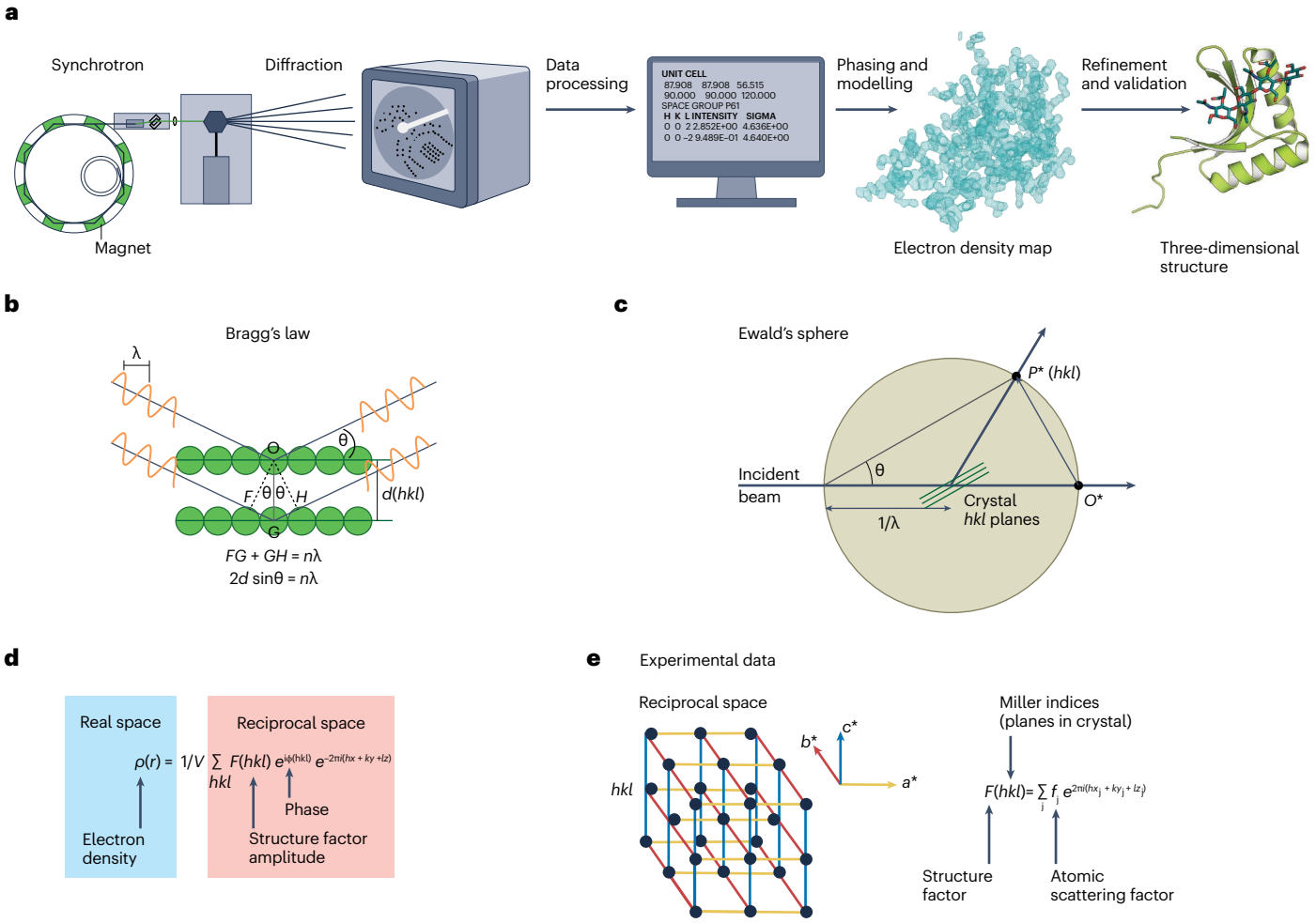


Fig. 1 | Overview of the macromolecular crystallography experiment and key fundamental concepts. a, The pathway from macromolecular crystal to structure solution in macromolecular crystallography, for typical X-ray diffraction experiments. **b**, Bragg's law for X-rays diffracting from a crystal lattice where *d* is the spacing between two crystal planes, *n* is the order of reflection, an integer number, λ is the wavelength of the incident X-ray beam and θ is the Bragg angle, which is the angle between the incident X-ray beam and the reflecting crystal plane; *FG* and *HG* segments mark the difference of the path travelled by the wave fronts O*F* and O*H*. **c**, Ewald sphere for visualizing X-ray diffraction geometries from crystal planes where θ is the Bragg angle, λ is the wavelength of the incident X-ray beam, *O** is the origin of reciprocal space and *P** is a point of the reciprocal lattice (a reflection produced by crystal planes *hkl*). **d**, The relationship between real and reciprocal space in a Fourier transform where $\rho(r)$ is electron density at position *r* (with coordinates *x*,*y*,*z*) in real space, $F(hkl)$ is the structure factor for the reflection (*h*,*k*,*l*) in reciprocal space and *V* is the volume of the unit cell in real space. **e**, The relation between the structure factors in reciprocal space and the crystal structure in real space is given by an inverse Fourier transform, where variables have the same meaning as in **d** and f_j is the atomic scattering factor of atom *j* describing the efficiency with which that atom scatters X-rays at a given angle.

Sample preparation

The quality of the crystal sample determines the quality of the diffraction pattern measured. Deviations from periodic ordering lower the signal obtained for each reflection, whereas internal disorder causes poor resolution. Crystals contain a large proportion of solvent, an aqueous solution containing inorganic salts, polyalcohols, detergents, organic compounds, and redox or pH stabilizing buffers. Solvent reduces crystal contacts and facilitates mobility and disorder of the macromolecule, leading to lower resolution.

Purification protocols that preserve protein structural and chemical homogeneity favour crystal nucleation, internal crystal homogeneity and sustained quality in X-ray diffraction[21]. Typical protocols require large amounts of protein (5–25 mg ml^{-1}, totalling at least 0.5–2 mg), highly pure samples (estimated >95%, single band on SDS–PAGE), chemical homogeneity (for example, disulfide bridges should not be partly oxidized so reducing agents are frequently added or to avoid polypeptide degradation protease inhibitors are added) and structural homogeneity (for example, there is no aggregation and the sample is monodisperse). Establishing the purification protocol requires considering molecular features such as the isoelectric point, which suggests that the pH of buffer solutions should differ by at least one point, or the presence of disulfide bonds, which indicates reducing agents. Equally important are the protein source (such as the organism, extremophile or mesophile and aerobic or anaerobic) and the natural context of a

protein (such as the cytoplasm, an organelle, at the interstitial space, in membranes, in blood or secreted to the medium). The context can suggest the need of cofactors (such as ions or organic molecules), protein partners, stabilizing molecules such as lipids, or compensatory factors such as high NaCl concentration for DNA-binding proteins.

Protein expression. The target protein can be either extracted from the natural source or expressed in dedicated bacterial or eukaryotic systems (such as mammalian or insect cells). The *Escherichia coli* (DE3) strain is widely used owing to its inducible T7 RNA polymerase system, triggered by isopropyl β-D-1-thiogalactopyranoside. The same promoter can be used for milder autoinduction, which might be convenient for some proteins[22]. Modified DE3 strains can include those with tight control of protein expression to avoid basal expression (such as pLysS strains), chaperones to assist protein folding (Origami and Origami2-pLysS strains), codons of tRNAs uncommon in *E. coli* (Rosetta, Rosetta Gami2) or that are resistant to toxic proteins (C41 and C43). Further, the plasmids carrying the protein target may fuse it to a 6-histidine tag or a second protein with affinity to specific chromatographic resins (maltose binding protein (MBP), glutathione *S*-transferase (GST), small ubiquitin-like modifier (SUMO) and guanine nucleotide-binding protein subunit-β (GB1)), conferring stability (NusA) or a reducing environment (thioredoxin). Although less versatile, more expensive and laborious, expression in mammalian and insect cells might be needed to approximate the natural environment for proper folding and the addition of post-translational modifications such as glycosylation.

An efficient cell lysis buffering solution for harvesting the protein in the soluble fraction includes the appropriate pH, salts, reducing agents or other additives. Upon lysis, purification includes an affinity column step and possibly a final size-exclusion chromatography polishing step. Intermediate steps include ion exchange or cation exchange, and/or pseudoaffinity purification with heparin columns for nucleic-acid-binding proteins.

Crystallization. Exhaustive combinatorial testing of all crystallization parameters (chemical species, concentrations, temperature, pH and set-up) would require an astronomical number of experiments. Instead, sparse subsets are trialled in a microlitre or nanolitre scale at room temperature or lower temperatures, with various combinations of pH buffers, inorganic salts, polyalcohols, ligands and detergents[23,24]. Conditions can only be effectively varied if concentrating the protein does not already require a rich buffer[25]. Screening identifies promising conditions to be scaled up and optimized[26,27].

Further modification of the crystals through co-crystallization or soaking might be necessary. In particular, for typical data collection at 100 K, optimal treatment with cryoprotectant agents is key to avoid the formation of ice crystals during sample mounting and diffraction data collection as ice leads to disruptions in the crystal lattice[28]. Most commonly, incorporating ions and chemical molecules into the crystal is done for experimental phasing[29]. In high-pressure crystallography, gases such as krypton and xenon are introduced to map hydrophobic tunnels and cavities[30], track molecular oxygen[31], carbon dioxide or methane[32], and identify lipid binding sites[33]. Even other proteins have been soaked into crystals, preserving the order and capacity to diffract[34].

Troubleshooting crystallization. If crystallization fails, alternative constructs can be designed using predictive approaches to define domain boundaries and avoid disordered or flexible regions. Experimentally, stable protein fragments can be obtained by limited proteolysis. In this procedure, serial dilutions of two or three sequence-specific proteases are incubated with the sample at different concentrations, temperatures and times, and the reaction products are analysed by SDS–PAGE[35]. The length of each fragment is determined by combining N-terminal sequencing with precise molecular weight determination. Surviving fragments are set for crystallization or cloned for expression. Rational modification of constructs can also be pivotal. For example, the excision of flexible loops or tails that interfere with crystal packing was useful in the case of crystallizing the human mitochondrial transcription factor B2 (ref. 36). In another example, the addition of a fragment of a partner to stabilize the contact interface that otherwise drives aggregation or precipitation was crucial for the crystallization of *E. coli* s70 subdomain 4, which required fusion to a helix from the RNA polymerase[37].

Sample stability can be optimized with additives at any step during protein production, purification or crystallization. Screening different additives using a thermal stability assay allows identification of unexpected stabilizers[38]. A common assay for this purpose involves measuring the binding of Sypro Orange to the protein under increasing temperatures, with binding indicating externalization of hydrophobic residues on the protein (and thus denaturing). This technique identified citrate as a stabilizer for mitochondrial mTERF[39].

Biophysical methods can detect dynamic conformational equilibria that might impair crystal growth. These techniques include dynamic light scattering to measures molecular polydispersity and aggregation, or SAXS to measures protein flexibility. Using dynamic light scattering and SAXS under different pH conditions, specific additives, salt concentrations and temperatures provides insight to the dynamics and homogeneity of the sample.

A useful principle to navigate all these options is to start from the simplest, most cost-efficient method and increase the complexity to overcome obstacles (such as failure to express, aggregation or failure to crystallize), following the routine acquired by your own research group. Additional tips for sample preparation and crystallization can be found in Supplementary Note 2.1.

Sample mounting. Crystals are commonly extracted from the crystallization drop, mounted in a standardized data collection pin and stored in a cryogenic container, a process that can be fully automated[40,41]. Diffraction can also be measured in situ, on the microplates or microfluidic chips where crystals are grown. This is useful for crystallization screening or to complement cryogenic structure determination and ligand screening at room temperature[42]. The size of the support might limit oscillation range and completeness, whereas measurements conducted at room temperature might limit X-ray dose or require serial approaches[43].

Equipment needed for data collection

The MX experiment requires an X-ray source, a goniometer, a detector, and a cryogenic cooling device. MX experiments are generally conducted at synchrotron beamlines, which provide intense radiation and allow data to be collected at a higher resolution than with bench-scale instruments. However, high flux can induce radiation damage, and is typically mitigated by cryogenically cooling the crystals using liquid nitrogen[44]. Lower temperatures also reduce atomic motion, enhancing crystal order and the resulting signal. Reducing detector noise and readout time represented a major technological jump, allowing more accurate measurement of the diffraction signal than previous technology.

A more rigorous statistical data analysis enables linking crystallographic model and data quality and correctly assess resolution[45,46].

Balancing the amount of radiation. Ideally, the experimental set-up should irradiate the crystal with as many photons as possible, concentrated in a volume as small as the crystal without degrading the sample. In addition, multiple, complete datasets that encompass every possible unique reflection for the crystal are desirable for increased accuracy. However, some experimental requirements collide, requiring compromises.

X-ray radiation can cause primary damage (such as photoreduction) and secondary effects owing to about 500 low-energy secondary electrons per primary absorption event, which are able to diffuse and induce further ionization and excitation events. These can be mitigated by reducing the mobility of the radicals produced as primary damage through cryo-cooling[47,48]. Radiation damage disrupts the 3D order and swiftly alters the intensity of reflections, altogether reducing resolution. To protect the 3D order and homogeneity, most crystal determinations are performed at 100 K. An exception is required to study dynamics, where changes caused by reactions are the object of the experiment and room temperature is used.

The larger the crystal the stronger the diffraction signal, but large volumes do not equilibrate as fast, limiting size as it creates inhomogeneity during derivatization and vitrification. If complete data to the diffraction limit can be obtained from a single crystal, all other datasets are often discarded. Multiple macromolecular crystals are not identical owing to slight differences in crystallization conditions, crystal growth, manipulation and cryo-cooling, causing these alterations to be evident in unit cell constants. In practice, achieving completeness might require combining data from several crystals. As an extreme, as little as one frame might be all that can be extracted from a single crystal. In that case, the instability associated to scaling such sparse data needs to be overcome through the availability of multiple reflections that are shared between frames.

Source and goniometer. An X-ray diffraction experiment can use a laboratory-scale X-ray source of fixed wavelength (determined by the anode element) with a modest flux, orders of magnitude lower than synchrotrons, and a beam diameter in the order of half a millimetre. Synchrotron optics are better suited for small and inhomogeneous crystals as they provide a higher flux and a smaller divergence beam that is concentrated and precisely located at the point where the crystal is rotated. Nevertheless, the availability of measurement time, not requiring transport that may damage crystals, the data confidentiality of industrial projects or the contention of hazardous samples may advise the use of a home source in some cases.

Regardless of the source, the beam needs to be stable and the crystal needs to remain centred within the beam. The mechanical demands to minimize the positional inaccuracy, termed the sphere of confusion, and the space taken up by other equipment (such as cooling devices or cameras) limit the available experimental geometry. While home-built goniometers are typically equipped with circles providing several degrees of freedom, it is rare to find more than one rotation axis at synchrotrons as radiation damage prevents the collection of several datasets on the same sample. However, flexible Kappa, Euler or PRIGo geometries[49] are in use.

Detectors. Most beamlines have photon-counting pixel detectors that shorten readout delay, reduce noise, increase the speed of data collection and eliminate the need of shutter synchronization. The high frame rate (>100 Hz), minimal dead time and high quantum efficiency of pixel detectors enable the acquisition of datasets with high multiplicity, where multiple measurements of each reflection increase the precision and allow a better error model and therefore more accurate data for experimental phasing based on the anomalous signal from sulfur, phosphorus or native metal ions. The small pixel size and sharp point-spread function at pixel detectors fully leverage the small focal size of state-of-the-art X-ray sources and optics, allowing accurate measurement from crystals[50].

Wavelength choice and soft X-rays. MX beamlines allow tunability of the X-ray wavelength, typically between 0.7 and 2.0 Å, incorporating fluorescence detectors for precise wavelength selection and identification of elements in the sample. A particular wavelength may be desirable to induce element-dependent scattering effects, typically at their absorption edges, and single out a few sites within the structure. Beamlines are designed with optimal wavelengths in mind, so flux may vary across the accessible range. Wavelengths around 1 Å are appropriate to determine structure in atomic detail, even on smaller detectors. X-rays around 1 Å induce less absorption than longer wavelengths but there is also a lower diffraction cross-section at higher-incident X-ray energies. Long wavelengths enhance anomalous signal from sulfur and phosphorus, natively present in proteins and nucleic acids, and ensure spatial separation of reflections for large unit cells. Additional advantages of long wavelengths include the identification and location of lighter atoms of biological relevance, such as Cl, K, Mg and Ca, within the structure[51]. Controlling air absorption with a helium path is convenient for wavelengths between 1.5 and 3 Å.

For wavelengths beyond 3.0 Å, specialized beamlines handle the higher absorption and wider diffraction angles associated with soft X-rays. A large, curved detector for the simultaneous collection of low- and high-resolution data keeps the scattered X-rays in focus over the entire surface, which is crucial for obtaining high-quality data. Additionally, long-wavelength X-rays are more prone to scattering and absorption by air, so an in-vacuum sample environment is used to ensure the X-rays reach the sample without interference. Conductive cooling of the sample is necessary to prevent radiation damage and maintain stability during the experiment. This pioneering set-up has been implemented for beamline I23 at Diamond Light Source[52].

Basic experimental set-up and variations. Monochromatic diffraction experiments differ in the way in which reciprocal space is sampled (Table 1). A single-crystal rotation experiment records many consecutive exposures during the smooth rotation of a single crystal over a large angle, thus covering a complete sphere in reciprocal space. The unique part of reciprocal space is usually sampled with high multiplicity given by the cell, symmetry and orientation of the crystal along with the geometry of the experiment (that is, the size and position of the detector in relation to the beam and rotation axis). Typically, a reflection dataset obtained from such an experiment is close to complete (with >95% of all unique reflections within the resolution limit). A few such experiments can be combined to reach 100% completeness.

Multicrystal rotation experiments combine data from partial sampling of reciprocal space to give 100% completeness with high multiplicity. In this mode, about 10–10,000 crystals in different orientations are rotated through a few degrees (for example, 1°–20°). The resulting incomplete datasets need to be scaled relative to each other.

Table 1 | Types of different diffraction experiments and their relevant properties

Type of experiment	Radiation Type	Number of crystals/crystal dimensions necessary	Rotation angle	Advantages	Disadvantages
Single (or few) crystal(s)	X-ray	1–10/0.5–0.05 mm	180°–360°	Visualize electron density, high data quality	Needs large crystals >5 µm
	Neutron	1–10/>0.5 mm	180°–360°	Hydrogen atom location	Deuterated sample, needs spallation source or reactor
	Electron	1–10/<1 µm, bidimensional	Stage limits rotation	Visualize electrostatic potential	Dynamic effects, absorption
Serial synchrotron crystallography	X-ray, electron	10–10,000/1–20 µm	1°–20°	Mitigating radiation damage	Complicated experiment; most data quality indicators are unsuitable for deselecting bad datasets
Serial femtosecond crystallography	X-ray	10,000–1,000,000/1–20 µm	0° ('still')	Overcoming radiation damage	Difficult to index; imprecise partiality estimates; difficult crystal preparation, requires large amounts of sample; low beamtime availability

These experiments are suitable for small, radiation-sensitive crystals that only deliver partial datasets.

