Seven quick tips for beginners in protein crystallography

Katarzyna Kurpiewska1✉, Ewa Kot2 and Tomasz Borowski2✉

1Department of Crystal Chemistry and Crystal Physics, Faculty of Chemistry, Jagiellonian University, Kraków, Poland; 2Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Kraków, Poland

The aim of this brief review is to provide a roadmap for beginning crystallographers who have little or no experience in structural biology and yet are keen to produce protein crystals and analyze their 3D structures to understand their biological roles. To achieve this goal it is crucial to perform crystallization, structure determination, visualization and analysis of the protein's structural features related to its biological function. Keeping that objective in mind, tips presented herein cover the most important steps in a crystallographic endeavor and present a selection of databases and software which can aid and accelerate the whole process. We hope that this short overview will help novices coming from different disciplines to navigate a protein crystallography project and, hopefully, allow avoiding some costly mistakes, even though being a crystallographer means learning by trial and error.

Keywords: biocrystallography, crystallography software, structure determination, bioinformatic tools, database

Received: 10 January, 2021; revised: 28 February, 2021; accepted: 31 March, 2021; available on-line: 11 August, 2021

✉e-mail: katarzyna.kurpiewska@uj.edu.pl (KK); tomasz.borowski@ikifp.edu.pl (TB)

Acknowledgements of Financial Support: T.B. acknowledges partial financial support of the project by the statutory research fund of ICSC PAS.

INTRODUCTION

X-ray macromolecular crystallography serves a variety of scientific disciplines and significantly accelerates discoveries in many research areas, including studies on protein biological function, drug screening and design, and human health and disease. Through decades, biocrystallography has evolved together with developments in computer science allowing faster structure determination. As a result, a spectacular growth in the number of new software, advanced databases and bioinformatic servers can be observed. The scale of constantly growing interest in structural biology, and hence also biocrystallography, is proven by the traffic on the central database Protein Data Bank (Berman et al., 2000). Worldwide, more than 1 million users visit the Protein Data Bank every year, as judged by counting unique IP addresses. They perform more than 1.5 million downloads of structures every day, or more than 500 million per year (Bruno et al., 2017). Beyond the final outcome
from the crystallographic experiment, one can feel lost
in the diverse and thriving ecosystem of software ap-
plied during the process of structure determination and
analysis. Here, we attempt to create a roadmap (Fig. 1)
for beginners in protein crystallography in a form of a
set of tips comprising a modest selection of macromo-
lecular crystallography software and bioinformatic tools
essential for a crystallographer’s journey. However, it is
not a comprehensive review on freely available software
packages, services, or commercial products. This subjec-
tive overview made by the authors comprises resources
that are well-known among the community and that are
currently available. It is hoped that the “seven tips” can
serve as a starting point especially for young researchers
and will act as a catalyst for the readers to deepen their
crystallographic knowledge.

TIP 1: LEARN WHAT IS KNOWN ABOUT YOUR PROTEIN

A good starting point for gathering information about
the chosen protein is the UniProt server (https://www.
uniprot.org/) (Bateman, 2019). It offers an advanced
search engine which accepts, inter alia, name of the pro-
tein, EC number or, via a BLAST (Zhang & Madden,
1997) search option, its amino acid sequence. UniProt
entry provides key information about the protein func-
tion, names and taxonomy, subcellular location, post-
translational modifications, its interactions with itself or
other proteins, similarity to other proteins and domains
present in the protein and its amino acid sequence. Lit-
erature references to information sources are provided,
as well as a rich selection of cross references to other
databases. One of very useful features of the server is
the “Add to basket” option available in the Sequence
section. With its use one can gather a set of protein
sequences, which one can later align using the Clustal
Omega program (Sievers et al., 2011) available at the
server.

Information on protein domains and their organization
within a chosen protein, as well as on the whole
protein family to which the protein belongs can be re-
trieved from the Pfam database (http://pfam.xfam.org/)
(El-Gebali et al., 2019). That database has a user-friendly
search engine that accepts, inter alia, UniProt ID and
PDB IDs. Entry for each protein family provides very
useful information on protein architectures, available
structures deposited in PDB, species and phylogenetic
trees and, importantly, it allows one to view or down-
load stored sequence alignments for the family. An avail-
able profile logo for the family aids in identifying con-
served residues and variable positions.

Protein solubility in E. coli can be predicted based
on the protein amino acid sequence with the use of the
SoluProt server (https://loschmidt.chemi.muni.cz/solu-
prot/) (Hon et al., 2020), which employs machine learn-
ing techniques trained on curated databases of exper-
imental data.

Before ordering or cloning a gene for the protein to
be crystallized, it is advisable to inspect the results of
the XtalPred server (https://xtalpred.godziklab.org/Xtal-
Pred-cgi/xtal.pl) (Slabinski et al., 2007). Using the amino
acid sequence as input, the server predicts a range of
physico-chemical properties and based on them, by us-
ing machine learning (random forest) method, predicts
protein crystallizability. Detailed reports on values of
computed target features vs. distributions of crystalliza-
tion successes and failures allow one to judge which fea-
ture can potentially be a major obstacle to crystallization.
The predicted sequence features, along with the amino
cacid sequence, can aid in construct design, e.g. suggest
removing a signal peptide or a long disordered fragment

Figure 2. Crystal structure determination.
at the protein’s termini. The XtalPred results, combined with information retrieved from the Pfam database, can also aid in deciding if the whole multi-domain protein should be crystallized as a whole or as separate fragments. The “homologs” section of the results provides, among others, a list of homologs with known structure deposited in PDB, which is a valuable information with respect to the feasibility of solving the structure by a molecular replacement method, which is currently the most common means of solving the crystal structure. To be useful for this purpose, the homolog should have an amino acid sequence that is at least 20–25% identical with the target protein. Homologs can be also directly searched at the PDB server (https://www.rcsb.org/) using an advanced search option and amino acid sequence as input. If close-enough homologs with a known structure are not available, then experimental phasing needs to be considered, among which the single wavelength anomalous diffraction (SAD) method is the most popular. SAD uses anomalous signal from either natural components of the investigated protein (e.g. Zn, Cu, Fe, Mn, Ni, or in favorite cases S) or from selenomethionine (SeMet), introduced into the protein (through bacterial or yeast growth medium) during protein production.

