UNIVERSITÄT ZU LÜBECK

From the Institute of Biochemistry
of the University of Lübeck
Director: Prof. Dr. Thomas Krey

Overcoming barriers in intracellular crystallography

Dissertation
for Fulfillment of Requirements for the Doctoral Degree
of the University of Lübeck

from the Department of Natural Sciences

Submitted by

Juliane Boger, née Meyer
from Lübeck

Lübeck, 2023

First referee: Prof. Dr. Lars Redecke

Second referee: Prof. Dr. Robert Huber

Date of oral examination: 08.11.2023

Approved for printing: Lübeck, 10.11.2023

Abstract

Intracellular crystallography is a method for elucidating the three-dimensional structure of proteins. It relies on the spontaneous crystallization of a protein in the producing cells, followed by the collection of X-ray diffraction data and structure elucidation. It is a promising method that allows not only the crystallization of proteins that could not be crystallized *in situ* (Gati et al., 2014; Redecke et al., 2013), but also the identification of cofactors bound in the quasi-native environment of the cell (Nass et al., 2020).

Nevertheless, most of the published intracellular crystallization workflows lack a systematic methodology for the collection and processing of diffraction data. Here, two strategies have been implemented to process diffraction data from two model proteins crystallized in insect cells. Diffraction data was produced by helical line scans over a MeshMount at a synchrotron source under cryo-conditions. The first strategy uses a multiple crystal approach (MSX), where a tailor-made script identifies successive frames belonging to the same crystal and then processes them individually with *XDS*. The second approach is a classical serial approach using *CrystFEL*, a software designed for still images from serial femtosecond X-ray diffraction (SFX) at room temperature. The data processing was adversely affected by the presence of salt and ice crystals, which interfere with the protein crystal diffraction, thus a filtering script was developed to enable data processing with *CrystFEL*.

Further, protein homology was used to identify potential candidates for *in cellulo* crystallization. Nine of the ten tested orthologs of the natively crystallizing protein HEX-1 from *Neurospora crassa* spontaneously crystallized when produced in living insect cells using a baculovirus expression system. Four protein structures were solved and showed a high degree of sequence and structural conservation among the different species. Additionally, *Trypanosoma brucei* GMP reductase, a paralog of the IMP dehydrogenase, formed needle-shaped crystals with a large unit cell but diffracted X-rays only at low resolution, hindering structure solution and cofactor identification.

To overcome the limited resolution of the *Trypanosoma brucei* GMP reductase, either the radiation intensity or the diffraction volume had to be increased. Since access to X-ray free electron lasers (XFELs) that provide radiation with higher brilliance is limited, the crystal volume was increased by fusing adjacent, baculovirus-infected cells. This proved to be a powerful method, and a resolution shift of 0.7 Å was obtained, allowing the identification of bound cofactors, namely a GTP moiety in the canonical binding site II of the Bateman domain and the presence of a substrate moiety in the active site of the enzyme.

The establishment of the pipeline using the cellular compartments as screening parameters (Schönherr, 2021), the introduction of the established strategies for diffraction data processing and structure solution, as well as the developed solutions for resolution improvement will extend the usability of *in cellulo* crystallography. The selection of proteins to be crystallized based on their sequence similarity will increase the number of structures solved with this method and lead to a better understanding of the processes of spontaneous intracellular protein crystallization.

Zusammenfassung

Die intrazelluläre Kristallographie ist ein Verfahren zur Aufklärung der dreidimensionalen Struktur von Proteinen. Dieses Verfahren basiert auf der spontanen Kristallisation eines Proteins in den produzierenden Zellen und der anschließenden Sammlung von Röntgenbeugungsdaten und der Strukturlösung. Es handelt sich um eine vielversprechende Methode, die nicht nur die Kristallisation von Proteinen ermöglicht, die bisher nicht *in situ* kristallisiert werden konnten (Gati et al., 2014; Redecke et al., 2013), sondern auch die Identifizierung von Kofaktoren, die in der quasi-nativen Umgebung der Zelle an das Zielprotein binden (Nass et al., 2020).

Dennoch fehlt in den meisten publizierten Arbeitsabläufen der *in cellulo* Kristallographie eine systematische Methodik zur Diffraktionsdatenaufnahme und -verarbeitung. Im Rahmen dieser Arbeit wurden zwei entsprechende Ansätze implementiert, um Beugungsdaten von zwei Modellproteinen zu verarbeiten, die durch sogenannte helikale Linienscans auf einem MeshMount an einer Synchrotronstrahlungsquelle unter Kryobedingungen aufgenommen wurden. Der erste Ansatz verwendet einen Mehrfachkristall-Ansatz (MSX), bei dem ein maßgeschneidertes Skript aufeinanderfolgende Beugungsbilder identifiziert, die zum selben Kristall gehören, und sie dann einzeln mit *XDS* verarbeitet; der zweite Ansatz ist ein klassischer serieller Ansatz unter Verwendung von *CrystFEL*, einer Software, die für Standbilder aus serieller Femtosekunden-Röntgenbeugung (SFX) bei Raumtemperatur entwickelt wurde. Die Datenverarbeitung wurde von Salz- und Eiskristallen in den Kristall-tragenden Zellproben beeinträchtigt, deren Beugung mit der Beugung der Proteinkristalle überlagert. Daher wurde ein Filterskript entwickelt, das kontaminierende Beugung aus dem Datensatz entfernt und somit die weiterführende Datenverarbeitung mit *CrystFEL* ermöglicht.

Die Homologie von Proteinen wurde genutzt, um potenzielle Kandidaten für die *in cellulo*-Kristallisation zu identifizieren. Neun von zehn getesteten Orthologen des nativ kristallisierenden Proteins HEX-1 aus *Neurospora crassa* bildeten spontan Kristalle, wenn sie in lebenden Insektenzellen mit Hilfe des Baculovirus-Expressionssystems produziert wurden. Vier HEX-1 Proteinstrukturen wurden gelöst und zeigten ein hohes Maß an Sequenz- und Strukturkonservierung zwischen den verschiedenen Spezies. Die GMP-Reduktase aus *Trypanosoma brucei*, ein Paralogon der IMP-Dehydrogenase bildete ebenfalls nadelförmige Kristalle mit einer großen Einheitszelle, beugte jedoch nur zu einer geringen Auflösung, was eine schnelle Strukturlösung verhinderte.

Um die Auflösungsbeschränkung der GMP-Reduktase aus *Trypanosoma brucei* zu überwinden, muss entweder die Strahlungsintensität oder das Beugungsvolumen der Kristalle erhöht werden. Da der Zugang zu freien Elektronenlasern (XFELs), die Röntgenstrahlung mit deutlich höherer Brillanz liefern, begrenzt ist, wurde das Kristallvolumen durch Fusion benachbarter, Baculovirus-infizierter Zellen vergrößert. Dies erwies sich als leistungsfähige Methode zur Optimierung der

Beugungsdaten. Es wurde eine Auflösungsverschiebung von 0,7 Å erzielt, was neben der Strukturaufklärung auch die Identifizierung von gebundenen Kofaktoren ermöglichte. So wurde in der kanonischen Bindungsstelle II der Bateman-Domäne ein GTP-Molekül identifiziert, sowie außerdem die Anwesenheit einer Substratkomponente im aktiven Zentrum des Enzyms nachgewiesen.

Die Etablierung der Pipeline, welche die zellulären Kompartimente als Screening-Parameter nutzt (Schönherr, 2021), sowie die Einführung der hier etablierten Strategien zur Diffraktionsdatenverarbeitung und Strukturlösung, sowie die entwickelten Lösungen für die Verbesserung der Auflösung werden die Nutzbarkeit der *in cellulo* Kristallographie erweitern. Die Auswahl von zu kristallisierenden Proteinen anhand ihrer Sequenzähnlichkeit wird die Anzahl an Strukturen, welche mit dieser Methode gelöst werden, vergrößern und dazu führen, dass die Prozesse der spontanen intrazellulären Proteinkristallisation besser verstanden werden.

Table of Content

1. Introduction

Scientists have been trying to unravel the mysteries of proteins since the beginning of the 20th century. Anfinsen was one of the first scientists to focus on how proteins fold into their functional structure. In 1896, Röntgen's discovery of X-rays as a powerful tool for revealing what lies beneath the surface of matter made the experimental investigation of crystals possible. A few pioneering findings later, such as those from scientists like Laue and the father and son duo Bragg, the diffraction of waves on a crystal lattice was understood. This was the basis of X-ray crystallography.

The 3D structure of a protein is closely related to the molecular mechanism of its function. Proteins have evolved to perform a specific function in the life cycle of a cell through their specific, unique and stable structure (Anfinsen, 1973). They cluster into subgroups with similar functions and folds. These subgroups may be similar in sequence, but sometimes evolution has driven different amino acid sequences into similar folds.

Structural information of proteins is required in many areas of research, such as drug development or for understanding the pathogenesis of diseases. In the last 15 years, with the advent of structure-based design of drugs and therapeutics, structural information has become more valuable than the cost of determining the three-dimensional structure of the protein, which decreased thanks to streamlined high-throughput methods (Stevens, 2003). Structural biology has begun to attract attention and is becoming increasingly important for rational design approaches (Stevens, 2003).

Structural information can be obtained by a variety of methods. More than 80 % of all structures in the protein database (pdb) were solved using X-ray crystallography (Berman et al., 2000). This method is based on the diffraction of photons at a crystal lattice. However, other diffraction-based techniques like neutron or electron diffraction are also well established. The second most common method of obtaining structural information is nuclear magnetic resonance (NMR) spectroscopy. This requires the protein to be in aqueous solution. In the last years, another non-crystalline approach, the single-particle cryo-electron microscopy (cryo-EM) has revealed more and more structures of multimers and larger assemblies of proteins. Additionally, KI-based structure prediction programs such as AlphaFold (Jumper et al., 2021; Varadi et al., 2022) or RoseTTAFold (Baek et al., 2021) have emerged during the last years. These methods rely on machine learning to compute the most likely three-dimensional structure of a given protein sequence within minutes or hours, challenging the need for experimental structural biology in general. However, the mentioned computational methods reach their limits, when it comes to cofactors, ligands, bound ions and posttranslational modifications as well as protein dynamics and interactions (Perrakis and Sixma, 2021). Therefore, experimental studies remain still important.

By exploiting intracellular crystallization of proteins followed by X-ray diffraction, previous studies have shown that it is possible to identify the native cofactors of an enzyme in the natural environment of the cell (Nass et al., 2020), something that AlphaFold cannot do yet. This makes *in cellulo* crystallography (i.e., the crystallization of recombinant proteins and subsequent X-ray diffraction inside living cells), a powerful tool for identifying potential drug interaction sites.

1.1 Brief introduction into X-ray diffraction on protein crystals

The diffraction of X-rays on a protein crystal was first described by Bernal and Crowfoot in 1934. They collected diffraction data of a pepsin crystal using the characteristic copper Kα radiation, revealing unit cell parameters of axis a and b. They broke with the believe of long fibrous proteins and speculated on solvent content and a globular shape as well as the structure of soluble proteins (Bernal and Crowfoot, 1934), which nowadays seem to be surprisingly correct.

Nevertheless, this work would not have been published without Wilhelm Henry and Wilhelm Lawrence Bragg (1913), who derived the relation between the scattering angle θ, the incoming wavelength λ and the relative spacing of adjacent diffraction planes d as shown in Figure 1.1. n is an integer and reflects to the order of the observed interference maximum. The resulting Bragg's law (Equation 1.1) is the key in understanding diffraction physics.

$$n\lambda = 2d \sin\theta$$

Equation 1.1: Bragg's law (Bragg, 1913)

Bragg's law relies on the optical path difference Δs of two reflected waves on different diffraction planes and can be derived as follows. If the electromagnetic wave of an X-ray photon hits a plane of a crystal lattice, the wave is elastically scattered on the lattice plane, where the angle of the incident wave equals the angle of the scattered wave. Waves reflected at different lattice planes have an optical path difference Δs which can be described by the sine function of the angle θ (Figure 1.1):

$$\Delta s = \overline{FG} + \overline{GH} = 2\overline{FG} = 2d \sin\theta$$

Equation 1.2: Relation of the optical path difference Δs as depicted in Figure 1.1 and the interplanar spacing d in dependence of the angle θ.

Overlapping waves interfere. In the extreme cases, the waves of the same frequency and direction either interfere constructively or destructively depending on their path difference. Destructive interference means that if two waves overlap and a wave crest meets a wave trough, the resulting wave is extinguished. Constructive interference on the contrary means that, if two waves overlap and a wave crest meets a wave crest, the resulting wave is enhanced. Constructive interference is observed when the path difference Δs is equal to an integer multiple of the wavelength λ, resulting in Braggs law (Equation 1.1).

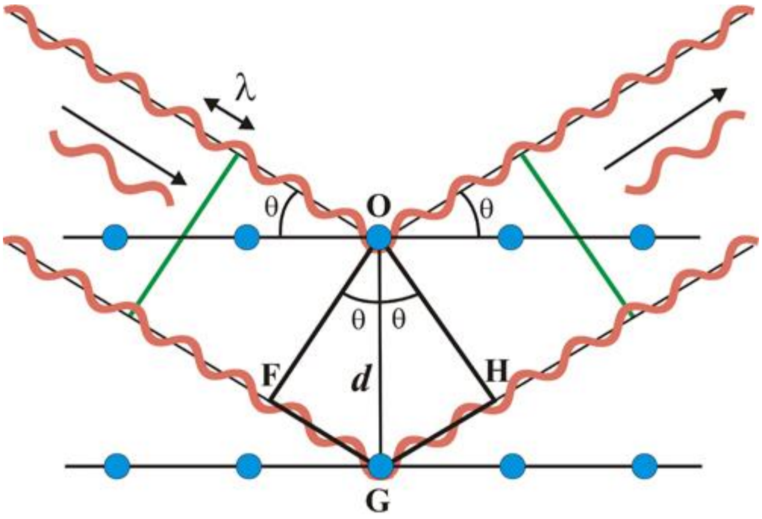
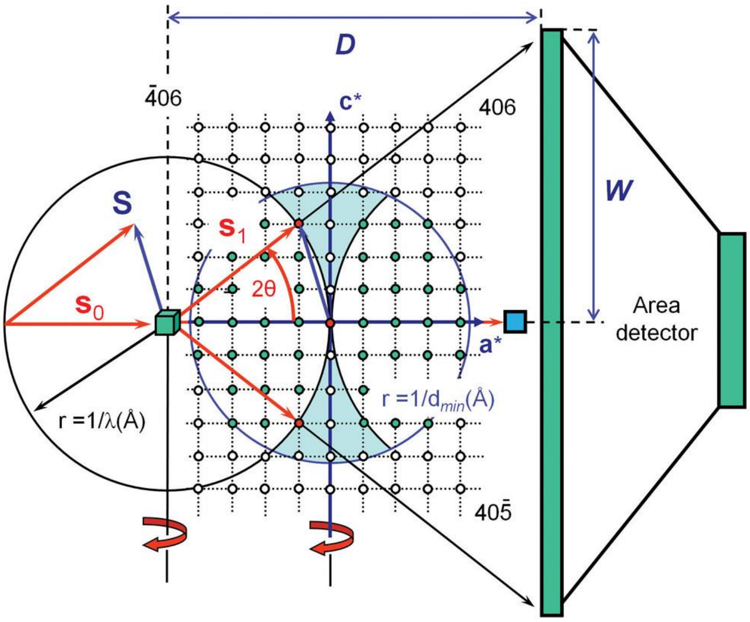


Figure 1.1: Graphical description of Bragg's law (Martínez-Ripoll, 2022). The incident waves are reflected on the different lattice planes of the crystal. The angle θ of the incoming wave equals the angle of the outgoing, reflected wave. The path difference between the reflected waves equals twice the distance between F and G.

The relation of the diffraction physics can also be depicted by the Ewald's sphere (Ewald, 1913; Rupp, 2010). The Ewald's sphere is a representation of the diffraction of X-rays with a wavelength λ on a crystal in the reciprocal space. The reciprocal space helps to understand and visualize the relations of the different parameters and thus enables the graphical description of diffraction physics (Rupp, 2010). The incoming wave (shown as wave vector s_0 with the length $1/\lambda$) is diffracted by the crystal depicted by the green cube in Figure 1.2 (Rupp, 2010). The X-ray detector, depicted as green rectangle, sets the resolution limit, by limiting the angle of the scattered wave that can be collected by the detector. The resolution sphere depicted here as blue circle around the origin of the reciprocal lattice depicted in red (Figure 1.2), with a radius of $1/d_{min}$ gives the minimal spacing of interplanar distances, that can result in a detectable Bragg spot. The intersection of the Ewald's sphere and the resolution sphere gives the amount of reciprocal lattice points, that cannot be observed by rotation of the crystal, depicted as blue shaded zone (Figure 1.2) (Dauter, 1999; Rupp, 2010).

To fully record the lattice points intersecting with the Ewald's sphere, crystals are oscillated (rotated) during data collection at synchrotron sources. This results in the formation of lunes of the Ewald's sphere and the resulting diffraction patterns (Rupp, 2010). The bandwidth of radiation from an X-ray free electron laser (XFEL) is larger than at a synchrotron source (Hwu and Margaritondo, 2021) resulting in an broader Ewald's sphere without a per-frame oscillation. However, reflections are sometimes only partially recorded and their partiality needs to be modeled (Rupp, 2010).



© Garland Science 2010

Figure 1.2: Ewald's sphere as a graphical description of diffraction physics using the reciprocal space. The incoming wave s_0 with the length 1/λ, is diffracted by the crystal (green cube) and results in constructive interference on the detector surface, as indicated by the reciprocal lattice point (red dot) intersecting the Ewald's sphere (black circle). The blue circle displays the resolution sphere with a radius of $1/d_{min}$. The blue shaded zone marks the reciprocal lattice points which will not intersect with the Ewald's sphere and thus cannot be collected. Reproduced with permission from Biomolecular Crystallography by Bernhard Rupp, © 2009-2014 Garland Science/Taylor & Francis LLC.

Molecules are arranged periodically in a crystal defined by the unit cell, which has identical content throughout the crystal. The combination of the lattice and the symmetry defines the space group of the crystal. Since proteins are chiral, the protein itself can never be mirrored by a mirror plane, but only by simple rotational axes or screw axes only. This limits the number of possible space groups for proteins to 65. The space group can usually be defined in different ways and still explain the same crystal, so it was decided that the maximum possible symmetry should be applied in each case (Rupp, 2010).

However, the description of the protein crystal as a lattice is a theoretical approach for describing diffraction behavior. X-rays are discretely scattered on the electrons of the atoms of the protein, which are periodically arranged in a crystal (Rupp, 2010). The scattering function of the entire unit cell is the summation of all single scattering functions of the specific atoms in the unit cell (Rupp, 2010). The scattering function gives the amplitude, phase, and position of the events of constructive interference that are depicted as Bragg spots in a diffraction image (Rupp, 2010). The distance of scattering atoms correlates with the distance of the peaks in the scattering function. The peak intensity is closely related to the amplitude of the scattered wave (Rupp, 2010). One single molecule still scatters X-rays the same way; however, the intensity of the scattering function is not detectable (Rupp, 2010). The crystal lattice serves as a 'signal amplifier', as more copies of the motif, i.e., the unit cell, contribute to obtain a strong and detectable signal. The signal amplification factor is proportional to the number of repeating motifs (Rupp, 2010). This emphasizes that to generate a detectable signal the protein of interest needs to be in a crystalline state.

1.2 Advances in data collection strategies and radiation sources enable the use of small crystals for structure elucidation

The establishment of improved radiation sources tremendously helped to promote X-ray crystallography for structure elucidation (Johansson et al., 2017). The establishment of the first synchrotron beamline for structural studies of proteins in Stanford in 1974 enabled recording the diffraction patterns of smaller crystals in a shorter time frame at higher resolution than with lab-sources (Phillips et al., 1976). When the usability and applicability of the X-rays as a byproduct of particle accelerators became clearer, synchrotron storage rings were built solely for that purpose (Johansson et al., 2017).

A synchrotron typically consists of an injector (the electron source) and a booster accelerator that increases the energy of the electrons, before eventually injecting these bunches of electrons into a storage ring i.e., the actual synchrotron (Lindley, 1999). When the electrons are then accelerated by, for example, undulators, X-rays are emitted and directed into the experimental hutch. The performance of a radiation source is defined by three parameters: the brilliance, the coherence, and the polarization. For macromolecular crystallography (MX) the most important parameter is the brilliance or brightness of a radiation source. It is defined by the concentration of light, measured as the emitted photon flux, in the zone of utilization and depends inversely on the solid angle of emission and the source area and (Hwu and Margaritondo, 2021). It has been increased by 22 orders of magnitude since the 1970s by using free electrons instead of solid matter as electron sources, by a smaller source area, and the use of undulators (Hwu and Margaritondo, 2021). Coherence is required to observe interference events and it was improved by the advent of undulators leading to longitudinal coherence without the use of monochromators (Hwu and Margaritondo, 2021).

X-rays are emitted whenever electrons are accelerated. Thus, acceleration of electron bunches by bending magnets or insertion devices like wigglers (second generation) or undulators (third generation) generate very intense X-ray radiation (Lindley, 1999). Insertion devices apply a periodic magnetic field on the electrons resulting in a cone-shape emission of photons. The angular spread of the beam gives the5egreee of lateral coherence. Undulators lead to emissions in a tight cone by emitting photons in a small deflection angle, while wigglers and bending magnets lead to a wider spread photon emission and a larger deflection angle. The wider cone of emitted photons of wigglers lead to a reduction in photons that reach the zone of utilization by simply being too far spread. The narrower photon emission cone emitted from undulators results in a photon bunch which is of higher longitudinal coherence or ‘in-phase’ for a particular wavelength. Thus, undulators produce the highest brilliance (Lindley, 1999). Additionally, advances in the focusing optics resulted in the implementation of microfocus beamlines, centering a high photon flux (photons per seconds) on a small sample area of e.g. 3 x 7 μm^2 (Flot et al., 2010).

In parallel, new sources of X-ray radiation were developed and resulted in the implementation of X-ray free electron lasers (XFELs). These radiation sources emit X-rays with a photon energy of up to 25 keV (see beamline specifications European XFEL) and a brilliance increased by a factor of one to ten billion (Johansson et al., 2017). Radiation from XFELs is characterized by very short pulse length (femtoseconds) and high coherence of the emitted photons achieved by optical amplification (Hwu and Margaritondo, 2021; Johansson et al., 2017). An exposure time in the range of femtoseconds allows data collection free of radiation damage as it is below the time scale of ionization and onset of radiation damage (between 10 and 100 femtoseconds). This is termed the diffraction-before-destruction principle (Neutze et al., 2000). However the applied energy destroys the crystal and inhibits further data collection from the same crystal (Johansson et al., 2017). Consequently, tens of thousands of crystals in random orientation are required to collect a complete data set covering the full rotational space. Accordingly, serial femtosecond crystallography (SFX) was born (Boutet et al., 2012; Chapman et al., 2011).

The massive increase in intensity and brilliance, which arose by the implementation of these improved radiation sources, allowed a decrease in crystal size that is sufficient to reach a comparable intensity of the Bragg spots to what was observed before these technical improvements. The intensity of a Bragg spot in a practical experiment depends not only on the structure factor as stated in chapter 1.1, but on various factors like exposure time, crystal volume and diffraction geometry (Holton and Frankel, 2010). The scale factor which scales the fully recorded reflection intensity to the structure factor is given by Darwin’s Formula (Equation 1.3) (Darwin, 1914; Havighurst, 1926; Holton and Frankel, 2010). It yields the integrated spot intensity I $[photons\ per\ spot]$ from facility parameters like the initial beam intensity I_{beam} $[photons\ s^{-1} m^{-1}]$, the polarization factor P, the Lorentz factor L, the X-ray wavelength λ and the angular velocity of the crystals ω with crystal parameters like the structure factor F and the illuminated X-ray volume V_{xtal}, the unit cell volume V_{cell} with the classical electron radius r_e, and the transmittance A through the traveled path of the X-rays to the detector (Holton and Frankel, 2010). This equation suggests that an increase in the intensity of the incoming beam by a factor of 9 orders of magnitude (like changing from a synchrotron source to an XFEL source) results in a comparable reflection intensity of a crystal having a volume smaller than 9 orders of magnitude or is shorter by a factor of 1000 in all three dimensions.

$$I = I_{beam} {r_e}^2 \frac{V_{xtal}}{V_{cell}} \cdot \frac{\lambda^3 L}{\omega V_{cell}} P \cdot A \cdot |F|^2$$

Equation 1.3: Darwin's Formula connecting the spot intensity with influencing factors of the radiation source and the crystal (Darwin, 1914; Holton and Frankel, 2010).

The concept of serial data collection is also applied at synchrotron sources, denoted as SSX, and is conducted by either collecting still images (snapshots) or by partial rotation of several smaller crystals and combining their partial data into a full data set. By analyzing each of the collected serial detector frames individually, the amount of data required for a full data set is higher for serial than for rotational diffraction data collection. When analyzing rotation data, all frames are connected by the specified rotation angle resulting in the requirement of less data.

As X-rays are ionizing radiation, they induce bond breaks or free radical formation, which leads to a decay of the crystal. Serial diffraction methods also reduce the influence of this radiation damage, because the transferred energy is applied to multiple crystals, whose diffractions are combined to obtain a full data set (Martiel et al., 2019; Schlichting, 2015).

Serial diffraction at XFELs require newly designed detectors, equipped for highly intense radiation, short pulses and fast repetition rates, as well as innovative algorithms for data analysis (Johansson et al., 2017). Not only the amount of data produced has increased, also the lack of information on the relative orientation of the single patterns harbor difficulties, for example in the determination of the partiality of each recorded reflection.

The crystal size determines the optimal strategy for data collection. Data from crystals in the upper micrometer range (min. 50 µm) is typically collected by the rotation method (Lindley, 1999). In this case, one single crystal is sufficient for a full data set and is rotated by increments of 0.1 to 1.5° during data collection (Dauter, 1999). The rotation during data collection leads to a higher number of reciprocal lattice points fulfilling Bragg's law. A larger increment (~1°) rotation is applied to generate full reflections, collected in one frame. A smaller increment (0.1 – 0.3°) is applied to collect a full three-dimensional spot profile by collecting the same reflection with different partialities on different frames (fine slicing) (Dauter, 1999; Pflugrath, 1999). However, data collection with small increments requires a certain crystal size in order to rotate a single crystal by 360° without it being affected by radiation damage. Thus, data from smaller crystals in the nanometer to small micrometer range is usually collected serially as described above (Chapman et al., 2011).

1.3 From detector images containing Bragg peaks to a three-dimensional model of the protein

To reduce the data collected in serial diffraction experiments, several programs, like *cctbx.xfel* (Hattne et al., 2014), *nXDS* (Kabsch, 2014) or *CrystFEL* (White et al., 2012), were developed or adapted to suit the needs of serial crystallography. Recent developments focused on the refinement of the detector geometry and of the unit cell parameters before integration and modelling of partial reflections (Johansson et al., 2017).

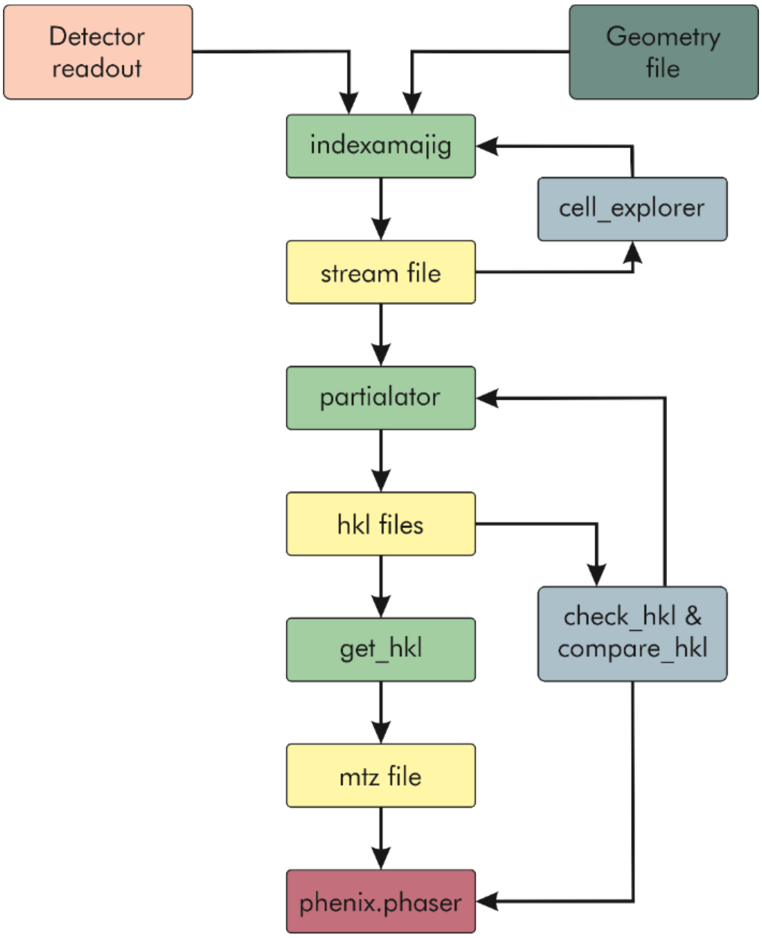


Figure 1.3: Schematic description of the data processing pipeline using CrystFEL. *The detector images together with a file describing the detector geometry are analyzed for Bragg peaks using* indexamajig, *which also indexes and integrates the peaks for a specific crystal lattice. The result is written into a stream file, which can be analyzed by the program* cell_explorer *to generate a histogram of the observed lattice and unit cell parameters.* Patialator *then merges and scales the reflection intensities written in the stream file and outputs three hkl-files, one with all reflections, and two with each harboring a half data set. Figures of merits can be calculated from these files using the scripts* check_hkl *and* compare_hkl, *while* get_hkl *generates an mtz-file, which is required for phasing and model building. In this step the resolution cut-off, defined by the figures of merits, is applied on the obtained data set.*

Diffraction data processing involves identifying, indexing, and integrating Bragg peaks, followed by merging and scaling of the reflections. In this study, serial diffraction data are mainly analyzed using the software suite *CrystFEL* (White, 2019; White et al., 2012). The workflow of *CrystFEL* is illustrated by the flowchart in Figure 1.3. The detector output is stored in an image format such as cbf or hdf5. To understand the geometry of the detector, a geometry file is needed for *CrystFEL* that provides parameters like wavelength, detector distance, beam position, and untrusted regions.

First, Bragg peaks must be identified from the background by one of the various peak detection algorithms. The peak detection algorithm used here is *peakfinder8* originating from *Cheetah* (Barty et al., 2014; White et al., 2012, 2016). Peak-finding algorithms require manual tuning to best detect the peaks for the measured protein crystal. The identified peaks are the basis for indexing by algorithms like *MOSFLM* (Battye et al., 2011), *XDS* (Kabsch, 2010), and *xgandalf* (Gevorkov et al., 2019), which assign the correct reciprocal basis vectors spanning the reciprocal lattice, corresponding to the reflections or, in direct space, the correct unit cell vectors to a given lattice. To identify unknown lattice and unit cell parameters *MOSFLM* (Battye et al., 2011) must be used. The assignment of the unit cells maximizes the symmetry to reduce the amount of data needed to cover the complete reciprocal space. *CrystFEL* then superimposes the expected reciprocal lattice on the detector and, after local background subtraction, integrates all positions where diffraction spots are expected. *Partialator* then models the partialities of the identified reflections to scale and merge all reflections for a complete data set in a given symmetry (White, 2019). *Partialator* provides different partiality models suiting the need of the explicit experiment. To define the high-resolution cut-off, the reflection data, the completeness, the signal-to-noise ratio, and the correlation coefficient of the merged data and the true intensity (CC*) are considered. Here, a minimal completeness of 95 %, a CC* value above 0.5, and a signal-to-noise ratio above 0.5 in the highest shell was used to define the high-resolution cut-off.

To generate an electron density map from the reflection data, phase information must be introduced. While the position and the intensity of Bragg spots can be read directly from the detector, the phase of the diffracted wave is lost. This is known as the phase problem in crystallography. Phases can either be derived from phases of a highly similar (homologous) protein or can be measured experimentally. In this thesis, all structures were solved by molecular replacement using homologous, already known structures to derive the phase information for the data set of interest. The application of these phases to the collected data allows the calculation of electron density maps by Fourier transform. The electron density map provides the basic landscape for model building by guiding the placement of main chain and side chain atoms in the correct position. Model building is performed in successive cycles between automated, global model refinement and manual, local refinement of the structure until the model fits the electron density perfectly, resulting in an estimate of the three-dimensional protein structure.

1.4 Intracellular crystallization of proteins

Protein crystallization is a native process occurring in all kingdoms of life. Protein crystals in living organisms (referred to here as *in vivo* crystallization) can either serve a specific function or be pathogenic. Functional *in vivo* crystallization is a highly regulated process (Schönherr et al., 2018). These intracellular protein crystals can function as protein storage, e.g. in seeds, as a protective mechanism, or as a matrix for solid-state catalysis (Schönherr et al., 2018). Examples include viral capsids, which form a protective, crystal-like shell (Henry and Atchison, 1971), as well as the alcohol oxidase in yeast cells or the uricase in rat hepatocytes, which are stored in a crystalline form (Hruban and Swift, 1964; Jakobi et al., 2016). The alcohol oxidase crystal performs solid-state catalysis, and the cell is protected from potentially harmful intermediates like H_2O_2. In case of insulin, the protein is stored in crystals placed in secretory vesicles prior to release (Dodson and Steiner, 1998). The advantage of crystals is the rapid release of a high concentrations of the molecule and the regulated disassembly. Disease-associated crystallization occurs for example in several forms of cataract, where crystallin proteins form crystals in the lens of the eye (Schönherr et al., 2018). The naturally occurring protein crystallization of some proteins indicates an intrinsic tendency to self-assemble into crystalline structures.

Proteins can be crystallized *in situ* (also referred to as conventional crystallization), experimentally using diffusion-based approaches. This requires the production, isolation, and purification of the protein of interest, followed by a trial-and-error approach to find a condition under which the protein will form crystals of a sufficient size and order. Influencing parameters include the pH, the temperature, the concentration of salts and precipitants. Many of these crystallization approaches rely on the diffusion of a supersaturated protein-precipitant solution into a precipitant reservoir to find the right condition, where crystal nucleation is initiated and subsequently reaches a phase of crystal growth. All steps involved in the protein production to crystallization are not only time consuming and labor intensive, but also do not necessarily result in a protein crystal. However, automated screening procedures to find the best crystallization conditions have largely increased the output (Stevens, 2003).

The crystallization of recombinantly produced proteins in cell culture (referred to here as *in cellulo* crystallization) was described as early as 1996 by Fan et al. However, the structure elucidation of the *in cellulo* crystallized, heterodimeric complex of *Neurospora crassa* calcineurin A and human calcineurin B was not possible at that time due to limitations in radiation brilliance. Since then, many developments in accelerator physics, data collection and processing (see chapter 1.2) have enabled the use of *in cellulo* crystallized proteins for structural studies. Since the discovery of Fan et al. other examples of proteins crystallized in insect cells have been published, like the fusion protein of EGFP and μNS of the avian reovirus, the firefly luciferase (Schönherr et al., 2015), as well as cathepsin B (Redecke et al., 2013), and IMP dehydrogenase (Nass et al., 2020) from the parasite *Trypanosoma brucei*. *In cellulo* crystals have not only been described in insect cells, but also in bacteria (Oeda et al., 1989), and mammalian cells (Hasegawa et al., 2011). These results suggest that intracellular crystallization of proteins can be developed into a powerful method to produce homogeneous, micrometer-sized crystals suitable for structural studies (Schönherr et al., 2018).

Several workflows for crystallization of proteins inside insect cells have been published. The first, published in 2016 by Boudes and coauthors, describes a workflow for protein crystallization in Sf9 insect cells utilizing a baculovirus-based expression system and diffraction data collection in intact crystal-carrying cells. They detect the crystals by bright-field microscopy and enrich crystal-bearing cells by cell sorting, exploiting a change in the light scattering behavior of cells containing crystals, to overcome the efficiency problem. Diffraction data were collected from many single crystals by rotation and subsequent merging of small rotational data sets (Boudes et al., 2016). Tang and co-authors published a similar approach for crystal generation in 2020, focusing on high-throughput by introducing a gateway-compatible cloning system. This offers the possibility of cloning many genes in parallel, combined with different fusion tags suitable for various scientific applications. The use of the BacMagic3 expression systems ensures prolonged survival and higher expression yields by deleting the viral cathepsin, chitinase, p10, p26 and p74 genes. They apply second-harmonic generation (SONICC) techniques as well as transmission electron microscopy (TEM) for crystal detection. This pipeline focusses on high-throughput and includes options for conventional protein purification and crystallization if intracellular crystals are not formed (Tang et al., 2020b).