Multicrystal still experiments combine 10,000–1,000,000 exposures taken without rotation, each from a fresh crystal in a random orientation[53]. This yields highly incomplete reflection datasets consisting of partially measured reflections. Estimating the full intensity from the partial measurements is challenging and the subject of ongoing research[54,55]. In addition, the incomplete reflection datasets must be scaled relative to each other. These experiments typically utilize an X-ray free electron laser (XFEL), with the advantage of avoiding radiation damage as the exposures are shorter than it takes to destroy the crystals[56]. Diffraction experiments vary in the type of radiation (such as X-rays, electrons and neutrons[57]; Table 1) and the wavelength range (monochromatic versus polychromatic). Polychromatic (also known as Laue) experiments are rare given the complications to factor the wavelength distribution into the intensity determination, and significant reflection overlap.

Practical considerations for synchrotron data collection

The basic steps of a single crystal X-ray diffraction experiment include mounting the macromolecular crystal in the X-ray beam, optical or raster-based centring, characterization of the diffraction limit, adjustment of the detector distance and collection of the reflections. Data collection seeks a compromise between minimizing the radiation damage and maximizing the accuracy of the intensity estimates, the resolution and the completeness. The size of the beam should match that of the crystal. A wider beam produces more background scattering whereas a smaller beam does not fully utilize the scattering potential of the crystal. Perfect control of crystal positioning is required to keep it in the beam while being rotated during data collection[58]. The distance between the crystal and the detector should be chosen such that the high-resolution limit of the diffraction pattern is within the edge of the detector. For the first dataset of an unknown crystal, a short distance that allows to reach about 1.5 Å resolution is convenient. Generally, the crystal is rotated around a single axis over the total range of 360°, with the detector providing a readout every 0.1° of the counts recorded by every pixel during this incremental rotation. The value of 0.1° is below the mosaicity of most crystals, leading to a fine sampling of reflection profiles, avoidance of overlaps, minimization of the background and reduction of overloads[59,60]. Using a smaller rotation width has no disadvantages but negligible benefits, and is only necessary for highly ordered crystals, such as those from viruses, that diffract to high resolution. As the total dose is thus distributed over 3,600 frames, these frames appear to be weakly exposed when viewed individually. However, owing to the high multiplicity of observations rendered by the total rotation, the information per unique reflection is maximized. The radiation dose that a crystal can tolerate before the diffraction power goes down by 30% is called the Garman limit and is in the order of 30 MGy (ref. 61), or lower if the crystal contains metal atoms that strongly absorb X-rays. To mitigate damage from radiation, the total dose for the experiment is chosen as a fraction of the Garman limit[62], such as 1/2 to 1/10, accomplished by adjusting the attenuation of the beam and the exposure time of each frame (radiation dose calculation). The choice of dose depends on the target resolution as high-resolution reflections suffer the strongest radiation damage.

The diffraction pattern

Figure 2 shows typical diffraction patterns, ranging from excellent in quality (Fig. 2a) to problematic (Fig. 2b–f). These images or frames show the projection of a slice through the reciprocal lattice, recording reflection intensities without phases. They provide useful information on lattice parameters, structure periodicity and diffraction properties, but it is generally not possible to read or understand a diffraction pattern in connection to the atomic structure. Figure 2 shows partial spots that need to be integrated to derive intensities for a reflection and scaled to have consistent data; only then can structure solution start.

Software for remote and automated data collection. Data collection can be integrated into a single workflow for high throughput. Full beamline automation enhances reproducibility and allows for the remote control of an experiment from the home laboratory using laboratory information systems such as ISPyB[63]. Graphical user interfaces such as MXcuBE[64], Blu-Ice[65] and MxDC[66] are used for adjusting beamline configurations, optimizing alignment, changing beam parameters and managing samples without physical intervention while allowing sample visualization. Standardized sample holders enable the seamless use of robotic equipment across beamlines worldwide, but one should prepare ahead.

Ligand and fragment screening. Automated beamlines with protein-to-structure pipelines[40,41] enable extensive experiments for structure-based drug design, providing insight into binding sites and modes. Beyond large-scale ligand screening[67], soaking or co-crystallizing a given protein crystal form with potentially active

chemicals, the fragment screening approach uses libraries composed of small and simple molecules with a variety of functional groups and chemical properties[68]. As an alternative to screening larger and more complex molecules, the use of fragment libraries allows efficient exploration of the accessible chemical space around a certain target with a limited set of probes. Initial fragment hits typically have low affinity and specificity, which is improved towards drug candidates by fragment growth or fragment merging. Some facilities support large-scale fragment screening (with over 1,000 fragments) such as the XChem facility at the Diamond Light Source or the HTX laboratory at the European Molecular Biology Laboratory, Grenoble[41,69]. Additional details about these facilities are provided in Supplementary Table 1. Using automated platforms benefits companies and academic groups in the development of chemical probes, such as the identification of highly potent small-molecule modulators targeting a specific macromolecule[70], accelerating costly preclinical development. Chemical probes can be applied to study function in vivo and to establish the therapeutic or biotechnological potential of certain targets, facilitating the transition from fundamental to applied research.

Time-resolved serial crystallography

Serial crystallography, initially developed at XFELs, is now used at synchrotron sources to capture structural snapshots of biological molecules in action. Time-resolved serial synchrotron crystallography (TR-SSX) is restricted to probing timescales of microseconds and above, whereas time-resolved serial femtosecond crystallography enables studies on timescales starting at hundreds of femtoseconds[53]. Most enzymes have turnover rates in the millisecond-to-second range[71], and the diffusion of small molecules into crystals in mixing experiments always exceeds microseconds, making them suitable cases for TR-SSX.

The most commonly used sample delivery platforms for TR-SSX are fixed targets (utilizing photoactivation or rapid mixing) and high-viscosity injectors (with photoactivation). Tape-drive methods, highly effective at XFELs, are now adopted at synchrotrons[72–74]. Microfluidics methods promise to improve TR-SSX[75,76].

Complementary spectroscopic techniques

Complementary techniques integrated into the experimental set-up allow simultaneous characterization from the same crystal sample. X-ray absorption, ultraviolet (UV)–visible absorption, Raman and infrared spectroscopies enrich our understanding of molecular structures and dynamics[77]. X-ray absorption and X-ray absorption near-edge structure spectroscopy reveal the oxidation state and local environment of metal centres, whereas UV–visible absorption verifies electronic states of chromophores. Raman and infrared spectroscopy add insights into molecular vibrations, chemical bonding and secondary structure, complementing macromolecular analysis beyond atomic coordinates. In time-resolved experiments, these techniques are adapted to capture molecular and electronic changes in real time within dynamic processes such as enzyme catalysis.

Safety and ethics considerations

Exposure to high-energy radiation is hazardous, therefore synchrotrons contain and monitor radiation maintaining a dose under

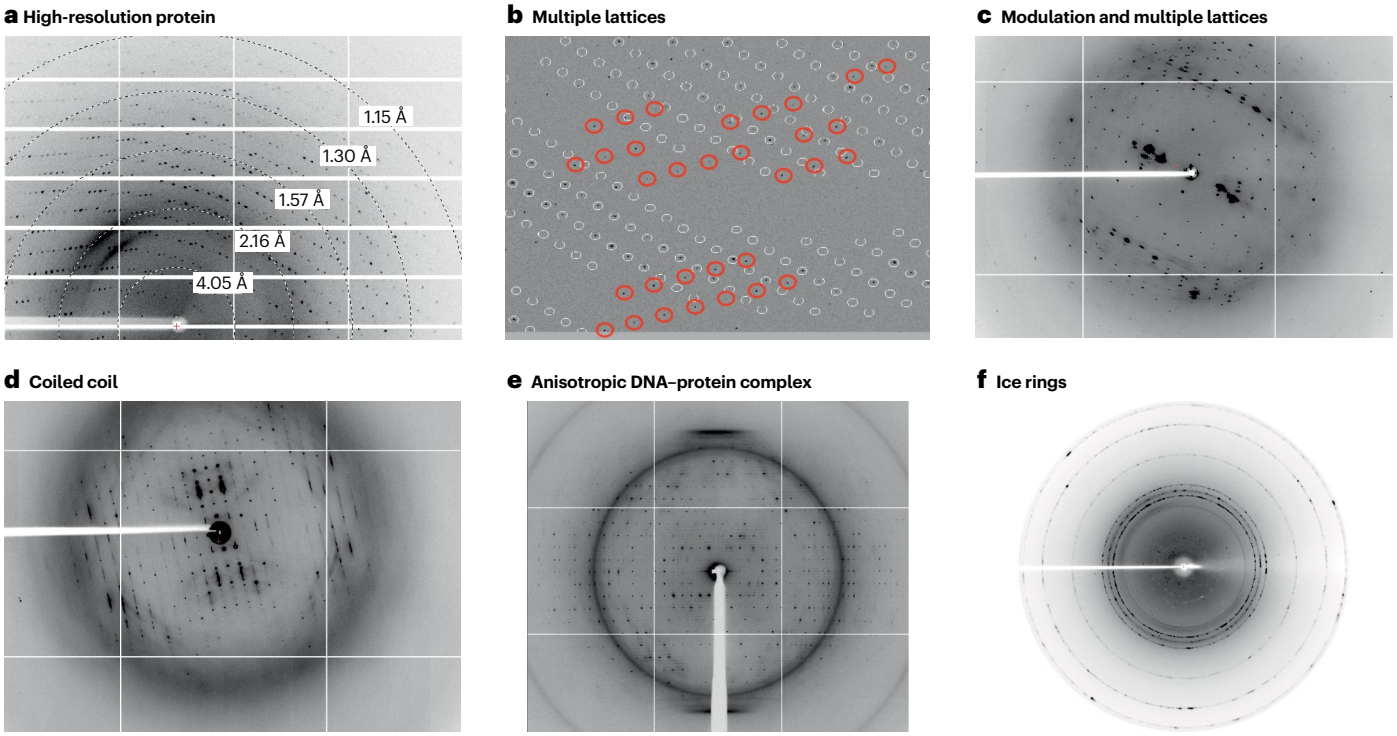


Fig. 2 | X-ray diffraction patterns with 1° crystal rotation. a, A single crystal of lysozyme diffracting to high resolution with low mosaicity. **b**, Although most reflections belong to the lattice in white, a second lattice in a different orientation can be indexed and is shown in red. **c**, Multiple lattices and diffraction modulation can be seen in layers of strong and weak reflections. **d**, A coiled coil, evidencing the underlying helical symmetry, is reminiscent of DNA patterns. **e**, The diffraction limit in the horizontal direction is much higher than in the vertical one. **f**, Ice crystals within or around the sample appear as rings at characteristic resolutions superposed to the pattern.

1–20 mSv year^{-1} for all staff and users. Multiple passive safety systems operate to avoid accidents. For example, opening the door of an experimental hutch while the X-ray shutter is open will halt not just the beamline, but the whole electron storage ring. Users are trained to actively collaborate by making everyone's safety their personal concern. Approval for each experiment needs to be requested in advance so that specific risks, such as an infectious hazard, may be assessed. Specific measures may be arranged to contain risks, or the experiment may be turned down.

Results

MX generates an experimental, atomic model of the species in the crystal. The way from the data to this result is computational, and can be tested, repeated and optimized. It starts with data processing and statistical analysis of the diffraction intensities, and requires a phasing method to establish an initial map and simple model, which will be further built and become increasingly detailed in refinement. Decisions on where to allow degrees of freedom or supply prior knowledge parameterizing this process are important owing to the phase problem, as errors in the model bias the determination. This is especially concerning at low resolution, where experimental data are limited and prior knowledge dominates.

Data processing

Extracting the scattered intensity of every reflection arising in a diffraction experiment is called data reduction as the raw diffraction data, stored in individual image files known as frames with many pixels, are condensed to a list of (h,k,l) triples with their associated values for intensity (I_{obs}) and standard error (σ). The reduction may be 100–1,000-fold. The statistical analysis of these 'observations' delivers precision estimates that, presented as overall values and as a function of resolution, are used to decide on the use of resulting data for calculations that elucidate the structure of the molecules that were crystallized.

Single-crystal data processing. The principles of monochromatic data processing are common for neutron, electron and X-ray experiments, and the same software is used for all three types of experiments. In 2023, 9,605 X-ray crystallography entries were released by the PDB. For 6,188 of these, the XDS data processing package[78,79], including the autoPROC[13] pipeline, was used; other data processing packages often used are HKL/DENZO[80], DIALS[62], including xia2, and MOSFLM[81].

The starting point for data processing is a series of diffraction images. Each reflection spreads out over several images and is therefore only partially recorded on any single image. Determining the intensity of a reflection requires precisely predicting which images each reflection will occur on and analysing the counts recorded in the corresponding pixels.

Steps in data processing. Data processing involves three steps: indexing, integration, and scaling and merging. Indexing determines the crystal parameters based on reflection positions on the detector and the corresponding rotation angle. Reflection positions are determined by spot finding where the local count maxima on each frame is identified. This process is computationally intensive as each pixel in a subset of images must be evaluated. The subsequent analysis establishes the unit cell parameters, probable Laue groups and crystal orientation. Prior knowledge of cell and symmetry facilitates the indexing step. Indexing can be complicated by weak diffraction signals, crystal splitting, radiation damage, anisotropic diffraction, diffuse scattering, ice rings/spots, multiple lattices, high mosaicity, unresolved or overloaded spots and shaded parts of the detector. Furthermore, instrument malfunction or poor parameterization of the diffraction geometry may disrupt this step. If the causes are identified, tuning the parameters for indexing allows the processing to be adapted to specific situations while incorporating previous knowledge and experience.

In integration, the raw intensity of reflections on the diffraction images is evaluated while simultaneously refining crystal orientation, reflection profiles and detector parameters. This step requires meticulous record keeping as the counts stored in the pixels must be correctly assigned to their nearest reflection profile or to the background underlying the reflections, and the contributions from all images over which a reflection is recorded must be weighted according to the model for the reflection's 3D profile. This time-consuming step determines the throughput of the computational pipeline.

In scaling and merging, symmetry equivalent reflections, which should have the same intensity, are used to bring measured intensities on the same scale. This corrects differences in absorption and the reflection-specific path through the Ewald sphere. Outliers are discarded at this step. Then, the remaining differences of the scaled intensities of symmetry-related reflections provide an estimate of the precision of the individual reflections, and statistics as a function of rotation angle and resolution are calculated (Table 2). Finally, a weighted average for the scaled integrated equivalent reflections yields the intensity of the unique reflection and its standard error (see formulas $\langle I/\sigma \rangle$ in Table 2). The resulting dataset contains merged reflections.

Error types and error estimation. During scaling, the relationship between the intensities of the reflections and their estimated intensity errors is iteratively updated, establishing a self-consistent error model. The estimated errors should, on average, match the observed differences between the intensities of symmetry equivalents. Random intensity errors are estimated from photon-counting statistics whereas systematic errors are attributed to three main sources: crystal imperfections (including damaged or multiple lattices), hardware instability (such as varying beam flux, vibrations, nonlinear detector responses or shadows) and software inadequacies (regarding processing parameters, corrections or models for reflection profiles).

The error model uses two empirical parameters and updates the error estimates accordingly. One of these parameters estimates the effect of systematic errors on the intensities. Furthermore, the indicator ISa (Table 2) estimates the reciprocal of the fraction of systematic error[82], with typical values between 3% and 10%. Values of ISa below 5 are often a sign that data processing has failed; in that case, the error model signals an excessive systematic error.

Limits of automatic data processing. Data are automatically processed at most beamlines on site-specific computational pipelines, often based on the XDS and DIALS software packages. These systems allow visualization of the results and even initial structural analysis. Additional metadata such as space group, cell parameters and PDB model, if available, are usually collected. It is convenient to build upon these preliminary results, but the interpretation of statistical data quality indicators is not straightforwards, and automated choices may be suboptimal. Processing software can miss the issues detailed below that require manual processing.

Choosing the optimal compromise between completeness and avoidance of radiation-damaged data requires image inspection, including the masking of beam stop or other shadows cast on the

detector, which alter reflection intensity. Ice rings from water crystallized in and around the sample due to improper cryoprotection or sample handling need to be excluded if measurement of a less defective sample is not possible. This can be checked visually or automatically with AUSPEX[83], a data diagnostics tool that uses machine learning to estimate the background[84].

Space group issues are often connected to incomplete datasets or aberrant diffraction derived from translational non-crystallographic symmetry (tNCS), twinning, split crystals and lattice translocation disorder[85–87]. Twinning, where several lattices are simultaneously diffracting, is very frequent and unavoidable in some samples. Some space group issues can be visually recognized from the spot positions and others should be identified by statistical *H* tests and *L* tests, which are calculated by most scaling programs as well as the program phenix.xtriage[88,89]. In these cases, the correct lower Laue symmetry space group should be chosen. However, the *L* test can be biased by tNCS[90] as well as spots that come close together because of a large unit cell axis[91]. Visualization of processing results is available in program packages such as the CCP4 suite[11], XDSGUI[92] and the HKL suite[80].

Making sense of crystallographic statistics

Crystallographic statistics provide numerical indicators to judge the diffraction quality of crystals, the stability of the experimental set-up and the success of data processing. Correlation coefficients ($CC_{1/2}$ and its derived quantity CC*) allow deeper insight into the relation of data and model quality than the residual R_{merge}[45,93]. Published tables of crystallographic data typically incorporate a multitude of related, but subtly different, indicators. We have compiled the most relevant statistics in Table 2. It is important to avert major points of confusion, which prompt wrong decisions when indicators are used outside of their domain, for example, about suitable resolution cut-offs. Precision is the difference between independent measurements of the same quantity (reproducibility), whereas accuracy is the difference between a measurement of a quantity, or the average of such measurements, and its true value. Indicators estimate precision and overestimate the accuracy of the resulting intensities because they neglect the systematic error.

A traditional group of ad hoc precision indicators are the *R* values, where *R* stands for reliability or residual[93]. Data *R* values measure the agreement, within a dataset or across several scaled datasets, of the measured intensities of symmetry-related reflections and include the term R_{sym} (also referred to as R_{merge}), R_{meas} or R_{pim}. Model *R* values (R_{work} and R_{free}) measure the agreement of diffraction amplitudes (calculated from measured intensities) with amplitudes calculated from the model and will be introduced in refinement. Textbooks and structure deposition requirements perpetuate the traditional role of data *R* values, despite their known weaknesses. For example, R_{sym} is multiplicity-biased and needs a correction factor, which then yields R_{meas} (refs. 94,95). The general understanding that R_{sym}/R_{meas} is related to R_{work}/R_{free} is unfounded, not only because the former are based on intensities whereas the latter are calculated from amplitudes, but more so because their asymptotic behaviour (for weak data and poor data/model agreement, respectively) is completely different. Additionally, R_{sym}/R_{meas} measures the precision of unmerged data, whereas refinement (which reports R_{work}/R_{free}) is conducted against merged data in macromolecules.