Since the whole process of structure determination starts from a protein sample (Fig. 2), after gathering all available information about the macromolecule of interest the researcher should answer a few more questions regarding the protein source or its production. A good planning procedure at this step that includes decisions about working with multidomain versus single domains of the studied molecule and possible truncation of the flexible parts, as well as the awareness of a wide range of methods used for solving the structures based on intrinsic features of the macromolecule (i.e. metal content, SeMet derivatives etc.) can save a great amount of time at the next stages. It is good to remember that storage of the protein sample can be critical. Not all proteins tolerate freezing at −20°C, thus most samples are kept at 4°C or −80°C, but the activity and stability must be regularly checked. In addition, as a general rule, it is better to store proteins that are concentrated than diluted.

TIP 2: CRYSTALLIZATION IS AN ART, YET PLANNING IS ADVISABLE

Before planning the crystallization experiments, it is important to realize which factors influence crystal growth (Table 1). All of these variables, categorized as physical, chemical or biochemical, can heavily impact the crystal formation process. The details of these various parameters have been largely described in the literature (Abdalla, 2016; Bhat et al., 2018). In practice, the purity, homogeneity and stability of the protein sample are the very first factors that should be considered. Protein concentration is always case dependent, but for the initial experiments concentration of 1–25 mg/ml is recommended (10–15 mg/ml is typical). Eukaryotic proteins tend to be less soluble than bacterial proteins. All further approaches strongly depend on the amount of the protein sample, the equipment available and resources. As already mentioned, searching for optimal crystallization conditions is still a try and error process, enhanced by usage of commercially available screens. However, there are a couple of evidence-based rational approaches that are very likely to improve a chance of obtaining protein crystals. Apart from purity of the protein sample, the second very important aspect is based on observations that pH of the crystallization solution has a significant impact on crystal growth. It has been suggested that pH should deviate from pI of the protein by up to 3 pH units and that pH of the protein solution should be “as low, as high or as divergent from the pI as possible for basic, acidic or neutral proteins, respectively, within their stable pH range” (Zhang et al., 2013). Thus, initial screening for crystallization conditions should explore the widest set of pH/precipitants/buffers/additives, which can be easily conducted with the use of crystallization kits provided by many suppliers. The best way to increase the success in macromolecular crystallization is to initiate a collaboration with a structural biology group or with dedicated core facilities equipped with crystallization robots, cold rooms and/or crystal hotels. A number of such initiatives supporting their users through the entire crystallization process has rapidly multiplied in recent years all over the world. What is encouraging in robotic handling of crystallization plates is the substantially smaller amount of the sample used by crystallization robots in comparison to the traditional path with manually setup crystallization drops where a sample volume between 1 µL and 5 µL is used. For example, to set up 10 screens (96 conditions each) 150–200 microliters of protein at the proper concentration should be prepared. Discussion of the theoretical principles behind crystallization (McPherson & Gavira, 2014; Russo Krauss et al., 2013), description of the strategies regarding the experiments (Cheraghian Radi et al., 2021) and how to proceed with optimization (He et al., 2020) is beyond the scope of this review. However, as an extension of this tip we would like to point to one more rational approach enhancing crystallization chances – protein surface entropy reduction, which can be planned with the use of the SERp server (http://services.mbi.ucla.edu/SERp/) (Goldschmidt et al., 2007). The server identifies regions on the protein surface characterized by a high side chain conformational freedom (and hence, entropy) and based on the secondary structure prediction results (coil regions are preferred) and sequence alignment to homologous proteins (amino acid conservation analysis) suggests the best candidates (up to three consecutive residues) for mutation. The resulting mutant is expected to have low-

| Physical factors | Chemical factors | Biochemical factors |
|------------------|------------------|---------------------|
| Temperature      | Sample concentration | Sample purity          |
| Pressure         | Buffer type       | Sample homogeneity   |
| Time             | pH                | Sample pI            |
| Viscosity        | Ionic strength    | Sequence modifications |
| Magnetic or electric fields | Precipitant type   | (protein surface entropy reduction, usage of fast- and slow-translating codons, interface mutants) |
| Vibrations and sound | Precipitant concentration | Posttranslation modifications |
| Method of crystallization | Additives: heavy ions, metal ions, polymers, detergents | Chemical modifications |
| Surface of crystallization drop | Ligands, cofactors, inhibitors | Proteolysis |
| Nucleants        |                   |                     |
| Equilibration rate |                   |                     |

### Table 1. Parameters affecting protein crystallization.
er entropy penalty for the transition from the in solution to the crystal state.