In the research group of Lars Redecke, another pipeline for *in cellulo* crystallization of proteins in insect cells was developed (Schönherr, 2021). This pipeline, includes a screening approach that targets the protein of interest to different compartments of the cell to find optimal conditions for crystallization, similar to conventional buffer screening. Additional to Boudes and coauthors, Schönherr introduced fluorescence-assisted cell sorting based on the fluorescence of the infection marker EYFP in the DH10EmBacY-based expression system as well as the use of High Five cells for crystal growth (Schönherr, 2021). High Five cells have an increased gene expression compared to Sf9 cells (Schönherr, 2021; Wilde et al., 2014). Improvements established by this pipeline include the steps of titration, virus amplification, infection, and crystal detection, but also the process of crystallization has been studied in detail, revealing the optimal time frame and MOI for best yields (Schönherr, 2021). However, Tang and Schönherr do not cover the data collection, processing, and model building steps in their studies (Schönherr, 2021; Tang et al., 2020b).

A general overview of the workflow developed by Schönherr, the so-called *InCellCryst* pipeline, extended by the structure elucidation part, applied in this thesis is presented in Figure 1.4. At the beginning, the gene of interest is PCR-amplified to enable ligation-dependent cloning. Three cloning approaches are established. First, the cloning of the gene of interest without additional amino acids using *Bam*HI and *Hind*III as restriction enzymes (referred to here as *ori* variants). Second, the cloning system I, which relies on blunt-end cloning using *Ehe*I for vector restriction to minimize the extend of additional amino acids to one at each terminus but allows the encoding of localization signals for different compartments of the cell. The latest approach, cloning system II, uses the more efficient sticky-end cloning by introducing *Kpn*I and *Nhe*I restriction sites that result in the addition of two amino acids at each terminus. For all these approaches, the amplified gene is then cloned into pFastBac1 variants of choice. The vector library contains on the one hand cellular translocation signals for the mitochondria, the endoplasmic reticulum, the peroxisomes as well as the secretory pathway and the nucleus. On the other hand, different tags for purification or antibody recognition can be fused to the protein of interest (Schönherr, 2021). The vectors used in this study are *cyto* (cytosolic localization, additional glycine residue at N-terminus and additional alanine residue at C-terminus from cloning system I), *cyto v2* (cytosolic localization, additional glycine and threonine residues at the N-terminus and additional alanine and serine residues at the C-terminus from cloning system II), *NLS v2* (additional glycine and threonine at the N-terminus and an additional alanine and serine at the C-terminus followed by the nuclear localization signal of human c-myc from cloning system II) and *SS* (signal sequence for the endoplasmic reticulum, additional glycine residue at the N-terminus as well as an additional alanine residue at the C-terminus from cloning system I). A list of all sequences and encoded tags can be found in Schönherr, 2021.

After successful cloning DH10EmBacY cells (Geneva Biotech) are transformed with the pFastBac1 vector encoding the protein of interest to generate a recombinant bacmid via transposition, which is subsequently used to transfect Sf9 insect cells. Sf9 cells produce and amplify the recombinant baculovirus for two passages to enable harvesting of a high titer baculovirus stock. The recombinant baculovirus stock (rbv) is then used to infect High Five insect cells to induce high protein production using the strong, very late polyhedrin promoter hopefully resulting in intracellular crystal growth of the target protein. Intact, crystal containing High Five cells are then prepared for X-ray diffraction data collection, and the three-dimensional structure is elucidated.

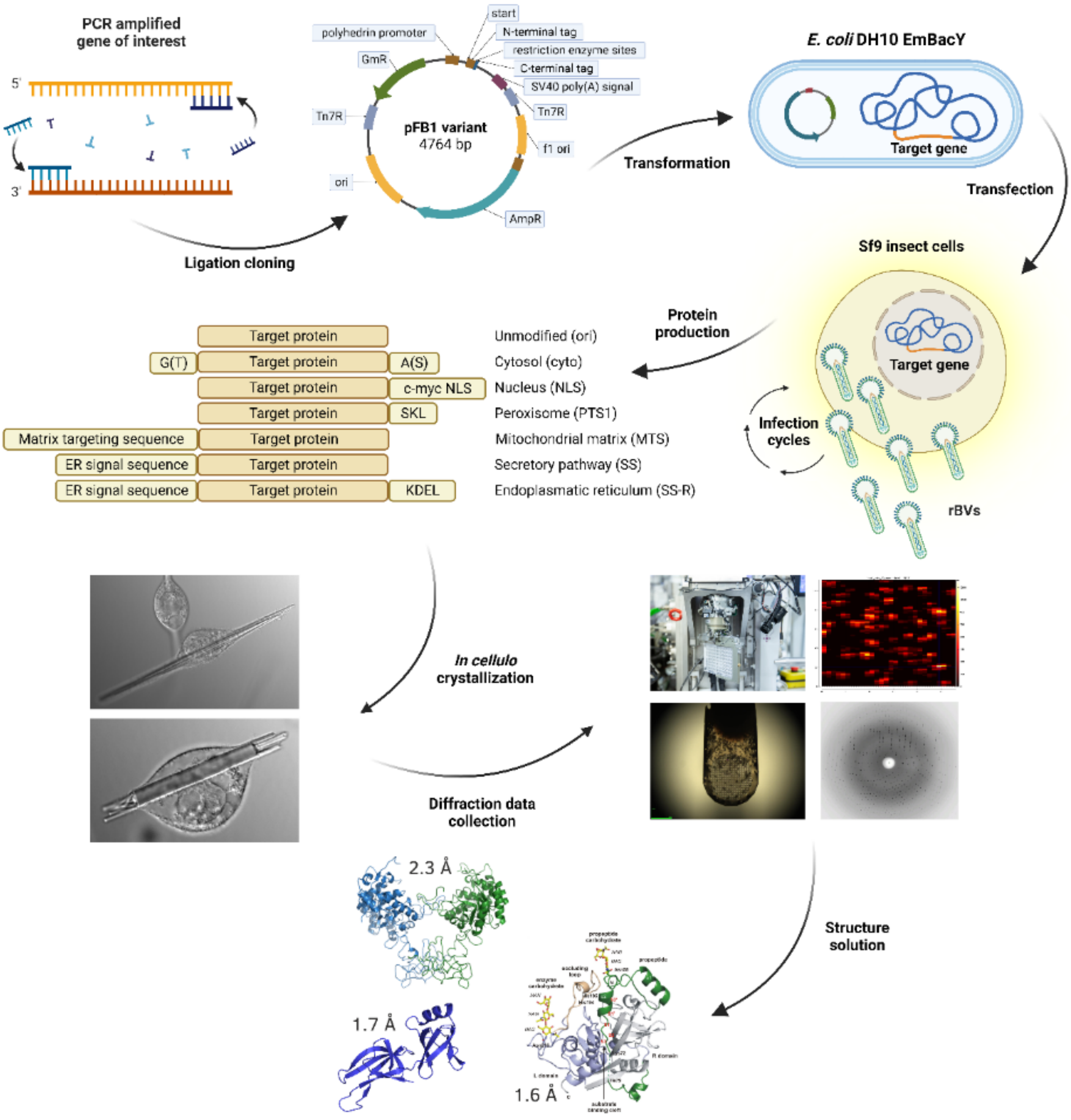


Figure 1.4: InCellCryst *Pipeline for protein crystallization in living insect cells (Schönherr et al., 2023). The gene encoding the protein of interest is amplified by PCR and inserted into a pFastBac1 variant of choice via ligation dependent cloning. The vector variants include for example tags which direct the protein of interest into a specific cellular compartment or enables antibody recognition.* E. coli *DH10EmBacY cells are then transformed by the cloned vector to generate a recombinant bacmid by transposition. Then recombinant baculoviruses (rbvs) are generated after transfection of Sf9 insect cells with the recombinant bacmid and amplified by 2 passages of virus amplification. Then High Five insect cells are infected with these rbvs. If spontaneous growth of ordered structures can microscopically be observed, X-ray diffraction data can be collected from crystals in intact, rbv-infected High Five cells to enable structure solution.*

For diffraction data collection, third-generation synchrotron sources or X-ray free-electron lasers (XFELs) can be used (Boudes et al., 2016; Lahey-Rudolph et al., 2021). For synchrotron diffraction, cells are either loaded on a MeshMount and flash-frozen in liquid nitrogen for cryo-crystallography (Boudes et al., 2016) or measured at room temperature using CrystalDirect™ plates or different chips which place the crystal-containing cells or isolated crystals in the beam (Norton-Baker et al., 2021). For XFELs, fixed-targets such as silicon microchips are preferred in terms of hit rates over liquid- or viscous-jets that spray the intact cells or isolated crystals into the beam (Lahey-Rudolph et al., 2021). The data collection strategy is chosen according to the application. MeshMounts are mounted on a goniometer and can therefore be rotated, enabling the collection of small rotational data sets, while the targeted rotation/oscillation of a crystal in a liquid- or viscous-jet is not possible. Then still images of the crystal are collected, requiring the use of recently developed software like *CrystFEL*, as no information on the crystal orientation can be derived from rotation of the crystal. If data is collected serially and one crystal is measured during rotation, conventional software like *XDS* can be applied to integrate data sets from multiple crystals which can then be merged. After obtaining the merged and scaled reflection file the model building for rotational and still images is comparable. To sum this up, streamlined approaches facilitate the accessibility of *in cellulo* crystallization for new users and results in homogeneously sized microcrystals suitable for structural studies.

In cellulo crystallization not only reveals the three-dimensional structure of the protein studied, but also allows the identification of native cofactors in the natural environment (Nass et al., 2020). The octameric structure of *Trypanosoma brucei* IMP dehydrogenase (*Tb*IMPDH) was crystallized in the inactive state in living Sf9 insect cells. The crystals were isolated, and their diffraction data later collected at an XFEL source. Without artificially added cofactors or soaking approaches, the structure revealed an adenosine triphosphate in the canonical binding site I and a guanosine monophosphate in the canonical binding site II, indicating these as the native regulators bound to IMPDH in the inactive state (Nass et al., 2020). The identification of native cofactors is unique to *in cellulo* crystallization making it a valuable tool for understanding the mechanisms of enzyme function and activity.

1.5 Model proteins

1.5.1 *Neurospora crassa* HEX-1 (NcHEX-1), a natively crystallizing model protein

Woronin bodies consist of hexagonally crystalline HEX-1 protein (Tang et al., 2020b; Woronin, 1864). This is a good example of native and functional intracellular crystallization. Woronin bodies are specialized organelles derived from peroxisomes in filamentous fungi of the genus *Ascomycota*. *Ascomycota* form multicellular syncytia which are connected by septal pores. When the hyphae are either physically damaged or under osmotic stress, the Woronin bodies act as a plug to prevent the loss of cellular contents like cytosol or organelles from neighboring cells (Maruyama et al., 2005). This process is not the result of uncontrolled bleeding and clogging of the Woronin bodies at the septal pore, but is tightly regulated and controlled instead (Beck et al., 2013). The difference between the Woronin body and an organelle is the dense packing of the contents. While an organelle can be deformed and thus be sucked through the septal pore, the HEX-1 crystal inside the Woronin body cannot be deformed and is therefore able to fulfill its plugging function. Woronin bodies are also involved in maintaining the heterogeneity of adjacent hyphal cells (Tang et al., 2020b).

The crystallization of *Nc*HEX-1 *in situ* after recombinant gene expression in *E. coli* has already been described, and the structure was solved by Yuan et al. (2003). Its spontaneous crystallization inside living insect cells after application of the *InCellCryst* pipeline has already been published (Lahey-Rudolph et al., 2020, 2021; Schönherr, 2021). *Nc*HEX-1 crystallized in almost all compartments tested, including the cytosol, the nucleus, the endoplasmic reticulum (ER), and the mitochondria (Schönherr, 2021). Depending on the localization of the protein, the crystal shape varied from spindle-like to cubic, both with hexagonal cross-section (Figure 1.5). Bipyramidal shapes were also observed (Schönherr, 2021). The structure of HEX-1 cyto, which was crystallized and its diffraction data collected at an XFEL *in cellulo*, was elucidated by Lahey-Rudolph et al. in 2021 and shows a ß-sheet-rich fold like the *in situ* crystallized protein,, almost superimposable with the conventionally crystallized HEX-1 structure (Yuan et al., 2003).

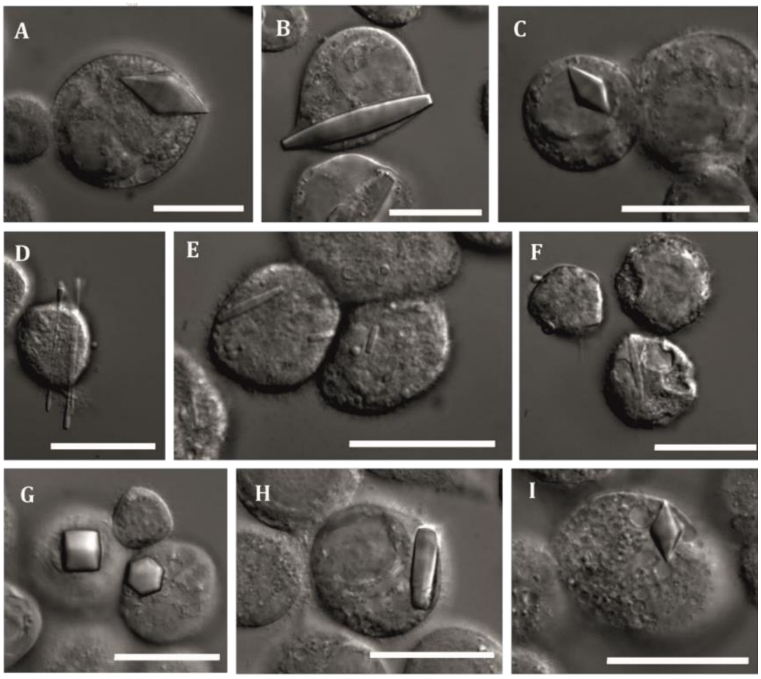


Figure 1.5: Microscopic images of High Five cells infected by different rbvs encoding **Nc*HEX-1*. *The cells display various crystal shapes and morphologies. A: Infection with rbv* Nc*HEX-1 NLS v2 reveals the growth of bipyramidal-shaped crystals, while after rbv* Nc*HEX-1 cyto infection two different morphologies of the crystals are observed (B and C), namely a spindle-like form in B and the bipyramidal-shaped morphology in C. D-F:* Nc*HEX-1 SS crystallizes either into a needle-shape (D), a rectangular shape (E) or into a small spindle-like morphology (F). G: The expression of* Nc*HEX-1 ori, without any additional amino acids added as a consequence of cloning, cubic crystals with a hexagonal cross-section are formed. H & I: After infection with rbv* Nc*HEX-1 cyto v2 two morphologies are observed, a rectangular shape with hexagonal cross-section (H) and a bipyramidal shape (I). Modified from Franke, 2021.***

1.5.2 *Trypanosoma brucei* guanosine-monophosphate reductase (*Tb*GMPR) and Inosine-5′-monophosphate dehydrogenase (*Tb*IMPDH) as a recombinantly expressed, *in cellulo* crystallizing model protein

The African sleeping sickness, caused by *Trypanosoma brucei*, is a fatal yet neglected disease in sub-Saharan Africa. The protozoan parasite is inoculated into the bloodstream of the host by the bite of the tsetse fly and develops a haemolymphatic stage followed by a meningoencephalitic stage. Depending on the subspecies of *T. brucei*, humans or other hosts may be chronically or acute infected. If left untreated or treated inappropriately regarding the specific disease stage, African sleeping sickness is fatal in almost all cases. Unfortunately, only early- and late-stage treatments are currently available. The late-stage drugs are highly toxic, with a risk of developing a post-treatment reactive encephalopathy, resulting in a fatality rate of 6% for late-stage treatment. Although several new drugs have been developed and are currently under evaluation, further understanding of the

Trypanosoma life cycle and structural investigation of potential drug targets is still mandatory to enable rational design approaches (Kennedy, 2019).

Inosine-5′-monophosphate dehydrogenase (*Tb*IMPDH)

*Tb*IMPDH is involved in the purine metabolism and catalyzes the oxidation of IMP to XMP by reduction of NAD^+ to NADH (Wang and Hedstrom, 1997). This reaction is the rate limiting step in the conversion of IMP to GMP, which is followed by the reaction of the GMP synthase that subsequently catalyzes the amination of XMP to GMP (Patton et al., 2011). Its activity controls the intracellular levels of the purines and thus plays a key role in cell proliferation and replication (Jayaram et al., 2022), making this enzyme a suitable drug target for the treatment of African sleeping sickness.

*Tb*IMPDH was the third structure solved using intracellularly grown protein crystals in the group of Redecke (Nass et al., 2020). The 55.7 kDa monomer forms a biologically active octamer by the formation of two ring-like tetramers. *Tb*IMPDH cyto forms needle-shaped crystals of about 60 µm in length that exceed the cell diameter in around 60 to 80 % of High Five cells in a culture infected at an MOI of one (Schönherr, 2021). The structure of *Tb*IMPDH N-His revealed a dimer of two monomers in the asymmetric unit that is formed by interaction of the Bateman and finger domains, displaying the interaction of the two tetrameric rings, while the intra-tetramer contacts are stabilized by interactions of the catalytic domains (Nass et al., 2020). The diffraction data collection of isolated *in cellulo* crystals revealed co-factors bound to the allosteric binding sites, namely an ATP that is bound at the first canonical site and a GMP that is bound at the second canonical binding site. This study suggests that natively bound co-factors can be identified by crystallizing the protein in its native environment without the need of soaking or engineering of the protein (Nass et al., 2020).

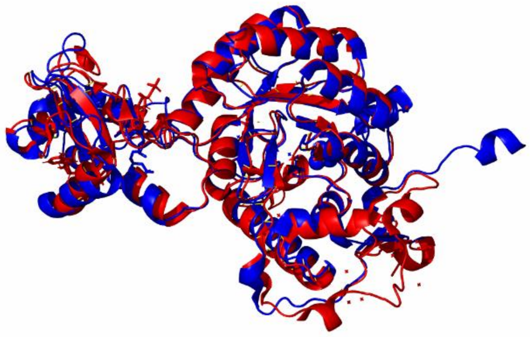


***Figure 1.6: Overlay of the structural models of conventional crystallized* Tb*GMPR (pdb: 6lk4) depicted in blue and in cellulo crystallized* Tb*IMPDH (pdb: 6rfu) depicted in red both in cartoon representation. Despite a low sequence identity of 32 % both structural models display a high structural conservation indicated by 85 % of matched secondary structural elements as calculated by* PDBeFold.**

Guanosine-monophosphate reductase (*Tb*GMPR)

As the already described for *Tb*IMPDH, also *Tb*GMPR is involved in the purine salvage pathway catalyzing the reduction of GMP to IMP while NADPH is oxidized to NADP$^+$ (Imamura et al., 2020). *Tb*GMPR is a structural homolog of *Tb*IMPDH (Imamura et al., 2020; Nass et al., 2020) as displayed by a structural superposition in Figure 1.6. The sequence identity is 32,4 %, while the sequence similarity is 52.2 % (calculated by Ident and Sim of the Sequence Manipulation Suite). GMPR catalyzes the reverse reaction of IMPDH and is also regulated oppositely (Patton et al., 2011). The GMPR of trypanosomatids like Trypanosoma and Leishmania are the only GMPRs known so far, which encode not only the catalytic but also the ~130 aa long Bateman domain (Bessho et al., 2016; Smith et al., 2016). The Bateman domain is usually part of IMPDHs and enables an additional level of regulation (Bessho et al., 2016) indicating the purine biosynthesis to be crucial for the survival of these parasites.

The structure of *in situ* crystallized *Tb*GMPR recombinantly expressed in *E. coli* was already elucidated by Imamura et al. in 2020. The structure was solved in the apo form as well as in complex with GMP and GTP. *Tb*GMPR either forms a tetrameric structure or an octameric structure depending on its activity and regulating co-factors bound (Imamura et al., 2020). The overall composition and interaction of the subunits of the octameric structure are similar to those described above for *Tb*IMPDH (Imamura et al., 2020; Nass et al., 2020). Nevertheless, Imamura et al. showed a twisted octameric structure of GMP- and GTP-bound GMPR indicating its active state and a relaxed structure when no guanosine nucleotide is present. GMPR cyto v2 also forms needle-shaped crystals of up to 140 µm length, starting to grow 44 hours post infection (hpi) in recombinant baculovirus (rbv) infected High Five cells (Kardoost et al., 2023; Schönherr, 2021). So far, neither the structure of *in cellulo* crystallized GMPR nor the native cofactors could be elucidated.

1.6 Challenges in intracellular protein crystallization

In cellulo crystallization is facing some limitations. First, not all proteins successfully produced inside the cells form crystals. Secondly, sometimes crystallization efficiency, i.e. the number of crystal-containing cells within a cell culture, meaning the number of cells, in which conditions for nucleation and crystal growth were successfully met, is too low to enable the collection of a full diffraction data set in due time. Thirdly, sometimes the crystals and their diffractive volume are too small, resulting in low resolution diffraction.

Like in conventional crystallization approaches not all proteins crystallize spontaneously inside the cell. Even though every protein has a natural tendency to crystallize, evolution has managed to decrease the self-assembly of proteins, which should not crystallize in the living organism/cell to prevent cellular damage, to a minimum. But this in turn also prevents its crystallization in experimental studies (Doye and Poon, 2006). As described before, conventional approaches rely on testing different buffer conditions to find suitable conditions for successful crystallization (Rupp, 2010). In intracellular crystallization the different milieus of various compartments of the cell are exploited to find an environment that supports crystal growth (Schönherr, 2021). The amount of protein present in the cell was identified as critical parameter driving the crystallization of proteins (Koiwai et al., 2019; Schönherr, 2021). This is in accordance with conventional approaches where the rate-limiting step is also the nucleation, which requires high levels of supersaturation (Bergfors, 2003). If nucleation has occurred, the crystals grow when the metastable phase was reached (Bergfors, 2003).

Whereas for many proteins indications for intracellular crystallization have been observed, the overall very low efficiency frequently prevents diffraction data collection (Koiwai et al., 2019). However, enrichment of crystal-containing cells by cell sorting is limited by the lack of survival of sorted cells (Schönherr, 2021). Human neuraminidase 1 crystallized in HEK cells reveal a crystallization efficiency of 3.4 % (Koiwai et al., 2019). The crystal concentration per volume and the hit rate could be improved by crystal isolation and concentration (Koiwai et al., 2019). The automated crystal identification in living cells established by Kardoost et al. (2023) is a step forward to overcome the detection limits for proteins that crystallize with a too low efficiency and may be a starting point for cell sorting as a prerequisite to collect data of these systems in due time. Even though upscaling is comparably easy, if a protein does not crystallize spontaneously, driving it to a crystalline state in the context of the living cells is hard to obtain. An approach of adding crowding agents like nucleotides or small RNA molecules, acting as precipitants *in cellulo,* is discussed to improve crystallization of the target protein (Mudogo et al., 2020). Cell reactive compounds, like Brefeldin A which inhibits the protein secretion, showed positive effects at least on endoplasmic protein crystals in insect and mammalian cells (Harms, 2019; Hasegawa et al., 2014).

Like in conventional crystallization, not all *in cellulo* crystallized proteins, diffract X-rays to high resolution. The higher the periodicity and homogeneity inside the crystal, the clearer and stronger the Bragg peaks resulting from its scattering function (Rupp, 2010). Furthermore, the more copies of the unit cell correspond to the scattering function, the better the signal to noise ratio (Rupp, 2010). Hence, the better the order and/or the larger the crystal, the better the intensity of the Bragg reflections. For proteins like GFP-µNS, intracellular crystallization is described, however structure elucidation is prevented by its low-resolution diffraction (Nagaratnam et al., 2020; Schönherr et al., 2015). The application of XFEL radiation of higher brilliance and intensity than synchrotron radiation is one possible solution to improve the signal to noise ratio, however structure elucidation was still not possible (Nagaratnam et al., 2020). Access to XFELs is restricted and rare, thus this approach is not suitable for high-throughput applications. Up to now no other option to improve the resolution of *in cellulo* crystals is published.

Finally, the data collection and processing of the *InCellCryst* pipeline described by Schönherr is not implemented yet. Even if suitable software is available for processing serial synchrotron data, the data collection inside living cells still requires further adjustments. Problems occur e.g., during the loading of intact, crystal-carrying cells where salts and other metabolites are present, which can crystallize during the sample preparation procedure and result in additional Bragg diffraction overlaying the diffraction of the protein.

To summarize, the broad application of *in cellulo* crystallization as an additional approach to crystallize proteins is still limited by some challenges. Improvements are needed in the data collection and processing stage next to solutions for difficult samples, that do not diffract well, crystallize rarely or in a diffractive volume too small for structure elucidation. If these limitations are overcome, the intracellular crystallization approach can be implemented as supplement to other structure biological methods by a broader user community due to its streamlined workflow.

2. Objective

Due to advances in the accelerator physics and the diffraction data collection and processing, intracellular crystallized proteins can be successfully used for X-ray crystallography (Gati et al., 2014; Redecke et al., 2013). The quasi-native environment of the cell provides the full range of regulating molecules and substrates, allowing co-crystallization and identification of the naturally bound cofactors without the selection bias of conventional co-crystallization experiments (Nass et al., 2020).

The aim of this study is to implement and optimize a strategy to process diffraction data from crystals grown in living cells as well as to overcome common challenges of intracellular crystallography. Namely, the statistical problem of finding a suitable candidate for *in cellulo* crystallization as well as cellular characteristics of the producing cell limiting crystal size, diffractive volume, and the achievable resolution of the protein structure that is to be elucidated.

Two approaches are developed to optimize diffraction data processing. Firstly, data processing using a semi-serial approach, relying on processing many small rotational data sets using the software *XDS*, which are subsequently merged into one complete data set using *XSCALE*. And secondly, the optimized application of a software for serial snap shots, namely *CrystFEL*.

Additionally, the influence of the amino acids added by the cloning systems developed by Schönherr (2021) on the structure of *Nc*HEX-1 is to be investigated. For this purpose, three variants of *Nc*HEX-1 are subjected to structure elucidation and model building.

To overcome the statistical problem of finding a protein that crystallizes *in cellulo*, homologs of proteins, that are already crystallized in living insect cells are tested for their ability to crystallize. For this approach not only orthologs, i.e., homologs from different species encoding a protein with the same function, but also a paralog pair, i.e., homologs in one species resulting from gene duplications with different functions, are selected and recombinantly expressed using the pipeline developed by Schönherr (2021).

Structure elucidation of different homologs of the Woronin body major protein is interesting especially because there is only limited literature available concerning the structures of various HEX-1 proteins other than *Nc*HEX-1. The structural comparison of multiple, natively crystallizing HEX-1 proteins may lead to a deeper understanding of the strong, intrinsic crystallization tendency and the identification of essential areas involved in the crystal formation.

The purine salvage pathway is a good target for drug development against African Sleeping Sickness. Especially treatments against the *Tb*GMPR, a paralog of *Tb*IMPDH, may be effective. This protein encodes a Bateman domain, which is unique to Leishmania and other Trypanosomatids, and displays an additional regulatory element known from IMPDHs. The crystallization and structure solution of *Tb*GMPR will deepen the knowledge on its regulation and natively bound cofactors to develop effective drugs against Trypanosomiasis.

Cell fusion is used to increase the cell size and to combine the protein production capabilities of the individual cells to overcome resolution limits caused by a limited diffractive volume of the crystals. A strategy for PEG-mediated cell fusion will be established and the optimal conditions for cell fusion will be investigated to maximize the diffractive volume of different model protein crystals. The ability to increase the crystal volume is a milestone in intracellular crystallography because it opens a way to collect diffraction data from the many proteins which form small or bad diffracting crystals in cells.

Overcoming these challenges can open up *in cellulo* crystallography to a wide community of users, further consolidating its importance in structural biology by exploring the natively bound cofactors of the respective protein, which cannot be conducted in the same way by KI-based structure prediction software or conventional soaking experiments.

3. Material

Manufacturers of materials used in this study are listed below together with their place of incorporation. In the following only the manufacturer is named.

BDH Laboratory Supplies (Poole, UK), **Becton Dickinson** (Franklin Lakes, NJ, USA), **BIO-RAD** (Hercules, CA, USA), **Bioair Instruments** (Siziano, I), **Biolegend** (San Diego, CA, USA), **Biozym Scientific GmbH** (Hessisch-Oldendorf, DE), **Brunel Microscopes Ltd.** (Chippenham, UK), **Carl Roth GmbH** (Karlsruhe, DE), **Carl Zeiss AG** (Oberkochen, DE), **Consort BVBA** (Turnhout, BE), **Corning Inc.** (Coning, NY, USA), **Eppendorf AG** (Hamburg, DE), **Expression Systems LLC** (Davis, CA, USA), **Fluka Chemie GmbH** (Buchs, CH), **GE Healthcare Life Sciences** (Marlborough, MA, USA), **Gelaire Ltd.** (Sydney, AUS), **Gerbu Biotechnik GmbH** (Heidelberg, DE), **Greiner Bio-One** (Frickenhausen, DE), **Hassa Laborbedarf** (Lübeck, DE), **Heraeus Holding GmbH** (Hanau, DE), **Andreas Hettich GmbH & Co. KG** (Tuttlingen, DE), **highQu GmbH** (Kraichtal, DE), **Invitrogen** (Carlsbad, CA, USA), **LAUDA DR. R. WOBSER GmbH & Co. KG** (Lauda-Königshofen, DE), **Leica Microsystems GmbH** (Wetzlar, DE), **LGC Genomics** (Berlin, DE), **Marienfeld Superior** (Lauda-Königshofen, DE), **Merck** (Darmstadt, DE), **Metabion GmbH** (Planegg/Steinkirchen, DE), **Millipore** (Billerica, MA, USA), **MiTeGen LLC.** (Ithaca, NY, USA), **Nalge Nunc Int.** (Rochester, NY, USA), **NET New Electronic Technology GmbH** (Finning, DE), **New Brunswick Scientific** (Enfield, CT, USA), **New England Biolabs** (Ipswich, MA, USA), **Nikon Instruments** (Melville, NY, USA), **NIPPON Genetics Europe** (Düren, DE), **Promega** (Madison, WI, USA), **Sarstedt** (Nürmbrecht, DE), **Sartorius AG** (Göttingen, DE), **SERVA Electrophoresis** (Heidelberg, DE), **Sigma-Aldrich** (St. Louis, MO, USA), **Starlab** (Hamburg, DE), **Techne** (Cambridge, UK), **Thermo Fisher Scientific** (Waltham, MA, USA), **VWR International** (Erlangen, DE)

3.1 Laboratory devices

Table 3.1: List of all used laboratory devices by specification of the exact notation and the manufacturer.

Device	Article description	Manufacturer
Agarose gel chamber	Easy Phor Mini or Midi	Biozym Scientific
Power supply	EV 1450	Consort BVBA
Agarose gel documentation		
Transilluminator	Blue/Green LED Transilluminator XL	Nippon Genetics
Camera	FOculus IEE1394	NET New Electronic Technology GmbH
Centrifuge for 5- or 15-ml tubes	Rotina 380	Andreas Hettich
Centrifuge for 1.5- or 2-mL tubes	Centrifuge 5424	Eppendorf
Centrifuge for 1.5- or 2-mL tubes	Heraeus Pico 21 Centrifuge	Thermo Fisher Scientific
Freezing Container	Nalgene Cryo 1 °C Freezing Container	Nalge Nunc Int.
Fluorescence microscope	Nikon Eclipse Ts2R-FL	Nikon Instruments
Camera	Nikon Qi2	Nikon Instruments
Light source	505 nm LED unit	Nikon Instruments
Software	NIS Elements BR 5.2	Nikon Instruments
Light microscopy	DM IL LED	Leica Microsystems

Device	Article description	Manufacturer
Camera	MC 170 HD	Leica Microsystems
Neubauer counting chamber	Improved Neubauer	Marienfeld Superior
Microscope for MeshMounts	Axioskop 20	Carl Zeiss
Pipets	Ergo One Eppendorf Research Plus	Starlab Eppendorf
Shaking incubator	Innova 4230	New Brunswick Scientific
Sterile hood	MSC-Advantage	Thermo Fisher Scientific
Spectrophotometer	Nanodrop 2000c Nanodrop One^{C} Ultrospec 3100 pro	Thermo Fisher Scientific Thermo Fisher Scientific GE Healthcare Life Sciences
Thermocycler	Techne Genius	Techne
Thermomixer	Mixing Block MB-102	Biozym
Water bath	Hydro H8	LAUDA Dr. R. Wobser

3.2 Consumables

Table 3.2: List of the used consumables by specification of the manufacturer.

Article	Manufacturer
Cover slip Ø 25 mm	VWR
Pipette tips	Starlab Sarstedt
Cryo-Tubes 2 mL	Hassa Laborbedarf
Micro screw-cap tubes 2ml, Caps	Sarstedt, Biozym Scientific
Multi Well plates	Greiner Bio-One
Reaction tubes: 0,2 mL, 0,5 mL, 1,5 mL und 2 mL	Sarstedt
Reusable Goniometer Base B1A	MiTeGen
Tubes, 15 mL, 50 mL	Sarstedt
Sterile filters, 0,2 µm	Sartorius
Cell culture flasks 25 cm^2, 75cm^3 for suspension cultures	Sarstedt

3.3 Enzymes, markers, and commercially available kits

Table 3.3: List of used enzymes, markers, and commercially available kits by specifying the manufacturer.

Name	Manufacturer
ALLin HiFi DNA polymerase	highQu
DpnI	New England Biolabs
ESCORT IV Transfection Reagent	Sigma-Aldrich
FastAP Thermosensitive Alkaline Phosphatase	Thermo Fisher Scientific
FastDigest *Bam*HI	Thermo Fisher Scientific
FastDigest *Hind*III	Thermo Fisher Scientific
FastDigest *Kpn*I	Thermo Fisher Scientific
FastDigest *Nhe*I	Thermo Fisher Scientific
GeneJET Gel Extraction Kit	Thermo Fisher Scientific
GeneJET PCR Purification Kit	Thermo Fisher Scientific
GeneJET Plasmid Miniprep Kit	Thermo Fisher Scientific
GeneRuler DNA Ladder Mix	Thermo Fisher Scientific
T4 DNA ligase	Thermo Fisher Scientific

3.4 Chemicals

Table 3.4: ***List of used chemicals by specifying the source of supply.***

Name	Source of supply
10x FastDigest Buffer	Thermo Fisher Scientific
10x FastDigest Green Buffer	Thermo Fisher Scientific
100x Anti-Anti	Sigma-Aldrich
5-Brom-4-chlor-3-indoxyl-β-D-Galactoside (X-Gal)	Carl Roth
6x TriTrack DNA Loading Dye	Thermo Fisher Scientific
Agar	Becton Dickinson
Agarose NEEO Ultra-Qualität	Carl Roth
Ampicillin	Sigma-Aldrich
ALLin Buffer (5x)	HighQu
Calcium chloride ($CaCl_2$)	Sigma-Aldrich
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich
Ethanol 100 % denatured	Merck
Ethanol 100 %	Carl Roth
Gentamycin	Gerbu Biotechnik
Glucose	Carl Roth
Glycerol	Sigma-Aldrich
Glycine	Sigma-Aldrich
Yeast extract	Becton Dickinson
Immersion oil Type N	Nikon
Insect cell medium ESF921	Expression Systems
Isopropyl-β-D-thiogalactopyranoside (IPTG)	Gerbu Biotechnik
Kanamycin sulfate	Sigma-Aldrich
Potassium chloride (KCl)	BDH Laboratory Supplies
Magnesium chloride ($MgCl_2$)	Sigma-Aldrich
Magnesium sulfate ($MgSO_4$)	Fluka Chemie
Midori Green Advance	NIPPON Genetics Europe
Sodium chloride (NaCl)	Sigma-Aldrich
PEG200	Sigma-Aldrich
PEG400	Sigma-Aldrich
PEG4000	SERVA Electrophoresis
Penicillin/Streptomycin (100x konz.)	Sigma-Aldrich
Tetracycline	SERVA Electrophoresis
Tris Base	Promega
Triton-X 100	Carl Roth
Tryptone	Becton Dickinson
Tween-20	SERVA Electrophoresis
T4 DNA Ligase Buffer	Thermo Fisher Scientific
YT medium	Carl Roth

3.5 Buffers and solutions

Table 3.5: ***List of used buffers and solutions by specification of their composition.***

Name	Composition
Bacmid-Agar	TY medium with 1,5 % agar, 50 µg/mL kanamycin, 7 µg/mL gentamycin, 10 µg/mL tetracycline, 100 µg/mL X-Gal, 40 µg/mL IPTG
$CaCl_2$ solution	0,1 M $CaCl_2$, 10 % glycerol

Name	Composition
YT medium	5 g/L yeast extract, 10 g/L tryptone, 0,5 g/L NaCl, pH 7,0
YT bacmid medium	5 g/L yeast extract, 10 g/L tryptone, 5 g/L NaCl, pH 7,0, 50 μg/mL kanamycin, 7 μg/mL gentamycin, 10 μg/mL tetracycline
Insect cell culture medium (1)	ESF921 medium, 100 U/mL penicillin, 100 μg/mL streptomycin
Insect cell culture medium (2)	ESF921 medium, 1x Anti-Anti
Modified TAE buffer (50x)	2 M Tris-Acetate, 5 mM Na_2EDTA, pH 8,0
PBS (10x)	137 mM NaCl, 2,7 mM KCl, 10 mM Na_2HPO_4, 1,8 mM KH_2PO_4, pH 7,4
SOB medium	YT medium, 20 mM $MgCl_2$, 20 mM $MgSO_4$
SOC medium	YT medium, 20 mM $MgCl_2$, 20 mM $MgSO_4$, 25 mM glucose

3.6 Cell lines and bacterial strains

Table 3.6: List of used bacterial strains by specification of genotype and source of supply.