Resolution cut-off and anisotropy

A resolution cut-off is required at the point where noise is recorded rather than diffraction. This has traditionally been based on heuristics, for example, data with R_{sym} or R_{merge} >60% should be excluded, or by optimization of R_{work} and R_{free}. However, the goal of refinement is not to minimize *R* values, but rather to obtain the best model consistent with the data. This principle is evaluated in paired refinement where R_{free} values are compared for a common set of reflections, and between models obtained at different high-resolution cut-offs[45,96]. A directional fall-off of the mean reflection intensity is frequent and reflects anisotropic order in the crystal. With strong anisotropy, a single resolution cut-off would either discard sound data or incorporate noise as data (Fig. 2e). Possible solutions include using the anisotropic high-resolution cut-off

Table 2 | The main precision indicators for evaluating the quality of crystallographic data

Indicator	Meaning and property	Formula
Indicators using unmerged (individual) observations[a]		
$\langle I_i/\sigma_i \rangle$	The mean signal-to-noise ratio for *N* observations in a resolution shell. I_i is the measured intensity and σ_i is the estimated standard error	$\langle I_i/\sigma_i \rangle = \frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\sigma_i}$
ISa	The asymptotic upper value of I_i/σ_i limited by the level of systematic error in a given dataset. Only given as overall value	ISa = $1/\sqrt{ab}$ with the error model using the numerical parameters a and b, obtained by a least-squares fit, for a given dataset as $\sigma^2_{i,corrected}(I_i) = a(\sigma_i^2(I_i) + bI_i^2)$
R_{meas}	The relative deviation of symmetry-related observations from each other. *n* is the multiplicity of observations of unique *hkl* reflections. $R_{meas} \approx 0.8/\langle I_i/\sigma_i \rangle$	$R_{meas} = \frac{\sum_{hkl}\sqrt{\frac{n}{n-1}}\sum_{i=1}^{n}\lvert I_i - \bar{I}\rvert}{\sum_{hkl}\sum_{i=1}^{n} I_i}$
R_{merge}=R_{sym}	Similar to R_{meas}, but biased towards low values for low-multiplicity *n*. Still in use but deprecated	$R_{merge} = \frac{\sum_{hkl}\sum_{i=1}^{n}\lvert I_i - \bar{I}\rvert}{\sum_{hkl}\sum_{i=1}^{n} I_i}$
Indicators for merged (averaged) reflections[b]		
$\langle I/\sigma \rangle$	The mean (in resolution range) signal-to-noise ratio for the intensity (*I*) and standard error (σ) values of unique reflections that result from averaging *n* symmetry-related observations	$I = \frac{\sum_{i=1}^{n}\frac{I_i}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}$ with $\sigma = \sqrt{\frac{1}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}}$
$CC_{1/2}$	The Pearson correlation coefficient between internally averaged half datasets (x and y), each comprising half of the observations of each unique reflection The second formula gives approximate relationship of $CC_{1/2}$ to $\langle I/\sigma \rangle$	$CC_{1/2} = \frac{\sum_{hkl}(I_x - \bar{I}_x)(I_y - \bar{I}_y)}{\sqrt{\sum_{hkl}(I_x - \bar{I}_x)^2\sum(I_y - \bar{I}_y)^2}}$ $CC_{1/2} \approx \frac{1}{1 + \frac{2}{\langle I/\sigma \rangle^2}}$
CC*	Estimates the correlation coefficient against (the unobserved) true intensities, assuming that no systematic error exists	$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}$
R_{pim}	Estimated relative error of average of symmetry-related observations. $R_{pim} \approx 0.8/\langle I/\sigma \rangle$	$R_{pim} = \frac{\sum_{hkl}\sqrt{\frac{1}{n-1}}\sum_{i=1}^{n}\lvert I_i - \bar{I}\rvert}{\sum_{hkl}\sum_{i=1}^{n} I_i}$

[a]To support decisions between space groups or to detect radiation damage. [b]To assess data suitability for downstream calculations in phasing and refinement.

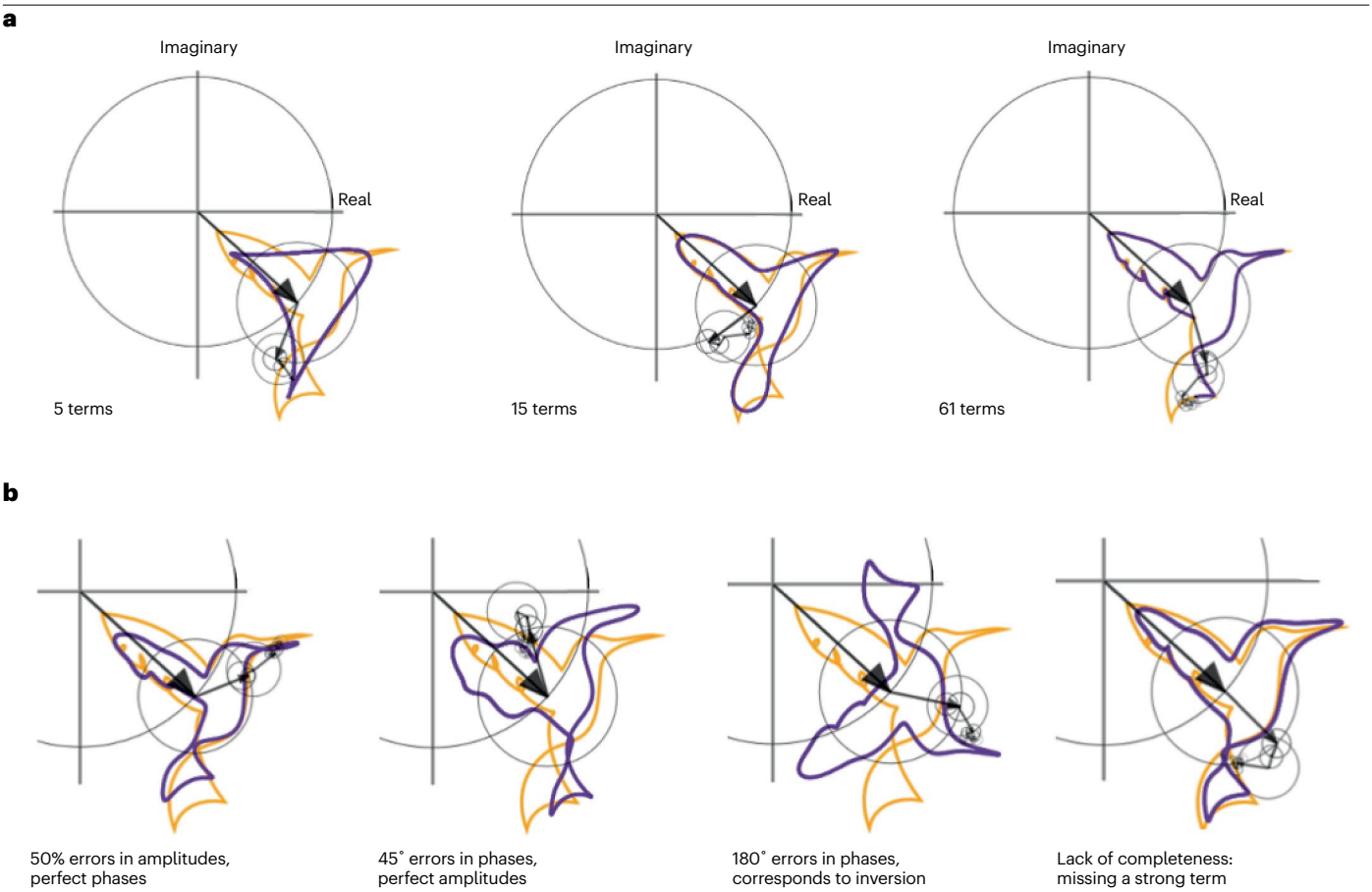


Fig. 3 | Properties of the Fourier transform and implications for resolution and errors in crystallography. The calculation of a path in time space (using equation (3)) approximates a line drawing from its Fourier coefficients in (reciprocal) frequency space (related by equation (4)). Each position (violet) along the example hummingbird drawing (gold) is given in coordinates in the complex plane (where x is the real component and y the imaginary) and represents a point in time, t. Each frequency term is depicted by a black arrow and adds a complex vector. The frequency terms approximate the path of the hummingbird as the arrows rotate in the grey circles but only with the right angular offset (phase). **a**, The effect of the number of terms used in the Fourier transform. The inner, slower rotating vectors (low resolution) give the overall shape, the outer, faster vectors (high resolution) add the detail. **b**, The effect of errors. See Supplementary Fig. 1 for more details.

server STARANISO, with mitigation of anisotropy by upscaling weak directions and downscaling strong directions, or careful selection of reflections according to their information content[97].

The crystallographic phase problem and the structure solution

The electron density in real space and the structure factors expressing diffraction in reciprocal space are inversely related by the Fourier transforms in equations (1) and (2)

$$\rho_{xyz} = \left(\frac{1}{V}\right) \sum_{hkl} F_{hkl} \exp[-2\pi i(hx + ky + lz)] \quad (1)$$

$$F_{hkl} = \int_V \rho_{xyz} \exp[+2\pi i\,(hx + ky + lz)]dV. \quad (2)$$

The electron density ρ is real and positive but the structure factor F is a complex number. Knowledge of the electron density in a crystal allows the scattering to be calculated, but the diffraction experiment does not yield the structure factors required to calculate the electron density map from which an atomic model can be interpreted. If it did, the Fourier transform could be trivially used to compute the electron density. The recorded intensities provide the squares of the scalar moduli of the structure factors but not the phases needed. This is known as the phase problem and is central to crystallography. Obtaining approximate phases is necessary to solve the structure and build an initial model, later improved and completed in refinement[98]. Model bias is a consequence of the impossibility of measuring phases as in subsequent calculations, the experimental diffraction amplitudes are used, whereas the phases calculated from the current model are adopted. As phases bear more information than amplitudes, errors in the model enter the calculated electron density map and compromise the determination.

The phase problem and its implications are visualized in Fig. 3 using an animated line drawing (Fourier transform animations) where the pair of Fourier transforms depends on the single variables time (t) and its reciprocal frequency (n).

$$\text{Position}_t = \sum_n \text{Coef}_n \exp[-2\pi i(nt)] \quad (3)$$

$$\text{Coef}_n = \int_L \text{Position}_t \exp[+2\pi i(nt)]dL \quad (4)$$

Figure 3 captures how the path at each time, t, depends not only on the moduli of the frequency term vectors, but more decisively on their relative angles (phases). The diagram also highlights how every term impacts the path and how every point in this path participates, informing the frequency coefficients. Likewise, every structure factor contributes to each atom and every atom contributes to each structure factor.

Figure 3a shows the effect of resolution in the reconstruction of the positions along the path. If the resolution is too low, even with perfect data and phases but few terms, the structure is not revealed. At low resolution, fine details are missed and truncation by the Fourier series induces artefactual features. For instance, the bulge between the tail tips in the case of 15 terms is an artefact in the example image. If the artefact was interpreted as a feature of the model and incorporated into the reconstruction, the new calculated phases would reinforce the error. With a large number of terms, corresponding to high resolution, the approximation approaches the ground truth.

In practice, both amplitudes and phases are impaired by errors. As illustrated in Fig. 3b, the reconstruction has a higher tolerance towards errors in the amplitudes. Given perfect phases, even large deviations in the amplitude (up to 50%) allow the original image to be recognized. By contrast, severe distortion arises from errors in the phases. If all phases are shifted by 180°, the identical image appears inverted. When a test phasing solution is compared with the known reference, an incorrect trial with random phases has a mean phase error of 90°, not 180° (orthogonal, not anticorrelated). A lack of completeness, especially for large amplitudes, also deteriorates the reconstruction of the map. Missing one large term out of the 61 used in Fig. 3a (right), shows remarkable degradation in Fig. 3b (right). For this reason, extrapolating unmeasured reflections, even beyond the experimental resolution limit, is frequent in phasing, and can be determinant to correctly solving the structure[99,100].

Phasing is a multidimensional search rather than an optimization problem, complicated by a small radius of convergence. In general, determining phase angles within 60° of the true phase for 10% of the strongest reflections in each resolution shell suffices to render a starting solution. Phasing is aided by boundary conditions; for example, the electron density map can only adopt zero or larger values but positivity does not hold for the electron-diffraction electrostatic potential maps nor for neutron diffraction. Phases are subject to symmetry in reciprocal space, which poses constraints depending on the space group. Electrons are mainly bound in atoms, although the atomicity constraint loses its strength at lower resolutions when atoms are not resolved. Given these boundary conditions, different phasing methods are available for different situations.

Ab initio phasing. Chemical structures, typically containing less than 200 non-hydrogen atoms in the asymmetric unit, can be solved ab initio from the recorded intensities alone, relying on boundary conditions derived from atomicity, without the use of additional experiments or particular knowledge about the stereochemistry of the structure. Multiple combinations of random starting phases for the strongest reflections in each resolution shell are refined or expanded, subject to direct-methods probabilistic relationships[101]. At atomic resolution of 1 Å or better, this procedure allows a solution to be reached, identified and improved in less than a second. For larger structures, dual-space recycling methods enforce atomicity iteratively by using phase relationships in reciprocal space and interpreting the map calculated from the resulting phase set by selecting as many atoms as the structure should contain[102]. An improvement to escape model bias is discarding a fraction of the selected atoms at random[103]. In MX, substructures of anomalous scatterers or heavy atoms also need to be determined by the same ab initio methods.

In addition to direct methods, the Patterson function can be calculated by a Fourier transform using the recorded intensities as coefficients and setting all phases to zero[104]. Box 1 describes how to navigate between reciprocal and real space when using different coefficients for the structure factors (see Supplementary Fig 2 for additional context). This corresponds to a self-convolution of the electron density, which bears the physical interpretation that the value at each point of the Patterson function depends on the correlation between the electron density function and a copy displaced by a vector from the origin to this point. Hence, the Patterson function attains its maxima for a shift that either brings points of high electron density to coincide (vectors relating strongest scatterers) or due to the accumulation of multiple interatomic vectors relating pairs of light atoms that coincide in distance and direction. Trivially, it has a maximum for zero displacement, used for scaling all values.

The Patterson function can be used to directly locate heavy atoms in chemical structures, but as the number of peaks corresponding to intramolecular vectors in a structure composed of N atoms would be $N^2 - N$, accidental overlap prevents identification of its peaks in macromolecular structures. Nevertheless, whether modified or combined, the Patterson function is present and very useful in macromolecular applications[105].

Direct methods fail if the resolution does not reach at least 1.2 Å (ref. 106) and atoms are not resolved. At lower resolution, enforcing atomicity in reciprocal space (using direct methods) and in real space (through peak picking) is not effective. Furthermore, the figures of merit used to identify solutions at atomic resolution — primarily the correlation coefficient between the normalized observed intensities and those calculated from the atoms in the solution[107] — fail to discriminate whether a solution composed of unconstrained atoms is correct or incorrect. Within dual-space recycling methods, substituting random starting atoms with guided hypotheses, incorporating the Patterson function, or strategically placing randomly rotated and translated fragments can be advantageous[108,109]. At medium resolution (up to 2–3 Å), small but accurately located helical fragments can be expanded by density modification and the resulting map or its trace can be discriminated as correct by the correlation coefficient when the structure is solved[110]. This combination of fragment placement and density modification is the basis of ab initio Arcimboldo methods[111].

Density modification. Density modification encompasses a variety of methods to express boundary conditions on a map calculated with partially correct phases and alter it to better comply with known properties of macromolecular maps and the expected characteristics of the particular structure[112,113]. The modified phases, combined with the original ones to limit bias, should lead to an improved, more probable map from which the process can be iterated[114,115]. Solvent flattening exploits the known property that density has a higher variance in the map regions occupied by the macromolecule than in the solvent. Enforcing a flat electron density distribution in areas of solvent is very

Box 1 | From reciprocal to real space with different coefficients

Fourier transforms are used in crystallographic calculations to convert a description of the same phenomenon from reciprocal to real space. With structure factors as coefficients, the Fourier transform gives the electron density ρ within the volume of the unit cell V

$$\rho_{xyz} = \left(\frac{1}{V}\right) \sum_{hkl} F_{hkl} \exp[-2\pi i(hx_j + ky_i + lz_i)],$$

where ρ_{xyz} electron density at the position with coordinates x,y,z in real space, $F_{hkl} = A_{hkl} \exp(i\phi_{hkl})$ is the complex structure factor for the reflection (h,k,l) in reciprocal space and V is the volume of the unit cell in real space. Expressing the exponentials with equivalent trigonometric terms gives

$$\rho_{xyz} = \left(\frac{1}{V}\right) \sum_{hkl} A_{hkl}\{\cos(\phi_{hkl}) \cos[2\pi(hx + ky + lz)] + \sin(\phi_{hkl}) \sin[2\pi(hx + ky + lz)]\}$$

Common coefficients for substituting the amplitude, A_{hkl}, with specified phase values result in the real space functions provided in the table.

Amplitude (A_{hkl})	Phase (ϕ_{hkl})	Electron density function (ρ_{xyz})
$\lvert F_o^2 \rvert$	0	Patterson
$m\lvert F_o \rvert$	ϕ_{exp}	Weighted density calculated with experimental phases
$\lvert F_o \rvert$	ϕ_c	Observed density
$\lvert F_o \rvert - \lvert F_c \rvert$	ϕ_c	Difference density
$m\lvert F_o \rvert - D\lvert F_c \rvert$	ϕ_c	Weighted difference density
$2m\lvert F_o \rvert - D\lvert F_c \rvert$	ϕ_c	Weighted density
$\Delta F = \lvert F_{hkl} \rvert - \lvert F_{-h-k-l} \rvert$	$\phi_c - 90°$	Anomalous density maps

The Fourier transform of a squared function, like the Patterson function, corresponds to an auto-convolution of the form

$$P(\mathbf{u}) = \rho(\mathbf{r}) \times \rho(-\mathbf{r}) = \int_{\mathbf{r}} \rho(\mathbf{r})\rho(\mathbf{u} + \mathbf{r}) d\mathbf{r},$$

(where **u** and **r** are positional, fractional vectors in the unit cell)

In the table above, the values of m and D are designed to reduce model bias and depend on the agreement between calculated and observed amplitudes, F_c and F_o.

powerful even at moderate resolution, if such areas can be correctly identified in the preliminary noisy map. In practice, solvent flattening algorithms can be formulated in reciprocal space as well[116,117]. Image processing algorithms such as histogram matching, averaging the density for non-crystallographic symmetry-related molecules or sharpening density in the protein regions are other methods for modifying the electron density[118].

Model building also constitutes a way of density modification. Autotracing, where regions of the map are interpreted in terms of a polypeptide or amino-acid model, is very effective as the map calculated from an accurate trace is constrained by regular stereochemistry and will extend phase information throughout the full resolution range. On the other hand, an incorrect trace will be deleterious, and hence extending partial solutions is less effective at low resolution, where model building loses accuracy. As in the ab initio context, extrapolation of unmeasured data enhances the map and its autotracing[100].