Assuming that the protein crystals can be seen in the drop, now the question “what’s next?” should be answered. How to handle the crystals? How to prepare samples for their journey to the synchrotron and for data collection? Crystals that look good under the microscope are only a promising start. Working with protein crystals is not the easiest one. Before the measurements, they need to be harvested and protected from destruction. Since crystals are formed from solutions based on water, large part of their crystal lattice is composed of water (Chayen & Saridakis, 2008). Large amount of the mother liquor in the crystals ensures that the protein molecules adopt a native conformation that is similar to that observed under physiological conditions. Furthermore, the presence of water channels makes it possible to easily introduce low molecular weight components into the protein crystals, e.g. heavy element ions, inhibitors or activators (Gnesi & Carugo, 2017). On the other hand, the presence of water in the crystals also has a negative side. During diffraction experiments with intense X-ray, free radicals are produced by ionizing radiation. Unfortunately, the presence of water channels allows these very dangerous molecules to spread quickly, and when reaching protein molecules they cause destruction and degradation of the crystal. To mitigate this process, a cryoprotection method is applied (Pellegrini et al., 2011). This step is deeply connected with the next one – crystal handling. Finding the right cryoprotecting agent and its concentration is a crucial step for preserving good crystal condition. Cryoprotectant selection remains a trial and error exercise, where the first combination that “works” is accepted. During this step, one should remember that an efficient cryoprotectant solution should firstly stabilize the crystal, but the addition of a cryoprotectant should also prevent ice formation on the surface of the sample during flash-cooling. At this point we should mention that soaking in a cryoprotection solution is not the only method for protein crystals’ protection from damage. Dehydration, high-pressure cryocooling or crystal annealing can be also applied (Huang & Szegény, 2016). Crystals should be handled one by one and as fast as possible, otherwise the crystal and the drop can dry (in result, other crystals in the same drop will be lost). Many tools for crystal handling and mounting can be found on the market: a wide variety of loops (different shapes and sizes), microtools, sets for the room temperature measurements and capillaries. With the use of the loop that is a bit larger than the crystal, after fishing it out, the crystal can be stepwise transferred to cryoprotection solutions with gradually increased concentration of the cryoprotectant or can be immediately soaked in the already established final cryoprotecting solution (Vera & Stura, 2014). In both cases, the next step requires transfer of the crystal into liquid nitrogen. After flash-cooling, crystals should be directly mounted on the X-ray diffractometer or placed inside a dewar, where samples can be kept for as long as it is necessary. Once frozen, crystals are transported under cryogenic conditions, usually with the use of dry-shipper dewars.

**TIP 3: HAVE A PLAN FOR DATA COLLECTION**

The most important part of this tip could be enclosed in one sentence: data collection is the last experiment in the course of a structure determination and it requires compromises (Fig. 3). Collecting bad data can unfortunately ruin all previous efforts and substantially influence the expected outcomes. To avoid this situation, the diffraction experiment should be prepared and conducted after careful planning. The vast majority of X-ray crystal structures in the Protein Data Bank is based on synchrotron data. State-of-the-art synchrotron sites dedicated to structural studies of biological samples offer small and focused beams, which allow routine diffraction measurements for microcrystal samples. Furthermore, the X-ray diffraction data collections, including optimized anomalous dispersion element identification or phasing, experiments with crystals featuring large unit cells, as well as high resolution measurements are now possible at shorter measuring times. Intense in-house laboratory sources also serve as tools for collection of single-wavelength diffraction data, which even enable obtaining data suitable for the effective S-SAD phasing, however they are limited to the characteristic radiation of the X-ray anode material. The process of recording diffractograms relies on several principles that should be considered before data collection:

- The first important parameter is the wavelength of the X-ray that will hit the crystal. X-rays are of the same nature as visible light or radio waves, the only difference is their wavelength, which is very short (about 1Å). A phenomenon caused by the interaction of electromagnetic waves with the matter inside the crystal (particularly with the electrons) depends on radiation wavelength.
The choice of X-ray wavelength used during data collection depends on the strategy that will be used during structure determination, and here the most common approaches are molecular replacement and anomalous signal methods. In the case of molecular replacement, which is viable when the structure of the model protein is known, single wavelength without special consideration of anomalous data suffices. Anomalous signal phasing methods require collection of single wavelength anomalous data (SAD) for selected marker elements (also possible for native sulfur (S-SAD) and native phosphorus (P-SAD)) or multiple anomalous diffraction data (MAD). In those cases, after spectrum determination of the absorption edge for anomalous marker(s), data sets are collected at single or various selected wavelengths in order to obtain the maximal anomalous signal. Also, when metal atoms are already present in the structure, it is advisable to collect the X-ray fluorescence spectrum which can be collected at most synchrotron beamlines. Recording X-ray fluorescence spectra, and collecting diffraction datasets above and below the corresponding metal absorption edges, in most cases allow to gather sufficient evidence to unambiguously determine the identity and location of the metal of interest, as well as to accurately characterize the coordinating ligands in the metal binding environment within the protein.

– Keeping in mind that the crystal structure is encoded in the diffracted X-rays, where crystal orientation, shape and symmetry of the unit cell define the directions of the diffracted beams, whereas the positions of all atoms in the unit cell define their intensities, a few more important aspects should be considered for successful data collection, including inspection of the first diffraction image and strategy determination. By visual examination of the first diffractogram, salt and protein samples can be easily distinguished. Furthermore, for cryocooled samples, the inspection of the collected image for presence and strength of diffraction rings caused by ice, will reveal whether the choice of the cryoprotective agent was appropriate. Modern software can deal with regions that should be excluded from data processing in cases where such “ice rings” are present on the images. Observation of strong, well-shaped and resolved spots up to a high resolution region suggests that collection of good quality data is possible. Nevertheless, it is not uncommon that the diffraction power occurred, some reflections may not be measured at all, and the data may not be complete. A number of synchrotron sites for macromolecular crystallography in Europe operates with MXCuBE (Gabadinho et al., 2010) and the latest version MXCuBE3 (Mueller et al., 2017) (https://mxmctb.github.io/mxmctb/), which supports the users in making reasonable decisions during data collection. Another important aspect of making the most of the beam time is the opportunity to process your data during or just after data collection. Quick examination of the final statistics will be beneficial in situations when for some reason the measurements went wrong and data collection needs to be repeated.

– As a final remark to this tip, remember that making a good plan for data collection is an effort that will pay off at the structure determination step. Losing a chance of obtaining good data for crystals that were not easy to obtain or cannot be easily reproduced can be fatal for the scientific project.

Additionally, a good practice is to save the raw images and to keep the copy at least till the work with structural results has been accepted for publication. The processes of structure validation and reviewing the manuscript can require re-inspection of data or even collect data re-processing. Moreover, it is highly advisable to deposit raw images at some open data repository once the publication has been accepted (vide infra).