Name	Source of supply
E. coli DH5α	New England Biolabs
E. coli HB101	Institute of Virology and Cell Biology, University of Lübeck
E. coli DH10Bac	Invitrogen
E. coli DH10EmBacY	Institute of Physics, University of Lübeck

Table 3.7: List of used used cell lines by specification of their source of supply.

Name	Organism	Description	Source of supply
Sf9	*Spodoptera frugiperda*	Clonal isolate of Sf21 cell line, derived from pupa ovarian tissue of the fall armyworm	Institute of Virology and Cell Biology, University of Lübeck
High Five (BTI-Tn-5B1-4)	*Trichoplusia ni*	Clonal isolate from ovarian tissue of the cabbage looper	Institute of Virology and Cell Biology, University of Lübeck

3.7 Plasmids

Table 3.8: List of used plasmids with a short description, supplied by Prof. Dr. Lars Redecke (University of Lübeck).

Plasmid	Description
pFastBac1 (pFB1)	Plasmid for expression of recombinant genes in insect cells using the transposition-based Bac-To-Bac system.
pFB1 IMPDH ori	pFastBac1 plasmid encoding IMP dehydrogenase of *Trypanosoma brucei*
pFB1 GMPR cyto v2	pFastBac1 plasmid encoding GMP reductase of *Trypanosoma brucei* modified with Glycine and Threonine (N-terminal) and Alanine and Serine (C-terminal)

3.8 Oligonucleotides

Table 3.9: List of oligonucleotides and their sequenced used, supplied by metabion.

Name	sequence (5'→3')
*Af*HEX-1delN fwd	GATCGGATCCATGGGTTACTATGACGATGAC
*Af*HEX-1 rev	GATCAAGCTTTTACAGACGGGAACCG
*Mr*HEX-1delN fwd	GATCGGATCCATGGGTTACTATGATGAGGAC
*Mr*HEX-1 rev	GATCAAGCTTTTACAGGCGAGAGCCG
*Pl*HEX-1delN fwd	GATCGGATCCATGGGTTACTATGACGAGG
*Pl*HEX-1 rev	GATCAAGCTTTTACAGGCGCGAGC
*Pt*HEX-1delN fwd	GATCGGATCCATGGGCTATTATGACAATGAG
*Pt*HEX-1 rev	GATCAAGCTTCTACAAGCGGGAGCC
pFB1 fwd seq v3	GTTGGCTACGTATACTCCGGA
pFB1 rev seq v2	ACAAACCACAACTAGAATGCAGTG
pUCM13 fwd	CCCAGTCACGACGTTGTAAAACG
pUCM13 rev	AGCGGATAACAATTTCACACAGG

3.9 Recombinant baculoviruses

Table 3.10: List of all recombinant baculoviruses used. They were all supplied by Prof. Dr. Lars Redecke (University of Lübeck).

Name	Description
*Af*HEX-1	Produces *Af*HEX-1 (UniProt: A0A0J5PW54) of Aspergillus fumigatus
*An*HEX-1	Produces *An*HEX-1 (UniProt: Q9P8K9) of *Aspergillus nidulans*
*Ao*HEX-1	Produces *Ao*HEX-1 (UniProt: I8TQ26) of *Aspergillus oryzae*
*Cg*HEX-1	Produces *Cg*HEX-1 (UniProt: Q2H4T3) of *Chaetomium globosum*
*Mg*HEX-1	Produces *Mg*HEX-1 (UniProt: Q9UW16) of *Magnaporthe grisea*
*Mr*HEX-1	Produces *Mr*HEX-1 (UniProt:E9EN82) of *Metarhizium robertsii*
*Nc*HEX-1	Produces *Nc*HEX-1 (UniProt: P87252) of *Neurospora crassa*
*Nc*HEX-1 cyto	Produces *Nc*HEX-1 (UniProt: P87252) of *Neurospora crassa* elongated by an Alanine residue at C-terminus
*Nc*HEX-1 cyto v2	Produces *Nc*HEX-1 (UniProt: P87252) of *Neurospora crassa* N-terminal elongated by Glycine and Threonine and C-terminal by Alanine and Serine
*Pl*HEX-1	Produces *Pl*HEX-1 (UniProt: A0A179H7Q1) of *Purpureocilium lilacinum*
*Pt*HEX-1	Produces *Pt*HEX-1 (UniProt: B2VT43) of *Pyrenophora tritici-repentis*
*Vd*HEX-1	Produces *Vd*HEX-1 (UniProt: G2WVW3) of *Verticillium dahliae*
*Vg*HEX-1	Produces *Vd*HEX-1 (UniProt: A0A0D2A1I5) of *Verruconis gallopava*

3.10 Programs

Table 3.11: List of programs used.

Program	Developer
Adxv 1.9.14	https://www.scripps.edu/tainer/arvai/adxv.html
benchling	Benchling
blastp	(Johnson et al., 2008)
ccp4i 7.1.015	(Winn et al., 2011)
Clustal Omega 1.2.4	(McWilliam et al., 2013)
CrystFEL 9.0, 9.1, 10.0, 10.1,10.2	(White, 2019; White et al., 2012, 2013)
Coot (Linux) 0.9.5	(Emsley et al., 2010)

Program	Developer
Coot (Windows) 0.8.9.2	(Lohkamp et al., 2005)
CorelDRAW 21.0.0.638	Corel GmbH
FiJi/ImageJ 1.52b	(Schindelin et al., 2012)
GraphPad Prism 9	GraphPad Software
Ident and Sim	(Stothard, 2000)
Ligplot v2.2.5	(Wallace et al., 1995)
MXCuBE 2 & 3	(Oscarsson et al., 2019)
NIS Elements BR 5.2	Nikon Instruments
PDBeFold	(Krissinel and Henrick, 2005)
PDBePISA	(Krissinel and Henrick, 2007)
Phenix 1.19.2-4158, 1.20.1-4487	(Liebschner et al., 2019)
Pymol	Schrödinger Inc.
XDS, XSCALE, XDSCONV	(Kabsch, 2010)
WebLogo3	(Crooks et al., 2004; Schneider and Stephens, 1990)
Zanuda	(Lebedev and Isupov, 2014)

4. Methods

4.1 Molecular biological methods

4.1.1 Polymerase chain reaction (PCR) using ALLin HiFi DNA polymerase

Nucleic acids were amplified using polymerase chain reactions. For analytical purposes, 20 µL reaction volume was used and 50 µL for preparative purposes. In Table 4.1 the standard recipe for PCR reactions is displayed. Usually, a master mix was prepared and then aliquoted to individual reaction tubes.

Table 4.1: Composition of PCR reactions for analytical and preparative purposes.

	20 µL	50 µL
Polymerase	0.05 µL	0.1 µL
Buffer (5x)	4 µL	10 µL
Template	1 ng (plasmids) 10 ng (bacmids)	1 ng (plasmids) 20 ng (bacmids)
Primer fwd (10 µM)	1 µL	1 - 1.5 µL
Primer rev (10 µM)	1 µL	1 - 1.5 µL
H2O	ad 20 µL	ad 50 µL

For DNA amplification, the program depicted in Table 4.2 was executed by the thermocycler. The optimal annealing temperature for the used primers was calculated using the NEB-Tm calculator. Normally, the steps denaturation, annealing and elongation were repeated 30 to 35 times.

Table 4.2: Program used for the amplification of DNA by PCR.

	Temperature	Time
Initial denaturation	95 °C	2 min (plasmids) 3 min (bacmids)
Denaturation	95 °C	20 s (plasmids) 30 s (bacmids)
Annealing		20 s
Elongation	72 °C	30 s/kb
Final elongation	72 °C	2.5x elongation time

4.1.2 Agarose gel electrophoresis

To analyze the size of nucleic acids, 1%-Agarose gels were supplemented with the DNA-dye MidoriGreen Xtra (1:50.000). Loading buffer, either 6 x TriTrack or 10 x FastDigest Green Buffer, was added to the sample and after loading the samples, the separation was conducted by the application of 100 V for 20 min. The gel was then transilluminated by blue-green light to excite the fluorescence dye and a FOculus IEEE1394 Camera was used for documentation. If needed, the desired bands were cut using a scalpel.

4.1.3 DNA restriction

DNA was restricted using FastDigest-Enzymes of Thermo Fisher according to the manufacturer's protocol. After digestion, enzymes were heat inactivated and desired nucleic acids were gel purified using the GeneJet-Gel-Extraction Kit. When vectors were digested, producing cohesive ends, FastAP was added to reduce vector religation in further steps.

4.1.4 DNA ligation

DNA ligation was performed using T4-Ligase of Thermo Fisher according to the manufacturer's manual. Usually, a molar ratio of 1:3 (vector : insert) was used. The ligation reaction was conducted either within 1 h at RT (room temperature) or overnight at 16 °C.

4.1.5 Gel extraction and Ethanol precipitation

Nucleic acids were extracted from Agarose gels using the GeneJet-Gel-Extraction kit according to the manufacturer's protocol. If required, DNA was purified and concentrated by ethanol precipitation. For this purpose, 0.1 volumes of NaOAc (3 M) and 2.5 volumes of 100 % ethanol were added to the sample following incubation at -80 °C for at least 1 h. The DNA was then pelleted by centrifugation at min. 15.000 g for 30 min at 4 °C. The supernatant was removed and 500 µL of 70 % ethanol was carefully added to wash the pellet without resuspending it. The sample was again centrifuged at min. 15.000 g for 20 min at 4 °C and the supernatant discarded. Exceeding ethanol was evaporated by warming the sample to 42 °C with open lid and the resulting DNA was resuspended in the desired amount of water.

4.1.6 Determination of nucleic acid concentration

The concentration of nucleic acids was determined by spectrophotometric measurements using a NanoDrop 2000c or a NanoDrop One^{C}.

4.1.7 Sequencing of nucleic acids

Sanger sequencing was performed by LGC genomics using the Ready-2-run sample preparation. Usually, 1 µg of plasmid or 100 to 400 ng of PCR products were diluted in 12 µL water and 2 µL of the sequencing primer was added.

4.2 Microbiological methods

4.2.1 Generation of chemically competent *E. coli* cells

For a preculture, one single colony was inoculated into 5 mL bacteria medium and if required antibiotics were added. Medium for chemically competent *E. coli* DH10EmBacY contained Kanamycin and Tetracycline. This culture was incubated shaking (250 rpm) at 37 °C overnight. The next day, 200 mL bacteria medium supplemented with antibiotics as required was prewarmed to 37 °C for at least 30 min and subsequently inoculated with 0.5 % of the preculture. Bacteria were incubated at 37 °C until they reached an OD_{600} of 0.5-0.8. Bacteria suspensions were transferred to four 50 mL reaction tubes and cooled in ice water for 30 min, followed by centrifugation at 2500 g at 4 °C for 30 min. All following steps were performed at 4 °C. The supernatant was discarded, the four pellets were resuspended in 10 mL ice-cooled Calcium-chloride-solution each and transferred into one tube. After additional 15 min on ice, bacteria suspensions were centrifuged again for 30 min at 2000 g. The supernatant again was discarded, the pellet resuspended in 1 mL ice-cooled Calcium-chloride-solution and aliquoted into 50/100/200 µL portions. Finally, the aliquots were flash-frozen in liquid nitrogen and stored at -80 °C.

4.2.2 Transformation of chemically competent cells

20 µL of chemically competent cells were thawed on ice per transformation reaction. Either 1 ng plasmid, 10 ng bacmid or 2 µL of ligation reaction were added to the bacteria suspension (*E. coli* DH5α/HB101 for plasmids and ligation reactions and *E. coli* DH10EmBacY for bacmids), and incubated on ice for 10 - 30 min, followed by a heat shock at 42 °C for 30 to 45 sec. Cells were then again incubated on ice for 2 minutes and bacteria medium was added. To plasmid and ligation reaction transformations, 180 µL medium was added, bacmid transformations received 380 µL medium. Subsequently, transformed bacteria were recovered at 37 °C for either 1 h (plasmids and ligation reactions) or up to 16 h (for bacmids), before 100 µL of the suspension were plated on agar plates supplemented with the required antibiotic(s) using glass balls. Plates were incubated for 1 d (plasmids and ligation reactions) or 2 d (bacmids) at 37 °C.

4.2.3 Plasmid preparation from bacteria cells

To isolate plasmids from bacteria cells, an overnight culture was prepared by inoculation of 5 mL bacteria medium supplemented with the required antibiotics with a single clone of the bacteria carrying the required plasmid. The next day the culture was pelleted, and the plasmid isolated using the GeneJet Miniprep kit according to the manufacturer's protocol. DNA was eluted using either elution buffer or water, depending on downstream applications.

4.2.4 Bacmid preparation from bacteria cells

5 mL culture of the bacmid containing bacteria cells were spun for 3 min at 10,000 g at RT before the bacmid DNA was extracted from the cells using the ZR Bac miniprep kit according to the manufacturer's protocol. Bacmid DNA was eluted using 30 µL prewarmed elution buffer.

4.3 Cell biological methods

4.3.1 Maintenance of insect cell lines

Insect cell lines were cultivated at 27 °C in shaking cultures (110 rpm) in 15 to 20 mL ESF921 cell culture medium supplemented with Penicillin (100 U/mL) and Streptomycin (100 ng/mL). For insect cell cultivation, upright oriented 75 cm²-cell culture flasks were used. Cell densities were kept ideally between 0.5 and 4 million cells/mL. Cell densities were determined by using an improved Neubauer counting chamber.

4.3.2 Thawing of insect cells

Thawing of insect cells either stored at -80 °C or in liquid nitrogen was performed quickly in a 37 °C water bath. After prewarming two 75 cm²-cell culture flask containing 25 mL insect cell culture medium supplemented with Penicillin and Streptomycin, 500 µL of the thawed cell suspension was added. Cells were then incubated shaking upright until a density of 4 million cells/mL was reached. The cultures were then split to a density of 1 million cells/mL and further cultured as described above.

4.3.3 Insect cell transfection

For the transfection of insect cells 0.45 million Sf9-cells were seeded in antibiotic-free cell culture medium in a 12 well plate. To achieve semi-adherent cells, they were incubated for at least 30 min to let them settle down. Meanwhile 1 µg Bacmid-DNA was diluted in 50 µL antibiotic-free cell culture medium, as well as 3 µL ESCORT IV were added to 47 µL of antibiotic-free cell culture medium. To allow the formation of DNA-liposome complexes both preparations were mixed and incubated for 30 – 45 min at RT. Cells were washed twice with 1 mL antibiotic-free cell culture medium and then covered with 400 µL antibiotic-free cell culture medium. The DNA-liposome complexes were added in a dropwise manner covering the whole well. After 16-24 h incubation at 27 °C the supernatant medium was replaced by 1 mL cell culture medium supplemented with Penicillin and Streptomycin. Cells were incubated for additional 4 days and subsequently checked for fluorescence. A successful transfection was hallmarked by viral spread visible by the appearance of clusters of fluorescing cells. To harvest the viral P1 stock supernatant medium was removed and centrifuged at min. 13.000 g for 30 s at RT.

4.3.4 Virus stock amplification

To further amplify viral stocks, 6 million Sf9-cells were incubated in a total volume of 5 mL insect cell culture medium with the desired amount of virus stock (for P2 stock generation 1 mL of P1 virus stock was used) into a 25 cm^2-cell culture flask. The flask was incubated lying for 4-6 days until at least 85 % of the cells were fluorescing. Harvesting of the amplified virus stock was performed as described for P1 virus stocks in chapter 4.3.3.

4.3.5 Determining virus stock titers

Titers of virus stocks were determined using an endpoint-dilution method based on the infection reporter EYFP encoded in the DH10EmBacY bacmid. 0.45 x 10^4 High Five-cells were applied in each well of a 96-wellplate in a total volume of 180 µL cell culture medium per well. Six wells were used in parallel for 8 steps of serial dilution to determine the virus dilution required to infect half of the cultures ($TCID_{50}$ value).

20 µL of the pre-diluted (1:10) virus stock was added into each well of the first row. After intensive mixing serial dilutions were performed by transferring 20 µL into the next well. In the last row 20 µL of each well were discarded. After 4 days of incubation at 27 °C all wells were checked for fluorescence and a well was counted as positive if more than two distant cells showed fluorescence.

The titer ($TCID_{50}$/mL) was calculated according to Reed and Muench (1938) and the required amount of virus for following infections was calculated according to Equation 4.1.

$$\frac{MOI * cell\ count}{TCID_{50}\frac{1}{mL} * 0.7} = V_{virus}\ [mL]$$

Equation 4.1: Equation to calculate the volume of the respective, titrated virus stock to infect a given number of cells with a defined MOI.

4.3.6 Infection of insect cells

To recombinantly express proteins in living insect cells, High Five-cells were seeded to 0.5 to 1 million cells per well of a 6-wellplate in a total volume of 2.5 mL cell culture medium. Cells were infected by adding the calculated volume (Equation 4.1) of the desired virus stock corresponding to a MOI of one to the medium. Cells were then incubated at 27 °C for 4 to 7 days.

4.3.7 Fusion of infected insect cells

Insect cells were infected as described above in chapter 4.3.6 and 23 to 47 h later fused by slow addition of 10 µL PEG4000 at various concentrations as a drop into the center of the well of a 6-wellplate. The stock solution of 50 % PEG4000 dissolved in water was diluted in insect cell medium to prepare PEG4000 concentrations needed.

4.4 Microscopic methods

4.4.1 Sample preparation for microscopy

For live cell microscopy using 100x oil-immersion objectives, the sample preparation included the infection of insect cells on cover slides inside one well of a 6-wellplate. After at least 4 days of incubation the cover slide was then placed onto a tailor-made holder, again covered by insect cell medium and subsequently imaged using the Nikon Ti2R microscope.

4.4.2 Evaluation of microscopic pictures

Microscopic images were analyzed using the Fiji software.

4.5 X-ray diffraction

4.5.1 Sample preparation

Insect cells were infected using the desired rbv as described in chapter 4.3.6. After 4 to 6 days, cells were harvested into a 1.5 mL tube and allowed to settle down by gravitation. 0.5 µL cells from the loose pellet were then loaded onto a micro MeshMount under microscopic supervision. Excess liquid was removed using liquid wicks. 0.35 µL cryoprotectant (40 % PEG200) was added and the excess liquid again removed by liquid wicks. The micro MeshMount was then flash frozen and stored in liquid nitrogen until diffraction.

4.5.2 X-ray diffraction at beamline P14 at PETRA III

The synchrotron data was collected at beamline P14 operated by EMBL Hamburg at the PETRA III storage ring (DESY, Hamburg, Germany). MeshMounts were automatically mounted on the goniometer and diffracted using a helical line scan approach using an oscillation of 1° or 0.1° per frame, an exposure time of 0.02 s and 100 % transmission. The beam was double focused to a size of 3 x 7 µm. Photon flux was around 10^{13} photons per second.

4.6 Bioinformatic methods

4.6.1 X-ray diffraction data reduction and processing using *CrystFEL*

To process X-ray diffraction data using *CrystFEL*, a geometry file specifying information about the sample to detector distance (*clen*), the wavelength (*photon_energy*), the size of the detector (*max_fs* and *max_ss*) as well as the beam position on the detector relative to the detector corner (*corner_x* and *corner_y*) is required. Then peak detection parameters can be specified using the graphical user interface of *CrystFEL* version 10.0/10.1 or the *check-peak-detection* script of version 9.1 and older versions using peakfinding algorithm *peakfinder8*. In general, for this study, a *threshold* of 0, a *local_bg_radius* of 3, *min_res* of 50, a *max_res* of 2200 and a *max_pix* value of 50 were a suitable starting point. The *min_snr* and *min-pix* were adjusted to allow proper peak detection for each sample. Afterwards, all frames were indexed by moslfm, and the correct lattice and unit cell parameters were optimized, and cycles of beam position refinement were performed by executing the *detector_shift* script (version 9.1) on the obtained stream files. If the optimized beam position did not differ more than 0.1 mm from the input, the beam position was accepted and all frames were indexed invoking *mosflm-latt-cell*, *mosflm-latt-cell*, *xds* and *xgandalf*. Then the *peakogram-stream* script was executed to set the maximum allowed intensity for each resolution range to separate salt reflections from protein reflections applying the filtering script described in chapter 5.1.3. Finally, the filtered intensities were scaled and merged applying *partialator* modelling partialities according to the model *xspheres* and restricting the per-frame resolution by a *push-res* value between 1 and 2. The resolution limit was set where CC* is above 50 %, SNR above 0.5 and the completeness over 95 %. Mtz-files for subsequent modeling were generated using *get_hkl*. To set up peak detection parameters *CrystFEL* version 10.0 was applied, for indexing *CrystFEL* version 10.1 or 10.2, while for merging and scaling *CrystFEL* version 9.1 or 10.2 were used.

Table 4.3: Peak finding parameters for peakfinder8 for all data sets elucidated in this study.

	Threshold	Min-snr	Min-pix	Max-pix	Min-res	Max-res	Min. peaks	Int-radius
AfHEX-1delN	0	4	4	50	50	2200	14	3/5/7
GMPR cyto v2	0	4	4	50	50	2200	14	3/5/7
MgHEX-1	0	3.5	4	50	50	2200	14	3/5/7
MrHEX-1delN	0	4	4	50	50	2200	14	3/5/7
NcHEX-1ori	0	3	4	50	50	2000	20	3/5/6
NcHEX-1 cyto	0	3.5	3	40	50	2000	20	3/5/7
NcHEX-1 cyto v2	0	3	4	50	50	2000	20	3/5/7
*Pl*HEX-1delN	0	4	4	50	50	2200	14	3/5/7

4.6.2 X-ray diffraction data reduction and processing using *XDS*

A prerequisite for data processing by *XDS* is the definition of crystal wedges from the helical line scans. This is performed using the tailor-made script and then *XDS* is applied to the crystal wedges identified. After some optimization steps the successfully generated HKL-files were scaled and merged using *XSCALE*. The resolution limit was set where CC* is above 50 %, I/sigma(I) above 1.5 and the completeness over 95 %. Mtz-files were generated applying *XDSCONV* and a *ccp4* script.

4.6.3 X-ray diffraction phasing and model building

Phasing was performed by molecular replacement (MR) using *phenix.phaser* by testing all possible space groups in the named point group. Then refinement was conducted using *phenix.refine* followed by manual refinements cycles using *coot*. If resolution was sufficient individual B-factors, real-space and xyz coordinates, as well as TLS refinement was performed using the automatic optimization of target weights. Sometimes additional restraints (secondary structure, NCS as well as rotameric and Ramachandran restraints) were applied if necessary.

4.6.4 Analysis of crystal symmetry elements

To investigate the symmetry elements of the GMPR cyto v2 structure, the refined structure, processed in space group P1 was given to *Zanuda*, which analyses the symmetry elements of the different molecules to identify crystallographic and non-crystallographic symmetry elements.

4.6.5 Sequence alignment

DNA sequences were aligned using the inbuilt alignment software invoking MAFFT v7 (Katoh and Standley, 2013) of the cloud-based molecular biology platform *benchling*.

4.6.6 Primer design

Oligonucleotides were designed using the inbuilt primer design software of the cloud-based platform *benchling*.

4.6.7 Database search for HEX-1 homologs in different fungi species

A *blastp* search was conducted using the *Neurospora crassa* HEX-1 (UniProt-ID: P87252) as query, searching RefSeq database by applying the default parameters including up to 5,000 hits. All hits were then analyzed by their annotated sequence length.

4.6.8 Alignment of selected HEX-1 homologs

The amino acid sequences of the 10 selected HEX-1 homologs were compared using the alignment function of the webserver *UniProt* executing *Clustal Omega*. The sequence alignment was then introduced into *Ident and Sim* to generate the identity and similarity matrices. With the alignment of the highly conserved, C-terminal 158 amino acids as input to *WebLogo3* the sequence logo was generated.

4.6.9 Structure comparison and alignment

Protein structures were compared using either the webserver *PDBeFold* or the graphical software *pymol* and aligned with the program superpose of the software suite *ccp4i*.

5. Results

Intracellular crystallization of proteins is a complementary method to existing crystallization techniques. Proteins that could not be crystallized *in situ* so far have been crystallized in living insect cells, enabling the elucidation of their structure (Duszenko et al., 2015; Gati et al., 2014; Nass et al., 2020).

In order to successfully implement a strategy for diffraction data processing and model building of *in cellulo* crystallized proteins, two model proteins, *Nc*HEX-1 and *Tb*IMPDH, were selected and their structures were elucidated. In the process of data processing, problems arose regarding overlapping ice and salt reflections that needed to be separated from the protein reflection signal.

Furthermore, the influence of additional amino acids, which are cloning artifacts of the recently developed cloning systems (Schönherr, 2021), was investigated by structure elucidation followed by examination of the 3D model of different *Neurospora crassa* HEX-1 proteins with these additional amino acids.

Additionally, suitable candidates for structure elucidation were selected by sequence homology to the selected model proteins to investigate their ability to form intracellular crystals. Not only orthologs, but also paralogs were tested for this approach and were then subjected to structure elucidation.

The structure of *Tb*GMPR could not be solved without the improvements in the crystal volume. Thus, cell fusion was established as a powerful tool to overcome resolution limits in *in cellulo* crystallography.

5.1 Establishing a data processing workflow for protein crystals grown and diffracted in living insect cells

To extend Schönherr's *in cellulo* crystallization pipeline with a workflow for diffraction data processing and model building, two model proteins were selected: the Woronin Body Major Protein (HEX-1) from *Neurospora crassa* and the IMP dehydrogenase from *Trypanosoma brucei*. Both are known to crystallize inside living insect cells (Lahey-Rudolph et al., 2021; Nass et al., 2020). In addition to establishing a data processing and modeling strategy, the effects of additional amino acids, resulting from the cloning strategy developed by Schönherr, on the structure will be investigated for *Nc*HEX-1. Therefore, three versions of *Nc*HEX-1 were subjected to structure elucidation: *Nc*HEX-1 ori without additional amino acids, *Nc*HEX-1 cyto cloned using the first generation cloning system and *Nc*HEX-1 cyto v2 encoding two additional amino acids at each terminus (Schönherr, 2021). The results discussed here are mainly included in Schoenherr et al., (2023).

For both model proteins reference structures from *in cellulo* grown crystals that have been diffracted at XFEL sources are available. Nass and coworkers isolated crystals of *Tb*IMPDH N-His, which includes an N-terminal fused His-tag next to the protein sequence and crystallizes spontaneously in Sf9 cells after rbv infection. The

*Tb*IMPDH N-His crystals were sprayed into the pulsed XFEL beam using a liquid-jet and revealed native cofactor binding in both canonical binding sites of the Bateman domain (Nass et al., 2020). In case of *Nc*HEX-1 cyto, intact, crystal-containing High Five cells were loaded into the pores of a silicon chip for X-ray diffraction (Lahey-Rudolph et al., 2021). Both structures were solved using *CrystFEL* for data processing and served as references in this study.

Due to limited access to XFEL sources, the strategy developed here for diffraction data processing focuses on diffraction experiments at synchrotron sources with fixed targets, which allow for a high hit rate. At synchrotron sources, different approaches for crystal delivery are established. For example nylon loops (Gati et al., 2014) or MeshMounts (Boudes et al., 2016), as well as different chip systems (Lahey-Rudolph et al., 2021; Norton-Baker et al., 2021) can be mounted on a goniometer to place the diffraction sample into the beam. Additionally, crystals can be delivered by moving tape drives covered with crystal containing solutions (Zielinski et al., 2022) or CrystalDirect™ plates on which crystals can be grown *in situ* (Cipriani et al., 2012; Schoenherr et al., 2023). In this study, micro MeshMounts were chosen for data collection because they enable the collection of rotational data and stills by using the helical line scan approach with or without oscillation (Gati et al., 2014). The same is true for nylon loops, but sample loading is easier when using MeshMounts. Silicon chips and CrystalDirect™ plates have also been successfully tested and implemented for delivery of *in cellulo* crystals but will not be discussed here (Norton-Baker et al., 2021; Schoenherr et al., 2023).

The high energy of the XFEL radiation leads to the destruction of the crystal and thus XFEL sources allow only serial data collection, where just one image per crystal can be collected (Chapman et al., 2011; Doerr, 2011; Neutze et al., 2000). While synchrotron radiation allows the collection of more than one image per crystal, also semi-serial approaches for data collection are possible (Gati et al., 2014). These rely on the collection and merging of small rotational data sets from many individual crystals. In serially collected data, difficulties arise in identifying the crystal orientation as well as in collecting full reflections.

In this work, small rotational data sets were collected by using the helical line scan approach. In order to use the best out of both conventional single-crystal diffraction and serial data collection, these small rotational data sets need to be identified, individually processed and finally merged into a complete data set. In this study, a script was developed to identify subsequent frames corresponding to the same crystal, combined with a script that subsequently filters for duplicates. The so identified crystal wedges were then processed individually with *XDS*. A detailed description can be found in chapter 5.1.1. The developed scripts were applied to diffraction data from *Nc*HEX-1 ori and *Tb*IMPDH ori. Additionally, diffraction data collected from *Nc*HEX-1 ori, cyto and cyto v2 were processed using *CrystFEL* as described in chapter 5.1.2. To compare the two software suites, *XDS* was applied on the same raw data of the diffraction data from *Nc*HEX-1 ori and *Tb*IMPDH ori, respectively.

While *CrystFEL* is the software suite of choice for snapshot data, also called still images (White et al., 2012, 2016), *XDS* is frequently used to process rotational data (Kabsch, 2010). For both processing approaches, the data can be collected by helical line scans on a MeshMount. Rotational data sets have the advantage of requiring fewer diffraction patterns, as additional information is added by the rotation series (Rupp, 2010). Integrating diffraction data with *CrystFEL* has the advantage that this software is built to work with large amounts of data (White, 2019; White et al., 2012, 2016). Both software suites are able to integrate even low diffraction signals at the predicted spot positions up to the detector edges. When signals from thousands of crystals are accumulated a more reliable signal can be obtained, than if signals from one small crystal wedge of only several detector frames is integrated. By combining the low signals from many different crystals, the signal-to-noise ratio is improved, and the resolution of the resulting data set can be higher than the Bragg spot with highest resolution found by the peak finding algorithm per single image.

For data collection, crystal-containing cells were harvested from the cell culture dish and loaded onto micro MeshMounts, which were cryoprotected, flash-frozen and stored in liquid nitrogen until diffraction data collection. The meshes were loaded into the automatic sample changer installed at the synchrotron beamline P14, operated by EMBL in Hamburg (PETRA III, DESY). Data were collected by helical line scans comparable to the approach described by Gati et al., (2014). The oscillation was set to 1.0° per frame and the exposure time was 0.02 s at full transmission.

5.1.1 Development of a script for serial diffraction data processing using *XDS*

A prerequisite for data processing with *XDS* is the use of small rotational data sets within the helical line scan. In this work, this is performed by analyzing one vertical line of subsequent frames by their spot count given by *dozor* (Zander et al., 2015), which is implemented as a real time analysis tool at the P14 beamline, using a sliding window analysis. Here, the frame with the maximum spot count of 10 subsequent frames was defined as the crystal center. The adjacent frames in a crystal specific distance *half_length* before and after the crystal center were defined as part of the crystal if the detected spot count was above a protein specific threshold *spot_min*. The frames corresponding to one crystal wedge defined this way were individually indexed and integrated using *XDS*. Next, the window of 10 subsequent frames was shifted by two frames and analyzed again for the maximum spot count, which was again defined as the crystal center and thus analyzed as described above. For each crystal wedge, *XDS* requires an input file, containing information about the experimental setup, and the location of the data frames as well as the detector distance, the detector properties, crystal system information and the peak finding parameters.

This sliding window analysis resulted in duplication of crystals, but also allowed for the separate identification and processing of different crystals that were hit simultaneously. After indexing, integration, and optimizing the processing of each crystal wedge, crystal wedges were filtered by application of a second custom script and only unique crystals that were successfully indexed and integrated were scaled and merged using *XSCALE*. The filtering is based on the orientation of the crystal by analyzing the orientation matrix. If overlapping crystals have an orientation that differs less than 6°, they are considered as equal and only the crystal wedge described by most subsequent frames was kept. The script developed to filter for the desired crystals is presented in the Appendix, p. 140, along with a full list of the commands executed. *XSCALE* needs a similar input file as *XDS*, containing all the crystal wedges to be merged and the unit cell parameters as well as the space group.

The scripts were applied on the diffraction data sets of the two model proteins *Nc*HEX-1 ori and *Tb*IMPDH ori to verify their performance. The obtained MTZ-files were subjected to molecular replacement using the conventionally *in situ* crystallized *Nc*HEX-1 structure (pdb: 1khi) as a search model for the *Nc*HEX-1 ori data and the conventionally crystallized human IMPDH structure (pdb: 1jcn) to obtain phases for the *Tb*IMPDH ori data set. The structures are further analyzed in chapter 5.1.4 and 5.1.5, respectively.

5.1.2 Description of serial diffraction data processing using *CrystFEL*

As an alternative approach, *CrystFEL*, a software suite designed for processing of diffraction snapshots collected at XFELs, was applied to the diffraction data collected. This software treats all frames individually and does not use rotation or other prior information (Sawaya et al., 2014; White et al., 2012). It was applied to diffraction data from *Nc*HEX-1 ori, cyto and cyto v2 crystals. *CrystFEL* requires a list of frames as input, as well as an optional file describing the unit cell by the cell parameters and the lattice, and a geometry file describing the experimental setup, including parameters such as the detector size, beam center, untrusted regions, and sample to detector distance.

The information needed for the geometry file can be found in the image header of the detector frames. First, an appropriate detector file was generated manually, to allow *CrystFEL* to read the detector frames. An example detector file is shown in the Appendix (p. 143). Then the peak detection parameters were set manually. The geometry file and the peak-finding parameters were required to run the indexing program *indexamajig*, invoking the indexing algorithm without prior information on the unit cell and lattice (*mosflm-nolatt-nocell*), which then enabled the determination of the lattice and unit cell parameters with the program *cell_explorer*. When the unit cell parameters converged, they were combined into the unit cell file also shown in the Appendix (p. 143). Still using only *mosflm* but with the unit cell included in the indexing process, the refinement of the beam center was conducted by iteratively applying the *detector_shift* script and indexing again with the new beam center.

For the *Nc*HEX-1 cyto v2 data set, the beam center refinement did not converge neatly, indicating a tilt in the detector plane. In this sample, *cell_explorer* also showed a second unit cell population, indicating an imperfect geometrical setup. However, after separating the two unit cell populations with the *stream_grep* script, the unit cell histograms were clearer, but only one structure showed acceptable quality parameters and could be subjected to phasing and model building. These problems were not encountered in the other data sets shown here.

Of note, salt reflections and ice rings next to the Bragg spots corresponding to the protein of interest were observed in all data sets analyzed, as indicated by powder diffraction signals in the plots generated by the *peakogram_stream* script (Figure 5.1 A & B). To reduce the disturbance of the contaminating signals, all predicted spot positions needed to be filtered.

5.1.3 Filtering for salt and ice signals during data processing using *CrystFEL*

The occurrence of salt crystals was most likely a result of the sample preparation technique applied. MeshMounts were loaded with a dense suspension of crystal-containing cells in insect cell medium. The medium contains high concentrations of salt that can precipitate and form crystals when the excess liquid is blotted. Salt crystals produce additional strong Bragg reflections besides the desired spots corresponding to the respective protein crystal. Moreover, water within the sample forming ice crystals during the freezing process, due to insufficient cryoprotection, as well as ice crystals within the liquid nitrogen settling on the sample also result in the appearance of additional diffraction signals. These contaminant reflections were already observed during data acquisition and negatively affected data processing.