A number of implementations coexist, differing in their treatment of all outlined principles. For instance, solvent regions of the map can be defined globally with masks or assigned to each voxel depending on local map properties such as skewness or variance[119,120]. Non-crystallographic symmetry averaging for equivalent molecules can also be achieved by averaging the related map density values or replicating the trace across equivalent macromolecules[121]. Algorithms may perform equivalently or be better suited for a particular case depending on the data resolution and properties of the starting phases. Typically, experimental phases require noise reduction, whereas starting from a model depends on feature enhancement if partial and bias reduction if inaccurate[122].

Experimental phasing. Experimental conditions can be selected to differentiate a substructure of atoms within the total structure, for example, by choosing a wavelength that enhances the anomalous and dispersive contributions of a specific chemical element present in a limited number of sites. This element could be inherent to the macromolecule (for example, sulfur, phosphor or metals present in cofactors, which allow native phasing[123]) or be introduced into the crystallographic order. Differences in the diffraction intensities recorded among Friedel opposites (single-wavelength anomalous diffraction), for the same reflection at multiple wavelengths (multiple-wavelength anomalous dispersion) and for isomorphous crystals where heavy atoms are present or absent (single isomorphous replacement or multiple isomorphous replacement) can be related to the subset of atoms that modify the scattering. In a first step, the problem becomes that of solving a small molecule from the approximate difference data, which requires the more constrained dual-space recycling methods improving on direct methods[124]. Subsequently, the localized marker atoms and trigonometric relationships derived from the vectorial sum of atomic scattering factors can provide approximate phases for the complete structure. Maximum-likelihood methods might allow a more detailed model of the substructure, leading to better phase estimates[125,126]. In the case of multiple-wavelength anomalous dispersion phases, they may be accurate enough to keep as additional information in low-resolution refinement but, in general, they are improved through density modification[127]. In that case, the quality of the data being phased is limiting and the ancillary data providing the initial experimental phases are discarded and rarely deposited.

Molecular replacement. Given the availability of experimental models and accurate predictions, most determinations have a model hypothesis from the onset. In that case, phasing becomes a search problem, where rotation and translation have to be determined for one or more components of a preliminary model. With good models and data, methods based on the Patterson correlation implemented in programs

such as AMoRe and MolRep are fast and successful[128,129]. Geometrical deviations between the search model and the target structure, model incompleteness, and errors or modulation in the data complicate the solution and require sophisticated methods to account for these circumstances[130]. Supplementary Fig 4 illustrates the approximation to the maximum-likelihood function implemented to score placed models in Phaser[131]. This formalism calculates a score that a given model would reach if correctly placed, given the available experimental data. This expected log-likelihood gain can be used throughout the full range of model size and deviation, for any resolution, and explains how large, complete models such as a ribosome particle can be placed at low resolution, whereas models as small as a single atom require unusually high resolution to recognize a successful placement[132]. Accordingly, the expected log-likelihood gain can be used to guide decisions in model choice and parameterization[133], and models can be given internal degrees of freedom and refined against the intensity-based log-likelihood gain (LLGI) scoring for a given placement[134]. Once the model is placed in the unit cell, it can lend phases to the experimental amplitudes and give way to refinement or model building.

Phase information from different sources can be combined, taking care to refer it to a common origin[135]. Molecular replacement with single-wavelength anomalous diffraction, a method known as MRSAD, uses a partial model and experimental phases to solve difficult cases or establish a local feature in a protein, nucleic acid or ligand, pinpointed through an anomalous scatterer[136].

Phasing is self-validating when it reveals additional, unanticipated structural features. For a nearly complete starting model, considering model bias is most important[137]. Phasing molecules with a periodic structure such as coiled coils or DNA helices is typically deceptive as wrong solutions reach high figures of merit[138]. This introduces the need for verification, a process to actively attempt to disprove the best solution by creating close but different alternatives and probing both the correctness of the solution as well as the potential of the data to discriminate it. This is particularly relevant with noisy data and at low resolution[139,140].

Refinement

Model refinement optimizes the atomic model to fit the experimental data and structural principles while avoiding overfitting[141]. In practice, refinement is an iterative computational procedure interspersing model building, parameter optimization and validation[142,143]. This process requires a crystal structure model, experimental data, prior knowledge, refinement target functions that express the fit of the model to the data and prior knowledge, an optimizer that adjusts model parameters to improve the refinement score, and validation criteria.

Model. The crystal model includes an atomic component, disordered solvent, and scaling or corrections that are linked to the description of crystal properties affecting diffraction. Anisotropy can be corrected statistically before there is a model, but is better accounted for by including scaling parameters into the model[144]. When twinning occurs, diffraction data can be attributed to more than one crystal[145]. For merohedral twinning, there is an exact coincidence of the lattices in different orientations so that all recorded intensities are the sum of different scattering components[146]. Reticular or non-merohedral twinning involves total, partial or no overlap, and experimental intensity data need to be differentiated according to the lattices involved[147,148]. In refinement, previous anisotropy or twinning corrections used in phasing can usually be discarded.

Crystals of bio-macromolecules contain, on average, 50% disordered solvent, or much higher in some cases[149]. As every voxel of the unit cell contributes to every unique diffraction reflection, the disordered bulk solvent cannot be disregarded. Before an adequate treatment of the bulk solvent component was introduced[150], a common practice was to remove reflections where this contribution was substantial, typically in the 6–8 Å range or lower; however, missing low-resolution reflections leads to map distortions[151,152]. The disordered component of the model is almost universally accounted for using the efficient flat bulk solvent model, which requires few parameters[153]. Deviations from the flat bulk solvent may be considerable, motivating the development of new algorithms to account for them[154].

An atomic model consists of point scatterers defined by their chemical type (which determines their resolution-dependent scattering and anomalous signal), charge, coordinates and disorder. Disorder arises because atomic positions correspond to an average in the crystal and the distribution about this average may be continuous (dynamic) or discrete (static). Small-scale disorder, within the harmonic approximation, is described by atomic displacement parameters (known as ADPs or B-factors), whereas large-scale, static disorder is captured by occupancy. Displacement can either be simplified to isotropic motion or detailed to capture direction-dependent (anisotropic) behaviour, at the cost of six, rather than one, parameter per B-factor. A sophisticated compromise between both models distinguishes groups of atoms for which atomic vibration is considered as a hierarchy of motions, composed of the atom vibrating on its own, motions of a side chain about its torsional chi-angles, the side chain moving as part of the entire chain or a fraction thereof[155]. This motion description uses a translation, libration, screw model for each of the groups[156,157]. Correct identification of rigid groups is essential for success of the translation, libration, screw model refinement[158].

Static disorder is an inherent property of macromolecular crystals, which becomes apparent at resolutions of about 3 Å and better. The disordered sites are modelled with discrete instances of atoms that occupy different specific locations[159]. At resolutions better than 2 Å, most models in the PDB present instances of alternative conformations.

Parameterization. PDB models preserve a uniform expression linked to atomic properties to make them comparable. This description does not necessarily match the parameters that have been refined. Parameters are the degrees of freedom in the model, adjusted to improve the fit. For example, atomic coordinates are always present in the deposited structure, whereas the parameters refined in the model may be torsion angles within a polypeptide chain[160]. The level of structural detail a particular dataset can reliably model and the questions underlying the structure determination guide the parameterization choices. Given limited experimental data, degrees of freedom should be prudently controlled to establish the main points of interest (evaluating, for example, how the ligand is bound in the active site) rather than allowing large uncertainty for trivially known stereochemistry (such as covalent bond lengths and angles known from atomic resolution structures). At resolutions poorer than 2.8 Å, it is not uncommon to have more model parameters than data points, compensated by using prior knowledge formulated as restraints or constraints[151].

Data. Experimental data primarily consist of the scaled and merged diffraction intensities, but can also include experimental phase information formulated as probability distributions or NMR data[161–164].

Restraints. Restraints enter the target function formally as additional observations (for example, a given bond length should agree with equivalent occurrences) whereas constraints are exact mathematical conditions reducing the number of parameters (for example, riding hydrogen atoms are rigidly tied to another atom and cost no parameters). Whereas constraints are absolute, restraints are weighted based on the quality of the data and the model, and compensate for shortcomings in the experimental data. For example, the asymmetric unit frequently contains independent copies of the same molecule. This non-crystallographic symmetry can inform refinement through similarity restraints between non-crystallographic symmetry-related copies, or constraints considering them identical and dividing the number of parameters by the number of copies[165,166]. Restraints are preferred at high resolution, whereas constraints might be necessary at low resolution.

Restraints include terms that parameterize the local chemistry of the model through covalent bonds, angles, torsion angles, planes, chiral volumes and non-bonded repulsion[167]. Other properties, such as electrostatic attraction or secondary structure are usually not accounted for. For known chemical entities, such as amino-acid and nucleic-acid residues or usual ligands, standard restraints are embedded in the software[168]. A novel ligand or a chemically modified entity requires creating parameters and topology definitions, manually or with software[169–171].

The set of restraints with a unique target may not suffice to stabilize refinement against low-resolution data (3 Å or worse). To prevent distortions in secondary structure and unrealistic main and side chain torsion angles, additional ad hoc restraints are used, such as secondary structure, Ramachandran plot and side chain rotamer-specific restraints, reference model restraints or Cβ deviation restraints[172]. However, using additional restraints has drawbacks. The Ramachandran plot, Cβ deviations and side chain rotamer distributions are classic atomic model validation tools and using them as active refinement targets reduces their utility as independent validators. Also, defining these restraints relies on the annotation of secondary structure and assigning pairs of (φ, ψ) angles to the correct region of the Ramachandran plot, which is error prone because it depends solely on the geometric quality of the atomic model. Thus, these restraints may cause bias. An emerging forward-looking approach is the use of quantum chemical calculations to derive restraints[173]. The primary limitation of this approach is that the model must be chemically sensible (such as containing no colliding atoms) and atom complete (for example, including all hydrogen atoms). The computation time is now acceptable for using quantum mechanical approaches in refinement[174].

Rigid-bond and proximity restraints are introduced to enforce physical limitations on the magnitude and direction of atomic vibrations and to express that atomic motions cannot be independent[175]. For example, atomic vibrations cannot excessively stretch covalent bonds. Additionally, the differences between B values for covalently bonded atoms are tied to tabulated expectations or a set of ad hoc rules that incorporate physically meaningful behaviour of atomic motions within a local region[176].

Box 2 provides an overview of how the different elements in refinement come together in key formulas. Supplementary Fig 5 contains additional information regarding the key formulas. First, structure factors are calculated from the model for the measured reflections (Box 2, step 1). The model is scored against the data and prior knowledge, calculating a refinement goal or target function composed of two terms: one dependent on the data and another on the agreement to prior knowledge expressed through ideal stereochemistry or energy terms[177] (Box 2, step 2). The weight (w) between both terms is calculated by the refinement software such that it maximizes the contribution of experimental data while ensuring the geometric model properties are still within the accepted ranges. The fit to the diffraction data can be represented by a least-squares function but requires an accurate model and high-resolution data[178]. Better suited for macromolecules, maximum-likelihood functions statistically account for model incompleteness and errors in model and data[163,179,180]. The crystallographic R factor in Box 2, step 3, is not a refinement function but a standard indicator calculated for the data used in the optimization (R_{work}) and in a 5% subset of the reflections (R_{free})[181]. R_{free} reflections are excluded from the optimization function and solely used for cross-validation and in maximum-likelihood refinement for the error model. As the reliability of R_{free} and the omission of data becomes problematic for a small dataset, alternatives such as $R_{complete}$[182] or Free-kick[183] are used in some low-resolution, high-pressure or electron-diffraction scenarios.

The refinement target function can be optimized by various algorithms, differing in convergence radius, scalability and ability to escape local minima. For example, systematic grid searches have the largest convergence radius and can cross energy barriers, but are practical only for optimizing one or two parameters at a time. In contrast, conjugate gradient minimization efficiently optimizes all parameters at once, but has a limited convergence radius (~1 Å for coordinate refinement) and cannot escape local minima[184,185]. Simulated annealing combines a stochastic search to escape local minima, but it does not guarantee convergence to an improved minimum[186]. More directed algorithms such as morphing[187], jelly body restraints[172] or gradual resolution extension[188] are preferred and the choice of method depends on the refinement stage. Gradient minimization is used for correcting small errors in atomic model parameters. Morphing and jelly body are used when the initial model requires large adjustments to fit the map. Grid searches are used to fit residue side chains into the density by sampling torsion chi-angles to find the best fit for a favourable rotameric state. Inversion of the least-squares matrix is a costly algorithm and is only used as a final step within SHELXL[189], to calculate standard uncertainties on parameters and characterize structure precision[190]. Its use has revealed generalizable insights on parameterization[191], otherwise, in macromolecular refinement, precision is generally not determined for each parameter, but rather estimated.

The location of hydrogen atoms will depend on the type of radiation as electron density will appear closer to the bond than to the nucleus. Neutron crystallography can experimentally confirm hydrogen atoms positions at typical macromolecular resolutions of 2–3 Å (refs. 192,193), whereas X-ray or electron crystallography require data at atomic resolution (~1 Å) to be conclusive. However, hydrogens are included in atomic models because they contribute to scattering, which affects R factors and maps. Most hydrogen atoms have predictable positions, although biochemically relevant ones are usually mobile and require determination.

In reciprocal-space refinement, the diffraction term uses intensities or derived amplitudes to match the model to the experimental data. Alternatively, refinement can be formulated in real space to match an atomic model against the density map[194]. Because the available model provides the phases to compute the map, real-space refinement is more prone to model bias. However, in graphical programs for model building[195,196] or cryo-EM where the map itself represents the experimental data, real-space refinement is the natural choice[197,198].

Box 2 | Refinement steps and detailed elements

Step 1: structure factors representing the unit cell scattering are calculated from a model comprising atoms, solvent and corrections

$$F_{hkl} = \sum_j f_j \exp[2\pi i\,(hx_j + ky_j + lz_j)] \exp\left(-\frac{B_j \sin^2\theta}{\lambda^2}\right) + F_{solv},$$

where x,y,z describe the atomic coordinates and B is the atomic displacement. Each atom in the unit cell scatters accordingly to an element-specific form factor, f_j, times its site occupancy. λ is the wavelength and θ the diffraction angle, expressing the resolution of the reflection. F_{solv} is the solvent contribution to the structure factor. A scaled or corrected function (F_c) is used to model anisotropy or twinning. This approximates the observed structure factors corresponding to the density in the unit cell ρ_{xyz}

$$F_{hkl} = \int_V \rho_{xyz} \exp[+2\pi i\,(hx + ky + lz)]dV$$

Step 2: refinement compares observed and calculated data as well as stereochemistry and prior knowledge through minimization of least-squares differences

$$M = \sum_{hkl} w_x (F_0^2 - F_c^2)^2 + \sum w_r (T_{target} - T_c)^2,$$

where w_x and w_r are weights for data and restraints, respectively. F_0^2 and F_c^2 are the observed and calculated intensities (squared amplitudes) of the reflections, T_{target} and T_c are the ideal and computed values for the structural values under restraint and M is the least-squares refinement function.

Alternatively, the Bayesian probability of the model giving rise to the scattering can be maximized given prior knowledge of stereochemistry. Refinement uses logarithms for computational reasons

$$P(\text{model}|\text{data}) = P(\text{data}|\text{model}) \times P(\text{model})$$

$$P_{refinement} = P_{xray} \times P_{stereochemistry}$$

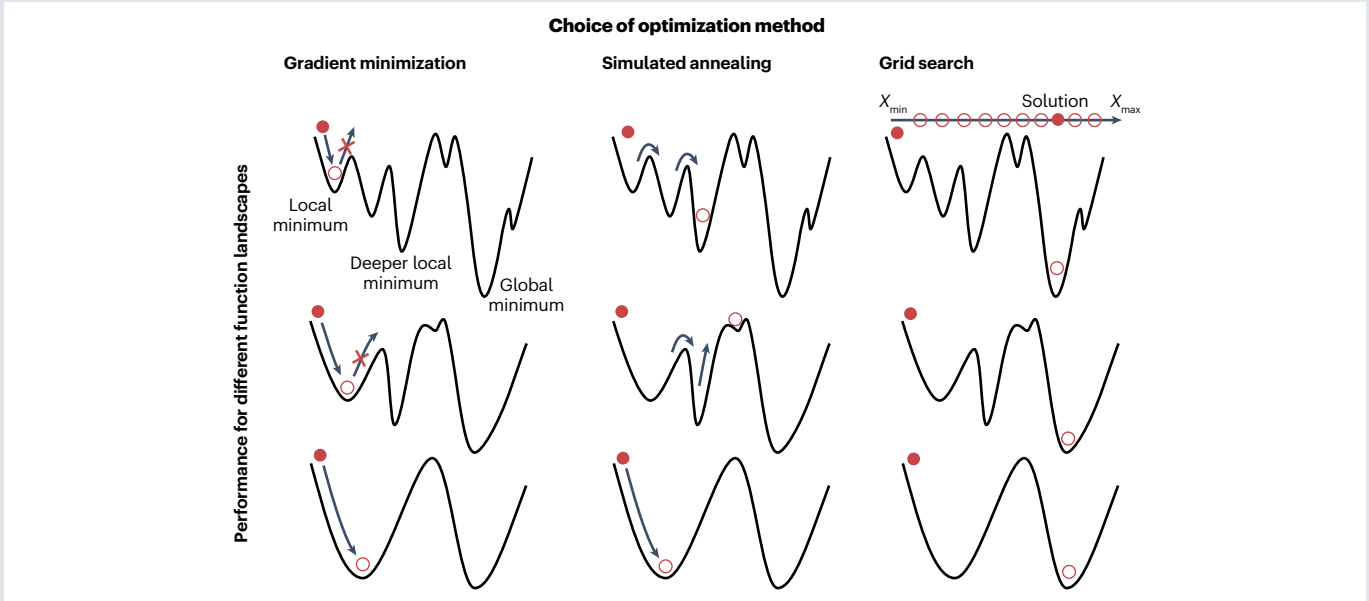
$$-\log(P_{ref}) = -w \sum \log(P_x) - \sum \log(P_{stereochemistry})$$

Step 3: a standard indicator is universally calculated for comparison purposes

$$R = \sum_{hkl} \frac{||F_o| - |F_c||}{|F_o|}$$

R_{free} is calculated with this same formula for a separate hkl subset.

Minimizers to optimize target functions are displayed in the figure. The landscape of a refinement function shows many local minima and varies depending on its large number of parameters. Reaching convergence into the global minimum requires an appropriate minimizer.



Density maps, central to structure determination, are approximations of electron (or nuclear or electrostatic potential) density distributions. Map quality will be affected by data completeness, especially with low-resolution data. In crystallography, a structural model is always involved in map calculation, contributing the phases. Therefore, errors and artefacts arise not only from the limitations and finite resolution of the experimental data, but also from inaccuracies in the models used to calculate the maps, thus perpetuating model bias. The implication for crystallographic refinement is that the map improves with the model. Predicted models have now been incorporated into model building by iteratively informing AlphaFold predictions with a partial model as template[199].