Finally, most European synchrotron beamlines dedicated to macromolecular crystallography offer some useful tips, access to management system (i.e. ISPyB) (Delegenière et al., 2011) and guidelines for data collection that can be found on the respective web sites:

- ALBA (Barcelona, Spain) https://www.cells.es/en/contact-info/
- BESSY (Berlin, Germany) https://www.helmholtz-berlin.de/forschung/quellen/bessy/index_en.html
- DESY (Hamburg, Germany) https://www.desy.de/
- Diamond Light Source (Oxfordshire, United Kingdom) https://www.diamond.ac.uk/
- Elettra (Trieste, Italy) https://www.elettra.trieste.it/
- The European Synchrotron Radiation Facility (ESRF; Grenoble, France) https://www.esrf.eu/
- SOLEIL (Saint-Aubin, France) https://www.synchrotron-soleil.fr/en
- Max IV (Lund, Sweden) https://www.maxiv.lu.se/
- Swiss Light Source (SLS; Villigen PSI, Switzerland) https://www.psi.ch/en/sls

However, this can be done for example with the BEST (Bourenkov & Popov, 2010) or RADDOSE (http://www.raddose.com) software packages. The final strategy applied in data collection also depends on the available geometry of the goniostat. The higher degree of freedom of crystal orientation the better. The most common synchrotron setups allow to rotate crystals around a single axis (phi), while 3- or 4-axes goniostats can be found as part of the in-house diffractometers, but increasing number of macromolecular crystallography beamlines also allows to rotate the sample around more than one axis. By using large area detectors, rotation around a single axis in most cases allows one to obtain complete data, regardless of the initial orientation of the crystal. The latest software available at synchrotron sites and in-house machines greatly helps to predict and collect data and it supports the most popular phasing methods, nevertheless the decisions need to be made by the crystallographer according to all available and previously gathered information. The data collection experiment should be conducted properly in order to obtain complete data. If the strategy was planned in a wrong way or a rapid decay of diffraction power occurred, some reflections may not be measured at all, and the data may not be complete. A number of synchrotron sites for macromolecular crystallography in Europe operates with MXCuBE (Gabadinho et al., 2010) and the latest version MXCuBE3 (Mueller et al., 2017) (https://mxmctb.github.io/mxmctb/), which supports the users in making reasonable decisions during data collection. Another important aspect of making the most of the beam time is the opportunity to process your data during or just after data collection. Quick examination of the final statistics will be beneficial in situations when for some reason the measurements went wrong and data collection needs to be repeated.

Paper in Press       5
**TIP 4: UNDERSTAND DATA PROCESSING**

Before we discuss the most important issues of data processing, a minimal portion of theory regarding the diffraction experiment should be recalled. Each ray behind reflections that can be seen on the collected images is characterized by its amplitude and phase. However, only reflection amplitudes, which are proportional to modulus of structure factor \( F \), which in turn is a sum of contributions of all atoms from the unit cell:

\[
F(hkl) = \sum_{i}^{N} f_{i} e^{2\pi i (hx_{i}+ky_{i}+lz_{i})}
\]

can be obtained from the measured intensities:

\[
I(hkl) \sim |F(hkl)|^2
\]

but no direct information about reflection phases is provided by the diffraction experiment. The function of electron density defined at every point in the unit cell, which is reconstructed from the measured structure factors' amplitudes and their phases has to be calculated:

\[
\rho(x,y,z) = \frac{1}{\sqrt{N}} \sum_{h,k,l}^{N} F(hkl) e^{-2\pi i (hx_{i}+ky_{i}+lz_{i})}
\]

Therefore, data processing that is aimed at extracting the relative intensities of the diffracted X-ray beams is a very important step in protein crystallography projects after diffraction data collection. First, recorded diffraction spots have to be indexed, next respective raw pixel intensities must be properly integrated and scaled after noise and background subtraction. Several different computer programs exist and can be used for this purpose. Among these are:

- XDS (http://xds.mpimf-heidelberg.mpg.de/) (Kabsch, 2010)
- HKL (https://www.hkl-xray.com/) (Owinskiowski & Minor, 1997)
- DIALS (https://dials.github.io/) (Winter et al., 2018)
- XIA2 (https://xia2.github.io/) (Winter, 2010)
- Mosflm (https://www.mrc-lmb.cam.ac.uk/mosflm/mosflm/) (Battye et al., 2011).

Special attention should be paid at the step of space group assignment. Wrong choice of the symmetry can lead to problems in finding the correct position of the model during molecular replacement, as well as can result in difficulties in phasing performed with the use of other methods. When refinement seems to be problematic, it is not an unusual procedure to search the solution after data reprocessing and select a different space group. If needed, this procedure can be performed with tools implemented in crystallographic software packages mentioned above.

As mentioned earlier, the collected data can be anisotropic. In case of anisotropic data it is now possible to address the statistical significance of the intensity data after merging with StarAniso (http://staraniso.globalphasing.org/cgi-bin/staraniso.cgi) (Tickle et al., 2018).

At this point we would like to encourage scientists who are new to protein crystallography to extend their knowledge about data collection statistics by reading dedicated literature. It is the author’s responsibility to collect and provide accurate information about the data quality that fulfill the standards established by the crystallographic community. Here, the most valuable metrics pertinent to results of data processing are mentioned. The first parameter is resolution that limits overall achievable information about the structure. Second is the signal-to-noise ratio, which addresses data quality. The ratio \( I/\sigma(I) \) is the most recognizable parameter that proves the signal strength, but a particularly informative indicator of the internal data consistency, apart from popular \( R_{merge} \) and \( R_{max} \) (Evans & Murshudov, 2013) used nowadays is the correlation coefficient between randomly chosen half data sets, \( CC_{1/2} \) (Karpplus & Diederichs, 2012). Also, Ios, an asymptotic \( 1/\sigma(I) \), the parameter used for identification of random and systematic errors associated with each dataset should be evaluated (Diederichs, 2010). In order to estimate the useful “resolution” of the data, \( CC_{1/2} \) is a better measure than \( R_{merge} \) or \( R_{max} \) (Evans and Murshudov, 2013). Another important issue is data completeness, defined as the coverage of all theoretically possible unique reflections within the measured data set. Data completeness remarkably influences the process of structure determination and shouldn’t be lower than 95% (Dauter, 2017). Keep in mind that the completeness can and often depends on the resolution range and can be lower in the highest resolution shell. If lower values are observed in the middle resolution ranges, the data should be carefully inspected. The last parameter to be mentioned in our roadmap is redundancy (multiplicity), which refers to the fact that every reflection is measured with a certain degree of random error (Bourenkov and Popov, 2006), therefore the higher the redundancy, the more precise the final estimation of the averaged reflection intensity.