Ice rings are usually excluded in the peak detection by setting a maximum number of pixels per peak, but integration on predicted peak positions also leads to integration of contaminant reflections. Salt reflections can be identified by their high intensity even in higher resolution shells and their larger spacing due to the smaller unit cells. Salt reflections also do not cover the low-resolution range and start to appear at around 10 Å due to their small unit cell (Rupp, 2010). These characteristics were used to exclude salt reflections when filtering for reflections corresponding to the protein by applying a custom script. The script filters all integrated intensities at the predicted spot positions, based on a protein-specific intensity limit in a defined resolution range, that removes unreliable high intensities at high resolution because of the high probability that this reflection overlaps with a contaminant signal. The developed script is presented in the Appendix (p. 142).

The thresholds for the different resolution ranges were defined using the *peakogram-stream* script of *CrystFEL*. The *peakogram-stream* script analyzes the stream file and plots the observed intensities at each resolution (Figure 5.1 A-C). Without salt or ice crystals present in the sample, the intensity decreases with resolution. With salt or ice crystals present, there are vertical lines in the *peakogram-stream* plot at certain resolutions, corresponding to the powder diffraction signal or Bragg spots of the salt or ice crystals present. Finally, to avoid the merging and scaling of these unusual high-intensity salt or ice reflections, resolution-dependent intensity cut-offs were applied, and the reflections of the stream-file were filtered accordingly, as shown in Figure 5.1 C for *Nc*HEX-1 cyto.

Compared to unfiltered stream files, application of the script revealed improved correlation coefficients with a smoother decrease as shown for $CC_{1/2}$ of the *Nc*HEX-1 cyto data set in Figure 5.1 D (Table 5.1). The calculation of CC* values failed in resolution shells affected by the powder diffraction signal. The established filtering step enabled the calculation of the correlation coefficient, CC*, and revealed reliable values. The signal-to-noise ratio as well as the completeness were unaffected, while the R_{split} values were lower in the shells above the resolution cut-off (Table 5.1).

The filtering for the contaminant reflections resulted in an overall higher resolution cut-off of the obtained data set with improved figures of merit. However, the Wilson B-factor increased from 27.6 Å^2 to 44.3 Å^2. The script for filtering salt and ice contamination was applied to all data sets except for that of *Tb*GMPR cyto v2, where, due to its low resolution, the overlap with the contaminant signal was not that strong.

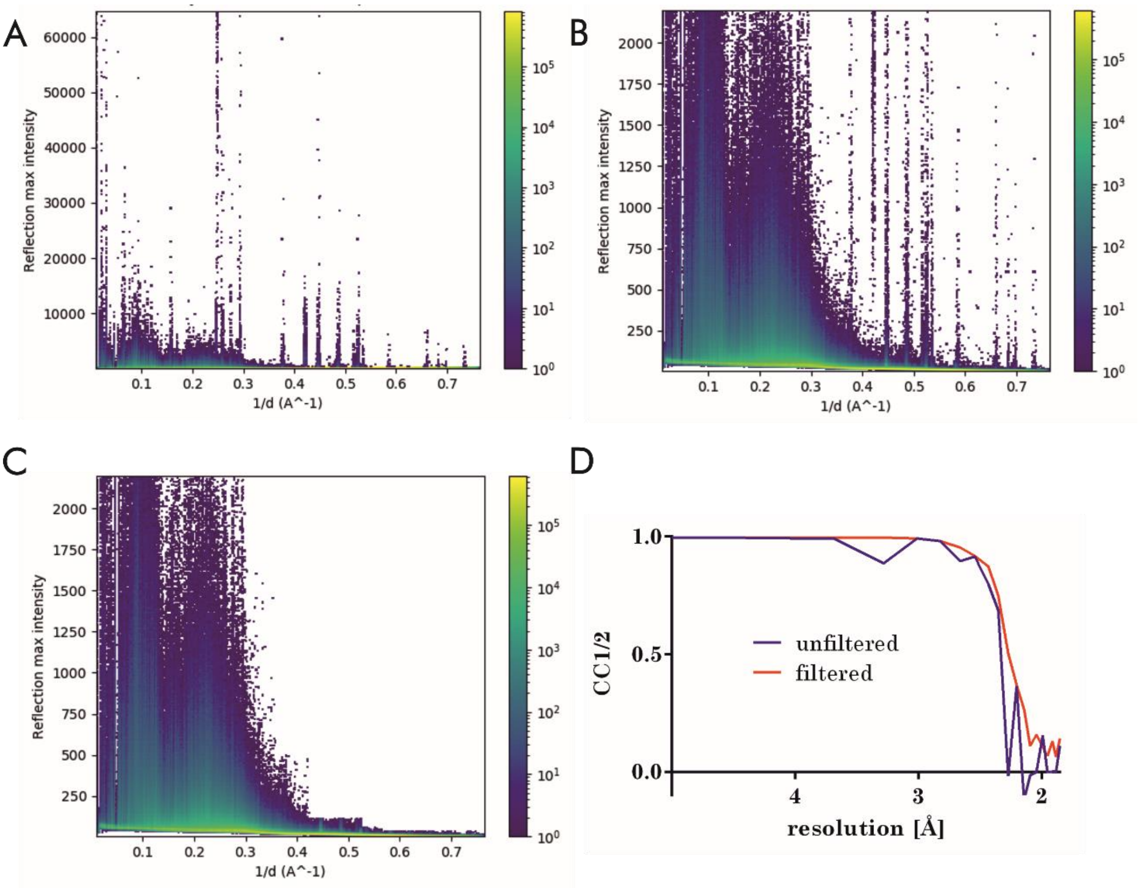


Figure 5.1: Analysis of the stream file of Nc*HEX-1 cyto before and after filtering for salt reflections and ice rings.* Peakogram_stream *output from the unfiltered stream file (A), from the unfiltered stream file restricted to an intensity maximum of 2,200 (B), compared to the filtered stream file (C). The reflection intensity is plotted against the reciprocal resolution in Å^{-1}. The unfiltered stream file shows powder diffraction signal from salt and ice crystals inside the sample (A & B), after filtering with the developed script, the obtained stream file shows a decrease of intensity along the resolution (C) and a smoother decrease of $CC_{1/2}$ values (D) as expected for a protein crystal. The $CC_{1/2}$ values are plotted against the resolution in Å.*

Table 5.1: Figures of merit of the unfiltered and filtered stream according to the resolution shell. The filtered stream file shows improved correlation coefficients, while the completeness and the SNR is unaffected. CC values could not be calculated for the unfiltered stream file in some resolution shells. The resolution cut-off is defined by the highest shell where SNR and CC* are above 0.5 and the completeness is above 95 %. The resolution cut-off is depicted as dashed line for the filtered data set.*

	Unfiltered					Filtered				
Resolution (Å)	R_{split} (%)	CC*	$CC_{1/2}$	Completeness (%)	SNR	R_{split} (%)	CC*	$CC_{1/2}$	Completeness (%)	SNR
9.45	2.85	0.9990653	0.9962697	100.00	42.01	2.51	0.9996535	0.9986154	100.00	42.23
4.40	3.44	0.9986719	0.9947052	100.00	37.10	2.77	0.9995033	0.9980156	100.00	38.57
3.68	4.38	0.9978531	0.9914583	100.00	27.36	3.73	0.9992751	0.9971058	100.00	27.90
3.28	6.34	0.9695712	0.8869139	100.00	19.42	4.87	0.9990285	0.9961234	100.00	19.69
3.01	7.18	0.9980937	0.9924108	100.00	13.44	7.40	0.9979452	0.9918230	100.00	13.46
2.82	11.71	0.9951728	0.9809217	100.00	8.80	11.59	0.9953245	0.9815144	100.00	8.83
2.66	19.55	0.9720771	0.8956157	100.00	6.19	16.26	0.9881320	0.9538975	100.00	6.34
2.54	23.50	0.9776022	0.9151696	100.00	4.59	23.18	0.9784502	0.9182163	100.00	4.60
2.43	33.85	0.9420120	0.7975695	100.00	3.38	29.84	0.9656482	0.8734950	100.00	3.46
2.35	51.47	0.9021191	0.6860833	100.00	2.20	47.02	0.9252314	0.7483331	100.00	2.26
2.27	148.01	nan	-0.0177005	100.00	1.37	75.79	0.8186323	0.5039391	100.00	1.48
2.20	113.43	0.7312103	0.3648789	100.00	0.94	111.75	0.7347476	0.3697261	100.00	0.96
2.14	194.34	nan	-0.1225313	100.00	0.57	187.80	0.6464223	0.2641121	100.00	0.56
2.09	361.18	nan	-0.0146618	100.00	0.28	325.81	0.4443764	0.1095518	100.00	0.34
2.04	363.71	nan	-0.0006223	99.88	0.18	466.93	0.5236477	0.1588875	100.00	0.22
1.99	776.12	0.5167239	0.1540705	100.00	0.12	967.18	0.4375711	0.1058696	100.00	0.12
1.95	415.68	nan	-0.0009013	100.00	0.11	741.85	0.3550181	0.0672574	100.00	0.12

5.1.4 Comparing structural models obtained by processing using either *CrystFEL* or *XDS*

The *Nc*HEX-1 ori data set was processed with *XDS* and *CrystFEL*, while the *Tb*IMPDH ori data set was only processed with *XDS*. The obtained mtz-files were then subjected to phasing with molecular replacement and subsequent model building and refinement. The same raw data of *Tb*IMDPH ori were *CrystFEL*-processed, phased, and modelled by Mia Lahey-Rudolph within the framework of the respective publication (Schoenherr et al., 2023) and is used here for comparison.

For both proteins, *Tb*IMPDH ori and *Nc*HEX-1 ori, the structure obtained with *XDS* was highly similar to that elucidated with *CrystFEL,* with the same biological information and shows only minor deviations to the reference structures solved from diffraction experiments at XFEL sources (Table 5.3). The *Nc*HEX-1 ori structures (Figure 5.3 A (*CrystFEL*) and B (*XDS*)) showed the same fold as the conventionally crystallized NcHEX-1 (Yuan et al., 2003). Analysis of structural similarity between the *Nc*HEX-1 ori structures obtained after processing with both software suites by *PDBeFold* revealed a very high similarity indicated by a Q-score of 0.913 and a mean RMSD of Cα atoms of 0.078 Å. The Q-score scores the quality of the match by considering the number of aligned residues, the total number of residues in the structures to be compared and the RMSD of Cα atoms. Its value ranges between 0 and 1. A Value of 1 indicates a perfect match, while a value close to 0 indicates low similarity.

The structure solution of *Tb*IMPDH ori crystals revealed two monomers in the asymmetric unit with a conserved fold of two domains, the catalytic domain, and the Bateman domain. Due to the symmetry elements this reveals the biologically unit of an octamer. Like the *Tb*IMPDH N-His structure (Nass et al., 2020), this octamer is in the inactive, closed conformation. The diffraction of *Tb*IMPDH ori crystals inside living cells provided additional information about the natively bound cofactors. Nass et al. (2020) isolated the *Tb*IMPDH N-His crystals prior to diffraction and obtained ATP and GMP bound to the canonical binding sites in the Bateman domain. In the *Tb*IMPDH ori structure presented here, ATP was bound in two different conformations to the canonical binding site I, while a GDP was found in the canonical binding site II (supp. Figure 10.1, Appendix p. 148). These results indicate that cofactor hydrolysis can occur if crystals are isolated from the protecting cells prior diffraction, and that further detail can be obtained when crystals are measured inside the intact insect cells. The same cofactors were identified in the *CrystFEL*-processed *Tb*IMPDH ori structure, elucidated by Mia Lahey-Rudolph. Both *Tb*IMPDH ori structures are highly similar, indicated by high Q-scores (mean 0.98) and small deviations (mean RMSD of Cα atoms 0.338 Å) between the structures (Table 5.2).

Table 5.2: Structural comparison of both chains of* Tb*IMPDH ori structures processed with* XDS *or* CrystFEL *in accordance with the Q-score, the percentage of matched secondary structure elements and the RMSD of Ca atoms in Å, revealing a high structural similarity.

Chain ID of the *XDS*-processed model	Chain ID of the *CrystFEL*-processed model	Q-score	% of matched secondary structure elements	RMSD (Å)
B	A	0.98	91	0.28
B	B	0.98	91	0.39
A	B	0.97	94	0.31
A	A	0.97	88	0.37

The resolution of the *XDS*-processed structures is worse than that of the *CrystFEL*-processed structures (Table 5.3). Nevertheless, for the *Nc*HEX-1 ori data set, the electron density obtained after processing with *CrystFEL* was in some parts, e.g., the termini, slightly weaker than in maps obtained from *XDS*-processing. Consequently, the amino acids 28 to 175 could be modeled in the *XDS*-processed *Nc*HEX-1 ori structural model, while interpretable electron density in the *CrystFEL*-processed structural model starts at residue 29. The number of detector frames used are comparable in both processing approaches.

The obtained refinement R-factors (all in the range of 0.21 to 0.24) as well as the atomic B-factors and Wilson B-factors (*Tb*IMPDH ori *XDS*: 53.73 Å^2 vs. *CrystFEL*: 53.06 Å^2, *Nc*HEX-1 ori *XDS*: 27.53 Å^2 vs. *CrystFEL*: 31.60 Å^2) were comparable in the structural models after processing with *XDS* and *CrystFEL* (Table 5.3), indicating that the structures are of comparable quality even if the resolution differs.

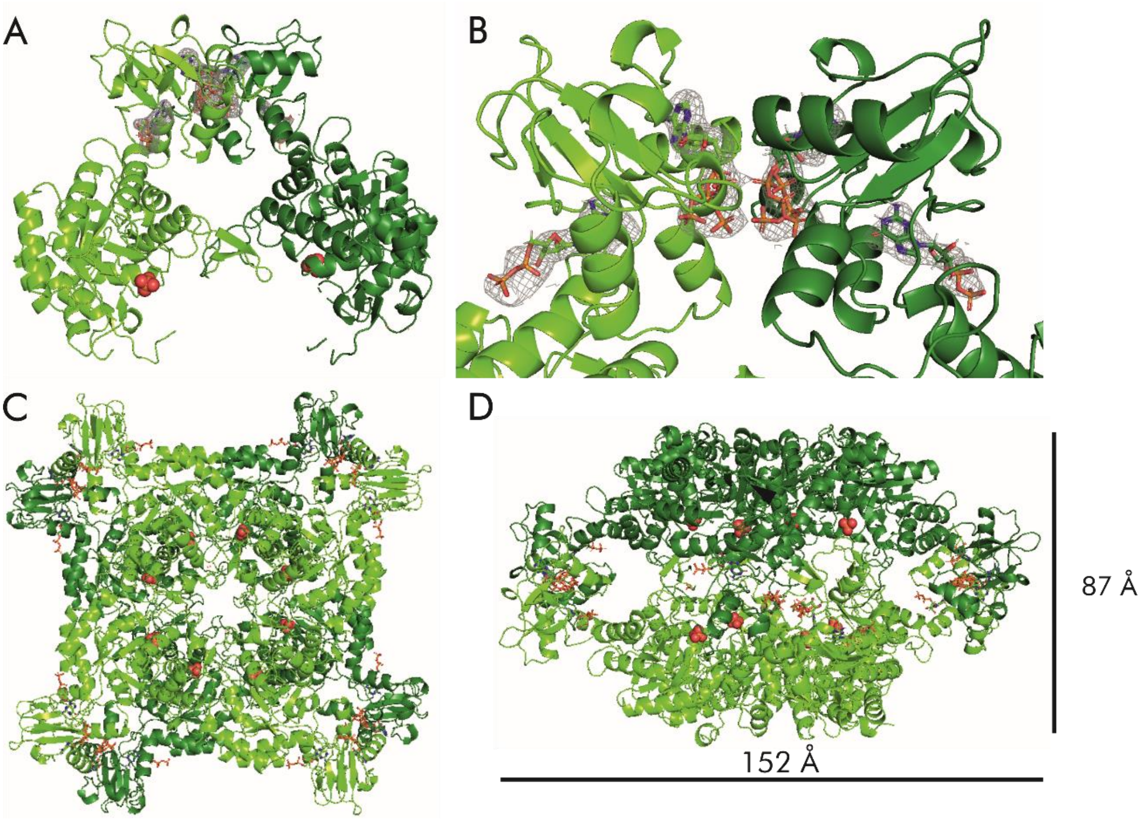


Figure 5.2: Cartoon representation of the Tb***IMPDH ori structure obtained after diffraction data processing with*** XDS***. Overview on the dimer in the asymmetric unit with chain A in dark and chain B in light green (A). The omit maps of the bound ligands, ATP (with alternative conformations) and GDP, are depicted as grey mesh. Detailed view on the regulatory Bateman domain with the identified ligands and their omit maps (Fo-Fc) contoured at 3.0 sigma carved at 2.0 Å (B). Due to the symmetry operators the biological unit, an octamer, is obtained (top view C). The octamer is shown from side view and reveals a size of 152 Å in width and 87 Å in height (D).***

Table 5.3: Data processing and refinement statistics of* Tb*IMPDH ori and* Nc*HEX-1 ori processed by CrystFEL and XDS. Values in parentheses refer to the highest resolution shell.* [a] *of up to 11 consecutive frames;* [b] *of up to 7 consecutive frames

Data set	*Tb*IMPDH ori	*Tb*IMPDH ori	*Nc*HEX-1 ori	*Nc*HEX-1 ori
Processing software	CrystFEL	XDS	CrystFEL	XDS
Pdb code	8c51	8cgy	8cd5	8cgx
Space group	$P4\,2_1\,2$		$P6_5\,2\,2$	
Unit cell (Å or °)	207.12 207.12 92.54 90 90 90	207.10 207.10 92.50 90 90 90	57.23 57.23 198.15 90 90 120	57.24 57.24 198.19 90 90 120
Resolution range (Å)	57.38 - 2.30 (2.330 - 2.299)	92.62 - 3.00 (3.107 - 3.000)	48.08 - 1.56 (1.616 - 1.560)	33.03 - 1.85 (1.916 - 1.850)
Total reflections	326,687,476	97,508,239	210,384,917	80,449,866
Unique reflections	89,070 (5,857)	40,798 (9,383)	28,553 (1,386)	17,305 (1,674)
Multiplicity	3,667 (1,246)	1,274 (1,135)	7,367 (610)	4,639 (4,197)
Completeness (%)	99.88 (99.09)	99.80 (99.20)	99.83 (98.82)	99.80 (98.30)
SNR/ $< I/\sigma(I) >$	10.06 (0.58)	20.37 (4.21)	31.47 (0.59)	35.15 (3.30)
Wilson B-factor	53.06	53.76	27.56	31.60
R_{meas} (%)	-	341.1 (8,831.4)	-	117.0 (56,232.3)
R_{split} (%)	8.62 (199.52)	-	1.95 (188.97)	-
CC1/2	0.9976 (0.1957)	0.9810 (0.8700)	0.9997 (0.2865)	0.9990 (0.9390)
CC*	0.9994 (0.5722)	-	0.9999 (0.6674)	-
No. collected images	96,949		56,891	
No. hits/ indexed lattices	75,741/82,831	6,001[a]	54,077/131,442	7,922[b]
Removed salt spots	7.60 %	-	0.03 %	-
Refl. used in refinement	88,932 (8,710)	40,788 (3,982)	28,419 (2,729)	18,624 (1,699)
Refl. used for R-free	1078 (107)	1,598 (156)	1,473 (139)	1,268 (117)
R_{work}	0.2132 (0.3772)	0.2014 (0.3415)	0.1993 (0.4334)	0.2061 (0.3500)
R_{free}	0.2431 (0.3386)	0.2388 (0.4156)	0.2159 (0.4060)	0.2352 (0.4580)
No. non-hydrogen atoms	7,081	6,852	1,226	1,202
Macromolecules	6,799	6,716	1,128	1,124
Ligands	182	182	0	0
Solvent	148	2	98	78
Protein residues	894	894	147	148
RMS bonds (Å)	0.004	0.005	0.004	0.006

Data set	*Tb*IMPDH ori	*Tb*IMPDH ori	*Nc*HEX-1 ori	*Nc*HEX-1 ori
Processing software	CrystFEL	XDS	CrystFEL	XDS
RMS angles (°)	0.62	0.63	0.71	0.86
Ramachandran favored (%)	97.4	95.46	98.62	97.95
Ramachandran allowed (%)	2.6	4.54	0.69	1.37
Ramachandran outliers (%)	0	0	0.69	0.68
Rotamer outliers (%)	0.97	0.14	0	0
Clash score	5.26	12.78	4.44	4.03
Average B-factor (Å^2)	59.98	52.12	41.25	42.06
Macromolecules	59.87	51.77	40.6	41.65
Ligands	61.64	58.01	-	-
Solvent	57.22	34.41	48.67	47.91
Mean RMSD (Å) to 6rfu for IMPDH; 7asx for HEX-1	0.3650	0.4143	0.4240	0.4226
Max. RMSD (Å) to 6rfu for IMPDH; 7asx for HEX-1	7.0810	7.0310	2.1740	2.1700

5.1.5 Effect of cloning artifacts on the structure of *Nc*HEX-1

To investigate the effect of cloning artifacts introduced by the different cloning systems used, the structure of three *Nc*HEX-1 variants was elucidated by processing diffraction data with *CrystFEL*. All analyzed *Nc*HEX-1 variants, namely *Nc*HEX-1 ori (Figure 5.3 A), *Nc*HEX-1 cyto (Figure 5.3 C) und *Nc*HEX-1 cyto v2 (Figure 5.3 D), displayed the two-domain structure described by Yuan et al., (2003) for conventionally crystallized *Nc*HEX-1.

No electron density was observed for the first 27 N-terminal amino acids and for one C-terminal amino acid. Superposition of the different structures of *Nc*HEX-1 obtained in this study reveals a high level of similarity (Figure 5.3 E & Table 5.4), indicated by an overall Q-score of 0.92 and a mean RMSD of Cα atoms of 0.330 Å, calculated by *PDBeFold*. The highest deviation from the reference structure 7asx (*in cellulo* crystallized *Nc*HEX-1 cyto) collected from intact cells at an XFEL (Lahey-Rudolph et al., 2021) was observed in the variable loop region (amino acids 60-66, shown by the arrowhead in Figure 5.3 A) connecting two ß-sheets of *Nc*HEX-1 cyto v2, displaying a maximum RMSD of the Cα atoms of 2 Å. Overall, the deviations of the Cα atoms of the *in cellulo* crystallized *Nc*HEX-1 variants from the reference structure were small, as shown in Table 5.4, indicating a high structural similarity despite the additional amino acids added by the cloning strategy used.

The crystal lattice and contact interfaces are also not affected since an interface analysis by *PDBePISA* revealed no differences between the structures. All models show two types of interfaces stabilized by a salt bridge. The first type, including salt bridges between Arg41 and Asp44 as well as His39 and Asp 44, is comparably strong and induces self-assembly, while the second type, forming a salt bridge between Arg57 and Asp 110, is a result of the assembly instead and the interactions are weaker.

Despite the different crystal morphologies observed after recombinant gene expression of the *Nc*HEX-1 variants, all protein structures were solved in the same space group and revealed a similar resolution cut-off. While *Nc*HEX-1 ori forms squared crystals with hexagonal cross-sections, *Nc*HEX-1 cyto and *Nc*HEX-1 cyto v2 form spindle-like crystals with hexagonal cross-sections or bipyramidal ones (Schönherr, 2021). At least for *Nc*HEX-1, an impact of the additional amino acids resulting from the cloning strategy (C-terminal Ala in *Nc*HEX-1 cyto; N-terminal Gly and Thr and C-terminal Ala and Ser in *Nc*HEX-1 cyto v2) on the protein structure was not detected.

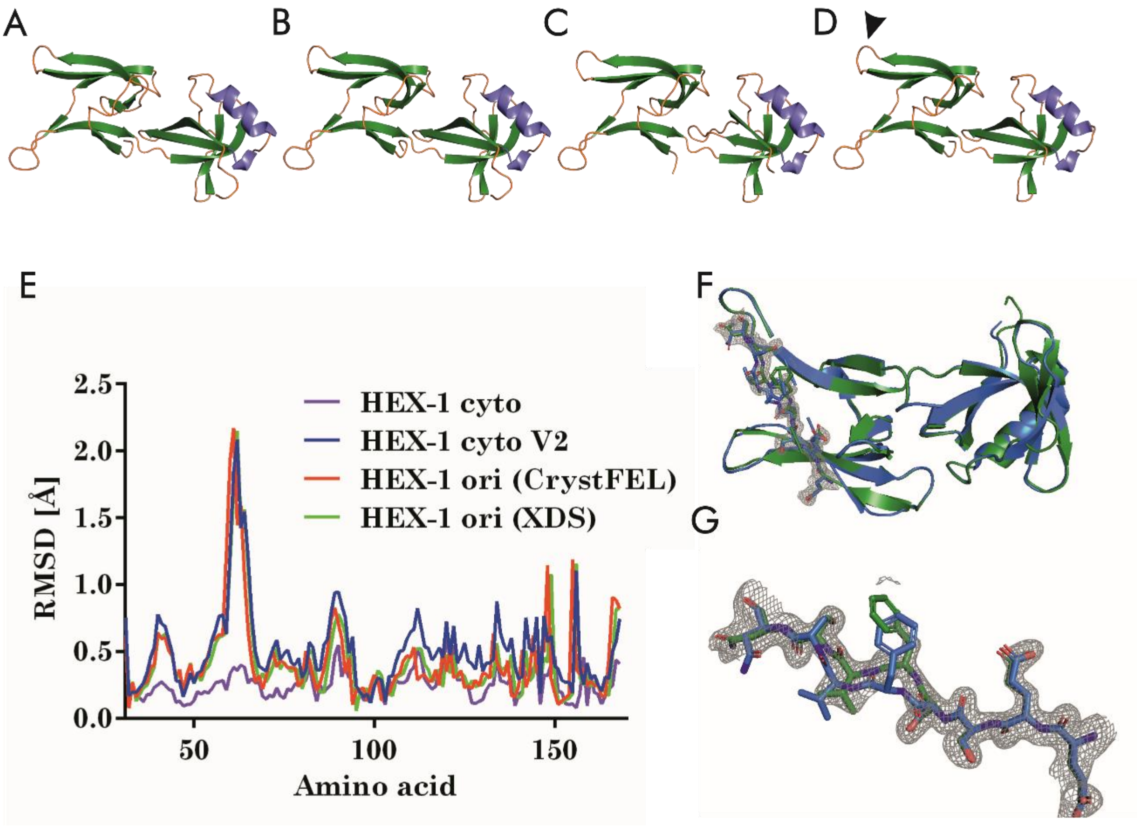
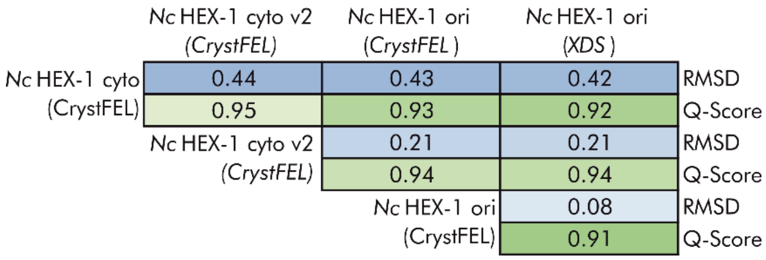


Figure 5.3: Overview of solved NcHEX-1 *structures. Cartoon representation of the structural models of (A)* NcHEX-1 *ori processed using* CrystFEL, *(B) using* XDS, *(C)* NcHEX-1 *cyto* (CrystFEL) *and (D)* NcHEX-1 *cyto v2* (CrystFEL). *β-sheets are displayed in green; a-helices are shown in blue. In (D) the highly variable loop region of residues 60 to 66 is marked by an arrowhead. (E) RMSD of Ca atoms calculated by* ccp4i superpose, *plotted per amino acid, numbered after the reference structure* NcHEX-1 *cyto (7asx), from which data were collected from intact cells at an XFEL. The differences observed are small, but larger differences were observed in the loop region e.g., for residues 60 to 66 indicated by a RMSD above 1.5 Å. Overview (F) and detailed view (G) of the structural overlay of* NcHEX-1 *ori* (CrystFEL) *displayed in green and the search model for molecular replacement, in situ crystallized* NcHEX-1 *(pdb: 1khi) displayed in blue. A representative section, residues 80 to 89, of the Fo-Fc map (omit map) contoured at 3.0 sigma of the* NcHEX-1 *ori* (CrystFEL) *data set is shown in mesh representation.*



	Nc HEX-1 cyto v2 (*CrystFEL*)	*Nc* HEX-1 ori (*CrystFEL*)	*Nc* HEX-1 ori (*XDS*)	
Nc HEX-1 cyto (CrystFEL)	0.44	0.43	0.42	RMSD
	0.95	0.93	0.92	Q-Score
Nc HEX-1 cyto v2 (*CrystFEL*)		0.21	0.21	RMSD
		0.94	0.94	Q-Score
Nc HEX-1 ori (CrystFEL)			0.08	RMSD
			0.91	Q-Score

Figure 5.4: Structural comparison of the different NcHEX-1 *variants using* PDBeFold *reveals high Q-scores and small RMSD values of Ca atoms (in Å), indicating a high similarity. A lighter coloring indicates a higher similarity of the respective structures.*

Table 5.4: Data processing and refinement statistics of NcHEX-1 variants. Values in parentheses refer to the highest resolution shell. The columns for NcHEX-1 ori are equal to Table 5.3 and are repeated here for comparison. [a] *of up to 7 consecutive frames.*

Data set	NcHEX-1 ori	NcHEX-1 ori	NcHEX-1 cyto	NcHEX-1 cyto v2
Processing software	CrystFEL	XDS	CrystFEL	CrystFEL
PDB code	8cd5	8cgx	8cd4	8cd6
Data collection date	26.11.2019	26.11.2019	24.08.2021	24.08.2021
Sample temperature [K]	100	100	100	100
Space group	$P6_522$	$P6_522$	$P6_522$	$P6_522$
Unit cell (Å or °)	57.23 57.23 198.15 90 90 120	57.24 57.24 198.19 90 90 120	57.39 57.39 189.96 90 90 120	56.70 56.70 196.80 90 90 120
Resolution range (Å)	48.08 - 1.56 (1.616 - 1.560)	33.03 - 1.85 (1.916 - 1.850)	50.19 - 1.83 (1.895 - 1.830)	47.64 - 1.85 (1.916 - 1.850)
Total reflections	210,384,917	80,449,866	24,754,589	4,735,984
Unique reflections	28,553	17,305	17,908	17,016
Multiplicity	7,367 (610)	-	1,382 (100)	278 (115.9)
Completeness (%)	99.83 (98.82)	99.80 (98.30)	99.92 (99.31)	99.81 (99.09)
SNR/ $< I/\sigma(I) >$	31.47 (0.59)	35.15 (3.30)	22.16 (0.64)	8.15 (0.53)
Wilson B-factor ($Å^2$)	27.56	31.60	31.70	34.14
R_{meas} (%)	-	117.0 (56,232.3)	-	-
R_{split} (%)	1.95 (188.97)	-	3.40 (168.07)	7.95 (184.98)
CC1/2	0.9997 (0.2865)	0.9990 (0.9390)	0.9994 (0.2745)	0.9953 (0.4226)
CC*	0.9999 (0.6674)	-	0.9998 (0.6564)	0.9988 (0.7707)
No. collected images	56,891	56,891	119,053	114,498
No. hits/ indexed lattices/ lattices after stream_grep	54,077/131,442	7,922 [a]	44,200/ 41,591	24,520/21,835/6,117
Removed salt intensities	0.03%	-	0.02%	0.01%
Reflections used in refinement	28,419 (2,729)	18,624 (1,699)	17,807 (1,718)	16,899 (1,631)
Reflections used for R-free	1,473 (139)	1,268 (117)	1,323 (127)	1,045 (102)
R_{work}	0.1993 (0.4334)	0.2061 (0.3500)	0.1923 (0.3548)	0.2034 (0.3964)
R_{free}	0.2159 (0.4060)	0.2352 (0.4580)	0.2148 (0.3489)	0.2369 (0.4258)
No. non-hydrogen atoms	1,226	1,202	1,185	1,174
Macromolecules	1,128	1,124	1,095	1,090

Data set Processing software	*Nc*HEX-1 ori CrystFEL	*Nc*HEX-1 ori XDS	*Nc*HEX-1 cyto CrystFEL	*Nc*HEX-1 cyto v2 CrystFEL
Solvent	98	78	90	84
Protein residues	147	148	143	143
RMS bonds (Å)	0.004	0.006	0.011	0.008
RMS angles (°)	0.71	0.86	1.02	0.91
Ramachandran favored (%)	98.62	97.95	97.87	98.58
Ramachandran allowed (%)	0.69	1.37	2.13	1.42
Ramachandran outliers (%)	0.69	0.68	0.00	0.00
Rotamer outliers (%)	0	0	0	0
Clash score	4.44	4.03	3.65	6.89
Average B-factor (Å^2)	41.25	42.06	39.19	40.88
Macromolecules	40.60	41.65	38.64	40.27
Solvent	48.67	47.91	45.94	48.73
Mean RMSD (Å) to ref. structure (pdb: 7asx)	0.424	0.423	0.260	0.520
Max RMSD (Å) to ref. structure (pdb: 7asx)	2.174	2.170	0.652	2.084

5.2 Overcoming statistical limitations by identifying promising candidates from sequence orthologs for *Nc*HEX-1

As in conventional crystallography not all proteins form crystals at given conditions, not all proteins spontaneously crystallize after recombinant expression inside living insect cells. To increase the chances of finding a protein that spontaneously crystallizes *in cellulo* to further investigate the crystallization process inside the living insect cells, homologs of proteins that are already known to crystallize were identified. A homolog of a specific protein can either be a protein that is found in a closely related species fulfilling the same function and can be traced back to a common ancestor (ortholog), or one that is found in the same species. The latter results from gene duplication within a species, and although both proteins have a high sequence similarity, they fulfill different functions (paralogs).

As described above, *Nc*HEX-1 forms crystals inside the Woronin bodies in *Neurospora crassa*, and thus represents an excellent target for *in cellulo* crystallization studies (Schönherr, 2021). Woronin bodies are a common feature of filamentous fungi functioning as a plug for the intercellular connections in case of cellular damage (Tang et al., 2020a; Tenney et al., 2000; Yuan et al., 2003). In this thesis, homologous genes encoding the Woronin body major protein (HEX-1) of different fungal species were identified, recombinantly expressed and crystallized in insect cells and subjected to structure elucidation.

5.2.1 Identification of suitable orthologs of *Nc*HEX-1

HEX-1 homologs are highly abundant in Ascomycota (Tang et al., 2020a). A UniProt search for “Woronin body major protein” yielded 390 proteins. A *blastp* search using the amino acids sequence of *Nc*HEX-1 as the query returned 1,782 results in the RefSeq database ranging from 30 % to 99 % sequence identity. The sequence length of the proteins found was highly variable, but two major groups were identified characterized by maxima in the histogram shown in Figure 5.5. Thus, the lengths converge either around 160 to 220 or around 400 to 500 amino acids (aa).

Structurally the best studied Woronin body major protein is that of *Neurospora crassa* (Tenney et al., 2000; Yuan et al., 2003). A literature review revealed that among others the HEX-1 orthologs from species listed in Table 5.5 have been further characterized. Most studies have focused on deletion mutants and the functional role of the HEX-1 protein and the Woronin body itself. Studies focusing on the three-dimensional structure (Yuan et al., 2003) or the mechanisms underlying its self-assembly (Juvvadi et al., 2007) are underrepresented.

Table 5.5: Overview of HEX-1 orthologs which are further characterized in the literature specifying their reference.