Convergence of the optimization process, meaning the absence of changes upon further iteration, should mark the end of refinement. Nevertheless, as correlations among parameters compromise convergence, refinement may yield models with stereochemical outliers and parts of the model may not be visible in the maps, model refinement may appear never ending[200]. Cross-validation provides an objective means to guide decisions in parameterization and model building[201], whereas tools such as POLYGON, which presents an overview of quality metrics in comparison with structures of similar resolution and alerts on issues related to incomplete refinement or model errors, can help to indicate when to stop refinement[202].

Applications

MX is broadly applicable across molecular biology to characterize biochemical processes. This section sketches a typical case in which unexpected results are revealed and describes a free application environment where all required calculations can be conducted without setting up equipment or installing software.

Case study on OBPs

ABC transport systems are vital for bacteria to acquire essential nutrients and maintain energy balance. *Streptococcus pneumoniae*, a pathobiont, relies heavily on these transporters for colonization and infection. Owing to its inability to synthesize certain amino acids, the Ami ABC transporter system, responsible for oligopeptide uptake, is crucial in amino-acid-deficient host environments. This system includes five oligopeptide-binding proteins (OBPs) and four proteins forming the transporter channel. X-ray crystallography provided atomic information about the rearrangement of OBPs upon substrate binding and substrate recognition.

The structural characterization of the OBPs in the Ami transporter system involved multiple X-ray crystallographic structures, capturing both open (apo) (Fig. 4a) and closed (holo) (Fig. 4b) conformations[203]. These complexes contained chemically synthesized oligopeptides that were solved at high resolution (Fig. 4c), providing insights into the molecular basis of their diverse peptide specificities. Some of the OBPs revealed an electron density at the substrate-binding site (Fig. 4d), pointing to the capture of oligopeptides during the expression and purification processes.

Contemporary approaches to crystallographic computing

Cloud-based crystallographic computing removes the need for laboratory-based facilities, offering a transformative solution for researchers of limited expertise or aiming to perform structure determination with minimal set-up. Automation has become essential as research increasingly addresses questions in structural biology, driving new methods and user-focused design. High-throughput, parallel computing is critical for efficiently analysing large numbers of datasets such as in fragment screening. Effective data management — including logistics, sharing, organization and archiving — is crucial for streamlined research. Crystallographic software must be user friendly, minimizing the need for specialized expertise while incorporating deep domain knowledge. Integration with bioinformatics and artificial intelligence tools, such as AlphaFold, is increasingly expected. Additionally, the software should be simple to deploy and maintain, adaptable to diverse scenarios (from budget laptops to high-performance computing facilities), and suitable for both academic and corporate settings.

CCP4 Cloud was designed to simplify access to CCP4 software, removing the need for local installation and maintenance while offering computational resources for advanced structure-solving methods[7]. Users can solve structures via a standard browser without installing software, regardless of the operating system or hardware. A publicly accessible instance is also available. CCP4 Cloud can be installed locally to control access and use local computational resources, or it can be run on a single laptop. It integrates with synchrotrons and data sources, streamlining workflows for handling large experimental datasets.

In CCP4 Cloud, work is organized into projects and tasks. User workspace is private, but projects can be shared with teammates for real-time collaboration. Users define the project scope and hierarchy, whereas task hierarchy emerges automatically to reflect the structure solution process. Tasks in CCP4 Cloud have coarse-grain functionality, often comprising multiple program components. For example, the structure refinement task includes Refmacat[204] and a validation module with BAverage, EDStats[205], PISA[206] and MolProbity analysis[207].

A typical structure determination project in CCP4 Cloud includes tasks such as image processing, scaling, merging, solving the phase problem, density modification, model building, refinement, water and ligand fitting, validation and PDB deposition. Currently, 99 tasks condense around 250 CCP4 program components. The project development system streamlines working with these tasks by checking data and task compatibility, resolving ambiguities and suggesting steps based on history, available data and user habits. It generates branched task trees that logically represent the structure solution process, simplifying understanding, revisiting and revising workflows.

CCP4 Cloud offers advanced automated tasks for key stages of structure determination including raw data processing, scaling and merging with Xia2, data analysis with Aimless, and various molecular replacement scenarios (with systems such as Molrep[208,209], Phaser[210], MrBump[211,212], MoRDA[213], Balbes[214,215], ARCIMBOLDO_LITE[216], ARCIMBOLDO_BORGES[217] and ARCIMBOLDO_SHREDDER[218,219], SIMBAD[220,221] and DIMPLE[222]). Additional tasks include experimental phasing (with SHELX[127,223] or Crank-2 (ref. 224)), model building (with Modelcraft[225] and CCP4Build), low-resolution refinement (via LoReSTR[226]) and validation (with PDB-REDO[227] or Zanuda[86]). These tasks integrate fundamental CCP4 algorithms with bioinformatics resources and services, streamlining complex workflows into a single pipeline. Additionally, an AlphaFold2 task is available for structure prediction. Seamless integration with interactive graphics for model building, such as Coot[228] and Moorhen (web-Coot), ensures that automation does not limit customization or manual intervention. Finished projects are archived and transferred to a searchable long-term storage. Archived projects are assigned a unique ID, to be cited in publications alongside the corresponding PDB code.

Reproducibility and data deposition

The PDB holds published, experimentally determined macromolecular models along with measured diffraction data that were used for refinement, in merged and scaled form. Measured diffraction images may be deposited in data banks such as IRRMC, Zenodo or SBGRID[229]. Crystallography warrants reproducibility according to the FAIR principles, ensuring that data are findable, accessible, interoperable and re-usable[230], with archiving formats designed to document key steps and software supplying reports. One shortcoming is that conditions for crystallization typically record the solutions that were mixed to produce crystals but the final pH is not recorded. Conditions may involve equivalent amounts of buffers at different pH and hydrolysis reactions are disregarded, so the pH reported for the buffer does not correspond to the pH in the crystal[231]. Otherwise, results vary justifiably as methods

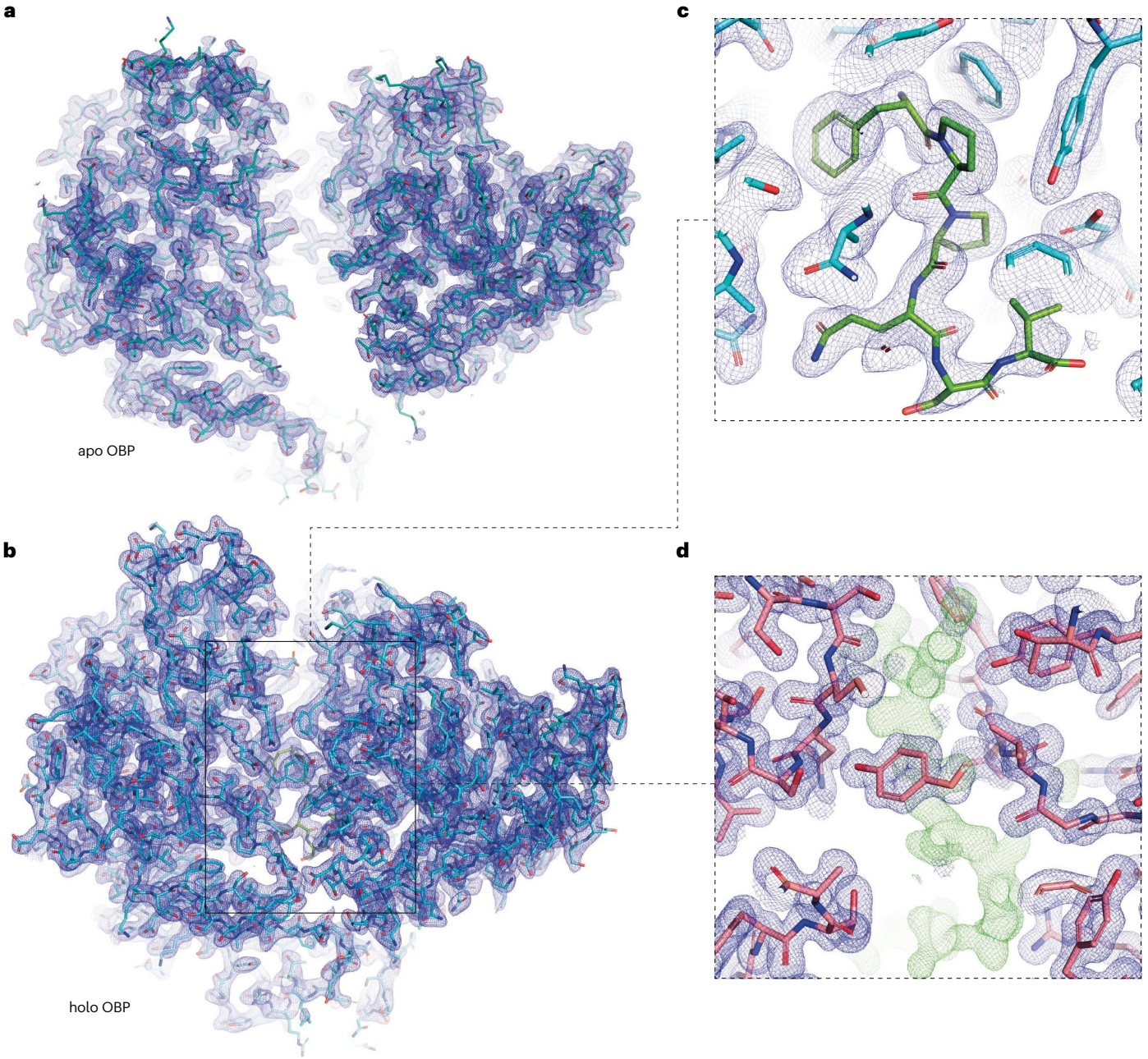


Fig. 4 | Crystallographic structure determination of oligopeptide-binding proteins. a, A $2F_o - F_c$ electron density map contoured at 1σ for the open (apo state) of AliD protein of *Streptococcus pneumoniae* (PDB ID 8QLC) solved at 1.80 Å resolution. **b**, A $2F_o - F_c$ electron density map contoured at 1σ for AliD in complex with FPPQSV (PDB ID 8QLG) at 1.98 Å resolution. **c**, A zoomed-in view showing the substrate-binding site of AliD (cyan sticks) in complex with the oligopeptide substrate (green sticks) (PDB ID 8QLG). **d**, A $2F_o - F_c$ electron density map contoured at 1σ (blue mesh) and $a\,F_o - F_c$ map contoured at 2.5σ (green mesh) for the AliB protein of *S. pneumoniae* (salmon sticks) in complex with unknown oligopeptides (PDB ID 8QLJ) at 1.65 Å resolution. An explanation of the map coefficients (such as $2F_o - F_c$) is provided in Box 1.

to analyse data improve and crystal specimens vary. More importantly, deposited data allow the quality of the structure and the plausibility of the hypotheses derived to be evaluated.

It is important to understand PDB or mmCIF records to use the data as some fields may appear contradictory. A complete structure matching the sequence given in a FASTA format is not necessarily found in the coordinates. If the map did not support atomic positions in a region, whether due to disorder or degradation, these atoms may be missing from the file, display comparatively high *B* values or have zero occupancies. In the latter case, atomic positions are meaningless.

Symmetry may be needed to generate the biological assembly and the coordinates represent an incomplete structure or there may be more than one biological assembly in the file. A dimeric structure may be found with one, two or more monomers represented in the asymmetric unit and hence in the file. Determination of small chemical entities (such as ions or ligands) often relies on an assumption rather than being established expelrimentally or unambiguous from their coordination[232]. Common misidentifications in ambiguous electron density include waters as magnesium (or the other way around), chloride as zinc, and zinc as a disulfide bond or poly(ethylene glycol).

Validation procedures are present throughout the crystallographic analysis and part of the PDB deposition mechanism. Final scores and report are available and can be consulted to establish structure quality for a molecular dynamics simulation or a structure comparison. Validation requires quality analysis of the diffraction data and the model, as well as an evaluation of how well the model and data fit together. Cross-validation with an independent subset of experimental data[201] has been established to avoid overfitting, and model stereochemistry is assessed within its statistical context[233–235].

Data quality

The availability of raw data is rare but enables the usage of improved procedures and the correction of errors. Inconsistency between the model and data or prior knowledge may prompt revisiting the data to investigate unnoticed issues regarding radiation damage, anisotropic diffraction, high mosaicity, space group assignment, twinning, ice rings or shadows on the detector. Beyond the statistics displayed in Table 2, resolution is the main quality indicator. The higher the resolution, the more independent observations support the structure. As a general guideline, at resolutions <1.7 Å, individual atomic positions (except hydrogens) can be determined, meaning that deviations from ideal bonds and angles are often chemically meaningful. Disorder is visible and can be modelled as alternative conformations. At resolutions between 1.7 and 2.6 Å, rotamers and conformations will mostly be correct, but ideal bonds and angles reflect external information. Disorder can still be modelled, but its occupancy can no longer be refined. From resolutions of 2.6–3.7 Å, whereas the fold is almost always correct, frequently side chains and the occasional peptide bond may have been incorrectly modelled and post-translational modifications may have been overseen. With no higher-resolution homologues as reference, there is a risk that some regions will be out of register, with amino acids shifted from their true position along the chain[236]. For natural proteins, the sequence may be uncertain[237]. If the protein is glycosylated, it is quite common to encounter sugar conformations that have been modelled backwards (flipped ~180° around the asparagine–sugar bond)[238]. Many rotamers will correspond to the ideal rotamer in the libraries. At resolutions poorer than 3.7 Å, individual atomic coordinates are meaningless, but the overall fold can perhaps be determined.

Model quality

Structural stereochemistry must be self-consistent and fit prior knowledge of characteristics auch as bond lengths, angles or van der Waals radii — the PDBe Knowledge base is a helpful resource for finding these characteristics. Protein geometry is constrained by the nature of its chemical bonds, resulting in each residue being found in a specific set of conformations inside a structure.

Stereochemistry has already been used in the process of structure solution. However, it is also used to evaluate model quality with respect to chemical prior knowledge as the chemical reason behind major deviations needs to be investigated. Outliers in mainchain torsion angles or van der Waals clashes are problematic, as is the total absence of outliers. MolProbity is a software and webservice used to evaluate the stereochemistry against other structures in the same resolution range[207]. The MolProbity score — for lower resolution structures, the percentage CaBLAM outliers (preferably <2%) — give a good first indication of data quality[239]. Comparison with a fold prediction can also uncover register errors, where a wrong sequence gets assigned to a stretch of electron density[236]. Modifications, such as glycosylation, and the refinement of RNA with its large conformational flexibility, warrant particular attention to detail as automatic geometry checks may not be sufficient in these cases.

Agreement of data and model

Refinement *R* factors, found in deposited coordinate files, express discrepancies between the model and the data, with the lower the *R* value expressing better agreement between the two. Cross-validation with R_{free} serves as a semi-independent (reflections in a dataset are not truly independent) criterion of overfitting. Typical *R* values are about 0.24, or 24% at 2–3 Å resolution, corresponding R_{free} values will be higher but should not diverge much. *R* values are generally lower with better resolution and their average values change in the presence of twinning, tNCS and anisotropy. As a tendency, twinning decreases *R* values, whereas tNCS increases them[240]. Furthermore, global statistics do not inform on the soundness of a local feature. PDB entries are re-refined with modern methods using the PDB-REDO procedure and provide an independent comparison to the authors' results[227]. Alternatively, any user may examine deposited density maps and even improve the models themselves[241].

The importance of visual inspection

Although automated tools aid in tasks such as discovering sequence shifts or allocating unknown sequences[237,242], there is no replacement for verifying the model by visual inspection. At a minimum, it is essential to examine each outlier in the validation reports, as well as all residues particularly pertinent to the scientific question at hand. Genuine outliers may be functionally relevant. Particular attention should be given to ligands, where expectation can often hinder good modelling: both chemical environment and electron density must support the claim of the ligand presence.

During the COVID-19 pandemic from 2020 to 2023, a group of crystallography students and method developers came together to systematically and rapidly check published structure related to SARS-CoV and SARS-CoV-2. This group, the Coronavirus Structural Task Force[243], selected the most important PDB depositions for manual evaluation, and checked hydrogen bonding configurations, chain conformations, ordered solvent compositions and electron density maps, among other parameters.