**TIP 5: PHASING MEANS THINKING, REFINEMENT NEEDS TIME, VALIDATION IS A MUST, DEPOSITION IS A GOLD STANDARD**

Several programs have evolved from the original concept of molecular replacement to allow faster and more sophisticated searches. The most popular, MOLREP (Vagin & Teplyakov, 1997) and Phaser (McCoy et al., 2007), are included in MrBUMP (Keegan and Winn, 2007) and BALBES (Long et al., 2007), two automated molecular-replacement pipelines. MoRDa is also an interesting choice regarding the available pipelines for automated molecular replacement protein structure solution based on its own domain database derived from the PDB (Vagin & Lebedev, 2015). The very distant models or even secondary structure elements can also lead to successful ab initio solution of macromolecular structures with Arcimboldo (Rodríguez et al., 2012). Several phasing methods are available (MIR, MAD, SAD and MR) and they all rely on the premise that phase information can be obtained if the positions of marker atoms in the unknown crystal structure are known. The SHELX (Sheldrick, 2010) module of SHELX ‘Suite’ (http://shelx.uni-ac.gwdg.de/SHELX), and SOLVE (Terwilliger & Berendzen, 1999) are widely used for locating the heavy-atom sites. Direct methods is a class of solution techniques that generates good starting phases using only experimental intensities as a source of phase information and here SnB (Miller et al., 1994), SHELXD and phenix.hyss implemented in PHENIX (Adams et al., 2002; Adams et al., 2010) can be applied. Often, starting phases can be improved by changing the phases by consideration of all available phase information that arise from a combination of the known structure factor magnitudes, the current phase estimates, and stereochemical information. For this purpose a wide range of software can be used: DM (Cowie, 2010), SOLOMON (Abrahams and Leslie, 1996), RESOLVE (Terwilliger, 2004) and PIRATE (Cowie, 2010).
### Table 2. Important crystallographic terms and parameters

| Term               | Definition                                                                 |
|--------------------|-----------------------------------------------------------------------------|
| **Unit cell**      | The unit cell is the parallelepiped built on the vectors, a, b, c, of a crystallographic basis of the direct lattice. Its volume is given by the scalar triple product, \( V = |a \times b| \) and corresponds to the square root of the determinant of the metric tensor. |
| **Space group**    | The symmetry group of a three-dimensional crystal pattern is called its space group. |
| **Waves**          | Waves diffracted by a periodic distribution of simple scatterers obey Bragg\'s law, which allows ready determination of interplanar distances and thus the easy recovery of a description of the crystal structure. Where the scattering objects are complex (e.g. in molecular crystals) the diffracted radiation suffers a phase shift arising from the spatial distribution of individual scatterers. The amplitudes of the resulting structure factors are directly derivable from the experimental measured intensities of the diffracted beams, but the phases are not. Without a knowledge of the phases, it is not possible to reconstruct the individual atomic positions. Estimating the phases is an essential step in successful structure determination. |
| **Structure factor**| The structure factor \( F_{\text{calc}} \) is a mathematical function describing the amplitude and phase of a wave diffracted from crystal lattice planes characterised by Miller indices \( h,k,l \). |
| **MAD**            | An approach to solving the phase problem in protein structure determination by comparing structure factors collected at different wavelengths, including the absorption edge of a heavy-atom scatterer. |
| **MR**             | An approach to solving the phase problem by concentrating on phase relationships that arise through X-ray diffraction from similar molecular components. The components can be molecular fragments related through noncrystallographic symmetry (e.g. icosahedral subunits of a virus) or a similar molecule such as a homologous protein with high sequence identity. |
| **SAD**            | The method of single-wavelength anomalous dispersion used for solving the phase problem, makes use of data collected at just one wavelength, typically at the absorption peak or high-energy remote. It minimizes problems of radiation damage and nonisomorphism, but requires very accurate measurements. |
| **SAD**            | In the method of multiple isomorphous replacement the interference effects on the intensities of the diffracted beams caused by the addition of heavy atoms to the protein provide the estimates of the phase angles. |
| **Resolution**     | In crystal structure determination, the term resolution is used to describe the ability to distinguish between neighboring features in an electron density map. By convention, it is defined as the minimum plane spacing given by Bragg\'s law for a particular set of X-ray diffraction intensities. The resolution improves with an increase in the maximum value of \([\sin \theta]/\lambda\) at which reflections are measured. |
| **Phase problem**  | The term phase problem in crystallography commonly taken to refer to the 'conventional' R factor is a measure of agreement between the amplitudes of the structure factors calculated from a crystallographic model and those from the original X-ray diffraction data \( (F_{\text{calc}}) \). The R factor is calculated \( (F_{\text{calc}}) \) during each cycle of least-squares structure refinement to assess progress. The final R factor is one measure of model quality. |
| **R (R work)**     | A residual function calculated during structure refinement in the same way as the conventional R factor (see above), but applied to a small subset of reflections that are not used in the refinement of the structural model. The purpose is to monitor the progress of refinement and to check that the R factor is not being artificially reduced by the introduction of too many parameters. |

*From Online Dictionary of Crystallography (International Union of Crystallography)*
Irrespective of the phasing method, the aim of crystallographic model building is to construct a model that explains the experimental data with the conditions that it should make a physical and chemical sense. The latest trend in computational tools in protein crystallography is the development of all-integrated pipelines. Examples of the latter are ARP/wARP (Macromolecular Model Building for Crystallography and Cryo-EM; http://www.embbl-hamburg.de/ARP/) (Chojnowski et al., 2019), RE-SOLVE (Terwilliger, 2001) and BUCCANEER (Potterton et al., 2004) (Cowtan, 2006).