Species	Reference
Arthrobotrys oligospora	(Liang et al., 2017)
Aspergillus flavus	(Yuan et al., 2019)
Aspergillus fumigatus	(Beck and Ebel, 2013; Beck et al., 2013; Yuan et al., 2019)
Aspergillus nidulans	(Beck and Ebel, 2013)
Aspergillus oryzae	(Beck et al., 2013; Juvvadi et al., 2007)
Colletotrichum orbiculare	(Kubo et al., 2015)
Fusarium graminearum	(Son et al., 2013)
Magnaporthe grisea	(Asiegbu et al., 2004; Soundararajan et al., 2004)
Metarhizium robertsii	(Tang et al., 2020a)
Verticillium dahliae	(Vangalis et al., 2020)
Trichoderma reesei	(Curach et al., 2004)
Zymoseptoria tritici	(Steinberg et al., 2017)

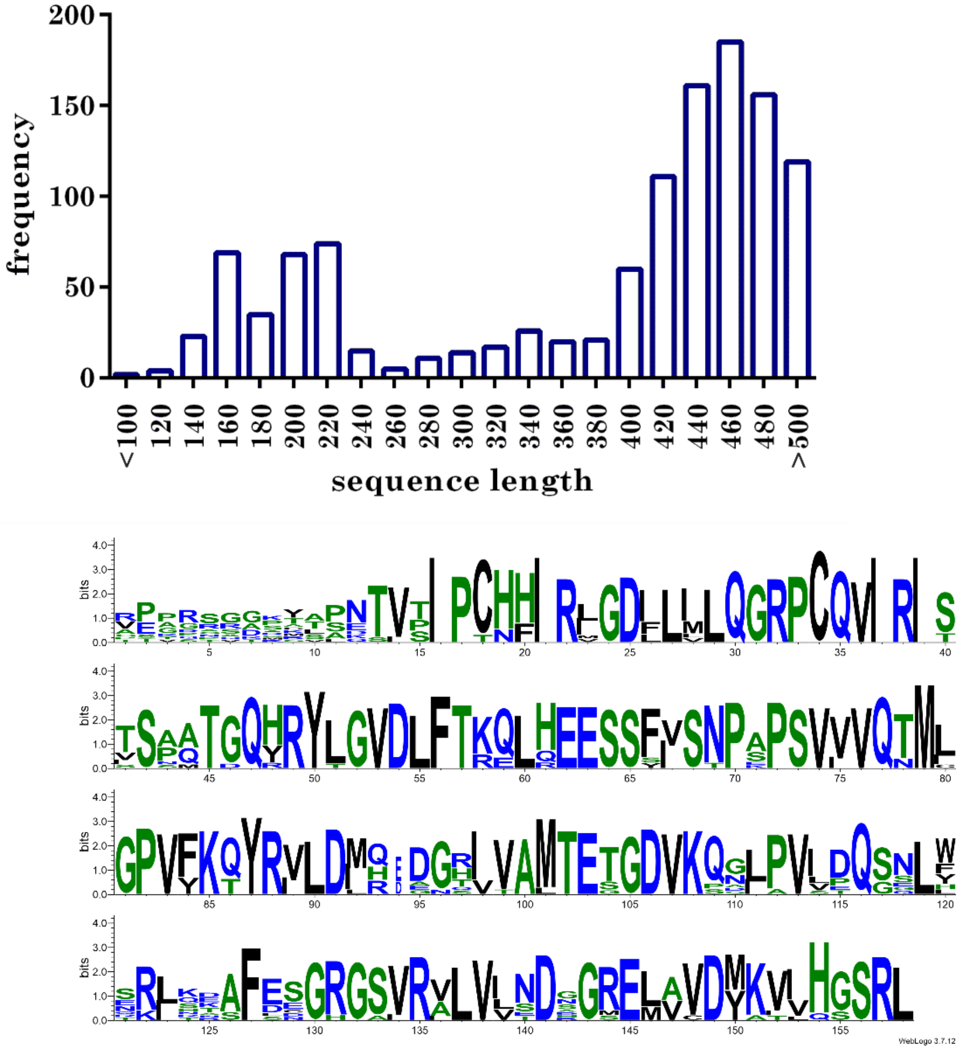


Figure 5.5: A: **Blastp** *search of the* **NcHEX-1** *protein sequence against RefSeq database revealed a high variety of sequence length of HEX-1 proteins. The most abundant sequence lengths cluster around 160 to 220 aa as well as*

around 400 to 500 aa. (B) The selected HEX-1 proteins, namely from **Aspergillus fumigatus** ***(*****Af*****),*** **Aspergillus nidulans** ***(*****An*****),*** **Aspergillus oryzae** ***(*****Ao*****),*** **Metarhizium robertsii** ***(*****Mr*****),*** **Verticillium dahliae** ***(*****Vd*****),*** **Chaetomium globosum** ***(*****Cg*****),*** **Magnaporthe grisea** ***(*****Mg*****),*** **Pyrenophora tritici-repentis** ***(*****Pt*****),*** **Purpureocilium lilacinum** ***(*****Pl*****) and*** **Verruconis gallopava** ***(*****Vg*****), and*** **Nc*****HEX1 were analyzed according to their sequence conservation in the C-terminal domain. Therefore, all 11 HEX-1 sequences were aligned using*** **Clustal Omega** ***and subsequently the last 158 C-terminal residues of each of the protein were analyzed using*** **WebLogo3** ***revealing a high sequence conservation in this domain.***

To cover the full spectrum of protein lengths and genera, HEX-1 proteins of the following 10 different species were selected to investigate their ability to crystallize spontaneously in living insect cells: *Aspergillus fumigatus* (Af), *Aspergillus nidulans* (*An*), *Aspergillus oryzae* (*Ao*), *Metarhizium robertsii* (*Mr*), *Verticillium dahliae* (*Vd*), *Chaetomium globosum* (*Cg*), *Magnaporthe grisea* (*Mg*) *Pyrenophora tritici-repentis* (*Pt*), *Purpureocilium lilacinum* (*Pl*) and *Verruconis gallopava* (*Vg*). The sequence length varies between 176 aa for *Ao*HEX-1, and 640 aa for *Pt*HEX-1 as shown in Table 5.6.

Table 5.6: Overview of 10 homologous HEX-1 proteins of different fungal species selected for this study, compared to **Neurospora crassa** ***HEX-1. Depicted is their sequence length, the origin, abbreviation and UniProt identifier.***

Organism	HEX-1 homolog	UniProt ID	Sequence length
Neurospora crassa	*Nc*HEX-1	P87252	176 aa
Aspergillus fumigatus	*Af*HEX-1	A0A0J5PW54	485 aa
Aspergillus nidulans	*An*HEX-1	Q9P8K9	221 aa
Aspergillus oryzae	*Ao*HEX-1	I8TQ26	176 aa
Chaetomium globosum	*Cg*HEX-1	Q2H4T3	173 aa
Magnaporthe grisea	*Mg*HEX-1	Q9UW16	182 aa
Metarhizium robertsii	*Mr*HEX-1	E9EN82	392 aa
Purpureocilium lilacinum	*Pl*HEX-1	A0A179H7Q1	416 aa
Pyrenophora tritici-repentis	*Pt*HEX-1	B2VT43	640 aa
Verticillium dahliae	*Vd*HEX-1	G2WVW3	197 aa
Verruconis gallopava	*Vg*HEX-1	A0A0D2A1I5	234 aa

The amino acid sequences were aligned using the *UniProt* alignment tool *Clustal Omega* for multiple sequence alignment. The resulting alignment (Appendix p. 145) shows a high degree of conservation in the C-terminal region ending with the conserved peroxisomal target sequence (PTS1) ‘SRL’, consisting of Ser-Arg-Leu. The 158 C-terminal amino acids were further analyzed for conserved motifs using the *WebLogo3* web application (Figure 5.5), confirming the high identity in the 146 C-terminal amino acids, referred to as common region, previously reported by Tang et al., (2020a). It further revealed a conserved motif, subsequently referred to as ‘heptapeptide’, including the amino acid sequence ‘M(G/A)YD(D/E/N)(D/E)’, which is encoded in 9 of the 10 investigated HEX-1 genes and have been described previously by Vangalis et al., (2020).

The analyzed sequences not only vary in their sequence but also in their length between the heptapeptide and the common region at the C-terminus. The four longest HEX-1 proteins selected here (AfHEX-1, *Mr*HEX-1, *Pl*HEX-1, and *Pt*HEX-1) even include additional amino acids N-terminal to the conserved heptapeptide. Another exception is *Vd*HEX-1, which encodes 9 additional amino acids after the peroxisomal target sequence and lacks the conserved heptapeptide. To allow proper localization of the target protein in the fungal cell, the PTS1 sequence must be C-terminal without any additional amino acids (Gould et al., 1989). Therefore, these 9 additional amino acids were not included in cloning of the respective gene.

5.2.2 Investigation of the intracellular crystallization of HEX-1 proteins from 10 different fungal species

In order to achieve high expression levels of the recombinant genes of the selected HEX-1 homologs, the gene sequences were codon-optimized for enhanced expression in *Trichoplusia ni*, commercially synthesized and inserted into pFastBac1 vectors without additional amino acids (denoted as ori) for subsequent gene expression. After generation of recombinant bacmids using DH10EmBacY cells, subsequent baculovirus production and amplification in Sf9 cells, High Five cells were infected at an MOI of one and incubated for 4 days. The infected cells were then examined using light microscopy for crystal formation. In infected cultures of six out of the ten HEX-1 homologues the formation of ordered structures was immediately visible (Figure 5.6): *Ao*HEX-1 (panel A), *An*HEX-1 (panel B), *Cg*HEX-1 (panel C), *Mg*HEX-1 (panel D), *Vd*HEX-1 (panel F), and *Vg*HEX-1 (panel G). *An*HEX-1 forms unique structures that are thin and elongated, reminding of a comet's tail (Figure 5.6 A). The morphology of *Mg*HEX-1 (Figure 5.6 D) is similar to that of *Nc*HEX-1 (Figure 5.6 E), which also forms square structures with a hexagonal base. In contrast, *Ao*HEX-1, *Cg*HEX-1, *Vd*HEX-1 and *Vg*HEX-1 form extended rodlike structures with a hexagonal cross section (Figure 5.6 B, C, F, G). Crystals of VgHEX-1 show severe growth defects (blue arrowheads Figure 5.6 G).

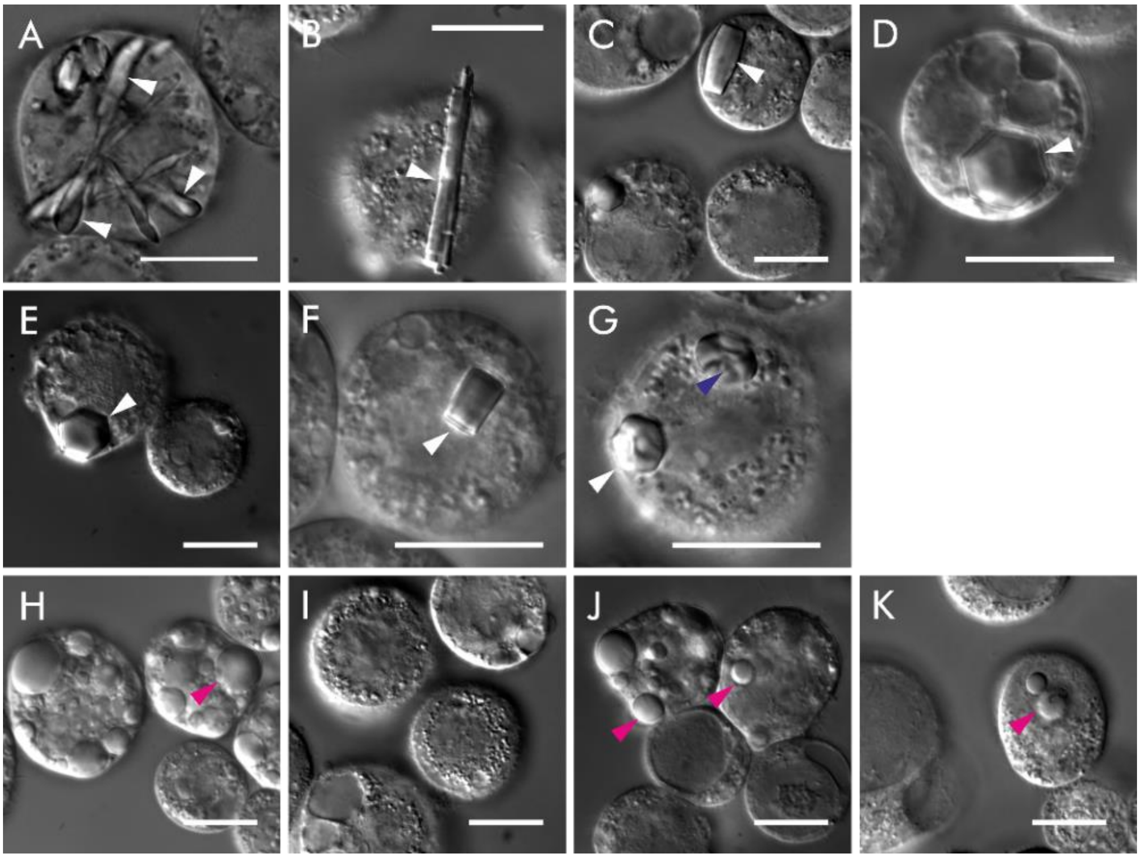


Figure 5.6: Investigation of the crystallization behavior of HEX-1 proteins from 10 different fungal species. High Five insect cells were infected at an MOI of one with the respective rbv. Cells were imaged 4 dpi using differential interference contrast (DIC) and a 100x objective on a Nikon Ts2R-FL. The scale bars correspond to 20 µm. High Five cells produce crystals looking like a comet's tail after infection with rbv **An*****HEX-1ori (A) and needle-shaped crystals after infection with rbv*** **Ao*****HEX-1ori (B). High Five cells produce spindle-like crystals with a hexagonal cross-section after infection with rbv*** **Cg*****HEX-1ori (C). High Five cells produce squared crystals with a hexagonal cross-section after infection with rbv*** **Mg*****HEX-1ori (D),*** **Nc*****HEX-1ori (E), while rectangular crystals with hexagonal cross-section are formed after infection with rbv*** **Vd*****HEX-1ori (F). High Five cells produce small hexagonal crystals with growth defects (blue arrowhead) after infection with rbv*** **Vg*****HEX-1ori (G). In High Five cells infected with rbv*** **Af*****HEX-1ori (H),*** **Mr*****HEX-1ori (I),*** **Pl*****HEX-1ori (J) or*** **Pt*****HEX-1ori (K) no ordered structures could be observed. However, larger vacuolar structures were detected in High Five cells expressing*** **Af*****HEX-1ori (H),*** **Pl*****HEX-1ori (J) or*** **Pt*****HEX-1ori (K) as indicated by magenta arrow heads.***

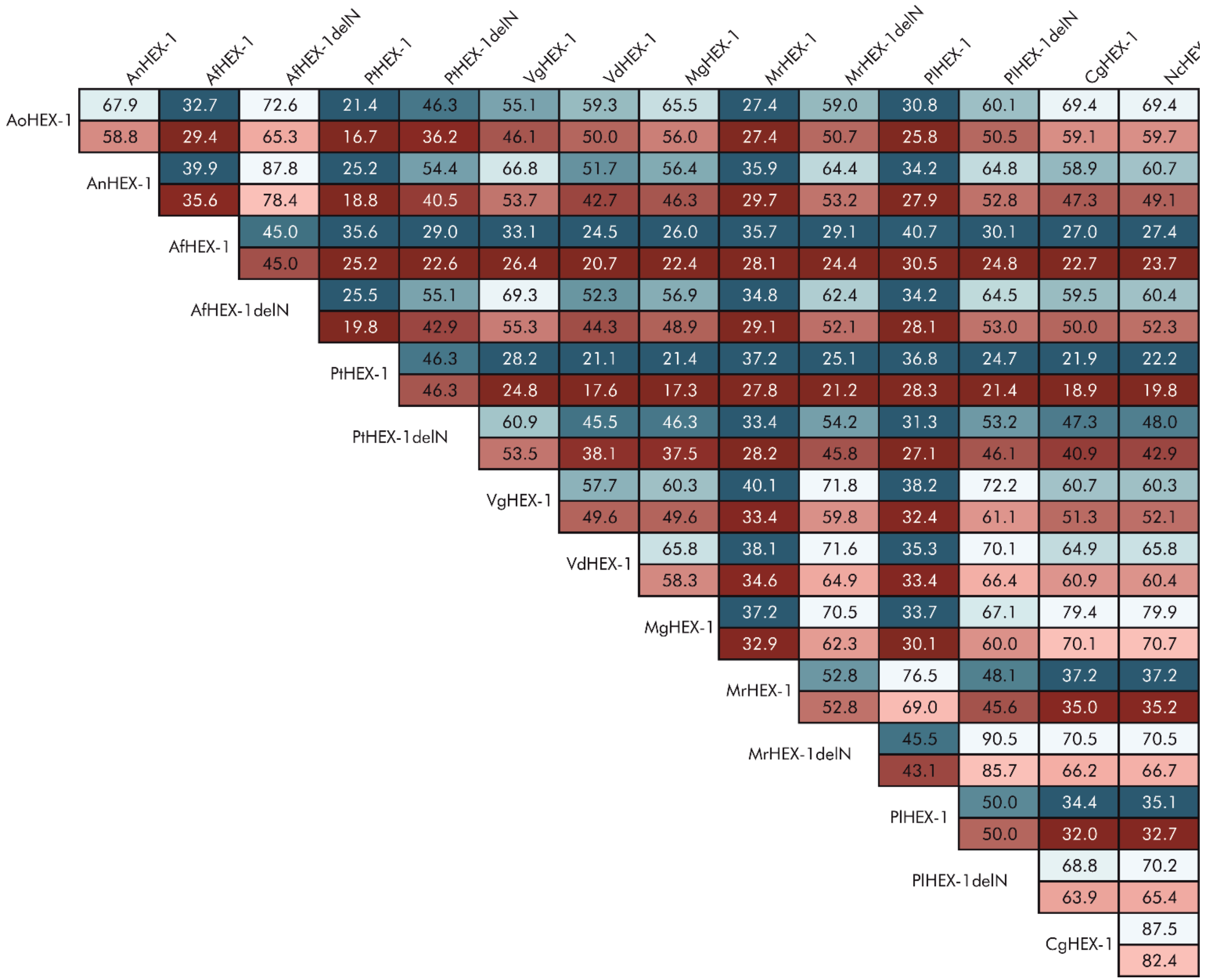
Four HEX-1 proteins, namely *Af*HEX-1, *Mr*HEX-1, *Pl*HEX-1, and *Pt*HEX-1, did not show any spontaneous growth of ordered structures inside the cells, despite a successful infection as evidenced by EYFP fluorescence. For *Af*HEX-1, *Pl*HEX-1 and *Pt*HEX-1 large, vacuolar structures were observed (Figure 5.6 H, J & K). The four homologs include an elongated N-terminus before the heptapeptide, which is discussed as a putative starting peptide for HEX-1 proteins (Vangalis et al., 2020). To test the hypothesis of incorrect translation start annotation in the database, the protein sequences lacking the N-terminus, referred to as delN-variant, were investigated regarding their spontaneous crystallization in insect cells. A truncation of the N-terminal extension results in an almost conserved length of the HEX-1-homologs of about 200 aa (Table 5.7). With a sequence length of 295 aa, *Pt*HEX-1 is still outstanding.

Table 5.7: To verify the correctness of the translation start at the conserved heptapeptide, N-terminally extended HEX-1 proteins were truncated at the N-terminus (delN). Sequence lengths of HEX-1delN proteins are compared to their original length.

Organism	HEX-1 homolog	N-terminal truncated sequence length	Original length
Aspergillus fumigatus	AfHEX-1	217 aa	485 aa
Metarhizium robertsii	MrHEX-1	206 aa	392 aa
Purpureocilium lilacinum	*Pl*HEX-1	208 aa	416 aa
Pyrenophora tritici-repentis	*Pt*HEX-1	295 aa	640 aa

Analysis of the sequence identity and similarity of the selected HEX-1 proteins and the N-terminal truncated versions led to the assumption that the translation start is indeed incorrectly annotated for the longest four of the selected HEX-1 proteins. A matrix of sequence similarity and identity is depicted in Figure 5.7. The four N-terminal extended proteins show less sequence similarity (mean sequence similarity 33.2 %) and identity (mean sequence identity 28.4 %) compared to the seven proteins that start with the conserved heptapeptide. However, the N-terminal truncation variants show a comparable sequence similarity (mean sequence similarity 62.9 %) and identity (mean sequence identity 54.5 %) to those which have been annotated to start with the N-terminal heptapeptide (mean sequence similarity 64.9 % and identity 55.9 %).

Briefly, the gene sequence between the encoded heptapeptide and the stop codon was amplified by PCR and restriction sites for *Bam*HI and *Hind*III were added to subsequently ligate the digested product into a similarly digested pFastBac1 vector. Bacmids and recombinant baculoviruses were prepared as described above. Microscopic examination of rbv-infected High Five cells revealed growth of ordered structures for three of the four HEX-1delN proteins (Figure 5.8): AfHEX-1delN, *Mr*HEX-1delN and *Pl*HEX-1delN. Only for *Pt*HEX-1delN still no ordered structures were observed. *Pl*HEX-1delN and AfHEX-1delN showed a squared morphology with a hexagonal cross section as observed for *Nc*HEX-1 and others, while *Mr*HEX-1delN produced an elongated shape with a hexagonal base comparable to *Ao*HEX-1.



	AnHEX-1	AfHEX-1	AfHEX-1delN	PtHEX-1	PtHEX-1delN	VgHEX-1	VdHEX-1	MgHEX-1	MrHEX-1	MrHEX-1delN	PlHEX-1	PlHEX-1delN	CgHEX-1	NcHEX
AoHEX-1	67.9 58.8	32.7 29.4	72.6 65.3	21.4 16.7	46.3 36.2	55.1 46.1	59.3 50.0	65.5 56.0	27.4 27.4	59.0 50.7	30.8 25.8	60.1 50.5	69.4 59.1	69.4 59.7
AnHEX-1		39.9 35.6	87.8 78.4	25.2 18.8	54.4 40.5	66.8 53.7	51.7 42.7	56.4 46.3	35.9 29.7	64.4 53.2	34.2 27.9	64.8 52.8	58.9 47.3	60.7 49.1
AfHEX-1			45.0 45.0	35.6 25.2	29.0 22.6	33.1 26.4	24.5 20.7	26.0 22.4	35.7 28.1	29.1 24.4	40.7 30.5	30.1 24.8	27.0 22.7	27.4 23.7
AfHEX-1delN				25.5 19.8	55.1 42.9	69.3 55.3	52.3 44.3	56.9 48.9	34.8 29.1	62.4 52.1	34.2 28.1	64.5 53.0	59.5 50.0	60.4 52.3
PtHEX-1					46.3 46.3	28.2 24.8	21.1 17.6	21.4 17.3	37.2 27.8	25.1 21.2	36.8 28.3	24.7 21.4	21.9 18.9	22.2 19.8
PtHEX-1delN						60.9 53.5	45.5 38.1	46.3 37.5	33.4 28.2	54.2 45.8	31.3 27.1	53.2 46.1	47.3 40.9	48.0 42.9
VgHEX-1							57.7 49.6	60.3 49.6	40.1 33.4	71.8 59.8	38.2 32.4	72.2 61.1	60.7 51.3	60.3 52.1
VdHEX-1								65.8 58.3	38.1 34.6	71.6 64.9	35.3 33.4	70.1 66.4	64.9 60.9	65.8 60.4
MgHEX-1									37.2 32.9	70.5 62.3	33.7 30.1	67.1 60.0	79.4 70.1	79.9 70.7
MrHEX-1										52.8 52.8	76.5 69.0	48.1 45.6	37.2 35.0	37.2 35.2
MrHEX-1delN											45.5 43.1	90.5 85.7	70.5 66.2	70.5 66.7
PlHEX-1												50.0 50.0	34.4 32.0	35.1 32.7
PlHEX-1delN													68.8 63.9	70.2 65.4
CgHEX-1														87.5 82.4

Figure 5.7: Similarity and identity matrix of the selected 11 HEX-1 orthologs including the N-terminal truncated variants. The matrix indicates the similarity (upper value, shaded in blue) and the identity (lower value, shaded in red) when comparing the protein sequences of the selected HEX-1 homologs. The darker the color, the less similar/identical are the sequences. The least identical and least similar protein sequences to all others is observed for N-terminal extended HEX-1 proteins like AfHEX-1, MrHEX-1, PlHEX-1, and PtHEX-1.

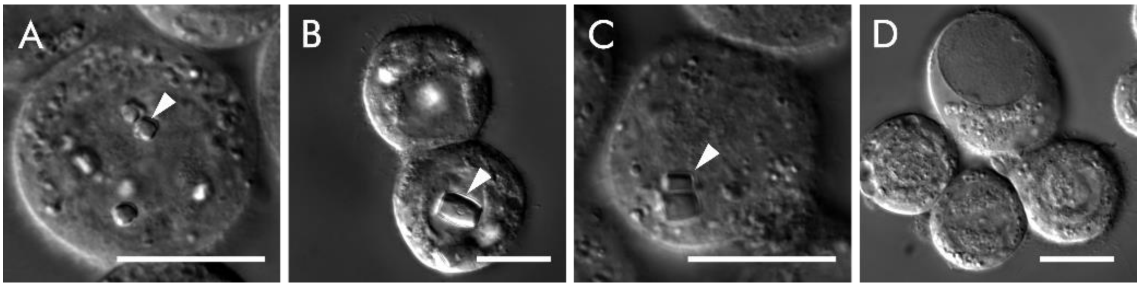


Figure 5.8: Investigation of the crystallization behavior of N-terminal truncated HEX-1 proteins from 4 different fungi species. High Five insect cells infected at an MOI of one of the recombinant baculoviruses were imaged 4 dpi using differential interference contrast (DIC) and a 100x magnifying objective at a Nikon Ts2R-FL. The scale bar corresponds to 20 µm. High Five cells produce small hexagonal crystals after infection with rbv AfHEX-1delN ori (A) and rectangular crystals with a hexagonal cross-section after infection with rbv MrHEX-1delN ori (B). Rbv PlHEX-1delN ori-infected High Five cells produce small, squared crystals with a hexagonal cross-section (C), while no indication of crystallization was observed in rbv PtHEX-1delN-infected High Five (D).

The analysis of the size and the estimation of the volume of the structures formed by all tested HEX-1 proteins displayed a high variability (Table 5.8). While *Af*HEX-1delN produces the smallest structures in terms of volume, expression of *Cg*HEX-1 leads to structures with the largest volume. All proteins formed crystals in 61 to 93 % of the cells, enabling the diffraction data collection in due time.

Table 5.8: Overview on crystal dimensions of 10 different HEX-1 proteins. The dimensions of the ordered structures were determined by measuring an average of 50 structures per protein on microscopic images taken by using a 100x magnifying objective.

Protein	Length [µm]	Width [µm]	Volume [µm³]
*Nc*HEX-1 ori	9.07 ± 2.53	8.82 ± 2.93	906.1 ± 891.8
*An*HEX-1 ori	12.31 ± 5.34	2.61 ± 1.04	114.0 ± 150.9
AoHEX-1 ori	24.70 ± 10.01	4.50 ± 1.87	578.6 ± 547.7
*Vd*HEX-1 ori	14.98 ± 5.59	5.78 ± 1.71	635.9 ± 554.3
CgHEX-1 ori	16.15 ± 5.53	6.78 ± 1.87	918.4 ± 823.5
*Vg*HEX-1 ori	7.59 ± 2.62	4.99 ± 3.11	358.6 ± 682.7
*Mg*HEX-1 ori	8.57 ± 2.48	8.21 ± 2.44	701.2 ± 536.2
*Pl*HEX-1delN ori	4.97 ± 1.46	4.89 ± 1.55	150.3 ± 140.8
*Mr*HEX-1delN ori	9.60 ± 4.22	4.85 ± 1.95	309.3 ± 370.3
AfHEX-1delN ori	3.29 ± 1.26	3.39 ± 1.42	60.3 ± 146.0

5.2.3 Diffraction data collection of HEX-1 proteins

To confirm the crystalline state of the HEX-1 proteins that formed ordered structures, they were tested for their ability to diffract X-rays. High Five cells were infected with recombinant baculoviruses encoding the crystallizing proteins at an MOI of one and harvested at 4 dpi. Crystal-containing cells were loaded onto MeshMounts, cryo-protected and immediately flash-frozen in liquid nitrogen for storage. Frozen MeshMounts were automatically placed at the goniometer at the PETRA III beamline P14 operated by EMBL Hamburg at DESY. Diffraction tests revealed visually observed resolution between 2 Å and 3 Å for all samples, allowing the collection of a complete, serial data set of each of the proteins by applying helical line scans. The diffraction data were processed with *CrystFEL*, which is described in detail in chapter 5.1.2.

While the *in cellulo* crystallization as part of the *InCellCryst* pipeline has already been optimized (Schönherr, 2021), the process of data collection, reduction, processing, and model building is still in its infancy. Thus, besides solving the problem of contaminant diffraction resulting from ice and salt crystals in the sample, optimizations regarding the data collection procedure were investigated.

Prior to the recent establishment of intracellular diffraction (Boudes et al., 2016), crystals had to be isolated before data collection (Gati et al., 2014; Nass et al., 2020). Spreading of isolated crystals on a MeshMount for data collection results in a preferred orientation of the crystals (Zander et al., 2015). Therefore, data collection was performed using helical line scans with 1.0° of oscillation per frame to ensure recording a complete data set. With the advent of *in cellulo* diffraction (Boudes et al., 2016), the diffraction parameters were not changed, but when the crystals are still inside the cells, they have a larger degree of freedom to orient themselves on the MeshMount, especially if the crystals do not exceed the diameter of the cell. This raised the question of whether the data collection parameters established so far were still optimal.

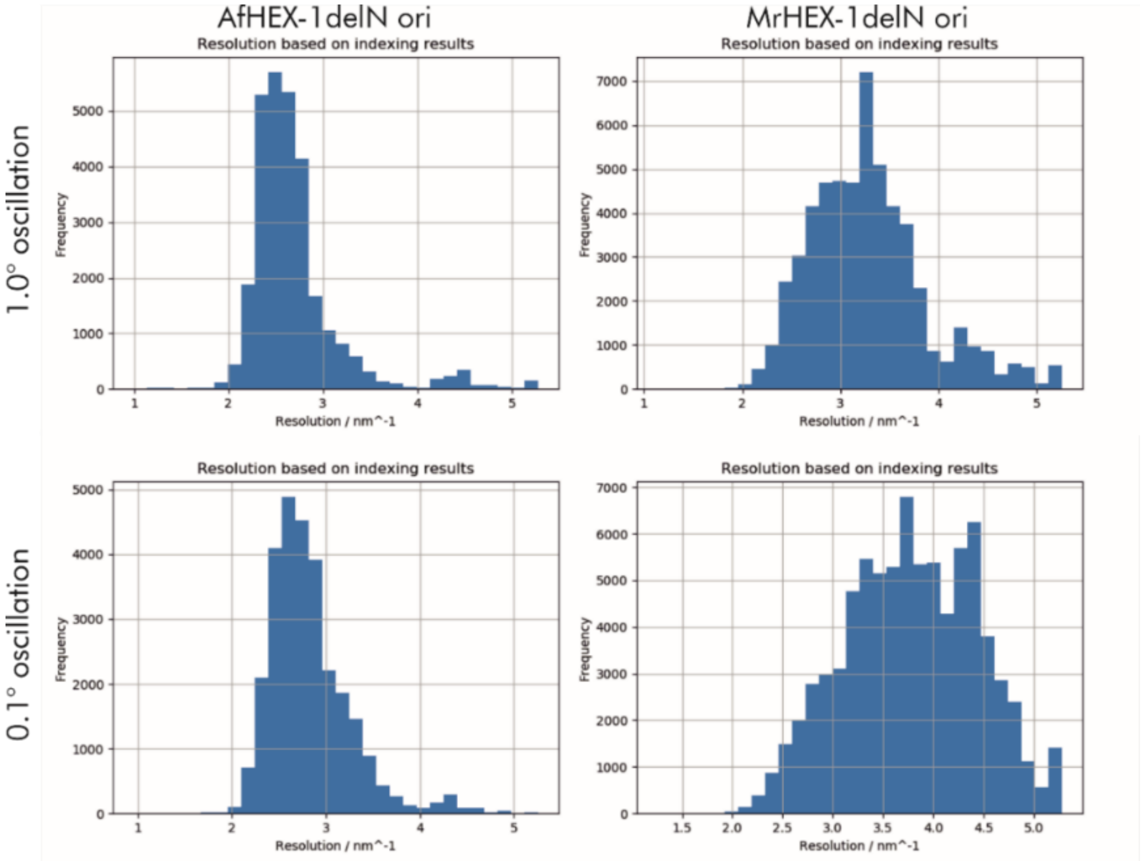


***Figure 5.9: Histograms of resolution distribution of successfully indexed crystals of* Af*HEX-1delN ori and* Mr*HEX-1delN ori measured at EMBL beamline P14 at PETRA III at DESY, Hamburg by applying different oscillation angles of the MeshMount during data collection using helical line scans. The y-axis shows the number of detector patterns against the maximum observed resolution on the x-axis in nm^{-1}. The histograms were prepared by the* CrystFEL *script* ave_resolution.**

A large oscillation range per image during data collection can lead to a loss of usable reflections due to overlapping reflections on the detector, and thus can lead to a loss of information (Rupp, 2010). On the other hand, too small oscillation during data collection may result in incomplete (partial) reflections or in a low number of reflections observed. The HEX-1 protein crystals, which were measured here, have small unit cell parameters, resulting in a larger spacing of the resulting Bragg spots on the detector. With a smaller unit cell, the risk of overlapping reflection lunes is higher. Consequently, *Af*HEX-1delN ori and *Mr*HEX-1delN ori were recorded with a reduced oscillation of 0.1° and a smaller spacing between adjacent frames was tested. The exposure time of 0.02 s per image was not changed.

***Table 5.9: Comparison of diffraction capabilities of different oscillation angles for* Af*HEX-1delN ori and* Mr*HEX-1delN ori. An 0.1° oscillation gave higher mean resolution values for both crystals,* Mr*HEX-1delN and* Af*HEX-1delN. However, changes in the values of the best and worst diffracting crystals are minor between the different oscillations. The statistics are exported by the* CrystFEL *script* ave_resolution.**

	AfHEX-1delN ori		MrHEX-1delN ori	
Oscillation	1.0°	0.1°	1.0 °	0.1°
Mean	3.7 Å	3.5 Å	3.1 Å	2.6 Å
Best	1.9 Å	1.9 Å	1.9 Å	1.9 Å
Worst	10 Å	11 Å	8.8 Å	8.0 Å
Std deviation	0.49 nm^{-1}	0.44 nm^{-1}	0.6 nm^{-1}	0.66 nm^{-1}

Analysis of the diffraction data recorded using intracellular crystals of two different proteins (*Mr*HEX-1delN and *Af*HEX-1delN) at 1° and 0.1° oscillation showed that an oscillation of 0.1° was the better choice. All four data sets were indexed using *CrystFEL* and analyzed using the *ave-resolution* script, which outputs a histogram of the observed per-image resolution (Figure 5.9) as well as some statistics on the highest resolution spots of the indexed patterns shown in Table 5.9. The resolution of the best diffracting crystal is equal for both data sets of the same protein. However, the number of diffraction patterns with higher resolution is higher when an oscillation of 0.1° is used. The data sets collected at 0.1° oscillation of both proteins were subjected to phasing and model building. Their structures are discussed in the following (chapter 5.2.4).

For both proteins analyzed, an oscillation of 0.1° seems to be the better choice if resolution is to be optimized. However, for structure elucidation, the redundancy and completeness should also be considered. If the oscillation is smaller, the rotational space of the data collected by the same number of images is not fully covered, the resulting data set may be incomplete. As shown in Table 5.10, the data sets are complete in all four cases, but the redundancy is slightly lower for an oscillation of 0.1° than for 1.0°, due to the smaller rotational space covered by the combined diffraction patterns. The high degree of symmetry of the space group $P6_522$ is favorable in this case. However, this should be considered in the case of a sample with low crystallization efficiencies or lower space group symmetry.

Table 5.10: Abstract of the data processing statistics of Mr**HEX-1delN **and Af**HEX-1delN **collected using either an oscillation of 1.0° or 0.1° per frame. Values in parentheses refer to the highest resolution shell.