Curating an accurate database is important for the increased use of structures for bioinformatics, in silico drug design, and the training of artificial intelligence or neural networks[244]. This can be illustrated by an example from the Coronavirus Structural Task Force in which the first structure of the viral RNA polymerase (that produces the RNA by which the virus infects new cells) had a misaligned nine-residue segment in the RNA-binding groove, probably due to low resolution (Fig. 5a–c). The error in this SARS-CoV structure (PDB ID 6NUS) propagated into all structural models of higher-resolution SARS-CoV-2 structures, including the one containing the antiviral drug remdesivir, which was relevant for drug development at the time[245]. The error was corrected with

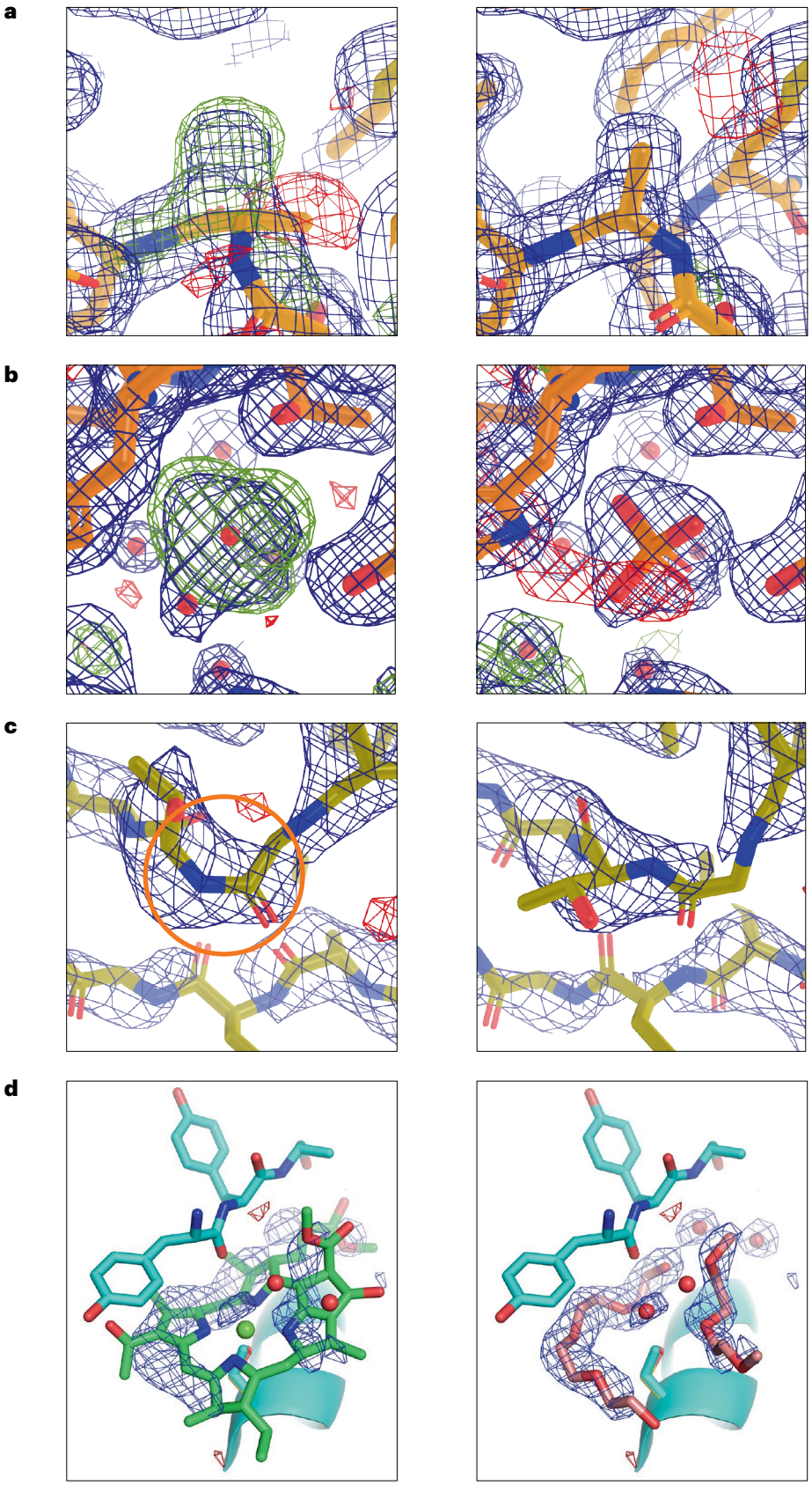


Fig. 5 | The importance in visual inspection of crystal structures. a, The alanine residue in this complex from SARS-COV (PDB ID 3R24, NSP10/NSP16 complex, A192) is in a wrong conformation (left), as indicated by the $2F_o–F_c$ map and the difference density. After correction in ISOLDE, an improved density is fit (right). Example from the Coronavirus Structural Task Force[243]. **b**, In the same structure, one water molecule sits in conspicuous density (left). A phosphate ion fits the density better (right). After re-refinement, the positive difference density vanishes. **c**, Loops can be particularly hard to build in low-resolution structures. This loop includes a *cis* peptide (left, red circle), which is unusual in non-proline residues (PDB ID 6YJT, SARS-CoV-2 NSP13, Gly A439). The *trans*-geometry after correction renders a better fit to the density (right). **d**, The contents of a binding site (PDB ID 3OEG) were incorrectly attributed an eighth bacteriochlorophyll A cofactor (green) as the ligand (left). The reinterpreted density (PDB ID 3VDI, right) was correctly attributed to two ethylene glycol molecules (red). In panels **a**–**c**, $2F_o - F_c$ maps (in blue) are contoured at 1σ and the difference maps (in red and green) are contoured at 3σ. In panel **d**, the electron density was calculated by removing the ligand and water molecules from the binding site and refinement was conducted with Buster[180]. Parts **a**–**c** images courtesy of Coronavirus Structural Task Force. Part **d** reprinted from ref. 246, Springer Nature Limited.

Glossary

Arcimboldo methods
Phasing approaches that use small model fragments (such as α-helices) combined with density modification and fragment expansion to solve structures.

Asymmetric unit
The simply connected smallest closed part of space from which, by application of all symmetry operations of the space group, the whole space is filled.

Co-crystallization
The process of crystallizing two or more molecules together, often a macromolecule with a ligand or inhibitor to view their interaction; with cryoprotectant to minimize stress on fragile crystals or with heavier elements to modify diffraction.

Conjugate gradient minimization
An optimization algorithm used in structure refinement to minimize the difference between observed and calculated structure factors by adjusting model parameters.

Cryogenic cooling device
An apparatus (usually using liquid nitrogen) to rapidly cryo-cool crystals and/or maintain them at the low temperature required to reduce radiation damage during X-ray exposure.

Crystallization drop
A (sub)microlitre droplet containing protein, buffer and precipitant that is used to grow protein crystals.

Diffraction limit
The maximum resolution (smallest detail) that can be resolved in a crystal structure, determined by the quality of the crystal and data.

Direct methods
A set of computational techniques exploiting statistical relationships among structure factors, used to solve the phase problem ab initio and in experimental phasing.

Electron density distribution
A 3D map of the asymmetric unit showing electron per cubic ångström levels, used to model atomic positions and establish structural features.

Friedel opposites
Pairs of reflections related by inversion in reciprocal space (for example, (h,k,l) and $(-h,-k,-l)$; their intensities are equal in the absence of anomalous scattering.

Goniometer
A precision device that holds the crystal and allows its controlled rotation during X-ray diffraction data collection.

Laue groups
Symmetry classifications based on the diffraction pattern, considering the point group of the crystal without translational symmetry.

Model bias
The electron density map is calculated with experimental amplitudes but phases derived from the current model. Hence, the model influences the map and errors can mask real structural features.

Monochromatic
Radiation of a single wavelength, typically used for high-resolution diffraction experiments.

Mosaicity
A measure of the spread of crystal plane orientations, given by the rotation angle over which the signal corresponding to a Bragg reflection is distributed.

Non-crystallographic symmetry
Occurs when copies of a molecule in the asymmetric unit are related by a rotation or translation that is not a symmetry operation of the crystal space group. Translational non-crystallographic symmetry causes aberrant diffraction and complicates structure solution.

Phase problem
The inability to directly measure the phase component of diffracted X-rays (neutrons or electrons), which is essential for reconstructing electron density maps. Phasing means providing approximate values for enough phases to be able to reconstruct an initial model of the structure in the crystal.

Polychromatic
Radiation composed of multiple wavelengths; often used in Laue diffraction experiments.

Real space
The physical, 3D coordinate system in which atoms and electron densities are located within a crystal structure.

Reciprocal space
An abstract space used in crystallography where diffraction data are represented; each point in a reciprocal space lattice corresponds to a set of planes in real space.

Structure factors
Mathematical complex quantities describing both the amplitude and phase of diffracted X-rays, calculated as a vector sum of atomic contributions within the unit cell.

Systematic grid searches
Automated exploration of crystallization conditions by systematically varying parameters such as pH, temperature and precipitant concentration. Also, automatic exploration of parameters in computational procedures like refinement and molecular replacement.

Voxel
A 3D pixel representing a small volume element in an electron density map.

ISOLDE[196], and the improved structure was propagated online, leading to almost all related PDB models being corrected and repetition of the error avoided in subsequent structural models. What is striking about this example is that despite the very large register error, the models appeared very good by traditional metrics, proving that direct visual inspection must remain a key step[245].

Figure 5d illustrates another case with an error caused by model bias and revealed in a map calculated with phases from a model lacking the region of interest, which is often used for ligand validation and called the OMIT map. The improved model was published alongside the full case[246].

Limitations and optimization

The quality of crystallographic data is mainly limited by the crystal sample. Optimizing purification is frequently the key to enhancing resolution. Failure at cryo-cooling may be due to temperature-dependent phase transitions, where mechanical stress shatters the crystal. Collecting room temperature data and testing beyond the standard 100 K allows this cause to be determined and maybe solved by collecting at temperatures above 100 K. However, exposure to cryoprotectant chemicals may greatly deteriorate the diffraction limit, even at room temperature[247].

A successful crystallographic determination may fail to answer the desired question. For example, a high-resolution structure may show no density in the region of interest due to local disorder or partial degradation. When investigating binding sites, the ligand may not have bound if packing blocks access to the binding site or insolubility of the ligand limits delivery. Conversely, a low-resolution structure may provide useful information. In the case of the 7.1 Å structure of the

transcription regulator TsaR in a tetragonal crystal form, a solvent content of 76% allowed to rule out that the identical conformation found in the monoclinic 1.8 Å structure could be a crystal packing artefact[248].

Salt crystals can be identified in crystallization assays using dyes, as only macromolecular crystals support their diffusion and turn coloured. On the contrary, crystallizing an undesired protein from the expression system will typically be recognized only after its solution, whereas checking for equivalent crystal space group and metric in the databases can save months of work. The CONTAMINER server allows to compare against crystal parameters from frequent contaminants[249].

Another hurdle in crystallography is aberrant diffraction, also called data pathologies, precluding structure solution or causing distortion of the maps. Coiled coils, for example, are a class of proteins containing coaxial associations of long helices, found in 5–10% of a proteome[250]. Crystallographic data from coiled coils are dominated by very strong reflections derived from their periodic repeat, which complicates structure solution. Normal data statistics do not hold and any false solution accounting for the periodicity will reach high figures of merit. Solutions involve establishing a case-dependent statistic by setting up competing hypotheses. The PDB contains numerous examples of twinning, tNCS and crystals with an order–disorder translocation[251]. There is foreseeably an even larger number of such datasets unreported, where the data pathology impeded structure solution. Characterizing the problem is a necessary step to solve it or it may advise to avoid it by searching for a different crystal form.

Owing to inherent model bias the map resembles the model, and what is not in the model may often be under-represented. Proteins are flexible, and with the crystal solvent content averaging ~46%, diffraction is due to an ensemble of similar structures rather than a single one[252]. *R* values average around 24% for the final model[253], suggesting that crystallography still has room for improvement as a method for modelling protein movement and solvation.

Outlook

Crystallography is evolving to meet the needs of integrative structural biology towards broader space and timescales. This will require combining crystallography with other techniques to bridge the space scale, such as with microscopy methods. In the timescale, XFEL will remain necessary to access the femtosecond range, with synchrotrons increasingly contributing to dynamic studies with larger timescales[254].

Integrating crystallography with electron microscopy started after its resolution revolution[255,256]. High-resolution microscopy has joined the same repositories as crystallography and it is now practised in many laboratories as a complementary technique. Both demand similar standards and procedures regarding the purification of high quantities of concentrated protein, and there is a correspondence among computational methods in both techniques. Also, as extracting information from a 3D atomic structure, seeing the function within the atoms is an important part of both techniques, still rooted in expert human experience rather than automated. Easy access, further automation, effective training and support to non-experts will remain key to mediate integration.

MX has incorporated accurate AlphaFold[257] and RoseTTaFold[17] models into all steps of a determination. From the prediction of structures for expression construct choices, to their use to solve the phase problem and act as guide during refinement, and to validation, where differences between experimental and predicted structures have pinpointed errors and led to their amendment. Artificial intelligence will furthermore have a central role in integrating data across techniques. Predicted models may render some crystallographic determinations superfluous. On the other hand, the effort of an experimental determination will be motivated by the purpose to add information beyond the prediction. Therefore, it will be important to differentiate the experimental results, shielding them from model bias. Furthermore, generative artificial intelligence is not inherently constrained by evolution or economy as living organisms are. Optimizing or designing new to nature protein functions may require protein sequences far from the training set currently informing predictions. MX will need to inform these new horizons.

As structural data from public repositories decisively inform the current prediction methods, ligand and fragment complexes obtained for drug design projects may come to inform predictions in this field. To that end, the standardization and curation of past experiments and their records as well as new crystallographic studies will be instrumental. The change in deposition standards made possible by the new .cif format should be appropriate to hold the relevant experimental context, absent in the PDB records. Specialized branches of MX are in expansion. For example, macromolecular applications of microED have started overcoming barriers that hindered high-quality data and high throughput[258]. The samples that will not provide suitable crystals for X-ray diffraction may reveal less-stable interactions and conformations, opening new views.

Timely applications and ease of access powered by increasingly integrated experiments and methods are expected to pave further growth of MX structures entering the PDB in the next several years.

Published online: 09 October 2025

References

1. Friedrich, W., Knipping, P. & Laue, M. Interferenz-erscheinungen bei röntgenstrahlen. *Sitzungsberichte Kgl Bayer. Akad. Wiss.* 303–322 (1912).
2. Smith, T. Early crystals. *Nat. Struct. Biol.* **6**, 411–411 (1999).
3. Bragg, W. H. & Bragg, W. L. The reflection of X-rays by crystals. *Proc. R. Soc. Lond. Ser. A* **88**, 428–438 (1913).
4. Ewald, P. P. Die Berechnung optischer und elektrostatischer gitterpotentiale. *Ann. Phys.* **369**, 253–287 (1921).
5. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
6. Helliwell, J. R., Hester, J. R., Kroon-Batenburg, L. M. J., McMahon, B. & Storm, S. L. S. The evolution of raw data archiving and the growth of its importance in crystallography. *IUCrJ* **11**, 464–475 (2024).
7. Krissinel, E. et al. CCP4 Cloud for structure determination and project management in macromolecular crystallography. *Acta Crystallogr. D* **78**, 1079–1089 (2022).
8. Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. *Acta Crystallogr. D* **61**, 449–457 (2005).
9. Usón, I., Ballard, C. C., Keegan, R. M. & Read, R. J. Integrated, rational molecular replacement. *Acta Crystallogr. D* **77**, 129–130 (2021).
10. Sheldrick, G. M. A short history of SHELX. *Acta Crystallogr. A* **64**, 112–122 (2008).
A highly cited article describing the comprehensive crystallography suite, SHELX, used in macromolecular, chemical, mineralogical and materials crystallography.
11. Agirre, J. et al. The CCP4 suite: integrative software for macromolecular crystallography. *Acta Crystallogr. D* **79**, 449–461 (2023).
This paper describes the CCP4 suite of macromolecular crystallography programs.
12. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* **75**, 861–877 (2019).
This paper describes the Phenix suite of macromolecular crystallography programs.
13. Vonrhein, C. et al. Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr. D* **67**, 293–302 (2011).
14. Schiltz, M. et al. Phasing in the presence of severe site-specific radiation damage through dose-dependent modelling of heavy atoms. *Acta Crystallogr. D* **60**, 1024–1031 (2004).
15. McCoy, A. J. Liking likelihood. *Acta Crystallogr. D* **60**, 2169–2183 (2004).
A tutorial for mastering the Bayesian statistics that govern modern crystallographic approaches to phasing and refinement.
16. Simpkin, A. J. et al. Predicted models and CCP4. *Acta Crystallogr. D* **79**, 806–819 (2023).
17. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