The model building is usually performed simultaneously with the process of refinement. In other words, after solving the crystallographic phase problem, the initial model is refined and accordingly the parameters of the model (geometry and B-factor values) are optimized to fit the observations using a refinement function. Different programs, provided by such crystallographic packages as CCP4 (Winn et al., 2011), SHELX (Sheldrick, 2008) or PHENIX (Adams et al., 2010) can be utilized for this purpose. Model refinement programs are coupled with the graphics display programs, for example with the most popular COOT (Emsley & Cowtan, 2004), that allow model rebuilding and interpreting regions of the difference Fourier map (unexplained by the model). The model is refined to the point when it is complete and further improvements to the structure are not possible. This is done in an iterative way until convergence is reached, monitored by the values of the R and Rfree factors (Table 2). The R-factors measure how well the simulated diffraction pattern matches the experimentally-observed diffraction pattern. Rfree is based on a test set consisting of a small percentage (usually 5–10%) of reflections excluded from a structure refinement. Another important aspect that should be kept in mind is the fact that the appearance of Fourier maps depends more on the phases than on amplitudes. Consequently, even if the correct amplitudes are known from a well-conducted diffraction experiment, inaccurate phases may introduce map bias, which may be difficult to eliminate during refinement and modeling process.

To perform automated crystal structure determination, sophisticated platforms can be used. By cascading execution of a number of macromolecular crystallographic programs, efficient pipelines are produced. A new version of HKL, HKL3000 (Minor et al., 2006) includes all the steps from data collection, processing and model determination within a single interface with the traditional graphical features of HKL. Similar functionality is offered by Auto-Rickshaw (Panjikar et al., 2005). Last years have brought more systems that facilitate the process of structure determination, for example XChemExplorer (XCE) provides an intuitive graphical user interface which guides the user from data processing, initial map calculation, ligand identification and refinement up to data dissemination (Krojer et al., 2017). Furthermore, the demand from a growing number of fragment screening experiments led to the development of PandDA (https://pandada.bitbucket.io/) (Pearce et al., 2017) that allows analysis of such data. Small molecules and ligands are abundantly represented in the PDB, nearly 80% of deposits contain chemicals that do not belong to proteins or nucleic acids. The quality of small molecule models can be improved by the use of geometrical restraints. This common technique for the refinement and validation of small molecule binding sites in protein–small molecule complexes benefits from geometrical parameters derived from the very high-resolution structures in the Cambridge Structure Database (CSD) (https://www.ccdc.cam.ac.uk/) (Groom et al., 2016) that can be used as restraints in small molecule refinement. The ligand binding-site identification, ligand description and conformation generation, ligand fitting, refinement and subsequent validation can be successfully performed with a set of dedicated software: eLBOW (part of the PHENIX suite) (Moriarty et al., 2009), JILgand (implemented in the CCP4 project) (Lebedev et al., 2012), and Grade (part of BUSTER) (http://grade.globalphasing.org).

It is the primary goal of structural databases to provide highly reliable data, where “reliability” is defined by rigorous validation strategies and quality indicators. Thus, for instance PDB actively works with journals and depositors to provide feedback at an early stage, often actually improving the quality of the data that is to be deposited. The latter was a motivation for an independent initiative, now running for many years, which is the PDB REDO project (https://pdb-redo.eu/) (Joosten et al., 2009). This server provides a re-refined structure with suggested improvements i.e. new coordinates’ set for each and every PDB deposit. It also offers a useful server to assist the depositors, before they deposit, to look at the PDB REDO version of their current cycle of model refinement.

Model validation on the protein polypeptide chain can be performed with several programs that provide a statistical evaluation of the geometrical parameters of the structure. For the purpose of validation, scientists can refer to MolProbity (Chen et al., 2010), PROCHECK (Laskowski et al., 1993), WHAT IF (Vriend, 1990) and SFCHECK (Vaguine et al., 1999). After careful inspection of the validation results, that can be also performed with wwPDB OneDep System (https://validate-rcsb-2.wwpdb.org/) and solving the pinpointed issues, the authors can deposit their structures to PDB. This last step, leading to the release of data via the public repository is a prerequisite for publishing structural reports and, by revealing experimental details, it also supports the idea of reproducible science.

Even though the validation system is nowadays an efficient procedure, one should remember that true and critical evaluation of macromolecule structures, in terms of quality and reliability, before referring to existing deposits (MR models, homologues, orthologs) and during submission is crucial (Dauter et al., 2014).

Furthermore, deposition and annotation tools implemented in PDB require from the depositors that atomic coordinates and primary experimental data plus associated metadata are submitted. The ease of archiving raw diffraction data sets is a remarkable development of recent years. In addition, the desire to maximize the availability of research data in accordance with the so-called FAIR principles – Findable, Accessible, Interoperable, and Re-usable (https://www.force11.org/group/fairgroup/fairprinciples) (Wilkinson et al., 2016), encourages crystallographers to deposit and share the raw data. The Integrated Resource for Reproducibility in Macromolecular Crystallography (https://proteindiffraction.org/) (Grabowski et al., 2019) and Macromolecular Xtallography Raw Data Repository (https://mxrd.icm.edu.pl/) are good examples of such initiatives that include a repository system.

**TIP 6: ANALYZE AND VISUALIZE WITH THE USE OF GRAPHICAL TOOLS**

A 3D protein structure model is a very rich information source which is best analyzed with the help of some advanced visualization software. There are currently many graphics programs that are suitable for displaying and analyzing protein structures, most of them...
with capability to: display various representations at once (cartoon, ribbon, ball-and-sticks, sticks, etc.), apply different coloring schemes (by: atom type, B-factor value, secondary structure, etc.), measure geometrical parameters of the model, identify steric clashes, display electron density maps, and save high quality graphics. Majority of the programs also have some scripting interface, which is very useful to automate routine procedures and also save and restore the work. A comprehensive review of the available graphical software packages is far beyond the scope of this brief review, hence here we just list some popular, freely available packages with links to their websites.