	AfHEX-1delN	AfHEX-1delN	MrHEX-1delN	MrHEX-1delN
Rotation	1.0°	0.1°	1.0°	0.1°
Resolution (Å)	63.35 - 1.97 (2.27 – 2.25)	63.35 - 2.00 (2.02 - 2.00)	66.10 - 1.90 (1.92 - 1.90)	66.10 - 1.72 (1.73 - 1.72)
Total reflections	6,091,067	3,304,618	26,479,476	19,214,102
Unique reflections	9,873 (464)	13,815 (649)	16,392 (810)	21,819 (1,065)
Multiplicity	617.0	239.2	1,615.4	880.6
Completeness (%)	100.00 (100.00)	99.64 (97.30)	100.00 (100.00)	99.98 (99.72)
SNR	13.74 (0.69)	12.57 (0.59)	19.12 (0.50)	21.92 (0.70)
Wilson B-factor (Å^2)	29.88	22.95	30.71	28.38
R_{split} (%)	6.65 (163.14)	8.97 (219.66)	4.30 (213.32)	3.38 (151.75)
CC1/2	0.9975 (0.4246)	0.9959 (0.2227)	0.9984 (0.1919)	0.9988 (0.2594)
CC*	0.9994 (0.7720)	0.9990 (0.6036)	0.9996 (0.5674)	0.9997 (0.6418)
No. collected images	85,665	75,544	78,380	82,410
No. hits/ indexed crystals	51,878/28,741	41,616/28,424	51,183/54,601	61,895/80,380

5.2.4 Structure elucidation of four HEX-1 proteins

For the N-terminal deletion variants the oscillation during data collection was optimized as described in further detail in chapter 5.2.3. However, structure elucidation was only possible for 4 of the HEX-1 homologs, namely *Af*HEX-1delN, *Mg*HEX-1, *Mr*HEX-1delN, and *Pl*HEX-1delN, due to a misalignment of beam and detector that could not be resolved during data processing. When the detector is tilted, the refinement for the position of the incoming beam does not converge and the data cannot be processed properly. Due to limited access to synchrotron sources, a second data collection was not possible.

After merging the data, molecular replacement with *phenix.phaser* was performed using the conventionally crystallized structure of *Nc*HEX-1 as search model (pdb: 1khi). Subsequently, cycles of automated refinement with *phenix.refine* and manual refinement using *coot* yielded models of the HEX-1 ori structures of *Af*HEX-1delN, *Mg*HEX-1, *Mr*HEX-1delN, and *Pl*HEX-1delN, (Figure 5.10). Complete data processing and refinement statistics are shown in Table 5.11 along with the residues for which interpretable electron density was obtained.

The overall fold is conserved among all HEX-1 proteins tested, especially in the so-called common region at the C-terminus (last 148 aa at C-terminus) and reveals a two-domain structure similar to the *in situ* crystallized *Nc*HEX-1 structure (pdb: 1khi). The N-terminal ß-barrel consists of six ß-sheets, while the C-terminal domain contains of five ß-sheets forming a ß-barrel and two α-helices (Yuan et al., 2003). *Pl*HEX-1delN and *Mr*HEX-1delN also have highly ordered N-termini with two additional ß-sheets formed by the first amino acids, for *Pl*HEX-1delN residues 2 to 19 and for *Mr*HEX-1delN residues 2 to 4 and 10 to 13 accordingly. The N-terminus also contains parts of the conserved heptapeptide motif. No electron density was observed for the 30 or 49 residues, respectively between the N-terminus and the common region at the C-terminus in these two HEX-1 structures indicating a high flexibility in this region. The electron density obtained for the other two HEX-1 data sets did only reveal interpretable electron density for the common region. In the following the numbering of residues refers to *Nc*HEX-1 to ease reading. For the other HEX-1 proteins, refer to the sequence alignment in the Appendix (p. 145).

The crystal lattice and contact interfaces were also found to be highly conserved in the structures of the HEX-1 homologs modeled in this study. Analysis of the interface between the monomers in a crystal using *PDBePISA* revealed, similar to what has been observed for *Nc*HEX-1, two types of interfaces: one which is responsible for the self-assembly and one resulting from the crystal packing. The first one encounters the stronger interactions and is observed in all HEX-1 structures presented here. This interface includes the formation of salt bridges between Arg41 and Asp44 as well as His39 and Asp44 (residue numbers refer to *Nc*HEX1 cyto but are conserved in the HEX-1 proteins investigated here). The interfaces of *Pl*HEX-1delN show two interfaces corresponding to the self-assembly function, the second one is characterized by salt bridges between Arg68 and Glu81.

Another type of interface encountered in all analyzed HEX-1 structures analyzed here is characterized by weaker interactions, mainly hydrogen bonds, but also includes a salt bridge between Arg57 and Asp110.

Superposition of the solved homologous structures with the *in cellulo* crystallized *Nc*HEX-1 cyto (pdb: 7asx) reveals a high degree of conservation in the overall motif, especially in the residues 31 to 170 as also observed by sequence analysis depicted in Figure 5.6. The structural deviations are minor with a maximum mean RMSD of 0.7 Å (Cα atoms) in all analyzed HEX-1 proteins and a maximum RMSD of 2.3 Å (Cα atoms) in *Pl*HEX-1delN ori in flexible loop structures connecting the individual ß-sheets. However, a higher degree of variability was observed in N- and C-terminal residues as well as in the loops connecting adjacent secondary structure elements.

The target protein selection based on sequence orthology was successful and revealed a high overall structural conservation of the HEX-1 proteins presented in this study (Figure 5.11), indicated by an overall Q-score of 0.7184 as calculated by *PDBeFold*.

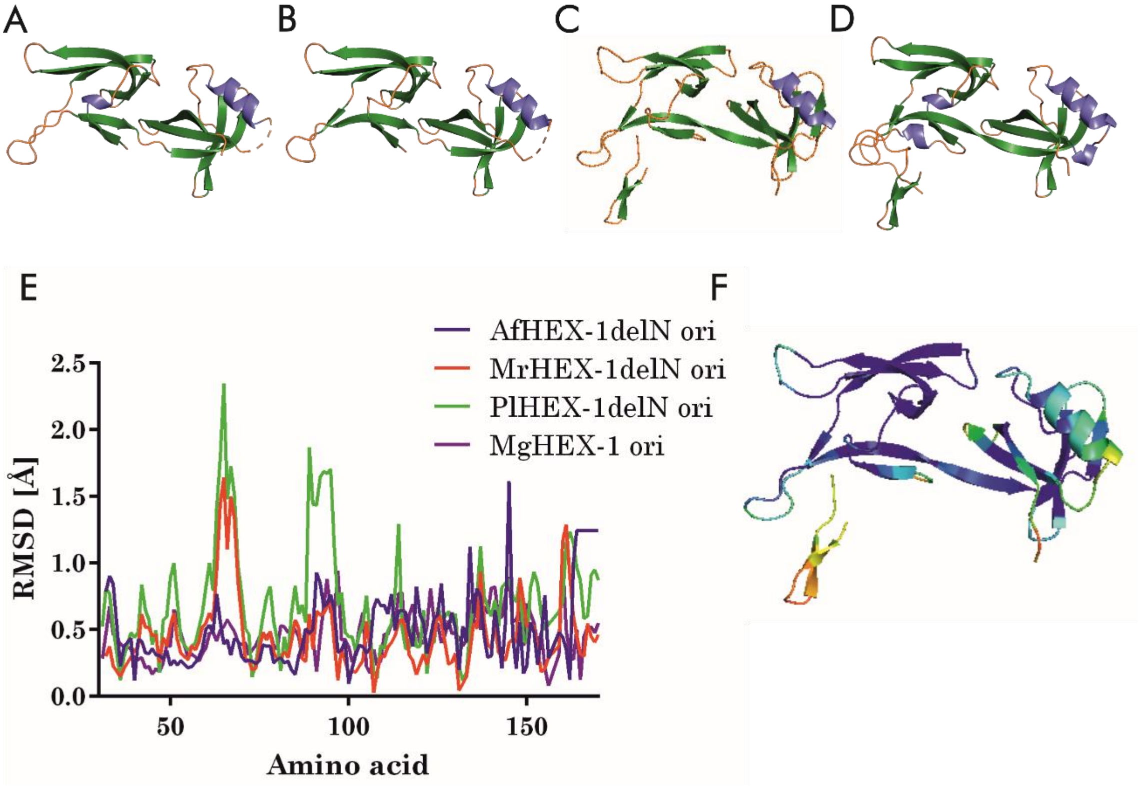
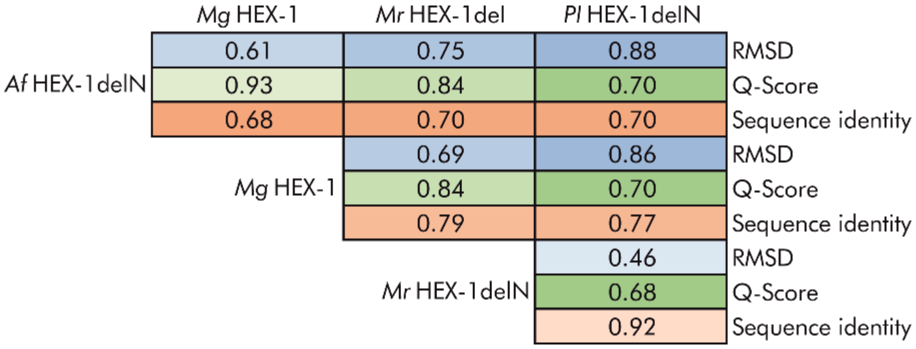


Figure 5.10: Overview of the solved HEX-1 protein structures of selected species: Af*HEX-1delN (A),* Mg*HEX-1 ori (B),* Mr*HEX-1delN ori (C) and* Pl*HEX-1delN ori (D). β-sheets are displayed in green, while a-helices are shown in blue. RMSD of Ca atoms plotted against the amino acid position of all HEX-1 structures determined by superposition to* Nc*HEX-1 crystallized* in cellulo *and measured at RT at a free electron laser (pdb: 7asx) (E). The differences observed are minor, but larger differences were observed in the loop regions e.g., for residues 60-66 indicated by a RMSD above 1.5 Å. 3D structural model of* Pl*HEX-*

1delN ori (F). The structure is colored according to the atomic displacement factor (B-factor). Blue indicates a low B-factor, while orange indicates larger flexibility by an increased B-factor. The distribution of B-factors was comparably observed in all other HEX-1 structures.



	Mg HEX-1	*Mr* HEX-1del	*Pl* HEX-1delN	
Af HEX-1delN	0.61	0.75	0.88	RMSD
	0.93	0.84	0.70	Q-Score
	0.68	0.70	0.70	Sequence identity
Mg HEX-1		0.69	0.86	RMSD
		0.84	0.70	Q-Score
		0.79	0.77	Sequence identity
Mr HEX-1delN			0.46	RMSD
			0.68	Q-Score
			0.92	Sequence identity

***Figure 5.11:** Comparison of the obtained structural models of four HEX-1 homologs, namely* Af*HEX-1delN,* Mg*HEX-1,* Mr*HEX-1delN, and* Pl*HEX-1delN. The matrix displays the mean RMSD of Ca atoms in Å, the Q-score of the respective structures and the similarity of their sequences as calculated by* PDBeFold *after multiple comparison. The Q-score, ranging between 0 and 1, considers the alignment length, the total number of residues in the structures to be compared and the RMSD of Ca atoms. A high value indicates a high structural conservation. The level of similarity is displayed by the color intensity, namely the higher the similarities of the structures the lighter the coloring of the cell.*

Table 5.11: Data processing and refinement statistics for the N-terminal truncated HEX-1 proteins originating from* Aspergillus fumigatus, Metarhizium robertsii, Purpureocilium lilacinum *and the full-length HEX-1 protein originating from* Magnaporthe grisea. *Values in parentheses refer to the highest resolution shell.

Data set	AfHEX-1delN ori	MrHEX-1delN ori	*Pl*HEX-1delN ori	MgHEX-1 ori
Data collection date	04.07.2022	04.07.2022	04.07.2022	24.08.2021
Sample temperature [K]	100	100	100	100
Processing software	CrystFEL	CrystFEL	CrystFEL	CrystFEL
Oscillation	0.1°	0.1°	1.0°	1.0°
Space group	$P6_522$	$P6_522$	$P6_522$	$P6_522$
Unit cell (Å or °)	58.36 58.36 190.05 90 90 120	57.65 57.65 198.29 90 90 120	57.10 57.10 200.34 90 90 120	57.66 57.66 187.89 90 90 120
Resolution range (Å)	63.35 - 2.00 (2.04 - 2.02)	66.10 - 1.72 (1.73 - 1.72)	66.78 - 1.97 (1.99 - 1.97)	93.95 - 1.80 (1.82 - 1.80)
Total reflections	3,301,177	19,214,102	17,726,635	31,949,243
Unique reflections	13,432 (633)	21,819 (1,065)	14,631 (684)	18,144 (855)
Multiplicity	245.8	880.6	1,211.5	176.9
Completeness (%)	99.64 (97.30)	99.98 (99.72)	100.00 (100.00)	100.00 (99.88)
SNR	12.92 (0.67)	21.92 (0.70)	18.27 (0.62)	25.86 (0.59)
Wilson B-factor	29.66	28.27	31.70	25.69
R_{split} (%)	8.54 (218.99)	3.38 (151.75)	5.07 (185.29)	3.34 (177.35)
CC1/2	0.9962 (0.2190)	0.9988 (0.2594)	0.9989 (0.1924)	0.9993 (0.2583)
CC*	0.9990 (0.5994)	0.9997 (0.6418)	0.9997 (0.5681)	0.9998 (0.6407)
No. collected images	75,544	82,410	102,476	110,030
No. hits/ indexed lattices	41,616/28,424	61,895/80,380	49,318/50,341	59,042/62,496
Refl. used in refinement	13,270 (1,206)	21,620 (1,988)	14,531 (1,409)	17,256 (1322)
Refl. used for R-free	1,323 (117)	1,673 (154)	1,453 (141)	1,401 (108)
R-work	0.2179 (0.3585)	0.1932 (0.3470)	0.2023 (0.3552)	0.2130 (0.4174)
R-free	0.2325 (0.3343)	0.2261 (0.3639)	0.2398 (0.3759)	0.2364 (0.4764)
No. non-hydrogen atoms	1,114	1,249	1,420	1,126
Macromolecules	1,045	1,156	1,343	1,033
Solvent	69	93	77	93

Data set	AfHEX-1delN ori	*Mr*HEX-1delN ori	*Pl*HEX-1delN ori	MgHEX-1 ori
Protein residues	137	150	175	139
RMS bonds (Å)	0.002	0.006	0.003	0.009
RMS angles (°)	0.45	0.84	0.66	1.05
Ramachandran favored (%)	100.00	98.56	98.83	98.52
Ramachandran allowed (%)	0.00	1.41	1.17	1.48
Ramachandran outliers (%)	0	0	0	0
Rotamer outliers (%)	0	0	0	1.77
Clash score	2.88	42.18	0.76	6.81
Average B-factor	38.60	38.05	39.58	40.54
Macromolecules	38.12	37.43	39.46	40.05
Solvent	45.91	45.83	41.63	45.92
TLS refinement groups	2	3	3	3
Mean RMSD (Å) to ref. structure (pdb: 7asx)	0.477	0.454	0.705	0.464
Max RMSD (Å) to ref. structure (pdb: 7asx)	1.614	1.643	2.348	1.046
Phasing model	1khi	1khi	1khi	1khi
Sequence identity to 1khi	69.38	67.79	66.83	71.98
Visible electron density was obtained for residues	73 – 190; 195 – 203; 205 – 215	2 – 4; 10 – 13; 62 – 205	2 – 19; 49 – 207	36 – 154; 157 – 177

5.3 Overcoming statistical limitations by identification of promising candidates from sequence paralogs for *Tb*IMPDH

Paralogous proteins result from gene duplication in the same species and may lead to a broad variety of functions of the resulting proteins which diverge during evolution and specify their functional properties. As the *Tb*GMPR structure of Imamura et al. (2020) displayed a high structural similarity to the *Tb*IMPDH structure this protein was subjected to investigation of its intracellular crystallization ability. While the similarity in structure is very high (mean RMSD of Cα atoms around 1.50 Å and 85 % of matched secondary structure elements calculated by *PDBeFold*), the sequence identity is low with 35 % identical amino acids (calculated by *PDBeFold*). Crystals of *Tb*IMPDH obtained by *in cellulo* methods revealed native cofactor binding (Nass et al., 2020), which have not yet been identified for *Tb*GMPR. Consequently, here, *Tb*GMPR was subjected to the established *InCellCryst* crystallization pipeline to investigate whether the *Tb*GMPR would also form similar crystals inside the cells, enabling the identification of natively bound co-factors. Revealing information about the native cofactors involved and their binding sites will broaden the understanding of enzyme regulation and function and open another possibility for structure-based drug design against African sleeping sickness.

5.3.1 Crystallization and first diffraction tests of *Trypanosoma brucei* GMPR

The intracellular crystallization of *Tb*GMPR has been previously described (Kardoost et al., 2023; Schönherr, 2021). High Five cells infected with rbv *Tb*GMPR cyto v2 produce long, needle-shaped crystals with a square base of 42.5 ± 14.8 µm in length and 4.7 ± 1.5 µm in width (corresponding to a crystal volume of 938.8 µm^3) in roughly 50 to 80 % of the cells 4 dpi (Figure 5.12 A). The morphology and size of the crystals is comparable to *Tb*IMPDH ori crystals that are 39.7 ± 11.9 µm in length and 4.6 ± 1.7 µm in width (corresponding to a crystal volume of 840.1 µm^3) (Figure 5.12 B).

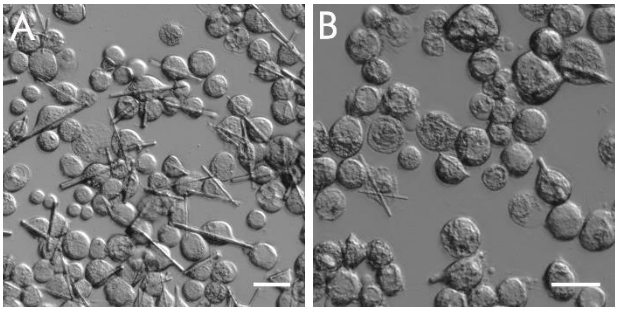


Figure 5.12: Detection of crystals in High Five insect cells infected at an MOI of one of the rbv* Tb*GMPR cyto v2 (A) and the rbv* Tb*IMPDH ori (B) 4 dpi applying Emboss contrast (EC) and a 25x magnifying objective at a Nikon Ts2R-FL. The scale bar corresponds to 40 µm.

Despite several attempts of X-ray diffraction data collection and subsequent processing, the three-dimensional structure of *Tb*GMPR cyto v2 remained to be solved. A data set collected at the beamline P14 operated by EMBL (PETRA III, DESY) under cryo-conditions in 2019 was processed using *CrystFEL* (Table 5.12). The unit cell dimensions derived, indicate a large tetragonal unit cell (a = b = 168.7 Å, c = 280.0 Å, α = β = γ = 90°), resulting in a narrow spacing of the Bragg peaks on the detector. A tetragonal space group was also observed for *Tb*IMPDH ori, with a smaller unit cell (a = b = 207.10 Å, c = 92.50 Å) containing two monomers of the protein in the asymmetric unit. The resolution of the obtained *Tb*GMPR cyto v2 data set was below 4.5 Å. This resolution is not sufficient for cofactor identification in a calculated structural model and can either be improved by using an X-ray source of higher intensity or by using larger crystals. Due to limited access to XFEL beamtime, a method for growing larger crystals had to be developed.

Table 5.12: Data processing statistics for* Tb*GMPR cyto v2 crystals measured in intact High Five cells at EMBL beamline P14 at PETRA III, DESY, Hamburg, at 100 K using MeshMounts, which were raster scanned by helical lines. Values in parentheses refer to the highest resolution shell.

Data set	*Tb*GMPR cyto v2
Data collection date	26.11.2019
Sample temperature (K)	100
Processing software	CrystFEL
Oscillation (°)	1
Space group	P422
Unit cell (Å resp. °)	168.7 168.7 280.0 90 90 90
Resolution range (Å)	280.00 - 4.51 (4.55 - 4.51)
Total reflections	58,271,456
Unique reflections	24,622 (1,206)
Multiplicity	2,366
Completeness (%)	100 (100)
SNR	3.24 (0.34)
R-split (%)	23.37 (339.27)
CC1/2	0.9901 (0.2482)
CC*	0.9975 (0.6306)
No. collected images	44,795
No. hits/indexed lattices	34,120/10,076

5.3.2 Development of a strategy to overcome crystal size limitations and resolution limits

As mentioned in chapter 5.3.1, intracellularly grown crystals of *Tb*GMPR cyto v2 have not been suitable for structure elucidation due to the comparatively low resolution of the diffraction data and the very large unit cells. Because of the close spacing of Bragg reflections, more constructive interference maxima must be collected to achieve comparable resolution to samples with smaller unit cells. The intensity of interference maxima decreases by their order (Rupp, 2010). Thus, to be able to detect further interference maxima the intensity of these reflections needs to be enhanced.

According to Darwin's Formula (chapter 1.2), the intensity of a reflection correlates with several parameters, some of which are related to the radiation source and the diffraction setup, while other parameters are defined by the crystal system (Darwin, 1914; Holton and Frankel, 2010). The ratio of unit cell to crystal volume, for example, directly correlates with the observed reflection intensity, and an increased crystal volume directly affects the signal-to-noise ratio, leading to an improved peak detection especially at high resolution. Thus, the crystal volume is a suitable parameter to be increased to enable structure elucidation of *Tb*GMPR cyto v2 at higher resolution.

The cell as a crystallization chamber has several limitations, and factors influencing the crystal growth are difficult to determine. The protein concentration inside the cell has been identified as a critical factor that was improved by choosing High Five cells over Sf9 cells for protein production (Koiwai et al., 2019; Schönherr, 2021; Wilde et al., 2014). The application of different substances, for example Brefeldin A, that leads to the accumulation of proteins in the ER by disassembly of the Golgi apparatus (Fujiwara et al., 1988), can also increase the crystal size (Harms, 2019; Hasegawa et al., 2011).

Although the cell diameter does not limit the size of the crystal (Schönherr et al., 2015), cellular resources are a limiting factor for the size a crystal can reach. Manipulation of cell size can be achieved by interfering with the regulatory factors involved or by fusing cells (Lentz, 2007; Lentz and Lee, 1999a), the latter was investigated in this study to increase crystal size. Cell fusion can be induced by applying electric fields (electrofusion) (Takamori et al., 2022; Zimmermann, 1982), by heating nanoparticles in the membrane, resulting in membrane melting (hot-particle fusion) (Rørvig-Lund et al., 2015), by inducing conformational changes in viral, fusogenic proteins involved in membrane-fusion (Gottesman et al., 2010; Harrison, 2015), or by adding high molecular weight (2000 to 4000 kDa) polyethylene glycol (PEG) (Lentz, 2007; Lentz and Lee, 1999b).

PEG-mediated cell fusion is highly reproducible and lacks immunogenicity (Lentz and Lee, 1999a). PEG acts as a dehydrating agent depleting the hydration shell of two adjacent membranes. This results in a closer contact of the membranes and, if the outer leaflet is either disrupted or highly curved, the membranes fuse. Consequently, lipids and cytosols mix and a syncytium, a cell of increased size with two or more nuclei is formed (Lentz, 2007; Lentz and Lee, 1999b).

By coincidence, a preliminary result had shown, that PEG-mediated cell fusion works with rbv-infected High Five cells, representing the basis for the subsequently described experiments (personal communication R. Schönherr).

5.3.3 Establishment of PEG-mediated cell fusion to increase the crystal volume

The size of the syncytia depends on the number of fused cells and defines the resources available for protein production and subsequent crystal formation. Thus, the syncytia size is tightly connected to the process of crystallization. For fusion induction PEG4000 was used in this study.

Possible factors influencing the syncytia size include the concentration of PEG4000 used, the cell density, the timing of PEG4000 addition, as well as the method of PEG4000 application. PEG4000 can be applied as a homogeneous solution to the entire well. However, preliminary results showed that rbv-infected High Five cells fuse, when a 50 % stock solution of PEG4000 is added as a drop in the center of the well (personal communication R. Schönherr).

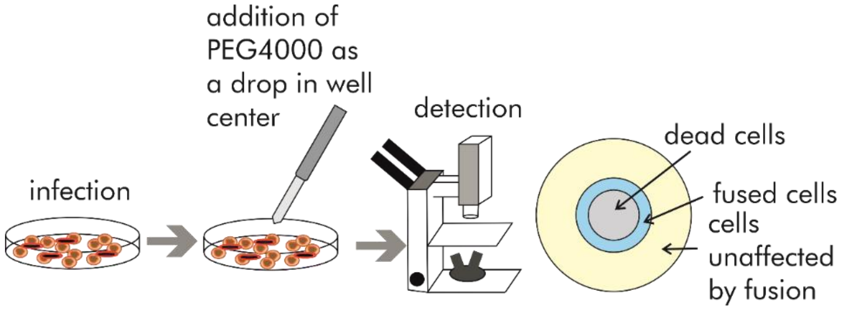


Figure 5.13: Workflow to induce cell fusion of baculovirus infected High Five cells. For fusion induction, 10 µL of the PEG4000 solution is added one to two days pi to the center of the well. Within several minutes, cell fusion can be observed in a ring with a specific distance around the inoculation site. Inside the ring the cells die due to high PEG4000 concentration and osmotic shock. Outside the ring, cells show no signs of cell fusion.

Both approaches were tested here, but after application of PEG4000 1 dpi as a homogeneous solution at various concentrations ranging from 50 % to 3.125 %, cells were reduced in size and showed a hypertonic morphology with no evidence of EYFP fluorescence and crystal growth 4 dpi. When PEG4000 was added in the same way at 2 or 3 dpi, crystals that had already formed were bent (Figure 5.14 A), indicating a negative effect of PEG4000 on crystal stability and integrity. Addition of PEG4000 as a drop establishes, due to its high viscosity, a diffusion gradient from the center to the edge of the well. Three distinct zones have been identified in the well, as depicted in Figure 5.13. In the outermost zone cells are unaffected by the addition of PEG4000 and behave as usual. Cells in the innermost zone appear either unaffected by fusion (applied stock solution concentration below or equal to 10 %) or, depending on the concentration of PEG4000 applied, exhibit the hypertonic morphology with a reduced cell diameter and granular appearance (applied stock solution concentration above 10 %) (Figure 5.14 B). The fusion zone was identified in between these zones and is characterized by the appearance of fused cells of various sizes (Figure 5.14 C). Therefore, PEG4000 was added as a drop in the center of the well in all further experiments.

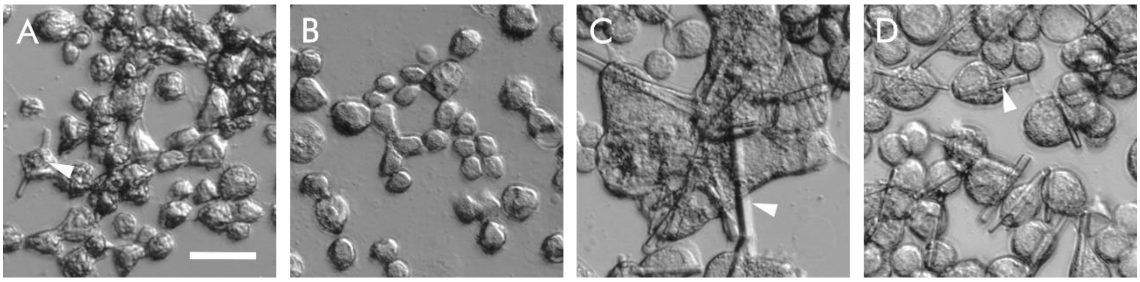


Figure 5.14: Light microscopic images of rbv* Tb*GMPR cyto v2 infected High Five cells with 25x magnifying object imaged 4 dpi. 25 % PEG4000 solution was added 2 dpi as a homogenous solution resulting in bending of already formed crystals and a small, granular phenotype (A). Cells in the innermost ring display a shrunken phenotype if 50 % PEG4000 was applied in the center of the well (B). When 25 % PEG4000 solution was added 2 dpi as a drop in the center of the well, a ring-like zone of fused, crystal-containing cells was observed (C). Control well of rbv* Tb*GMPR cyto v2 infected High Five cells without PEG4000 (D). The scale bar corresponds to 40 µm. Crystals are marked by an arrowhead.

In a first experiment 10 µL of 50 % PEG4000 was added in the center of the well 1, 2 and 3 dpi and the fusion zone was microscopically examined 4 dpi. The syncytia formed after fusion induction at 1 dpi are smaller than those after fusion induction at 2 dpi or 3 dpi, while the crystal size is maximized after fusion induction at 2 dpi (Figure 5.15 D-G).

To investigate the impact of the concentration of PEG4000 used, different concentrations of the stock solution were added to the cells at 1 dpi (23 hpi) and 2 dpi (47 hpi). At 4 dpi, samples were microscopically examined. Independent of the day of PEG4000 addition or the PEG4000 concentration, syncytia were formed in the fusion zone. All investigations described in this chapter focus on the behavior of this fusion zone.

The crystal size observed in cells that were fused by a stock solution of either 10 or 50 % PEG4000 is comparable (Figure 5.15 A, B & D-F). The timing of PEG addition is identified as a crucial parameter. However, regardless of whether fusion induction is performed 1 dpi or 2 dpi, the crystal volume of fused cells is higher than of unfused cells (Figure 5.15 H & Table 5.13).

Also, regardless of the concentration of the PEG4000 stock solution applied (between 10 and 50 %) to induce fusion, the crystal volumes of fused cells are comparable (Figure 5.15 H & Table 5.13). To minimize negative effects of PEG4000 on crystal stability and integrity and to achieve the same approx. 5-fold increase in crystal growth a 10 % PEG4000 stock solution was used in subsequent experiments.

Table 5.13: Tabular comparison of the average mean crystal volume enlargement when fusion is induced by PEG4000 stock solutions between 10 and 50 % 1 or 2 dpi of the three replicates. The crystal length and width were measured in microscopic images and the volume estimated by multiplying the length by the square of the width. In each replicate the relative crystal volume increase is standardized against the mean value of the respective control. The given mean values were calculated from the individual values of all replicates after standardization.

Concentration of PEG4000 stock solution	Time points of fusion induction	Mean crystal volume increase	Standard deviation
Control	-	1.11	0.77
50 %	1 dpi	4.63	2.98
50 %	2 dpi	5.83	4.90
40 %	1 dpi	4.16	2.92
40 %	2 dpi	5.27	4.20
30 %	1 dpi	4.55	3.54
30 %	2 dpi	4.85	3.71
20 %	1 dpi	4.35	2.92
20 %	2 dpi	5.72	4.48
10 %	1 dpi	3.96	3.31
10 %	2 dpi	5.73	5.09

Next, a closer look on the influence of cell density and the time point of PEG4000 addition was taken. Therefore, six different cell densities between 0.5 and 1.0 million cells per 6-well and four different time points between 23 hpi and 47 hpi were systematically analyzed as described above, again performed in three replicates. Confirming previous results, the higher the cell density and the later the cells were fused, the larger the resulting syncytia, however this was not true for the obtained crystal volume (Figure 5.16).

A cell density of 0.8 million cells per 6-well and PEG4000 application at 39 hpi were found to be the optimal conditions for growing *Tb*GMPR cyto v2 crystals of the largest dimensions (Figure 5.17 & Table 5.14). A maximum increase of crystal volume of more than 40-fold was observed at these conditions compared to the mean reference crystal size. The optimal conditions showed an average mean increase in crystal volume of approx. 7, which, according to Darwin's Formula, should result in an increase in intensity of the observed diffraction spots by the same factor and thus an increase in resolution should be observable.

However, it was observable, that due to the infective stress, cells tend to detach from the bottom of the well especially if they are also fused. This means that in some cases the cell in the fusion zone appears to be lower and thus the number of syncytia investigated in the respective replicate and condition is reduced.

Based on the observation of cells in the fusion zone that did not exhibit EYFP fluorescence and did not fuse, the conditions for cell fusion induction developed here were also applied to uninfected and rbv-infected Sf9 cells. These cells did also not fuse under the optimized cell fusion conditions, indicating that additional virus-associated factors are involved in the process of cell fusion.

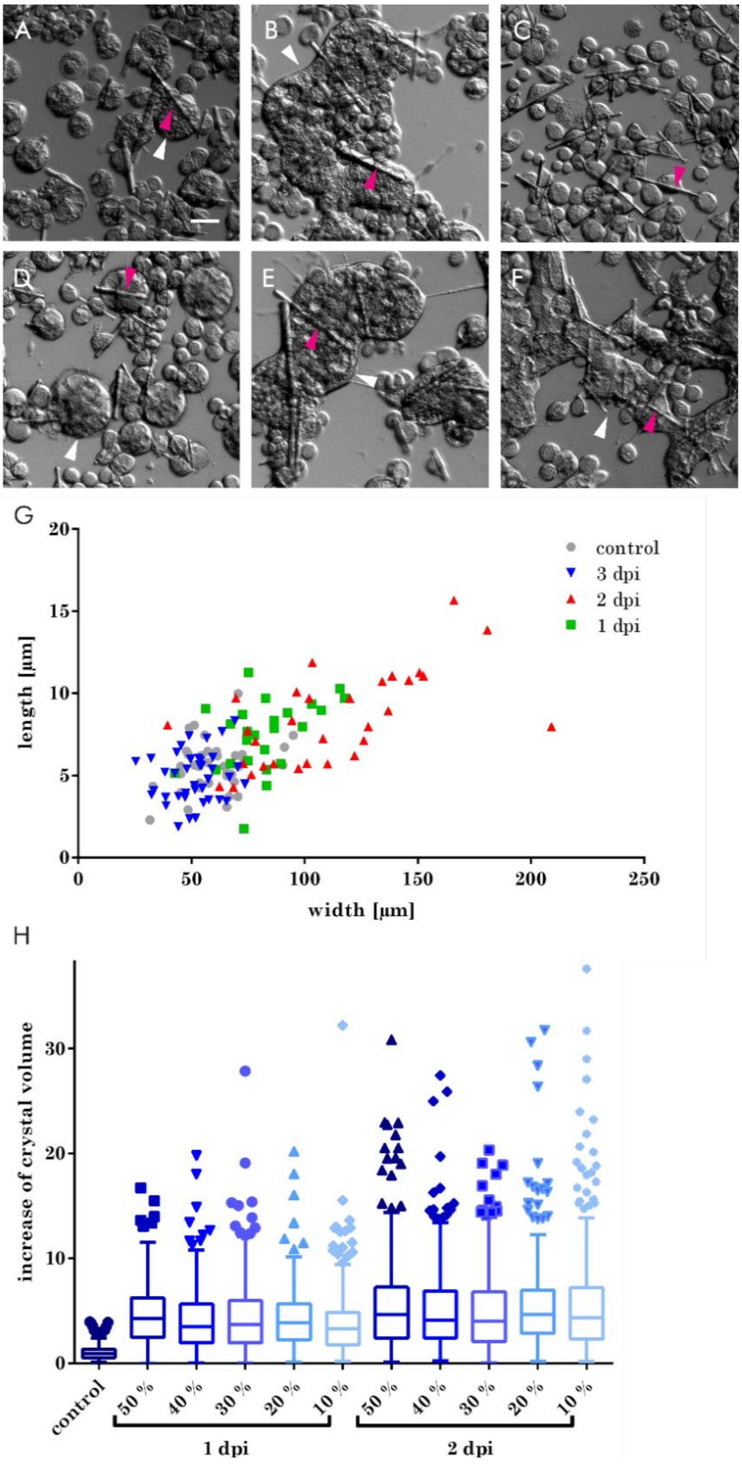


Figure 5.15: Investigation of the crystallization behavior of Tb*GMPR cyto v2 produced in High Five insect cells infected at an MOI of one of the rbv (A-F).*

Cells were imaged at 4 dpi applying Emboss contrast (EC) at a Nikon Ts2R-FL. The scale bar corresponds to 40 µm and applies to panels A-F. Fused cells are marked by white arrow heads, while crystals are marked by magenta arrow heads. PEG-mediated cell fusion was induced by adding 10 µL PEG4000 (10%) as a drop into the center of the well either 1 dpi (A) or 2 dpi (B). No PEG4000 was added and thus no cell fusion was induced (C). PEG-mediated cell fusion was induced by adding 10 µL PEG4000 (50%) as a drop into the center of the well 1 dpi (D), 2 dpi (E) or 3 dpi (F). The dimensions of the crystals in fused cells were measured and are shown as a scatter plot for the individual time points tested for fusion induction, when 50 % PEG4000 is added as a drop in the center of the well. When fusion is induced at 3 dpi, the resulting crystal dimensions are close to those of unfused cells (approximately 50 µm in length and 5 µm in width). The largest increase in crystal volume can be obtained by fusion induction at 2 dpi (resulting in between 100 and 200 µm in length and 5 to 15 µm in width). A moderate enlargement in the crystal volume can be seen when fusion is induced at 1 dpi (up to 120 µm in length and 12 µm in width). Fusion of rbv-infected High Five cells leads to an increase in length and width of the crystal (G). The diffractive volume was estimated by multiplying the length by the square of its width. The experiment was conducted in three replicates. To compare the replicates, the crystal volume of one replicate was normalized to the mean crystal volume of the respective control and the crystal volume increase of all individual crystals of the three replicates was combined. Box plots show the crystal volume increase relative to the control (H). Tukey outliers (McGill et al., 1978) are marked as single points. The box plot line marks the median of the measured crystal volume increase.