18. Terwilliger, T. C. et al. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat. Methods* **21**, 110–116 (2024).
An important view to the relationship between structural predictions initiated by AlphaFold and RosettaFold, and the experimental methods of structure determination.
19. Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
20. Banci, L. et al. Structural proteomics: from the molecule to the system. *Nat. Struct. Mol. Biol.* **14**, 3–4 (2007).
21. Kim, Y. et al. in *Advances in Protein Chemistry and Structural Biology* Vol. 75 (ed. Donev, R.) 85–105 (Elsevier, 2008).
22. Tahara, N. et al. Boosting auto-induction of recombinant proteins in *Escherichia coli* with glucose and lactose additives. *Protein Pept. Lett.* **28**, 1180–1190 (2021).
23. Newman, J. et al. Towards rationalization of crystallization screening for small- to medium-sized academic laboratories: the PACT/JCSG+ strategy. *Acta Crystallogr. D* **61**, 1426–1431 (2005).
24. Abrahams, G. & Newman, J. Data- and diversity-driven development of a shotgun crystallization screen using the protein data bank. *Acta Crystallogr. D* **77**, 1437–1450 (2021).
This paper combines crystallization expertise and modern tools.
25. Monferrer, D., Tralau, T., Kertesz, M. A., Panjikar, S. & Usón, I. High crystallizability under air-exclusion conditions of the full-length LysR-type transcriptional regulator TsaR from *Comamonas testosteroni* T-2 and data-set analysis for a MIRAS structure-solution approach. *Acta Crystallograph. F* **64**, 764–769 (2008).
26. Rosa, N., Watkins, C. J. & Newman, J. Moving beyond MARCO. *PLoS ONE* **18**, e0283124 (2023).
27. Berrow, N. et al. Quality control of purified proteins to improve data quality and reproducibility: results from a large-scale survey. *Eur. Biophys. J.* **50**, 453–460 (2021).
28. Moreau, D. W., Atakisi, H. & Thorne, R. E. Ice in biomolecular cryocrystallography. *Acta Crystallogr. D* **77**, 540–554 (2021).
29. Panjikar, S. & Tucker, P. A. Phasing possibilities using different wavelengths with a xenon derivative. *J. Appl. Crystallogr.* **35**, 261–266 (2002).
30. Colloc'h, N., Carpentier, P., Montemiglio, L. C., Vallone, B. & Prangé, T. Mapping hydrophobic tunnels and cavities in neuroglobin with noble gas under pressure. *Biophys. J.* **113**, 2199–2206 (2017).
31. Volkers, G. et al. Putative dioxygen-binding sites and recognition of tigecycline and minocycline in the tetracycline-degrading monooxygenase TetX. *Acta Crystallogr. D* **69**, 1758–1767 (2013).
32. Carpentier, P., Van Der Linden, P. & Mueller-Dieckmann, C. The high-pressure freezing laboratory for macromolecular crystallography (HPMX), an ancillary tool for the macromolecular crystallography beamlines at the ESRF. *Acta Crystallogr. D* **80**, 80–92 (2024).
33. Patel, O. et al. Crystal structure of the putative cell-wall lipoglycan biosynthesis protein LmcA from *Mycobacterium smegmatis*. *Acta Crystallogr. D* **78**, 494–508 (2022).
34. Sprenger, J. et al. Guest-protein incorporation into solvent channels of a protein host crystal (hostal). *Acta Crystallogr. D* **77**, 471–485 (2021).
35. Gheyi, T., Rodgers, L., Romero, R., Sauder, J. M. & Burley, S. K. Mass spectrometry guided in situ proteolysis to obtain crystals for X-ray structure determination. *J. Am. Soc. Mass Spectrom.* **21**, 1795–1801 (2010).
36. Hillen, H. S., Morozov, Y. I., Sarfallah, A., Temiakov, D. & Cramer, P. Structural basis of mitochondrial transcription initiation. *Cell* **171**, 1072–1081.e10 (2017).
37. Blanco, A. G., Canals, A., Bernués, J., Solà, M. & Coll, M. The structure of a transcription activation subcomplex reveals how σ^{70} is recruited to PhoB promoters: structure of a transcription activation subcomplex. *EMBO J.* **30**, 3776–3785 (2011).
38. Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N. & Nordlund, P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal. Biochem.* **357**, 289–298 (2006).
39. Jiménez-Menéndez, N. et al. Human mitochondrial mTERF wraps around DNA through a left-handed superhelical tandem repeat. *Nat. Struct. Mol. Biol.* **17**, 891–893 (2010).
40. Zander, U. et al. Automated harvesting and processing of protein crystals through laser photoablation. *Acta Crystallogr. D* **72**, 454–466 (2016).
41. Cornaciu, I. et al. The automated crystallography pipelines at the EMBL HTX facility in grenoble. *J. Vis. Exp.* **5**, 62491 (2021).
42. Skaist Mehlman, T. et al. Room-temperature crystallography reveals altered binding of small-molecule fragments to PTP1B. *eLife* **12**, e84632 (2023).
43. Nave, C. & Garman, E. F. Towards an understanding of radiation damage in cryocooled macromolecular crystals. *J. Synchrotron Radiat.* **12**, 257–260 (2005).
44. Garman, E. F. & Schneider, T. R. Macromolecular cryocrystallography. *J. Appl. Crystallogr.* **30**, 211–237 (1997).
45. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336**, 1030–1033 (2012).
Introduces CC1/2 and describes rational and practical solutions for deciding which measurements to keep and which to discard.
46. Karplus, P. A. & Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Curr. Opin. Struct. Biol.* **34**, 60–68 (2015).
47. Shelley, K. L. & Garman, E. F. Identifying and avoiding radiation damage in macromolecular crystallography. *Acta Crystallogr. D* **80**, 314–327 (2024).
This paper presents the state of the art solution for addressing radiation damage.
48. Dickerson, J. L., McCubbin, P. T. N., Brooks-Bartlett, J. C. & Garman, E. F. Doses for X-ray and electron diffraction: new features in RADDOSE-3D including intensity decay models. *Protein Sci.* **33**, e5005 (2024).
49. Waltersperger, S. et al. PRIGo: a new multi-axis goniometer for macromolecular crystallography. *J. Synchrotron Radiat.* **22**, 895–900 (2015).
50. Brönnimann, C. & Trüb, P. in *Synchrotron Light Sources and Free-Electron Lasers* (eds Jaeschke, E. J. et al.) 995–1027 (Springer, 2016).
51. El Omari, K. et al. Utilizing anomalous signals for element identification in macromolecular crystallography. *Acta Crystallogr. D* **80**, 713–721 (2024).
This paper presents the practical use of the extreme long-wavelength macromolecular crystallography in vacuum beamline.
52. Wagner, A., Duman, R., Henderson, K. & Mykhaylyk, V. In-vacuum long-wavelength macromolecular crystallography. *Acta Crystallogr. D* **72**, 430–439 (2016).
53. Barends, T. R. M., Stauch, B., Cherezov, V. & Schlichting, I. Serial femtosecond crystallography. *Nat. Rev. Methods Primer* **2**, 59 (2022).
54. Kabsch, W. Processing of X-ray snapshots from crystals in random orientations. *Acta Crystallogr. D* **70**, 2204–2216 (2014).
55. White, T. A. et al. Recent developments in CrystFEL. *J. Appl. Crystallogr.* **49**, 680–689 (2016).
56. Kupitz, C. et al. Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature* **513**, 261–265 (2014).
57. Meilleur, F. A beginner's guide to neutron macromolecular crystallography. *Biochemist* **42**, 16–20 (2020).
This paper presents an introduction to neutron macromolecular crystallography.
58. Flot, D. et al. The ID23-2 structural biology microfocus beamline at the ESRF. *J. Synchrotron Radiat.* **17**, 107–118 (2010).
59. Debreczeni, J. É., Bunkóczi, G., Ma, Q., Blaser, H. & Sheldrick, G. M. In-house measurement of the sulfur anomalous signal and its use for phasing. *Acta Crystallogr. D* **59**, 688–696 (2003).
60. Mueller, M., Wang, M. & Schulze-Briese, C. Optimal fine ϕ-slicing for single-photon-counting pixel detectors. *Acta Crystallogr. D* **68**, 42–56 (2012).
This paper presents modern data collection methods.
61. Owen, R. L., Rudiño-Piñera, E. & Garman, E. F. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proc. Natl. Acad. Sci. USA* **103**, 4912–4917 (2006).
62. Winter, G. et al. DIALS: implementation and evaluation of a new integration package. *Acta Crystallogr. D* **74**, 85–97 (2018).
63. Delagenière, S. et al. ISPyB: an information management system for synchrotron macromolecular crystallography. *Bioinformatics* **27**, 3186–3192 (2011).
64. Mueller, U. et al. MXCuBE3: a new era of MX-beamline control begins. *Synchrotron Radiat. N.* **30**, 22–27 (2017).
65. McPhillips, T. M. et al. Blu-Ice and the Distributed Control System: software for data acquisition and instrument control at macromolecular crystallography beamlines. *J. Synchrotron Radiat.* **9**, 401–406 (2002).
66. Fodje, M. et al. MxDC and MxLIVE: software for data acquisition, information management and remote access to macromolecular crystallography beamlines. *J. Synchrotron Radiat.* **19**, 274–280 (2012).
67. Murray, C. W. & Blundell, T. L. Structural biology in fragment-based drug design. *Curr. Opin. Struct. Biol.* **20**, 497–507 (2010).
68. Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**, 1531–1534 (1996).
69. Douangamath, A. et al. Achieving efficient fragment screening at XChem facility at diamond light source. *J. Vis. Exp.* **29**, 62414 (2021).
70. Arrowsmith, C. H. et al. The promise and peril of chemical probes. *Nat. Chem. Biol.* **11**, 536–541 (2015).
71. Bar-Even, A. et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).
72. Beyerlein, K. R. et al. Mix-and-diffuse serial synchrotron crystallography. *IUCrJ* **4**, 769–777 (2017).
73. Zielinski, K. A. et al. Rapid and efficient room-temperature serial synchrotron crystallography using the CFEL TapeDrive. *IUCrJ* **9**, 778–791 (2022).
74. Henkel, A. et al. JINXED: just in time crystallization for easy structure determination of biological macromolecules. *IUCrJ* **10**, 253–260 (2023).
75. Monteiro, D. C. F. et al. 3D-MiXD: 3D-printed X-ray-compatible microfluidic devices for rapid, low-consumption serial synchrotron crystallography data collection in flow. *IUCrJ* **7**, 207–219 (2020).
76. Stubbs, J. et al. Droplet microfluidics for time-resolved serial crystallography. *IUCrJ* **11**, 237–248 (2024).
77. Iglesias-Juez, A., Chiarello, G. L., Patience, G. S. & Guerrero-Pérez, M. O. Experimental methods in chemical engineering: X-ray absorption spectroscopy — XAS, XANES, EXAFS. *Can. J. Chem. Eng.* **100**, 3–22 (2022).
78. Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
79. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D* **66**, 133–144 (2010).
80. Otwinowski, Z., Minor, W., Borek, D. & Cymborowski, M. in *International Tables for Crystallography* (eds Arnold, E. et al.) Ch. 11.4 (Wiley, 2012).
81. Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D* **67**, 271–281 (2011).

82. Diederichs, K. Quantifying instrument errors in macromolecular X-ray data sets. *Acta Crystallogr. D* **66**, 733–740 (2010).
83. Nolte, K., Gao, Y., Stäb, S., Kollmannsberger, P. & Thorn, A. Detecting ice artefacts in processed macromolecular diffraction data with machine learning. *Acta Crystallogr. D* **78**, 187–195 (2022).
84. Parkhurst, J. M. et al. Background modelling of diffraction data in the presence of ice rings. *IUCrJ* **4**, 626–638 (2017).
85. Dauter, Z., Botos, I., LaRonde-LeBlanc, N. & Wlodawer, A. Pathological crystallography: case studies of several unusual macromolecular crystals. *Acta Crystallogr. D* **61**, 967–975 (2005).
86. Lebedev, A. A. & Isupov, M. N. Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program Zanuda. *Acta Crystallogr. D* **70**, 2430–2443 (2014).
87. Lovelace, J. J. & Borgstahl, G. E. O. Characterizing pathological imperfections in macromolecular crystals: lattice disorders and modulations. *Crystallogr. Rev.* **26**, 3–50 (2020).
88. Zwart, P. H., Grosse-Kunstleve, R. W., Lebedev, A. A., Murshudov, G. N. & Adams, P. D. Surprises and pitfalls arising from (pseudo)symmetry. *Acta Crystallogr. D* **64**, 99–107 (2008).
89. McCoy, A. J. et al. Phasertng: directed acyclic graphs for crystallographic phasing. *Acta Crystallogr. D* **77**, 1–10 (2021).
90. Read, R. J., Adams, P. D. & McCoy, A. J. Intensity statistics in the presence of translational noncrystallographic symmetry. *Acta Crystallogr. D* **69**, 176–183 (2013).
91. Knott, G. J. et al. A crystallographic study of human NONO ($p54^{nrb}$): overcoming pathological problems with purification, data collection and noncrystallographic symmetry. *Acta Crystallogr. D* **72**, 761–769 (2016).
92. Brehm, W., Triviño, J., Krahn, J. M., Usón, I. & Diederichs, K. XDSGUI: a graphical user interface for XDS, SHELX and ARCIMBOLDO. *J. Appl. Crystallogr.* **56**, 1585–1594 (2023).
93. Arndt, U. W., Crowther, R. A. & Mallett, J. F. W. A computer-linked cathode-ray tube microdensitometer for X-ray crystallography. *J. Phys.* **1**, 510–516 (1968).
94. Diederichs, K. & Karplus, P. A. Improved *R*-factors for diffraction data analysis in macromolecular crystallography. *Nat. Struct. Biol.* **4**, 269–275 (1997).
95. Weiss, M. S. & Hilgenfeld, R. On the use of the merging *R* factor as a quality indicator for X-ray data. *J. Appl. Crystallogr.* **30**, 203–205 (1997).
96. Diederichs, K. & Karplus, P. A. Better models by discarding data? *Acta Crystallogr. D* **69**, 1215–1222 (2013).
97. Read, R. J., Oeffner, R. D. & McCoy, A. J. Measuring and using information gained by observing diffraction data. *Acta Crystallogr. D* **76**, 238–247 (2020).
98. Hendrickson, W. A. Facing the phase problem. *IUCrJ* **10**, 521–543 (2023).
99. Caliandro, R. et al. Phasing at resolution higher than the experimental resolution. *Acta Crystallogr. D* **61**, 556–565 (2005).
100. Usón, I., Stevenson, C. E. M., Lawson, D. M. & Sheldrick, G. M. Structure determination of the *O*-methyltransferase NovP using the 'free lunch algorithm' as implemented in SHELXE. *Acta Crystallogr. D* **63**, 1069–1074 (2007).
101. Karle, J. & Hauptman, H. A theory of phase determination for the four types of non-centrosymmetric space groups 1*P* 222, 2*P* 22, $3P_1$ 2, $3P_2$ 2. *Acta Crystallogr.* **9**, 635–651 (1956).
102. Sheldrick, G. M. et al. *International Tables for Crystallography* Vol. F (eds Arnold, E. et al.) 413–429 (Wiley, 2012).
This paper is a primary reference for ab initio phasing.
103. Usón, I. & Sheldrick, G. M. Advances in direct methods for protein crystallography. *Curr. Opin. Struct. Biol.* **9**, 643–648 (1999).
104. Patterson, A. L. A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.* **46**, 372–376 (1934).
105. Tong, L. & Rossmann, M. G. The locked rotation function. *Acta Crystallogr. A* **46**, 783–792 (1990).
106. Morris, R. J. & Bricogne, G. Sheldrick's 1.2 Å rule and beyond. *Acta Crystallogr. D* **59**, 615–617 (2003).
107. Fujinaga, M. & Read, R. J. Experiences with a new translation–function program. *J. Appl. Crystallogr.* **20**, 517–521 (1987).
108. Usón, I. et al. The 1.2 Å crystal structure of hirustasin reveals the intrinsic flexibility of a family of highly disulphide-bridged inhibitors. *Structure* **7**, 55–63 (1999).
109. Nimz, O., Geßler, K., Usón, I. & Saenger, W. An orthorhombic crystal form of cyclohexaicosaose, CA26·32.59 H_2O: comparison with the triclinic form. *Carbohydr. Res.* **336**, 141–153 (2001).
110. Rodríguez, D. D. et al. Crystallographic ab initio protein structure solution below atomic resolution. *Nat. Methods* **6**, 651–653 (2009).
111. Millán, C., Sammito, M. & Usón, I. Macromolecular ab initio phasing enforcing secondary and tertiary structure. *IUCrJ* **2**, 95–105 (2015).
This paper generalizes fragment-based *ab initio* phasing without the need of atomic-resolution data.
112. Abrahams, J. P. & Leslie, A. G. W. Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr. D* **52**, 30–42 (1996).
113. Cowtan, K. D. & Zhang, K. Y. J. Density modification for macromolecular phase improvement. *Prog. Biophys. Mol. Biol.* **72**, 245–270 (1999).
114. Podjarny, A. D., Rees, B. & Urzhumtsev, A. G. in *Crystallographic Methods and Protocols* Vol. 56 (eds Jones, C. et al.) 205–226 (Humana,1996).
115. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **3**, 1171–1179 (2008).
116. Terwilliger, T. C. Reciprocal-space solvent flattening. *Acta Crystallogr. D* **55**, 1863–1871 (1999).
117. Wang, B.-C. in *Methods in Enzymology* Vol. 115 (eds Wyckoff, H. W. et al.) 90–112 (Elsevier, 1985).
This paper introduces density modification to macromolecules.
118. Cowtan, K. Recent developments in classical density modification. *Acta Crystallogr. D* **66**, 470–478 (2010).
119. Sheldrick, G. M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr. D* **66**, 479–485 (2010).
120. Sheldrick, G. M. Macromolecular phasing with SHELXE. *Z. Für Krist. Cryst. Mater.* **217**, 644–650 (2002).
121. Terwilliger, T. C. Using prime-and-switch phasing to reduce model bias in molecular replacement. *Acta Crystallogr. D* **60**, 2144–2149 (2004).
122. Urzhumtsev, A. G. Local improvement of electron-density maps. *Acta Crystallogr. D* **53**, 540–543 (1997).
123. Dauter, Z., Dauter, M., De La Fortelle, E., Bricogne, G. & Sheldrick, G. M. Can anomalous signal of sulfur become a tool for solving protein crystal structures? *J. Mol. Biol.* **289**, 83–92 (1999).
124. Usón, I. et al. Locating the anomalous scatterer substructures in halide and sulfur phasing. *Acta Crystallogr. D* **59**, 57–66 (2003).
125. Schiltz, M. & Bricogne, G. Exploiting the anisotropy of anomalous scattering boosts the phasing power of SAD and MAD experiments. *Acta Crystallogr. D* **64**, 711–729 (2008).
126. Hatti, K. S., McCoy, A. J. & Read, R. J. Likelihood-based estimation of substructure content from single-wavelength anomalous diffraction (SAD) intensity data. *Acta Crystallogr. D* **77**, 880–893 (2021).
127. Usón, I. & Sheldrick, G. M. An introduction to experimental phasing of macromolecules illustrated by SHELX; new autotracing features. *Acta Crystallogr. D* **74**, 106–116 (2018).
128. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
129. Vagin, A. & Teplyakov, A. MOLREP: an automated program for molecular replacement. *J. Appl. Crystallogr.* **30**, 1022–1025 (1997).
130. Read, R. J. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr. D* **57**, 1373–1382 (2001).
131. Read, R. J. & McCoy, A. J. A log-likelihood-gain intensity target for crystallographic phasing that accounts for experimental error. *Acta Crystallogr. D* **72**, 375–387 (2016).
132. McCoy, A. J. et al. Ab initio solution of macromolecular crystal structures without direct methods. *Proc. Natl Acad. Sci. USA* **114**, 3637–3641 (2017).
This papers explains the rational use of the phasing method, Phaser, spanning single atoms to ribosomes.
133. Oeffner, R. D. et al. The expected log-likelihood gain for decision making in molecular replacement. *Acta Crystallogr. A* **74**, e411–e411 (2018).
134. McCoy, A. J. et al. Gyre and gimble: a maximum-likelihood replacement for Patterson correlation refinement. *Acta Crystallogr. D* **74**, 279–289 (2018).
135. Millán, C., Jiménez, E., Schuster, A., Diederichs, K. & Usón, I. ALIXE: a phase-combination tool for fragment-based molecular replacement. *Acta Crystallogr. D* **76**, 209–220 (2020).
136. Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination. *Acta Crystallogr. D* **65**, 1089–1097 (2009).
This paper describes the integration of alternative phasing methods to molecular replacement and experimental phasing.
137. Medina, A. et al. Verification: model-free phasing with enhanced predicted models in ARCIMBOLDO_SHREDDER. *Acta Crystallogr. D* **78**, 1283–1293 (2022).
138. Caballero, I. et al. ARCIMBOLDO on coiled coils. *Acta Crystallogr. D* **74**, 194–204 (2018).
139. Richards, L. S. et al. Fragment-based ab initio phasing of peptidic nanocrystals by MicroED. *ACS Bio Med. Chem. Au* **3**, 201–210 (2023).
140. Caballero, I. et al. ARCIMBOLDO at low resolution: verification for coiled coils and globular proteins. *Protein Sci.* **33**, e5136 (2024).
This paper describes a method to conclusively solve difficult coiled coil proteins at low resolution.
141. Tronrud, D. E. Introduction to macromolecular refinement. *Acta Crystallogr. D* **60**, 2156–2168 (2004).
142. Afonine, P. V., Urzhumtsev, A. & Adams, P. D. Macromolecular crystallographic estructure refinement. *Arbor* **191**, a219 (2015).
143. Urzhumtsev, A. G. & Lunin, V. Y. Introduction to crystallographic refinement of macromolecular atomic models. *Crystallogr. Rev.* **25**, 164–262 (2019).
This paper provides an overview of macromolecular crystallography refinement methods.
144. Sheriff, S. & Hendrickson, W. A. Description of overall anisotropy in diffraction from macromolecular crystals. *Acta Crystallogr. A* **43**, 118–121 (1987).
145. Parsons, S. Introduction to twinning. *Acta Crystallogr. D* **59**, 1995–2003 (2003).
146. Herbst-Irmer, R. & Sheldrick, G. M. Refinement of twinned structures with SHELXL 97. *Acta Crystallogr. B* **54**, 443–449 (1998).
147. Herbst-Irmer, R. & Sheldrick, G. M. Refinement of obverse/reverse twins. *Acta Crystallogr. B* **58**, 477–481 (2002).
148. Sevvana, M., Ruf, M., Usón, I., Sheldrick, G. M. & Herbst-Irmer, R. Non-merohedral twinning: from minerals to proteins. *Acta Crystallogr. D* **75**, 1040–1050 (2019).
This paper addresses the problem of non-merohedrally twinned data introducing simultaneous refinement against multiple datasets.