Selected graphical tools for macromolecular crystallography (in alphabetical order):

- Coot ([https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/](https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/)) (Emsley & Cowtan, 2004)
- Jmol ([http://jmol.sourceforge.net/](http://jmol.sourceforge.net/)) [Jmol: an open-source Java viewer for chemical structures in 3D]
- MolMol ([https://sourceforge.net/projects/molmol/](https://sourceforge.net/projects/molmol/)) (Koradi et al., 1996)
- MolScript ([https://kraulis.se/MolScript/](https://kraulis.se/MolScript/)) (Kraulis, 1991)
- PyMOL ([https://pymol.org/2/](https://pymol.org/2/) or [https://github.com/schrodinger/pymol-open-source](https://github.com/schrodinger/pymol-open-source)) (DeLano, 2002)
- UCSF ChimeraX ([https://www.cgl.ucsf.edu/chimerax/features.html](https://www.cgl.ucsf.edu/chimerax/features.html)) (Petterson et al., 2020)
- VMD ([http://www.ks.uiuc.edu/Research/vmd/](http://www.ks.uiuc.edu/Research/vmd/)) (Humphrey et al., 1996)

**TIP 7: ANALYSIS OF STRUCTURAL FEATURES WILL PUT YOUR STRUCTURE IN A BROADER CONTEXT**

Analysis of protein structures and their interactions with other molecules is often very helpful in elucidating their cellular functions and mechanisms of action. Thus, XRD structural methods belong to the leading scientific strategies for identification of protein’s biological and biochemical relevance.

Analysis of macromolecular interfaces, including prediction of likely oligomeric state and generating its coordinates, calculations of interface area and estimation of free energy of assembly dissociation are only selected capabilities offered by the PDBePISA server ([https://www.ebi.ac.uk/msd-srv/prot_int/pistart.html](https://www.ebi.ac.uk/msd-srv/prot_int/pistart.html)) (Krissinel & Henrick, 2007). The server also lists amino acids making up the interfaces, evaluates significance of individual residues for macromolecular contacts and offers an advanced search engine for biological interfaces from among structures deposited at PDB.

The DALI server ([http://ekhidna2.biocenter.helsinki.fi/dali/](http://ekhidna2.biocenter.helsinki.fi/dali/)) (Holm, 2020) allows one to perform protein comparison based on the 3D structure. The server offers several options, including searching PDB for similar 3D structures, pairwise comparison between selected structures (or individual chains) and “all against all” structural comparison for up to 64 structures. Several modes of results visualization, including structural trees, structurally aligned sequence logos and 3D models with mapped structural or sequence variation aid in results’ analysis.

Structural studies on multi domain or multi chain proteins may yield structures corresponding to different conformational states of the macromolecule, e.g. closed vs open conformation. In such a case, the DynDom program or server ([http://dyndom.cmp.uea.ac.uk/dyndom/](http://dyndom.cmp.uea.ac.uk/dyndom/)) may turn out to be very useful to identify hinge residues and moving domains, as well as the axes by which the (components of) movement take place (Poornam et al., 2009). The DynDom website also hosts several browseable databases with results of protein domain movement analysis.

Structures of protein-ligand complexes provide valuable insights into interactions between a small molecule, which can be e.g. an inhibitor, a drug or a reactant, and the host macromolecule. Classification of these interactions is greatly enhanced by the Arpeggio program or server ([http://biosig.unimelb.edu.au/arpeggioweb/](http://biosig.unimelb.edu.au/arpeggioweb/)) (Jubb et al., 2017), which identifies the type of interaction between the ligand-protein atom pairs (above a dozen of different types) and generates a PyMOL session file, which can be used to visualize the results in 3D. For an example of such analysis see Fig. 4, where the overall structure of hyoscyamine 6β-hydroxylase (H6H, PDB: 6ttm) (Kluza et al., 2020) is depicted (panel A), the secondary structure is highlighted (panel B) and key interatomic interactions engaged in the enzyme-substrate recognition are presented (panel C).

PDBeSum server is also worth noticing here, as it provides succinct yet richly illustrated summary of protein structure (3D, secondary and primary), its interactions with ligands and metal ions analyzed and illustrated by LIGPLOT (Wallace et al., 1995), as well as 3D visualization of clefts and cavities within the protein molecule. Quality assessment report generated by PROCHECK is

![Figure 4. Visualization of the structure and key interatomic interactions for hyoscyamine 6beta-hydroxylase (H6H) in complex with its substrate – hyoscyamine (PDB: 6ttm).](https://www.cgl.ucsf.edu/chimerax/features.html)

(A) Overall structure of H6H complexed with Ni²⁺ (cyan sphere), hyoscyamine (sticks, C atoms in salmon), co-substrate mimes – N-oxalylglycine (sticks, C atoms in red). Two histidine and one aspartate that coordinate the metal are also shown (sticks, C atoms in green).

(B) An overview of H6H with secondary structure highlighted – helices in red, β-strands in yellow, loops and coil regions in green.

(C) Close-up view of hyoscyamine (sticks with C atoms in green) binding pocket with depicted key interactions between the substrate and protein that were identified by the Arpeggio server. Hydrogen bonds as red discs, C-H…π interactions as blue discs, weak polar interactions as orange discs. Graphics were generated with PyMOL.
CONCLUDING REMARKS

Protein crystallography together with Cryo-EM and NMR are the most powerful techniques for the structure determination of macromolecules, as well as for the analysis of mechanisms of protein actions and interactions at the atomic level. The algorithms and methods for structure determination initially formulated decades ago are now becoming more and more elaborate, but thankfully the computational tools wrapping around these advanced methods have evolved toward simpler and more user-friendly packages and web interfaces. This, combined with amply available tutorials, YouTube channels, manuals, data deposited at open repositories and other educational materials freely available in the internet lowers the “activation barrier” for a novice in the field eager to learn protein crystallography methods. We hope this short review will be a useful aid in this fascinating journey.