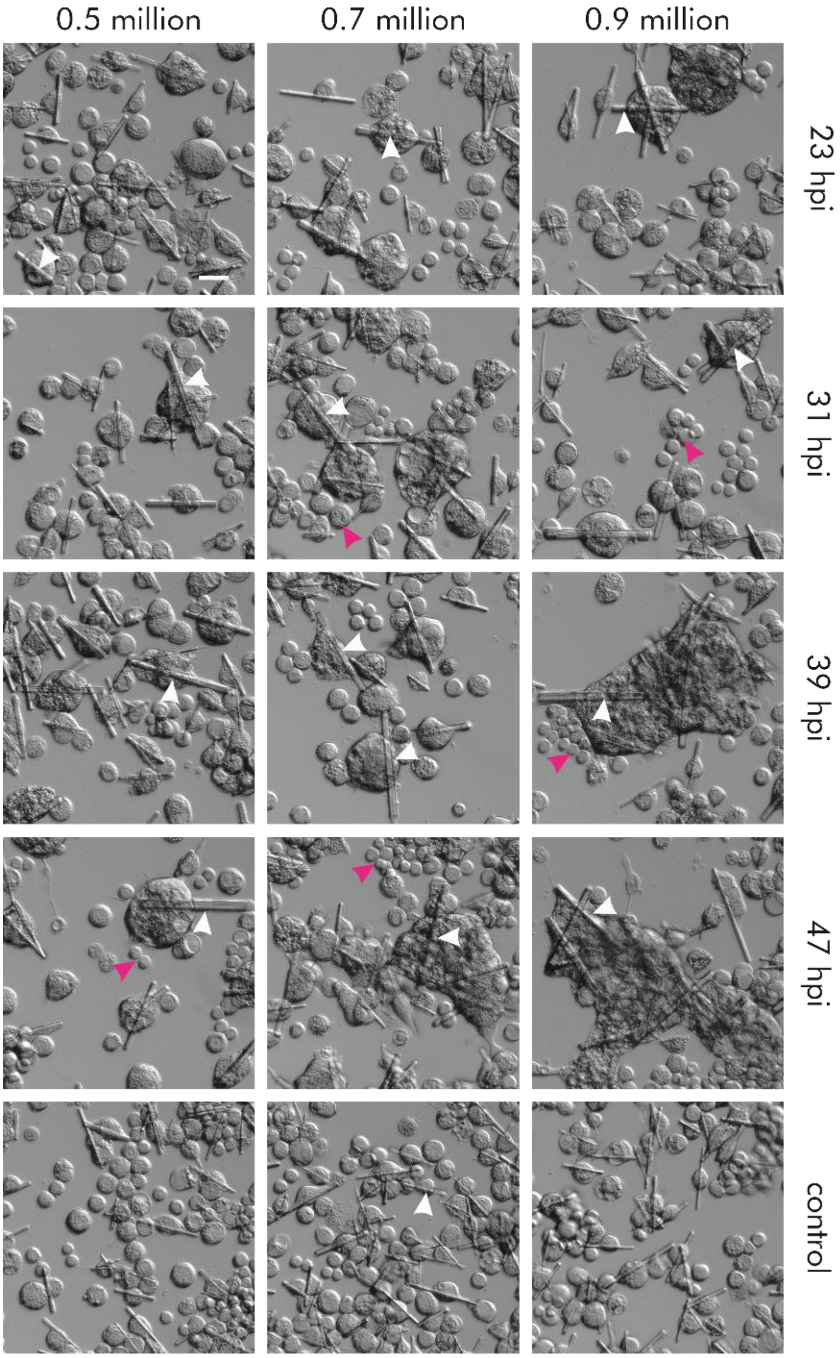


Figure 5.16: Selected bright-field microscopic pictures of a systematic screening for the optimal conditions of cell fusion induction to enable the largest growth of* Tb*GMPR cyto v2 crystals. High Five cells at different densities, ranging from 0.5 million to 1.0 million cells per six well, were infected at an MOI of one of the rbv* Tb*GMPR cyto v2. PEG-mediated cell fusion was induced at different time points ranging from 23 hpi to 47 hpi. Crystal growth was investigated 4 dpi using a 25x magnifying objective and applying Emboss contrast (EC). The scale bar corresponds to 40 µm. White arrow heads point to crystals inside fused cells, while arrow heads in magenta display residual unfused cells, which are also present in the fusion ring. The later infected High Five cells are fused and the higher the cell density, the larger the syncytia grow.

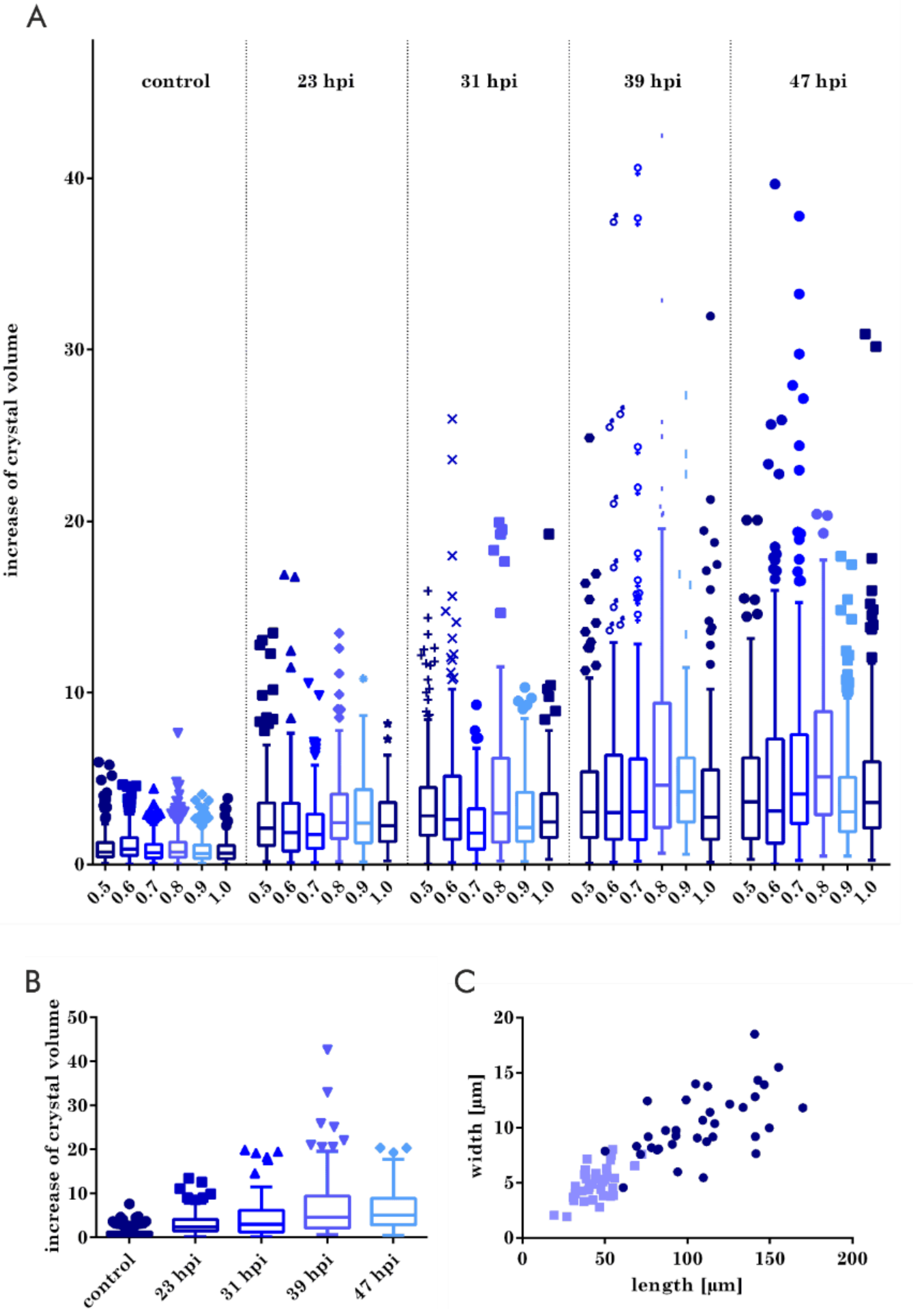


Figure 5.17: Fusion of baculovirus infected High Five cells increases the volume of Tb*GMPR cyto v2 crystals. Box plots, prepared as described above, display the crystal volume increase relative to the control at a cell density of 0.8 million cells per 6-well (A). Outliers and the median are marked as described above. Systematic screening for the largest crystal volume was performed by testing 0.5 to 1.0 million cells per 6-well and inducing fusion between 23 hpi and 47 hpi. A general tendency of increased crystal volume is*

observed when fusion is induced 39 hpi or 47 hpi, while a cell density of 0.8 million cells per well showed regardless of the timepoint of fusion induction the largest crystal enlargement. Therefore, the cell density of 0.8 million cells per well is inspected closer and the comparison of different timepoints of fusion induction revealed a maximum for 39 hpi (B). Crystal volume is enlarged on average by a factor of approx. 7 (mean) at the conditions, considered as optimal. The measured crystal dimensions are shown under optimal conditions (fusion induction at 39 hpi at 0.8 million cells per well) depicted in dark blue and control conditions (0.8 million cells per well without PEG4000) depicted in light blue (C). Crystals enlarged in their length are also increased in their width.

Table 5.14: Tabular presentation of the mean crystal increase of **Tb*GMPR cyto v2 crystals and its standard deviation (relative to the crystal volume observed in the respective replicate at 0.8 million cells) observed at 4 dpi for 6 different cell densities tested against four time points of PEG4000 addition between 23 hpi and 47 hpi. The same results are presented as box plots in Figure 5.17. The experiment was conducted in three replicates.***

Time points of PEG4000 addition	Cell density	Mean crystal volume increase	Standard deviation
-	0.5 million	1.00	0.90
-	0.6 million	1.17	0.90
-	0.7 million	0.87	0.68
-	0.8 million	1.01	0.94
-	0.9 million	0.85	0.72
-	1.0 million	0.81	0.65
23 hpi	0.5 million	2.77	2.41
23 hpi	0.6 million	2.59	2.66
23 hpi	0.7 million	2.25	1.82
23 hpi	0.8 million	3.22	2.55
23 hpi	0.9 million	3.00	2.31
23 hpi	1.0 million	2.56	1.61
31 hpi	0.5 million	3.67	2.99
31 hpi	0.6 million	3.98	4.02
31 hpi	0.7 million	2.32	1.80
31 hpi	0.8 million	4.23	4.00
31 hpi	0.9 million	3.16	2.47
31 hpi	1.0 million	3.22	2.57
39 hpi	0.5 million	4.04	3.62
39 hpi	0.6 million	4.68	5.02
39 hpi	0.7 million	4.98	5.96
39 hpi	0.8 million	6.99	6.83
39 hpi	0.9 million	5.10	4.50
39 hpi	1.0 million	4.50	5.12
47 hpi	0.5 million	4.63	3.77
47 hpi	0.6 million	5.21	5.80
47 hpi	0.7 million	6.33	6.28
47 hpi	0.8 million	6.40	4.36

Time points of PEG4000 addition	Cell density	Mean crystal volume increase	Standard deviation
47 hpi	0.9 million	4.06	3.25
47 hpi	1.0 million	4.84	4.44

Resuming from these experiments, the cell size has to be considered for further diffraction experiments. Cells are pipetted onto the MeshMount, requiring the large syncytia, which were observed to be stable in the cell culture dish, to survive the shear forces at the opening of the pipette tip. Because the crystal sizes were not maximized at the highest cell density (1.0 million cells) and the latest time point (47 hpi) studied, the cell size was still manageable and not limiting the size of the intracellular grown crystals.

To collect diffraction data of crystals from fused and unfused High Five cells infected with rbv GMPR cyto v2, they were harvested at 4 dpi by gently removing the cells from the entire well and transferring them to 1.5 mL tubes. Cells were allowed to form a light pellet by gravity and loaded onto a MeshMount, cryoprotected and flash-frozen in liquid nitrogen. Data acquisition was performed at the P14 beamline operated by EMBL, Hamburg, located at PETRA III, DESY, by applying helical line scans.

5.3.4 Diffraction data processing and model building of *Tb*GMPR

Processing of diffraction data was performed using *CrystFEL*. The derived unit cell parameters after indexing with *mosflm* were the same as for crystals from unfused cells (chapter 5.3.1), with 168.7 Å for the axes a and b and 280.0 Å for the c-axis (space group P422), determined by invoking the indexing settings *mosflm-nolatt-nocell*, followed by *mosflm-latt-nocell*. This implies that the cell fusion has no effect on the crystal lattice. Since all angles are 90° and axis a and b are almost identical, the spots were indexed in a tetragonal lattice (Figure 5.18 A). Intensities were merged, scaled and partialities modelled, resulting in a resolution cut-off of 2.75 Å, which is defined by the highest resolution shell where the signal-to-noise ratio and the CC* value are above 0.5 and the completeness is above 95 %. Although 11 different IMPDH monomers from various species were tested as search models (pdb: 6i0o, 6i0m, 6gjk, 6gjv, 4z87, 5tc3, 5mcp, 3usb, 3tsd, 3tsb and 6rfu) no molecular replacement solution was found that could be refined to a R_{free} below 45 %.

Therefore, the already published GTP-*Tb*GMPR structure (pdb: 6lk4, Imamura et al., 2020) was used as a search model. This revealed a model with 6 monomers of *Tb*GMPR in the asymmetric unit. Due to the symmetry of the crystal, this describes the biological unit of an octamer. However, after several cycles of manual and automatic refinement, R_{free} remained static above 32 % and atomic B factors above 100 Å^2, indicating problems in the space group determination and/or the possibility of twinning. The octamer composition was also unusual as shown in Figure 5.18 C, because chains B and C, as well as E and F, each formed dimers, which resulted in octamers due to the symmetry operation. Chain A, however, formed its own octamer, as did chain D (octamers in green and yellow Figure 5.18 C).

Therefore, the data were reindexed, merged, and scaled in space group P1 (Figure 5.18 B), resulting in a data set at a resolution of 3.19 Å. Molecular replacement using the published *Trypanosoma brucei* GMPR structure (pdb: 6lk4) enabled the refinement of a model consisting of 48 monomers building 3 full octamers and additional 16 partial octamers (completed by periodic repetition of the motif), as indicated by an R_{free} of 0.2335 (Figure 5.20 D & Table 5.15).

The refined model was loaded into *Zanuda* (Lebedev and Isupov, 2014) (part of *ccp4i*) which analyzes the symmetry of the molecules and performs refinements in all possible space groups. The algorithm then suggests a possible space group with its maximum symmetry considering the R-factors after refinement. Although the best R-factors were obtained for P121, *Zanuda* still suggested P422 as the optimal solution. Because of this uncertainty, the structure was finally refined and modeled in space group P1 (Table 5.15).

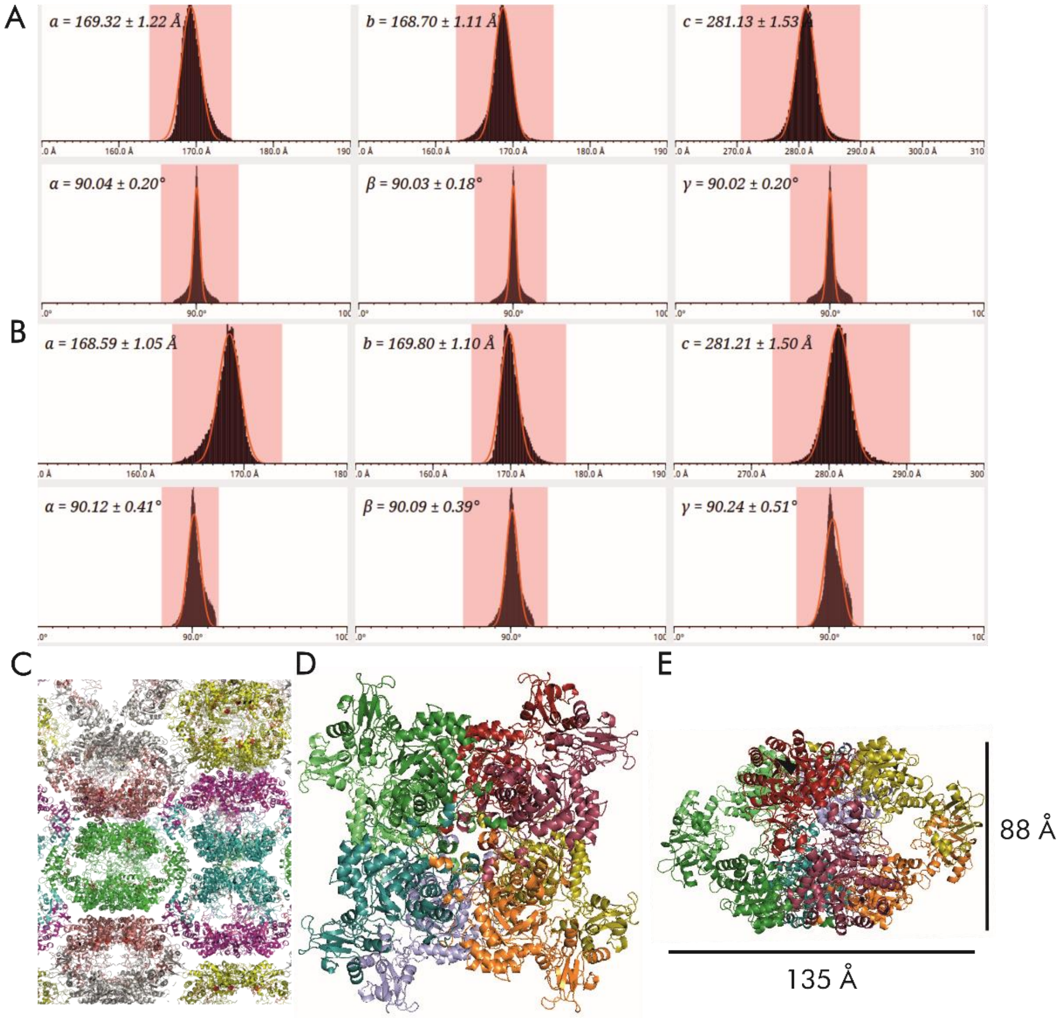


Figure 5.18: Histogram of unit cell parameters of the Tb*GMPR cyto v2 diffraction data after indexing with* CrystFEL *by applying a tetragonal lattice (A), indexed with* mosflm, xds *and* xgandalf, *or a triclinic lattice (B),*

indexed with* xgandalf*, reported by* cell_explorer*. Cartoon representations of the structural model of* Tb*GMPR cyto v2 (C-E). Crystal packing of* Tb*GMPR cyto v2 monomers in case of a tetragonal lattice colored by chain identity (chain A - green, chain B - blue, chain C - pink, chain D - yellow, chain E – pale pink, chain F - grey) (C). Top view (D) and side view (E) of the octamer containing chain A and C to I of the* Tb*GMPR cyto v2 structure, solved in space group P1.

The solved structure showed unusually high B-factors (Figure 5.21 E) despite a good electron density map (Figure 5.19 A). Following the optimization approach described by Sawaya et al. (2014), map sharpening improved the details visible in the electron density map (Figure 5.19 B) and real-space refinement against the sharpened map improved the B-factors to be in a reasonable range at around 40 Å^2 (Table 5.12). Superposition of the Cα atoms of the A chains of the real-space refined structure with the usually refined structure revealed a max. RMSD of 0.712 Å, excluding that significant structural changes were induced by the real-space refinement. Thus, both structural models are considered to be highly similar.

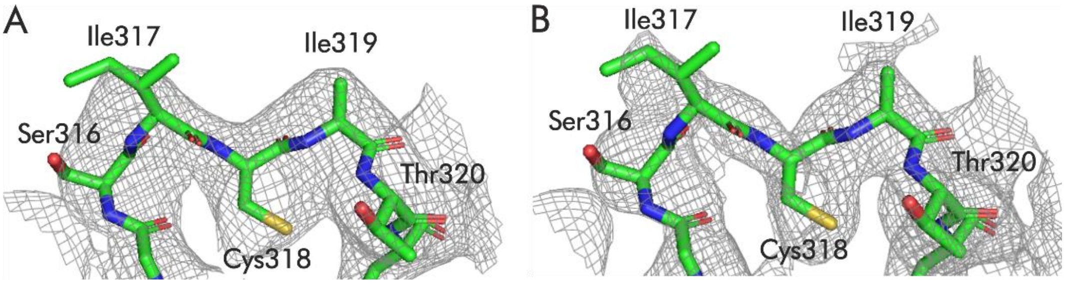


Figure 5.19: Chain A of* Tb*GMPR cyto v2, residues 315 to 321 shown as sticks overlaid with the map resulting from* CrystFEL *(A) and the sharpened map (B) in mesh representation. Both maps are contoured at 1.0 sigma and carved at 2.0 Å. The sharpened map shows a higher degree of details e.g., the electron density of the side chain of Ile319 is better resolved in the sharpened map.

As previously described, GMPR forms octamers in the biologically active state, as confirmed by the obtained structural model. Differences in between the Cα atoms of the different chains of *Tb*GMPR cyto v2 were small, with an average RMSD of 0.1 Å. No interpretable electron density was observed for the residues 1-2 (N-terminus), 402-412, 421-434 (part of the finger domain) and 487-495 (C-terminus). The octamer analyzed here consists of chain A and chains C to I. It is tightly packed and shows a twisted conformation which is 135 Å in width and 88 Å in height (Figure 5.18 E). The monomers exhibit the typical TIM barrel fold of the catalytic subunit (residues 1 to 97 and 226 to 484) as well as the tandem repeat of α-β-β-α folds in the Bateman or CBS domain (residues 98 to 225) and contains a finger domain (Imamura et al., 2020). The obtained model shows high structural similarity to the published, conventionally crystallized, twisted *Tb*GMPR solved by Imamura and coworkers with GMP bound in the canonical binding site II (pdb: 6jig, referred to as GMP-*Tb*GMPR). The electron density, especially in the Bateman domain, is less defined for some of the chains in the asymmetric unit and the obtained model shows increased B-factors indicating a high uncertainty in the position of the atoms (Figure

5.21 E). Based on the best-defined electron density, Chain A was selected for further investigation. Superposition of the Cα atoms of chain A and the GMP-*Tb*GMPR (pdb: 6jig) showed only minor differences in the overall structure indicated by an average RMSD of 0.816 Å. However, the finger domain region between residues 391 and 438, whose electron density is only partially visible in this structure, exhibits a different conformation, indicated by maximum RMSD values of up to 20 Å.

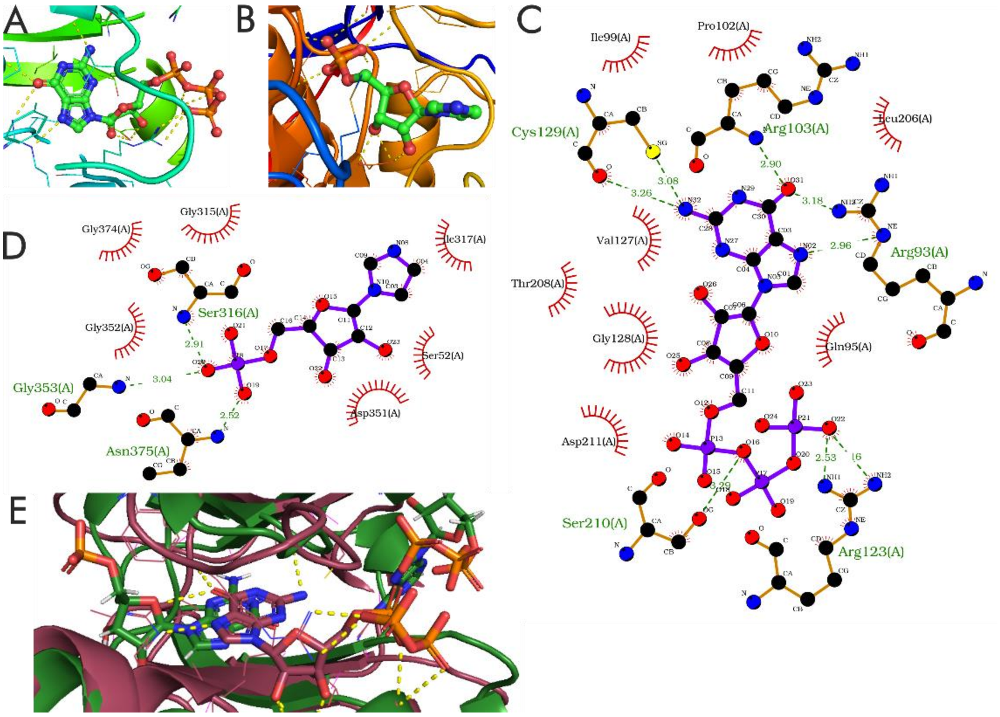


Figure 5.20. Binding sites of ligands in the* Tb*GMPR cyto v2 structure. Display of the interactions of GTP in the allosteric binding site II and the protein in chain A of the* Tb*GMPR cyto v2 structure model (A&C). Interactions of the substrate, a monophosphate ribonucleotide, close to the active center with the protein in chain A of the* Tb*GMPR structure (B&D). Ligand interaction plots in (C) and (D) were created using* LigPlot*. Cofactor orientation in the canonical binding site II in* Tb*IMPDH N-His structure (pdb: 6rfu) depicted in green and* Tb*GMPR cyto v2 structure depicted in berry (E).

The modeled structure reveals a GTP in the allosteric binding site II, close to Asp211, indicated by a well-defined electron density (Polder omit map in Appendix p.149, Figure 10.2). This suggests the formation of octamers instead of tetramers as the biological active unit, as reported by Imamura et al. (2020). Additionally, a monophosphate ribonucleotide was found to be bound to Asp351 in the active site, similar to GMP identified by Imamura et al.

In the following the structural models of *Tb*IMPDH ori and *Tb*GMPR cyto v2, that were elucidated in this study, are compared. In the structural model of *Tb*IMPDH ori, which was obtained after processing with *XDS* and model building, the position of the phosphate group of the ribonucleotide substrate is occupied by a single phosphate. The ribonucleotide bound to *Tb*GMPR cannot be further characterized due to the lack of electron density at the base. Following the enzymatic reaction this could either be GMP or IMP. Both ribose moieties are close to aspartic acid residues. The GTP interacts with the proteins with all the individual moieties. The base forms hydrogen bonds with residues Arg103, Arg93 and Cys129, while the phosphate groups form hydrogen bonds with residues Arg123 and Ser210 (Figure 5.20 A and C). The phosphate group of the ribonucleotide monophosphate in the active site forms polar interactions with the protein backbone at residues Ser316, Gly353 and the Asn375 (Figure 5.20 B and D).

While in *Tb*IMPDH crystals the relaxed octamers are linearly aligned with large cavities and high solvent content (72 %), in *Tb*GMPR crystals the twisted octamers are staggered, resulting in a tighter packing and a lower solvent content (61 %) as can be seen in Figure 5.21 A (*Tb*IMPDH N-His) and B (*Tb*GMPR cyto v2), respectively. The overall fold of *Tb*GMPR and *Tb*IMPDH is similar, but the exact orientation of the regulatory and the catalytic domains, which is a key element in regulating these enzymes, differs with respect to the angle between the two domains (Figure 5.21 F).

In *Tb*IMPDH and *Tb*GMPR the interactions between the monomers of a tetrameric ring and between adjacent octamers along the vertical axis are mediated via the catalytic domain, involved are, among others, the residue Thr2 (*Tb*GMPR) and the corresponding residue Asp13 (*Tb*IMPDH). Thr2 in the *Tb*GMPR is an amino acid added due to the cloning system II, used here. The tetrameric rings interact in both cases via their Bateman domain and the residues Arg190 and Arg199 in *Tb*GMPR and *Tb*IMPDH, respectively. In the *Tb*IMPDH structure the interactions are supported by the Finger domains. The interaction of octamers positioned side by side differs in the two structural models: The oblique placing of octamers in *Tb*GMPR is favorable, indicated by ΔG of -2.9 kcal/mol, while the interaction of adjacent *Tb*IMPDH octamers is unfavorable indicated by a positive ΔG divided on two interaction surfaces, one between chain A and chain A (2.0 kcal/mol) and one between chain B and chain B (0.9 kcal/mol). The lateral interaction interface for the eight monomers of an octamer of *Tb*GMPR is the product of the lateral interface area of two monomers (215 Å^2) and the number of monomers (8), yielding a lateral interface area of 1,720 Å^2. The lateral interaction interface area of *Tb*IMPDH is the sum of the single interaction interfaces (414 and 45 Å^2 respectively) multiplied by four. This yields a similar sized lateral interaction interface of 1,836 Å^2.

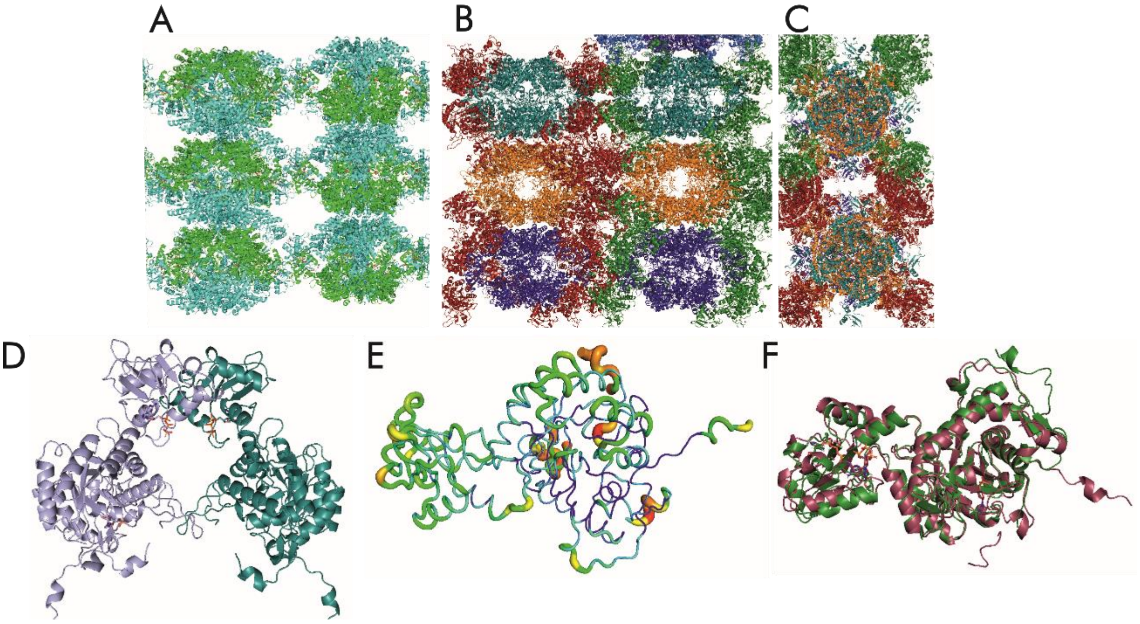


Figure 5.21: Cartoon representation of Tb*IMPDH N-His (pdb: 6rfu) (A) and* Tb*GMPR cyto v2 (B – F). Side view of the lattice of octamers of* in cellulo *crystallized* Tb*IMPDH N-His (pdb: 6rfu) (A) as well as side view (B) and top view (C) of the lattice of octamers of the structural model of* Tb*GMPR cyto v2. The complete octamers in one unit cell are consisting of chain A and C to I (represented in orange), the chains a, b, f, g, l, n, o and v (represented in blue) as well as chains c to e and h to k and m (represented in cyan). Cartoon representation of the dimer of chain A and F of the* Tb*GMPR cyto v2 structure (D). Monomer of chain A which displays the uncertainty in the atom position as atomic B-factor by a rainbow color code combined with a broadening of the cartoon representation (E). Red and orange indicate high B-factors while blue and green indicate low B-factors. Structural overlay of the catalytic domains of* Tb*IMPDH N-His (pdb: 6rfu) depicted in green and* Tb*GMPR cyto v2 depicted in berry (F). The orientation of the Bateman domain and the N-terminus differs in both structures.*

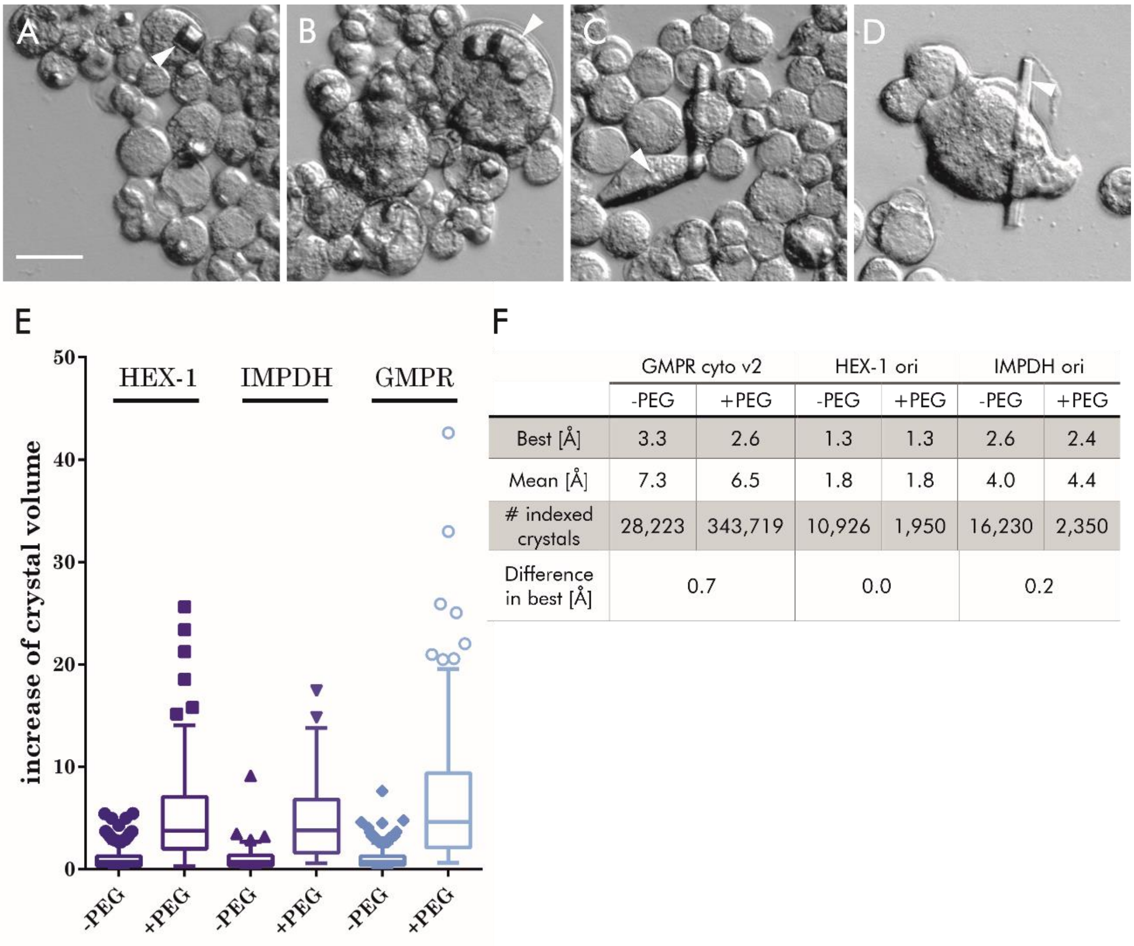
Table 5.15: Data processing and refinement statistics of the* Tb*GMPR cyto v2 structure modeled after data collection using rbv-infected, fused High Five cells. Values in parentheses refer to the highest resolution shell.