149. Weichenberger, C. X., Afonine, P. V., Kantardjieff, K. & Rupp, B. The solvent component of macromolecular crystals. *Acta Crystallogr. D* **71**, 1023–1038 (2015).
150. 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.* **91**, 201–225 (1975).
151. Sheldrick, G. M. & Schneider, T. R. in *Methods in Enzymology* Vol. 277 (eds Carter, C. W. Jr. & Sweet, R. M.) 319–343 (Elsevier, 1997).
152. Urzhumtsev, A. G. Low-resolution phases: influence on SIR syntheses and retrieval with double-step filtration. *Acta Crystallogr. A* **47**, 794–801 (1991).
153. Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D. & Urzhumtsev, A. Bulk-solvent and overall scaling revisited: faster calculations, improved results. *Acta Crystallogr. D* **69**, 625–634 (2013).
154. Afonine, P. V., Adams, P. D., Sobolev, O. V. & Urzhumtsev, A. G. Accounting for nonuniformity of bulk-solvent: a mosaic model. *Protein Sci.* **33**, e4909 (2024).
155. Schomaker, V. & Trueblood, K. N. On the rigid-body motion of molecules in crystals. *Acta Crystallogr. B* **24**, 63–76 (1968).
156. Winn, M. D., Isupov, M. N. & Murshudov, G. N. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D* **57**, 122–133 (2001).
157. Urzhumtsev, A., Afonine, P. V. & Adams, P. D. TLS from fundamentals to practice. *Crystallogr. Rev.* **19**, 230–270 (2013).
158. Merritt, E. A. To *B* or not to *B*: a question of resolution? *Acta Crystallogr. D* **68**, 468–477 (2012).
159. Dittrich, B. On modelling disordered crystal structures through restraints from molecule-in-cluster computations, and distinguishing static and dynamic disorder. *IUCrJ* **8**, 305–318 (2021).
160. Ginn, H. M. Vagabond: bond-based parametrization reduces overfitting for refinement of proteins. *Acta Crystallogr. D* **77**, 424–437 (2021).
161. Hendrickson, W. A. & Lattman, E. E. Representation of phase probability distributions for simplified combination of independent phase information. *Acta Crystallogr. B* **26**, 136–143 (1970).
162. Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. Incorporation of prior phase information strengthens maximum-likelihood structure refinement. *Acta Crystallogr. D* **54**, 1285–1294 (1998).
163. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
This paper introduces maximum-likelihood refinement in macromolecular crystallography.
164. Schirò, A. et al. On the complementarity of X-ray and NMR data. *J. Struct. Biol. X* **4**, 100019 (2020).
165. Usón, I. et al. 1.7 Å structure of the stabilized REI $_{\gamma}$ mutant T39K. Application of local NCS restraints. *Acta Crystallogr. D* **55**, 1158–1167 (1999).
166. Headd, J. J. et al. Flexible torsion-angle noncrystallographic symmetry restraints for improved macromolecular structure refinement. *Acta Crystallogr. D* **70**, 1346–1356 (2014).
167. Evans, P. R. An introduction to stereochemical restraints. *Acta Crystallogr. D* **63**, 58–61 (2007).
168. Lebedev, A. A. et al. JLigand: a graphical tool for the *CCP* 4 template-restraint library. *Acta Crystallogr. D* **68**, 431–440 (2012).
169. Long, F. et al. AceDRG: a stereochemical description generator for ligands. *Acta Crystallogr. D* **73**, 112–122 (2017).
170. Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. Electronic ligand builder and optimization workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr. D* **65**, 1074–1080 (2009).
171. Smart, O. S. et al. Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallogr. D* **74**, 228–236 (2018).
172. Murshudov, G. N. et al. REFMAC 5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
173. Bergmann, J., Oksanen, E. & Ryde, U. Combining crystallography with quantum mechanics. *Curr. Opin. Struct. Biol.* **72**, 18–26 (2022).
174. Zubatyuk, R. et al. AQuaRef: machine learning accelerated quantum refinement of protein structures. Preprint at *bioRxiv* https://doi.org/10.1101/2024.07.21.604493 (2024).
175. Thorn, A., Dittrich, B. & Sheldrick, G. M. Enhanced rigid-bond restraints. *Acta Crystallogr. A* **68**, 448–451 (2012).
176. Moriarty, N. W. et al. Improved chemistry restraints for crystallographic refinement by integrating the Amber force field into Phenix. *Acta Crystallogr. D* **76**, 51–62 (2020).
177. Brünger, A. T. et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
178. Sheldrick, G. M. in *International Tables for Crystallography* (eds Arnold, E. et al.) Ch. 18.9 (Wiley, 2012).
179. Lunin, V. Y., Afonine, P. V. & Urzhumtsev, A. G. Likelihood-based refinement. I. Irremovable model errors. *Acta Crystallogr. A* **58**, 270–282 (2002).
180. Blanc, E. et al. Refinement of severely incomplete structures with maximum likelihood in BUSTER–TNT. *Acta Crystallogr. D* **60**, 2210–2221 (2004).
181. Booth, A. D. LXXIV. An expression for following the process of refinement in X-ray structure analysis using fourier series. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **36**, 609–615 (1945).
182. Luebben, J. & Gruene, T. New method to compute $R_{complete}$ enables maximum likelihood refinement for small datasets. *Proc. Natl Acad. Sci. USA* **112**, 8999–9003 (2015).
This paper extends cross-validation to cases where R_{free} cannot be used.
183. Pražnikar, J. & Turk, D. Free kick instead of cross-validation in maximum-likelihood refinement of macromolecular crystal structures. *Acta Crystallogr. D* **70**, 3124–3134 (2014).
184. Konnert, J. H. A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units. *Acta Crystallogr. A* **32**, 614–617 (1976).
185. Tronrud, D. E. Conjugate-direction minimization: an improved method for the refinement of macromolecules. *Acta Crystallogr. A* **48**, 912–916 (1992).
186. Brunger, A. T., Adams, P. D. & Rice, L. M. in *International Tables for Crystallography* (eds Arnold, E. et al.) Ch. 18.2 (Wiley, 2012).
187. Terwilliger, T. C. et al. Model morphing and sequence assignment after molecular replacement. *Acta Crystallogr. D* **69**, 2244–2250 (2013).
188. Sheldrick, G. M. & Schneider, T. R. SHELXL: high-resolution refinement. *Methods Enzymol.* **277**, 319–343 (1997).
189. Sheldrick, G. M. Crystal structure refinement with SHELXL. *Acta Crystallogr. C* **71**, 3–8 (2015).
190. Cruickshank, D. W. J. Remarks about protein structure precision. *Acta Crystallogr. D* **55**, 583–601 (1999).
This paper addresses the standard uncertainties in the MX parameters.
191. Cowtan, K. & Ten Eyck, L. F. Eigensystem analysis of the refinement of a small metalloprotein. *Acta Crystallogr. D* **56**, 842–856 (2000).
192. Gruene, T., Hahn, H. W., Luebben, A. V., Meilleur, F. & Sheldrick, G. M. Refinement of macromolecular structures against neutron data with SHELXL2013. *J. Appl. Crystallogr.* **47**, 462–466 (2014).
193. Catapano, L. et al. Neutron crystallographic refinement with REFMAC5 from the CCP4 suite. *Acta Crystallogr. D* **79**, 1056–1070 (2023).
194. Diamond, R. A real-space refinement procedure for proteins. *Acta Crystallogr. A* **27**, 436–452 (1971).
195. Casañal, A., Lohkamp, B. & Emsley, P. Current developments in Coot for macromolecular model building of electron cryo-microscopy and crystallographic data. *Protein Sci.* **29**, 1055–1064 (2020).
This paper describes interactive model building and real-space refinement.
196. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D* **74**, 519–530 (2018).
197. Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71**, 136–153 (2015).
198. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).
199. Terwilliger, T. C. et al. Improved AlphaFold modeling with implicit experimental information. *Nat. Methods* **19**, 1376–1382 (2022).
200. Roversi, P. & Tronrud, D. E. Ten things I 'hate' about refinement. *Acta Crystallogr. D* **77**, 1497–1515 (2021).
201. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).
This paper proposed the now universally used cross-validation R_{free} method in macromolecular crystallography.
202. Urzhumtseva, L., Afonine, P. V., Adams, P. D. & Urzhumtsev, A. Crystallographic model quality at a glance. *Acta Crystallogr. D* **65**, 297–300 (2009).
203. Alcorlo, M. et al. Molecular and structural basis of oligopeptide recognition by the Ami transporter system in pneumococci. *PLoS Pathog.* **20**, e1011883 (2024).
204. Yamashita, K., Wojdyr, M., Long, F., Nicholls, R. A. & Murshudov, G. N. GEMMI and Servalcat restrain REFMAC 5. *Acta Crystallogr. D* **79**, 368–373 (2023).
205. Tickle, I. J., Laskowski, R. A. & Moss, D. S. Error estimates of protein structure coordinates and deviations from standard geometry by full-matrix refinement of γ B- and β B2-crystallin. *Acta Crystallogr. D* **54**, 243–252 (1998).
206. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
207. Davis, I. W. et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383 (2007).
208. Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. Model preparation in MOLREP and examples of model improvement using X-ray data. *Acta Crystallogr. D* **64**, 33–39 (2008).
209. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D* **66**, 22–25 (2010).
210. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
211. Keegan, R. M. & Winn, M. D. MrBUMP: an automated pipeline for molecular replacement. *Acta Crystallogr. D* **64**, 119–124 (2008).
212. Keegan, R. M. et al. Recent developments in MrBUMP: better search-model preparation, graphical interaction with search models, and solution improvement and assessment. *Acta Crystallogr. D* **74**, 167–182 (2018).
213. Vagin, A. & Lebedev, A. MoRDa, an automatic molecular replacement pipeline. *Acta Crystallogr. A* **71**, s19–s19 (2015).
214. Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. BALBES: a molecular-replacement pipeline. *Acta Crystallogr. D* **64**, 125–132 (2008).
215. Keegan, R. M. et al. Evaluating the solution from MrBUMP and BALBES. *Acta Crystallogr. D* **67**, 313–323 (2011).
216. Sammito, M. et al. ARCIMBOLDO_LITE: single-workstation implementation and use. *Acta Crystallogr. D* **71**, 1921–1930 (2015).
217. Sammito, M. et al. Exploiting tertiary structure through local folds for crystallographic phasing. *Nat. Methods* **10**, 1099–1101 (2013).
218. Sammito, M. et al. Structure solution with ARCIMBOLDO using fragments derived from distant homology models. *FEBS J.* **281**, 4029–4045 (2014).

219. Millán, C. et al. Exploiting distant homologues for phasing through the generation of compact fragments, local fold refinement and partial solution combination. *Acta Crystallogr. D* **74**, 290–304 (2018).
220. Simpkin, A. J. et al. SIMBAD: a sequence-independent molecular-replacement pipeline. *Acta Crystallogr. D* **74**, 595–605 (2018).
221. Simpkin, A. J. et al. Using Phaser and ensembles to improve the performance of SIMBAD. *Acta Crystallogr. D* **76**, 1–8 (2020).
222. Wojdyr, M., Keegan, R., Winter, G. & Ashton, A. DIMPLE — a pipeline for the rapid generation of difference maps from protein crystals with putatively bound ligands. *Acta Crystallogr. A* **69**, s299–s299 (2013).
223. Usón, I. & Sheldrick, G. M. Modes and model building in *SHELXE*. *Acta Crystallogr. D* **80**, 4–15 (2024).
224. Skubák, P. et al. A new MR-SAD algorithm for the automatic building of protein models from low-resolution X-ray data and a poor starting model. *IUCrJ* **5**, 166–171 (2018).
225. Bond, P. S. & Cowtan, K. D. ModelCraft: an advanced automated model-building pipeline using Buccaneer. *Acta Crystallogr. D* **78**, 1090–1098 (2022).
226. Kovalevskiy, O., Nicholls, R. A. & Murshudov, G. N. Automated refinement of macromolecular structures at low resolution using prior information. *Acta Crystallogr. D* **72**, 1149–1161 (2016).
227. Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallogr. D* **68**, 484–496 (2012).
228. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
229. Morin, A. et al. Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).
230. Helliwell, J. R. et al. Findable accessible interoperable re-usable (FAIR) diffraction data are coming to protein crystallography. *Acta Crystallogr. D* **75**, 455–457 (2019).
231. Wilson, J., Ristic, M., Kirkwood, J., Hargreaves, D. & Newman, J. Predicting the effect of chemical factors on the pH of crystallization trials. *iScience* **23**, 101219 (2020).
232. Zheng, H. et al. Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server. *Nat. Protoc.* **9**, 156–170 (2014).
233. Richardson, J. S., Williams, C. J., Chen, V. B., Prisant, M. G. & Richardson, D. C. The bad and the good of trends in model building and refinement for sparse-data regions: pernicious forms of overfitting versus good new tools and predictions. *Acta Crystallogr. D* **79**, 1071–1078 (2023).
This paper describes stereochemical validation.
234. Dodson, E. The role of validation in macromolecular crystallography. *Acta Crystallogr. D* **54**, 1109–1118 (1998).
235. Richardson, J. S. & Richardson, D. C. Amino acid preferences for specific locations at the ends of α helices. *Science* **240**, 1648–1652 (1988).
236. Sánchez Rodríguez, F., Simpkin, A. J., Chojnowski, G., Keegan, R. M. & Rigden, D. J. Using deep-learning predictions reveals a large number of register errors in PDB depositions. *IUCrJ* **11**, 938–950 (2024).
237. Borges, R. J. et al. SEQUENCE SLIDER: integration of structural and genetic data to characterize isoforms from natural sources. *Nucleic Acids Res.* **50**, e50–e50 (2022).
238. Dialpuri, J. S. et al. Online carbohydrate 3D structure validation with the Privateer web app. *Acta Crystallogr. F* **80**, 30–35 (2024).
239. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
240. Nicholls, R. A., Long, F. & Murshudov, G. N. Low-resolution refinement tools in REFMAC5. *Acta Crystallogr. D* **68**, 404–417 (2012).
241. Gao, Y., Thorn, V. & Thorn, A. Errors in structural biology are not the exception. *Acta Crystallogr. D* **79**, 206–211 (2023).
242. Chojnowski, G. Sequence-assignment validation in protein crystal structure models with checkMySequence. *Acta Crystallogr. D* **79**, 559–568 (2023).
243. Croll, T. I. et al. Making the invisible enemy visible. *Nat. Struct. Mol. Biol.* **28**, 404–408 (2021).
This paper describes a collaborative, macromolecular crystallography effort in the scope of the COVID-19 pandemic.
244. Thorn, A. Artificial intelligence in the experimental determination and prediction of macromolecular structures. *Curr. Opin. Struct. Biol.* **74**, 102368 (2022).
245. Croll, T. I., Williams, C. J., Chen, V. B., Richardson, D. C. & Richardson, J. S. Improving SARS-CoV-2 structures: peer review by early coordinate release. *Biophys. J.* **120**, 1085–1096 (2021).
246. Tronrud, D. E. & Allen, J. P. Reinterpretation of the electron density at the site of the eighth bacteriochlorophyll in the FMO protein from *Pelodictyon phaeum*. *Photosynth. Res.* **112**, 71–74 (2012).
247. von Bulow, R. et al. Defective oligomerization of arylsulfatase a as a cause of its instability in lysosomes and metachromatic leukodystrophy. *J. Biol. Chem.* **277**, 9455–9461 (2002).
248. Monferrer, D. et al. Structural studies on the full-length LysR-type regulator TsaR from *Comamonas testosteroni* T-2 reveal a novel open conformation of the tetrameric LTTR fold. *Mol. Microbiol.* **75**, 1199–1214 (2010).
249. Hungler, A., Momin, A., Diederichs, K. & Arold, S. T. ContaMiner and ContaBase: a webserver and database for early identification of unwantedly crystallized protein contaminants. *J. Appl. Crystallogr.* **49**, 2252–2258 (2016).
250. Liu, J. & Rost, B. Comparing function and structure between entire proteomes. *Protein Sci.* **10**, 1970–1979 (2001).
251. Caballero, I. et al. Detection of translational noncrystallographic symmetry in Patterson functions. *Acta Crystallogr. D* **77**, 131–141 (2021).
252. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **1**, e00311 (2012).
253. Holton, J. M., Classen, S., Frankel, K. A. & Tainer, J. A. The *R*-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. *FEBS J.* **281**, 4046–4060 (2014).
254. Banari, A. et al. Advancing time-resolved structural biology: latest strategies in cryo-EM and X-ray crystallography. *Nat. Methods* https://doi.org/10.1038/s41592-025-02659-6 (2025).
255. Amunts, A. et al. Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).
256. Kühlbrandt, W. The resolution revolution. *Science* **343**, 1443–1444 (2014).
257. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
258. Hofer, G., Wang, L., Xu, H. & Zou, X. Advances in protein electron diffraction (3D-ED/microED) sample preparation. *Acta Crystallogr. A* **79**, C391–C391 (2023).

Acknowledgements
The authors thank all the lecturers that have contributed to the Madrid Crystallography School, including M. Martínez-Ripoll, F. X. Gomis-Rüth, J. M. García-Ruiz, R. Kahn, J. Navaza, C. Giacovazzo, P. Emsley, J. M. Mancheño, P. Adams, T. Grüne, I. Muñoz, P. Bernadó, R. Nicholls, R. Marabini, J. M. Carazo, J. Martín-Garcia, R. Fernández-Leiro, R. Boer, A. J. McCoy, B. Herguedas, T. T. Terwilliger, M. Fando, F. Sánchez-Rodríguez, R. Keegan and L. Catapano. They also thank all students for their active participation in discussions and contribution of interesting crystallographic challenges. This work was supported by (Ministry of Science and Innovation/Spanish State Research Agency/European Regional Development Fund/European Union) grants PID2021-128751NB-I00 to I.U., PID2023-153118OB-I00 to J.A.H., PID2023-153108OB-I00 to A.A. and PID2021-129038NB-I00 to M.S.; grant 2021-SGR-00425 (AGAUR) to I.U. and M.S.; grant Horizon Europe ID 101094131 and 101046133 to J.A.M.; The National Institutes of Health (grants R01GM071939, P01GM063210 and R24GM141254), as well as support from the Phenix Industrial Consortium and the US Department of Energy under contract no. DE-AC02-05CH11231 to P.V.A. The German Federal Ministry of Education and Research (grant no. 05K19WWA), Deutsche Forschungsgemeinschaft (project TH2135/2-1) to A.T. The Collaborative Computational Project Number 4 in Protein Crystallography (CCP4) and Biotechnology and Biological Research Council (BBSRC UK) grants BB/Y009991/1, BB/V015591/1 and BB/S007040/1 to E.K.

Author contributions
Introduction (A.A., J.A.H., K.D. and I.U.); Experimentation (M.S., J.A.M., S.P., K.D. and I.U.); Results (P.V.A., K.D. and I.U.); Applications (J.A.H., E.K., K.D. and I.U.); Reproducibility and data deposition (A.T., K.D. and I.U.); Limitations and optimizations (K.D. and I.U.); Outlook (K.D. and I.U.); overview of the Primer (K.D. and I.U.).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43586-025-00433-8.

Peer review information *Nature Reviews Methods Primers* thanks Elspeth Garman, José Gavira and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Related links
CCP4 Cloud: https://cloud.ccp4.ac.uk
Crystallographic wiki: https://wiki.uni-konstanz.de/ccp4/index.php/Main_Page
CSIC Crystallography: https://www.xtal.iqfr.csic.es/Cristalografia/index-en.html
Fourier transform animations: http://chango.ibmb.csic.es/colibri
International Union of Crystallography (IUCr) dictionary: https://dictionary.iucr.org
MolProbity analysis: http://molprobity.biochem.duke.edu/
Protein Data Bank: https://www.wwpdb.org
Radiation dose calculation: https://raddo.se/
SBGRID: https://SBGrid.org
The PDBe Knowledge base: https://www.ebi.ac.uk/pdbe/pdbe-kb/
Zenodo: https://zenodo.org

© Springer Nature Limited 2025, corrected publication 2025