REFERENCES
from conventionally interpretable electron density. Nat. Commun. 8: 1–8. https://doi.org/10.1038/ncomms15123

Pellegrini E, Piano D, Bowler MW (2011) Direct cryocooling of naked crystals: Are cryoprotection agents always necessary? Acta Crystallogr. Sect. D Biol. Crystallogr. 67: 902–906. https://doi.org/10.1107/S090744491031210

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera – a visualization system for exploratory research and teaching. J. Comput. Chem. 25: 1694–1713. https://doi.org/10.1002/jcc.20084

Poonam GP, Matsumoto A, Ishida H, Hayward S (2009) A method for the analysis of domain movements in large biomolecular complexes. Proteins Struct. Funct. Bioinforma. 76: 201–212. https://doi.org/10.1002/prot.22339

Potterton I, McNicholas S, Kristinell E, Gruber J, Cowtan K, Emsley P, Murshudov GN, Cohen S, Perrakis A, Noble M (2004) Developments in the CCP 4 molecular graphics package. Acta Crystallogr. Sect. D Biol. Crystallogr. 60: 2288–2294. https://doi.org/10.1107/S0907444904023716

Powell HR (2017) X-ray data processing. [WWW document]. Bioinf. Res. Rev. 37: 20170227. https://doi.org/10.1093/bior/brx009

Rodriguez D, Sammito M, Meindl K, De Iardyna IM, Potratz M, Sheldrick GM, Uson I (2012) Practical structure solution with AR-CIMBODO. Acta Crystallogr. Sect. D Biol. Crystallogr. 68: 336–343. https://doi.org/10.1107/S1600536812010360

Russo Krauss I, Merlino A, Vergara A, Sica F (2013) An Overview of Biological macromolecule crystallization. Int. J. Mol. Sci. 14: 11643–11661. https://doi.org/10.3390/ijms140911643

Sheldrick GM (2008) A short history of SHELX. Acta Crystallogr. Sect. A Found. Crystallogr. 64: 112–122. https://doi.org/10.1107/S0001784307049391

Sheldrick GM (2010) Experimental phasing with SHELXC/D/E: Combining chain tracing with density modification. Acta Crystallogr. Sect. D Biol. Crystallogr. 66: 479–485. https://doi.org/10.1107/S090744490903860

Siers F, Wülm A, Discisin D, Gibson TJ, Karpulis KJ, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7: https://doi.org/10.1038/msb.2011.75

Slahinski I, Jaroszewski L, Rychlewski L, Wilson IA, Lesley SA, Godzik A (2007) XtalPred: a web server for prediction of protein crystallizability. Bioinformatics 23: 3403–3405. https://doi.org/10.1093/bioinformatics/btm273

Terwilliger T (2004) SOLVE and RESOLVE: Automated structure solution, density modification, and model building. J. Synchrotron Radiat. 11: 49–52. https://doi.org/10.1107/9074491030938

Terwilliger TC, Berendzen J (1999) Automated MAD and MIR structure solution. Acta Crystallogr. Sect. D Biol. Crystallogr. 55: 849–861. https://doi.org/10.1107/S090744499900839

Terwilliger TC (2001) Maximum-likelihood density modification using a pattern recognition approach to data reduction. Acta Crystallogr. Sect. D Biol. Crystallogr. 57: 1755–1762. https://doi.org/10.1107/907449101373

Tickle JI, Flensburg G, Keller P, Paciorek W, Shariff A, Voroncsh G, Bernocco G (2018) ST-IR-JNDO. Cambridge, United Kingdom: Global Phasing Ltd.

Vagunie AA, Richelle J, Wodak SJ (1999) SFCHECK: A unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr. Sect. D Biol. Crystallogr. 55: 191–205. https://doi.org/10.1107/907449101373

Vagunie AA, Micheal A, Yen S, Robinson D, Jaroszewski L, Richardson D (1997) Automated structure solution. Acta Crystallogr. Sect. D Biol. Crystallogr. 53: 912–917. https://doi.org/10.1107/907449101373

Vagunie AA, Richelle J, Wodak SJ (1999) SFCHECK: A unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallogr. Sect. D Biol. Crystallogr. 55: 191–205. https://doi.org/10.1107/907449101373

Vera I, Stura EA (2014) Strategies for protein cryocrystallization. Cryst. Growth Des. 14: 427–433. https://doi.org/10.1021/cg50131f

Vriend G (1990) WHAT IF: A molecular modeling and drug design program. J. Appl. Crystallogr. 23: 405–407. https://doi.org/10.1107/S0021889890000374

Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng. 8: 25–30. https://doi.org/10.1093/protein/8.1.25
Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3: 1–9. https://doi.org/10.1038/sdata.2016.18

Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr. D. Biol. Crystallogr. 67: 235–42. https://doi.org/10.1107/S0907444910045749

Winter G (2010) Xia2: An expert system for macromolecular crystallography data reduction. J. Appl. Crystallogr. 43: 186–190. https://doi.org/10.1107/S0021889809045701

Winter G, Waterman DG, Parkhurst JM, Brewster AS, Gildea RJ, Gerstel M, Fuentes-Montero L, Vollmar M, Michels-Clark T, Young ID, Sauter NK, Evans G (2018) DIALS: Implementation and evaluation of a new integration package. Acta Crystallogr. Sect. D Struct. Biol. 74: 85–97. https://doi.org/10.1107/S2059798317017235

Zhang CY, Wu ZQ, Yin DC, Zhou BR, Guo YZ, Lu HM, Zhou R Bin, Shang P (2013) A strategy for selecting the pH of protein solutions to enhance crystallization. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 69: 821–826. https://doi.org/10.1107/S1744309113013651

Zhang J, Madden TL (1997) PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. Genome Res. 7: 649–656. https://doi.org/10.1101/gr.7.6.649