Data set	GMPR cyto v2 fused
Data collection date	11.12.2022
Sample temperature [K]	100
Processing software	CrystFEL
Oscillation	0.1°
Space group	P1
Unit cell	168.75 169.09 280.66 90.03 90.02 90.02
Resolution range	109.93 - 3.19 Å (3.22 - 3.19 Å)
Total reflections	48,476,621
Unique reflections	1,033,341 (51,626)
Multiplicity	46.9
Completeness (%)	100 (100)
SNR/ < I/σ(I)>	5.14 (0.63)
Wilson B-factor	103.13
R-split	16.49 (191.01)
CC1/2	0.9763 (0.1656)
CC*	0.9940 (0.5330)
No. collected images	575,871
No. hits/ indexed lattices	247,691 /239,442
Reflections used in refinement	516,248 (51,625)
Reflections used for R-free	1,180 (116)
R-work	0.2147 (0.3819)
R-free	0.2335 (0.4628)
Number of non-hydrogen atoms	154,699
Macromolecules	152.299
Ligands	2,400
Solvent	0
Protein residues	22,031
RMS bonds (Å)	0.004
RMS angles (°)	0.66
Ramachandran favored (%)	96.13
Ramachandran allowed (%)	3.87
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	0.33
Clash score	5.18
Average B-factor	115.17
Macromolecules	114.80
Ligands	138.83
Protein after map sharpening	37.43
Ligands after map sharpening	41.99
TLS refinement groups	48
Mean RMSD (Å) to ref. Structure 6JIG	0.816
Max. RMSD (Å) to ref. Structure 6JIG	20.058

5.3.5 Application of PEG-mediated cell fusion on model proteins *Nc*HEX-1 and *Tb*IMPDH

The optimized parameters for PEG-mediated cell fusion induction were tested for their applicability to other model proteins. The crystallization properties of the tested model proteins differ from *Tb*GMPR cyto v2. While *Nc*HEX-1 ori crystallizes rapidly and shows an efficiency (number of cells containing a crystal) comparable to *Tb*GMPR cyto v2, *Tb*IMPDH ori crystallizes later after infection and shows a lower efficiency. *Nc*HEX-1 ori crystals only rarely exceed the diameter of the cell, whereas *Tb*GMPR cyto v2 and *Tb*IMPDH ori crystallize in needle-shaped crystals exceeding the cell diameter (Schönherr, 2021). As shown in Figure 5.22 A-D, PEG-mediated cell fusion also results in larger crystals for *Nc*HEX-1 ori (5.21-fold crystal volume increase) and *Tb*IMPDH ori (5.11-fold crystal volume increase). Without any further optimization, the crystal volume could be increased by a factor of about 5 for both proteins tested.

In cellulo crystals successfully enlarged by cell fusion showed an improved diffraction behavior. Diffraction tests of crystals from fused and unfused cells revealed a higher resolution of the data set obtained from *Tb*IMPDH ori and *Tb*GMPR cyto v2 crystals when cells were fused, as shown in Figure 5.22 E & F. The maximum resolution of *Tb*IMPDH ori was increased from 2.6 Å to 2.4 Å and for *Tb*GMPR cyto v2 from 3.3 Å to 2.6 Å. In contrast, no resolution increase could be achieved for *Nc*HEX-1 ori. However, the number of indexed patterns investigated largely differ between *Tb*GMPR cyto v2 and *Tb*IMPDH ori and *Nc*HEX-1 ori (Figure 5.22 F), impeding a quantitative comparison of the observed resolution shifts. The larger the number of indexed patterns, the higher the redundancy in the high-resolution shells resulting in a data set of increased resolution.



	GMPR cyto v2		HEX-1 ori		IMPDH ori	
	-PEG	+PEG	-PEG	+PEG	-PEG	+PEG
Best [Å]	3.3	2.6	1.3	1.3	2.6	2.4
Mean [Å]	7.3	6.5	1.8	1.8	4.0	4.4
# indexed crystals	28,223	343,719	10,926	1,950	16,230	2,350
Difference in best [Å]	0.7		0.0		0.2	

Figure 5.22: PEG-mediated cell fusion also leads to an increased crystal size for other model systems like Nc*HEX-1 ori and* Tb*IMPDH ori. Investigation of the crystallization behavior of* Nc*HEX-1 ori (A&B) and* Tb*IMPDH ori (C&D) recombinantly expressed in High Five insect cells infected at an MOI of one of the respective recombinant baculovirus. Cells were imaged 4 dpi applying Emboss contrast (EC) and a 25x magnifying objective at a Nikon Ts2R-FL. The scale bar corresponds to 25 µm. The cell density was 0.8 million cells/well. If fusion was induced in rbv-infected High Five cells by adding 10 µL PEG4000 (10 %) at 39 hpi enlarged crystals were observed (B&D). The increase in crystal volume under the optimal PEG-mediated cell fusion conditions can be observed for all three model proteins tested,* Tb*GMPR cyto v2,* Nc*HEX-1 ori and* Tb*IMPDH ori (E). After diffraction data collection using MeshMounts and helical line scans at the beamline P14, operated by EMB at PETRA III, DESY, Hamburg, data were processed using* CrystFEL*. Then the* ave_resolution *script was applied to all data sets, outputting the displayed resolution statistics (F). Comparison of crystals from fused and unfused cells evinces an improvement of maximum resolution for* Tb*GMPR cyto v2 and* Tb*IMPDH ori crystals. For* Nc*HEX-1 ori crystals no difference in the resolution statistics were observed.*

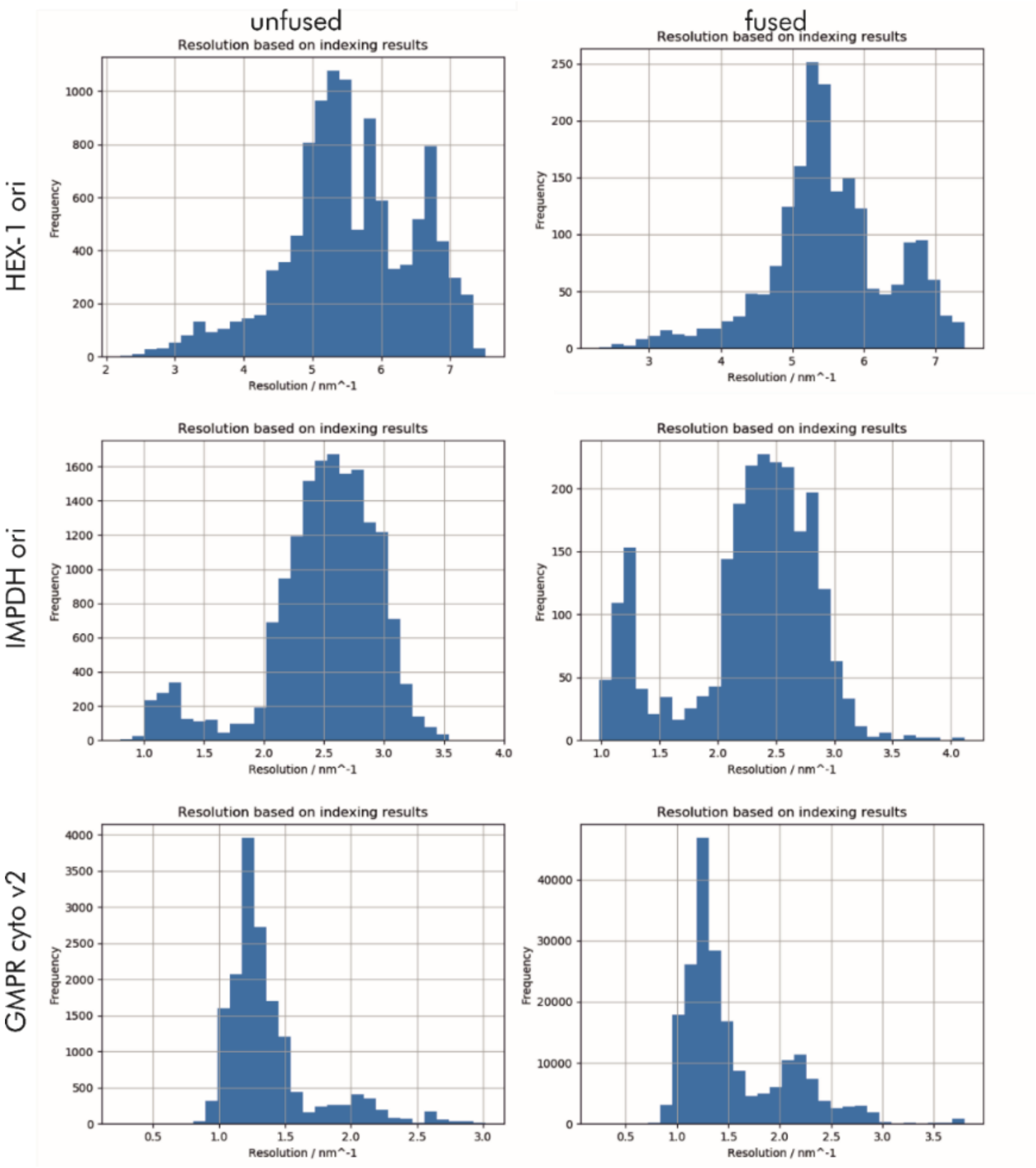


Figure 5.23: Comparison of resolution distribution histograms of the resolution distribution of diffraction data collected from fused and unfused cells, carrying Nc*HEX-1 ori,* Tb*IMPDH ori and* Tb*GMPR cyto v2 crystals. The y-axis shows the number of detector patterns against the maximum observed resolution on the x-axis in nm^{-1}. Resolution distribution of crystals from fused cells is broadened and shifted to higher resolution. For* Nc*HEX-1 both histograms show a comparable distribution along the resolution. The histograms are exported from the* CrystFEL *script* ave_resolution.

6. Discussion

The streamlined *in cellulo* crystallization and structure elucidation of proteins in insect cells is a promising and suitable approach to complement conventional protein crystallographic methods. Well-diffracting microcrystals of proteins grown in insect cells were used for serial diffraction experiments at synchrotron and XFEL sources, resulting in new bioinformation about natively bound ligands or native enzyme inhibition, enabled by the crystal formation in the quasi-native environment of the insect cell (Gati et al., 2014; Nass et al., 2020; Redecke et al., 2013).

Three groups have proposed streamlined workflows for intracellular protein crystallization in living insect cells (Boudes et al., 2016; Schönherr, 2021; Tang et al., 2020b). These approaches particularly focus on the generation of protein crystals, only Boudes et al. include the data collection and structure elucidation part in their strategy. They collect small rotational data sets *in cellulo* from individual crystals, which were then combined into a complete data set (Boudes et al., 2016). This data collection approach is called multiple crystal synchrotron X-ray diffraction (MSX) (Schönherr et al., 2018). Other techniques for collecting diffraction data from *in cellulo* crystals include single crystal X-ray diffraction (SXD) as well as serial synchrotron or serial free electron laser X-ray diffraction (SSX or SFX), as reviewed by Schönherr et al., (2018). All techniques can be applied either to isolated crystals or crystal-carrying cells. The inclusion of the data collection and processing part of X-ray crystallography in streamlined *in cellulo* crystallography approaches, a major focus of this thesis, is important to make this technique available to users. Automated processing pipelines at beamlines of synchrotron sources have frequently been implemented only for single crystal diffraction, and most users are not familiar with the processing of serial diffraction data, requiring specialized programs.

Due to the challenge of identifying protein crystallizing *in cellulo*, a strategy to select potential candidates of biological interest was developed in this study, which relies on the sequence homology to proteins already shown to crystallize in living cells. This strategy should be applied to other crystallizing proteins to further investigate its general applicability, as well as to better understand the underlying processes of crystal formation in cells and thus to increase the impact of *in cellulo* crystallography in the field of structural biology by elucidating new structures.

The resolution of a diffraction data set sets the level of detail visible in the electron density and is important for the biological interpretation of the obtained data. Thus, a strategy to enhance the resolution of a given protein crystal was developed by introducing cell fusion. Fusion of rbv-infected High Five cells results in enlarged protein crystals in giant syncytia. An increase in diffractive volume leads to an increase in signal and thus allows the collection of higher resolution diffraction data. This strategy enabled the structure elucidation and cofactor identification of *Tb*GMPR at 3.19 Å crystallized and measured in the quasi-native environment of living insect cells. A comparable approach to increase the resolution of X-rays diffracted at *in cellulo* crystals has not yet been published.

6.1 Successful implementation of a strategy for diffraction data processing

As shown in this study, the processing of diffraction data from conventionally grown crystals cannot be directly applied on *in cellulo* grown ones. Adjustments were necessary, e.g., regarding the filtering of diffraction signals resulting from salt and ice crystals. Moreover, in order to process serial diffraction data using processing software for single crystal diffraction data, the identification of subsequent detector images resulting from the same crystal is required.

Both software suites tested here, *XDS* (Kabsch, 2010) and *CrystFEL* (White, 2019), produced comparable results, which confirm the biological information obtained. Since a lot of manual input is still required for serial data processing, the success of the application of each of the programs still depends on the knowledge and experience of the user. Also, the programs internally differ in their algorithms and cannot be directly compared. A low crystallization efficiency, i.e., a low number of cells containing crystals in a culture, benefits from *XDS* by requiring fewer detector frames, when many frames connected by rotation are combined to obtain a complete data set, resembling the MSX approach (Schönherr et al., 2018). On the other hand, *CrystFEL* is optimized for separating lattices in detector frames containing multiple hits and is designed to work with large amounts of data (White, 2019; White et al., 2012, 2016). *XSCALE* (part of *XDS* software suite) was designed for merging and scaling of a few crystals resulting from single-crystal diffraction experiments, instead of 10s to 100s of wedges. The optimal oscillation during irradiation required for *XDS* and *CrystFEL* processing may vary due to their different algorithms but was not optimized here. Similarly, parameters such as the exposure time and the impact of radiation damage on the collected data was not part of this work.

The data processing strategy implemented here using *XDS* included all crystals that were successfully indexed and integrated. The filtering only focuses on the selection of individual crystals over redundant ones. An additional implementation of a hierarchical cluster analysis, as described in Zander et al., (2015) would complement the filtering step described here and crystals with a higher degree of homogeneity could be selected for merging to improve the obtained data set. If the merged crystals are more homogeneous, they contain less uncertainties, resulting in improved quality indicators and electron density maps (Zander et al., 2015). Exclusion of crystals, with large differences to the remaining crystals, is already implemented in *CrystFEL*'s *partialator* (White, 2019) by maximizing $CC_{1/2}$ when these crystals are rejected as described by Assmann et al., (2016).

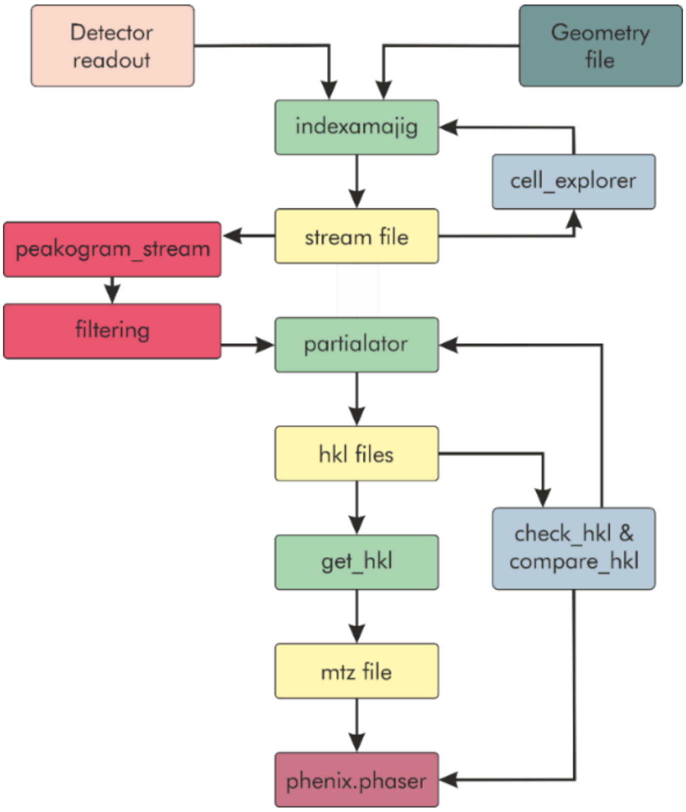


Figure 6.1: Optimized workflow for diffraction data processing with* CrystFEL. *The optimization implemented in this study cover the intensity-based filtering of the obtained stream file for contaminant reflections corresponding to salt or ice crystals formed during sample preparation. The filtered stream file is then further passed on to* partialator *to merge and scale the reflections into a complete data set.

According to the streamlined approach published by Boudes et al., (2016), for diffraction data collection of *in cellulo* crystals no differences in resolution could be observed when crystals were measured without cryoprotectant (50 % ethylene glycol) and no problems with salt or ice crystals were reported. On the contrary, our research group found that the use of cryoprotectant is mandatory to obtain high resolution diffraction patterns that can be used for successful data processing (personal communication Robert Schönherr). Even in the presence of 40 % PEG400 or PEG200 some cell-loaded MeshMounts still showed ice formation. According to Moreau and coworkers (2021) three types of ice are classified. First, ice that floats in the liquid nitrogen container and precipitates on the MeshMount surface. Second, ice formation within the solvent surrounding the sample due to low cryoprotectant or a thick liquid film on the MeshMount, and third ice formation within the cavities of the crystal. The third type is the most damaging and may result in the disruption of the crystal lattice of the protein (Moreau et al., 2021). Floating ice is typically monocrystalline and observed as discrete diffraction spots on the detector, while the ice inside the sample is polycrystalline and results in a ring-like powder diffraction signal, interfering with background subtraction (Moreau et al., 2021). The appearance of floating ice as well as ice formation from the solvent surrounding the cells has been frequently observed here, despite the use of cryoprotectant. Floating ice is difficult to avoid, but polycrystalline ice in the sample can be reduced by optimizing the cooling rates by increasing the blotting of excess liquid, and possibly by changing the cryoprotectant (Moreau et al., 2021). Optimizations regarding the sample quality, especially avoiding the formation of ice in the solvent during freezing, are certainly needed to improve the quality of diffraction data recorded from intracellular crystals in this approach.

6.1.1 About B-factors in serial crystallography

All structures solved in this study showed unusually high B-factors, despite reasonably resolved electron omit maps calculated from the diffraction data. The B-factor or atomic displacement factor describes the uncertainty of each atom in the structure, or in other words, it describes the degree of order in the crystal (Rupp, 2010). On the one hand, there is the Wilson B-factor, and on the other hand, there is the atomic displacement or (atomic) B-factor. The Wilson B-factor is derived from the Wilson plot, where the average reflection intensity is plotted against $\frac{(\sin\theta)^2}{\lambda^2}$, and describes an estimate of the B-factor derived from the raw data. The atomic displacement factor is calculated from the mean squared displacement factor u^2. It describes the attenuation of X-ray scattering by the thermal motion of the atoms (Rupp, 2010). Depending on the resolution, values of 50 Å^2 indicate increased disorder, while values below 20 Å^2 are considered to be normal for X-ray crystallographic structures determined at cryo conditions (Carugo, 2018).

Many factors introduce uncertainties that add up to high B-factors. As observed during data analysis of the structures presented here, the reported geometrical setup of the EMBL beamline P14 at PETRA III, DESY, in Hamburg is not perfectly accurate. Thus, not only the beam position, but also the detector to sample distance was not precisely specified in the metadata and sometimes the detector seemed to be tilted. While an incorrect detector to sample distance causes problems in indexing the reflections and in the correct assignment of the unit cell, a tilted detector complicates the beam position refinement, as there are two theoretical beam positions due to the tilt.

The distance between the sample and the detector is further affected by the thickness of the sample and the alignment of the sample holder. If the sample holder is not perfectly aligned or the sample is very thick, the sample cannot be properly focused and the distance between the diffracting crystal and the detector varies between different crystals diffracting from the same MeshMount and it deviates even more between crystals from different MeshMounts. An incorrect detector distance can be indicated by the presence of additional unit cell populations in the sample or by an asymmetric peak shape in the histogram of unit cell parameters. However, the occurrence of additional unit cell populations may have additional reasons like the presence of different crystal morphologies in the cells of the sample.

Moreover, contaminant diffraction causes problems in diffraction data processing. If a predicted spot position overlaps with either a salt or an ice signal, the signal integrated at that specific position will be incorrect (Moreau et al., 2021). This affects the correct determination of the structure factor by over- or underestimating the intensity of these reflections (Moreau et al., 2021). In general, the intensity of protein diffraction spots decreases with higher resolution. While ice rings occur at certain resolutions that are specific to ice (Moreau et al., 2021), salt reflections can be identified at resolutions starting at about 10 Å, due to the small unit cells of salt crystals, and are usually of higher intensity than Bragg reflections from proteins (Rupp, 2010). This has a direct impact on the Wilson plot and the determination of the Wilson B-factor. Filtering for potential salt and ice signals, as implemented in this thesis, does not only have beneficial effects. The removal of a certain amount of the strongest reflections results in a noisier electron density map (Rupp, 2010). The generation of a mask to selectively remove salt reflections based on their unit cell parameters, or ice rings based on their specific position, would be more beneficial to avoid the simultaneous removal of intense protein reflections.

Increased B-factors can also be an indication of radiation damage (Weik et al., 2000). However, the increase is not uniform throughout the protein. Especially side chains of Cys, Asp and Glu show higher B-factors when damaged, compared to an undamaged crystal (Weik et al., 2000). The impact of radiation damage on B-factors is considered to be less significant in this study because of the serial data collection approach, which prevents the accumulation of radiation damage by multiple irradiations of the same crystal. Moreover, no difference in the B-factors was observed between data sets collected at different grid spacings and oscillation settings. Applying smaller spacing and oscillation will result in the accumulation of radiation damage, since more frames of the same crystal are collected, which should result in increased B-factors compared to data sets recorded at larger spacing and oscillation.

Lower B-factors are expected when the solvent content of a protein crystal is low, because the fraction of the structure factor of the ordered protein is higher (Holton and Frankel, 2010). A high fraction of the ordered protein corresponding to the observed Bragg peaks results in a higher signal-to-noise ratio and a higher order of the crystal, which is reflected in a lower B-factor (Axford et al., 2014). The average B-factors of the structures analyzed in this study vary between 38 Å^2 (*Af*HEX-1delN ori) and 115 Å^2 (GMPR cyto v2), which is roughly in accordance with the solvent content of the associated crystals, which is 35 % for *Af*HEX-1delN ori and 61 % for *Tb*GMPR cyto v2. However, for *Tb*IMPDH ori, crystals have an increased solvent content (72 %) and a lower B-factor (52 Å^2), while for *Tb*GMPR, the crystal solvent content is reduced (61 %) and the B-factor of the structure is increased (115 Å^2). Consequently, the solvent content alone cannot explain the observed B-factors.

Table 6.1: Overview of* in cellulo *crystallized proteins sorted by the B-factors revealed. Information on crystal solvent content and diffraction data processing approach and software is given as well as the resolution of the data set and the respective reference.

Name	Resolution (Å)	Solvent content (%)	B-factor (Å^2)	Collection method	Processing software	Reference
Polyhedrin (*O. brumata* CPV18)	1.7	24	6.9	MSX	*FastDP (BLEND, POINTLESS, AIMLESS)*	(Axford et al., 2014)
Polyhedrin (*B. mori* CPV1)	1.98	22	17.3	MSX	*DENZO*	(Coulibaly et al., 2007)
Polyhedrin (*Wisena* NPV)	2.18	22	20.8	MSX	*Scalepack*	(Coulibaly et al., 2009)
HEX-1 (*N. crassa*)	1.80	50	24.0	SFX	*CrystFEL*	(Lahey-Rudolph et al., 2021)
Polyhedrin (G25D *Ac*MNPV)	2.3	22	24.1	MSX	*Scalepack*	(Coulibaly et al., 2009)
Polyhedrin (*Ac*MNPV)	3.09	19	28.8	MSX	*HKL2000/Denzo/ Scalepack*	(Ji et al., 2010)
Polyhedrin (G25W *Ac*MNPV)	1.84	19	34.1	MSX	*HKL2000/Denzo/ Scalepack*	(Ji et al., 2010)
Cry3A *(B. thuringiensis)*	2.9	57	38.4	SFX	*Cctbx.xfel*	(Sawaya et al., 2014)
HEX-1 (*N. crassa*)	1.56	50	41.3	SSX	*CrystFEL*	(This study)
HEX-1 (*N. crassa*)	1.85	50	42.1	SSX	*XDS*	(This study)
Cathepsin B (*T. brucei*)	2.1	61	49.0	SFX	*CrystFEL*	(Redecke et al., 2013)
Cathepsin B (*T. brucei*)	3.0	59	47.7	SSX	*XDS* (crystal wedge identification with *CrystFEL*)	(Gati et al., 2014)
IMPDH (*T. brucei*)	2.81	75	51.9	SFX	*CrystFEL*	(Nass et al., 2020)
IMPDH (*T. brucei*)	3.00	72	52.1	SSX	*XDS*	(This study)
HEX-1 (*N. crassa*)	2.50	50	55.6	SSX	*CrystFEL*	(Norton-Baker et al., 2021)
Cry3A *(B. thuringiensis)*	2.9	57	83.9	SFX	*CrystFEL*	(Sawaya et al., 2014)
GMPR (*T. brucei*)	3.19	61	115.1	SSX	*CrystFEL*	(This study)

A literature review revealed that high B-factors are a common characteristic of protein structures obtained from *in cellulo* grown crystals if the diffraction data were processed with *CrystFEL* (Table 6.1). From the examples listed, only the *in cellulo*-diffracted HEX-1 structure, which was collected by an SSX approach at the XFEL LCLS (Stanford, USA) revealed B-factors below 30 Å^2 (Lahey-Rudolph et al., 2021). Diffraction data of protein crystals with a lower solvent content, that were processed with other software than *CrystFEL* resulted in structural models with low to normal B-factors (Table 6.1). Comparison of other statistics revealed no correlation between the B-factors and the redundancy of the diffraction data, the method of data collection, or the radiation source. These findings indicate that the large B-factors observed in this study are introduced by the applied data processing strategy.

Sawaya et al. compared the processing of diffraction data of Cry3A toxin crystallized *in cellulo* using *CrystFEL* and *cctbx.xfel* (Sawaya et al., 2014). They processed the same raw data using both software suites and compared the results. Processing with *cctbx.xfel* yielded B-factors below 40 Å^2, while *CrystFEL* yielded B-factors above 80 Å^2 (Sawaya et al., 2014). The authors explain this difference by the different algorithms. While *cctbx.xfel* applies a per-frame resolution cut-off and scales the intensities according to a reference structure, *CrystFEL* sets a global resolution cut-off and determines scale factors without prior knowledge (Sawaya et al., 2014). The *CrystFEL* algorithm results in the integration of weak reflections especially in higher resolution shells, leading to a lower signal to noise ratio, as well as in a higher multiplicity and completeness, while in *cctbx.xfel* heterogeneities, caused by, *e.g.*, fluctuations in the beam or different crystal sizes, are reduced by setting a resolution limit for each frame (Sawaya et al., 2014). They observed differences not only in the B-factors, but also in the electron density maps. While water molecules were clearly visible in maps derived from *cctbx.xfel*, they were absent or poorly resolved in maps derived from *CrystFEL* (Sawaya et al., 2014), which is in accordance with the differences in the B-factors.

The set per-pattern resolution cut-off as implemented in *cctbx.xfel* is beneficial for the B-factors obtained. However, the application of a per-pattern resolution cut-off relative to the diffraction observed on the respective pattern is implemented in *CrystFEL* by using the *--push-res* function of *partialator* (White, 2019), which was used in this work. On the contrary, in the *XDS*-processed data set no per-frame or per-crystal wedge resolution limit was applied. For these datasets, one global resolution cut-off was applied after merging and scaling the intensities. Nevertheless, the resulting B-factors were comparably high as those obtained after processing the same raw data with *CrystFEL*. It cannot be excluded, that the per-pattern resolution cut-off has a beneficial impact on the B-factors, however the results of this study indicate this effect to be small. Nevertheless, a crystal-specific resolution cut-off before merging and scaling the intensities of the *XDS*-processed data with *XSCALE* could improve the B-factors compared to what was observed for processing with *cctbx.xfel*.

More importantly, the per-pattern refinement strategy implemented in *cctbx.xfel*, is based on the time-dependent variations occurring during data collection (Brewster et al., 2016). This means, *e.g.*, if during data collection a slight shift in the beam center is observed, this shift is refined in user-defined slices of the complete data set, reducing the impact of heterogeneities that influence the B-factor as discussed above. In *CrystFEL*, only one beam position is averaged from one data set, although this could easily be implemented in *CrystFEL* by manually splitting the data set into smaller portions and subsequent concatenating the individual stream files.

As discussed by Sawaya and coworkers, the application of prior knowledge always carries a risk of bias. The *cctbx.xfel* algorithm includes a reference structure to calculate the scale factor during intensity scaling. Nevertheless, they were able to show that the polyhedrin structure obtained after *CrystFEL* processing is more closely related to the *cctbx.xfel* structure than to the reference structure used to determine the scale factor (Sawaya et al., 2014). The application of a resolution-dependent scale factor to the amplitudes is comparable to imposing a uniform B-factor on the obtained map and results in a map with reduced extend of blurriness caused by computational procedures in a given map (Liu and Xiong, 2014). This process is called map sharpening. Real-space refinement of the structural models obtained from *CrystFEL*-processing against the sharpened map revealed a model with B-factors comparable to models obtained after *cctbx.xfel* processing (Sawaya et al., 2014).

Sawaya and colleagues, however, did not discuss the influence of the correct sample-to-detector distance. The determination of this distance is crucial because the effect of this uncertainty is larger for high resolution data than for low resolution data, which can be explained simply by the trigonometric function. Accuracy of the position is required especially for high resolution data to generate sharp electron density maps and models with reliable B-factors. The correct determination of the detector distance is required for the correct and precise determination of the unit cell parameters. In *XDS*, the sample-to-detector distance is refined for each crystal wedge before the diffraction of all crystals are merged to obtain one complete data set. *CrystFEL* lacks the refinement of this parameter, while for *cctbx.xfel* this distance is also refined in a time-dependent manner (Brewster et al., 2016). Nass et al. developed a strategy to refine this distance in *CrystFEL* by minimizing the standard deviation of the unit cell parameters (Nass et al., 2016). An implementation into the *CrystFEL* software suite and general application would be useful.

The large B-factors for the protein structures elucidated in this study are explained as systematic errors induced by the data processing algorithm, not as a marker of an unreliable protein structure. The lack of reproducibility of B-factors with different processing programs as well as many factors contributing to the atomic uncertainty question the significance of B-factors as a measure of atomic displacement and as a quality parameter of a protein structure. The use of processing programs other than *XDS* or *CrystFEL*, like *cctbx.xfel*, seems to be a promising way to improve the outcome and should be studied in the future.

6.2 Manipulation of the cell size leads to an increased crystal volume

Some protein crystals are too small to reveal detectable Bragg spots at high resolution, when they are diffracted at a synchrotron source. Diffraction data sets can then be collected either at an XFEL source, yielding a higher intensity of the observed reflections due to the substantial increased beam intensity, or by enlarging the crystal. The recording of high-resolution reflections during X-ray diffraction data collection enables the investigation of biological details of the protein under investigation. A data set of a crystal that diffracted X-rays to a resolution of 2.5 Å or better is set to be of good resolution and should reveal a good level of detail in the calculated electron density maps. Whereas the electron density obtained from a crystal that diffracted X-rays to 3 to 4 Å displays a sausage-like tube and is considered to be of low resolution (Rupp, 2010). The resolution of the diffraction data collected from a protein crystal is limited by the order and the illuminated volume of the crystal, the number of unit cells under illumination as well as radiation source specifications like radiation power and wavelength (Holton and Frankel, 2010).

An alternative to crystal enlargement is the application of radiation of higher intensity for diffraction data collection, e.g., by using an XFEL source (Nagaratnam et al., 2020). However, beamtime at XFELs is largely restricted and nowadays difficult to obtain, and the aim of this study was to implement simple, widely applicable approaches to improve the resolution of *in cellulo* grown crystals at easily accessible synchrotron sources into the *InCellCryst* pipeline.

One possibility to generate crystals of extended size is microcrystal seeding. Either small crystals or fragments of larger crystals are used as nucleation seeds to grow crystals *in situ* (Bergfors, 2003). This method is applied if increased mosaicity is observed or if crystallization conditions that allow the growth of a reasonably sized crystal, with dimensions in the 10 to 50 µm size range, have not been identified (Bergfors, 2003). *In cellulo*-grown microcrystals can also be isolated from the cells and used for seeding experiments to produce larger crystals. However, this approach is more labor-intensive than cell fusion, because in addition to growing and isolating intracellular crystals, soluble target protein needs to be purified while remaining stable outside the cellular environment.

6.2.1 Factors influencing the crystal dimensions after cell fusion

Cell fusion, as a strategy to improve the resolution of diffraction data, which in turn increases the biological information obtained from an elucidated structure, was successfully established in this study. According to Darwin's formula (Equation 1.3), the increase of crystal volume leads to an increase in reflection intensity in higher resolution shells (Darwin, 1914; Holton and Frankel, 2010).

The cell density was found to be a critical factor in cell fusion. Close contact of adjacent membranes is required for successful fusion. However, uninfected High Five cells and rbv-infected Sf9 cells did not show PEG4000-mediated cell fusion, indicating that additional factors are involved in the fusion process. PEG as dehydrating agent removes the hydration shell around the membrane resulting in tight contact of the adjacent membranes. However, an additional trigger, like membrane curvature induction or small perturbations in the membrane packing, are required to finally fuse the membranes (Hui et al., 1985; Yang and Shen, 2006).

For enlarged crystal growth not only the size of the syncytium plays a role, also the timing of cell fusion induction relative to the rbv infection is important. In this work, fusion was induced earliest 24 hpi to allow the infection of the cells without interfering agents. The target gene is under the control of the polyhedrin promoter, a strong promoter, which is activated in the very-late phase of infection, starting at 24 hpi (Rohrmann, 2013). The optimal time point for fusion induction was systematically screened to be between 23 hpi and 3 dpi. The later fusion is induced in this time frame, the larger the crystals grow, reaching a maximum crystal size, before this changes and the crystal size becomes comparable to that of non-fused cells 3 dpi. This highlights the importance of considering the timing of fusion induction and the time point of nucleation, when optimizing parameters. Nucleation is the first and rate-limiting step in crystallization (Bergfors, 2003; Rupp, 2010). It requires a local protein concentration above the saturation level (supersaturation); by potentially fusing two cells with unequal protein levels, the protein is diluted, preventing the nucleation in the syncytium.

In case of fusion induction after nucleation has occurred in all cells, all crystal nuclei are subject to crystal growth, resulting in many small crystals, comparable to the crystal size in unfused cells. If only a few of the fused cells show nucleation, the amount of protein, which can be incorporated into the crystal during growth phase is increased and the crystals can grow to larger size. This is the desired situation, where nucleation is separated from crystal growth, allowing to better meet the requirements for both phases (Bergfors, 2003). This separation reflects conventional crystal seeding, characterized by the introduction of either microcrystals or of larger crystals as a crystallization nucleus into a protein solution of high concentration (Bergfors, 2003).

The nucleation behavior also strongly depends on the protein investigated. While *Nc*HEX-1 and *Tb*GMPR form crystals in almost all infected cells, starting at 24 hpi and 44 hpi, respectively (Kardoost et al., 2023; Schönherr, 2021), *Tb*IMPDH crystallizes with a lower efficiency and forms first crystals 72 hpi (Schönherr, 2021). A comparable crystallization behavior was observed in fused cells. While after rbv *Tb*IMPDH infection many syncytia did not show protein crystals, cells produced many crystals after rbv *Nc*HEX-1 infection, indicating an early onset of nucleation; earlier fusion induction may be beneficial for *Nc*HEX-1 due to its fast crystallization. Evidence from rbv *Tb*IMPDH-infected cells suggests that a certain crystallization / nucleation efficiency is required to obtain an adequate number of cells with crystal nuclei for the production of larger crystals by cell fusion.

No improvement in diffraction behavior was observed for enlarged *Nc*HEX-1 crystals. Due to the overlapping salt reflections, which were not filtered prior to the resolution analysis performed on fused and unfused cells, the average and maximum resolution values reported by the *ave_resolution* script may be incorrect for crystal systems diffracting to high resolution by including the detected ice signal in the analysis. An alternative interpretation of these results is that the resolution limit has already been reached for *Nc*HEX-1 crystals.

For *Tb*IMPDH ori a relative crystal volume increase by a factor of 5.11 results in a resolution shift 0.2 Å of the diffraction data collected. The resolution shift is small, however, the differences in the process of crystal formation between *Tb*IMPDH and the *Tb*GMPR, for which the fusion conditions were optimized, require small customization of the protocol, e.g., in the time point of fusion induction. Importantly, the number of indexed diffraction patterns were low as well as the number of crystal-carrying syncytia. If the redundancy of enlarged crystals increases, the number of crystals diffracting to high resolution increases and so does the observable resolution shift.

6.2.2 What induces the syncytia formation?

High osmotic pressure, which can be introduced for example by PEG concentrations of 50 %, in the environment of a cell leads to a shrinking of the cell (Herrmann et al., 1985). Shrinking of the cell could favorably influence intracellular protein crystallization by increasing the local protein concentration required for crystal nucleation. Removal of the PEG would then restore the normal osmotic pressure and could favor crystal growth. However, the high osmotic pressure applied can have severe effects on cell viability (Herrmann et al., 1985).

In other studies, polyethylene glycol is added at higher concentrations (50 %) for cell fusion induction than in this study, and it was removed after a short incubation period (seconds to minutes). The removal of PEG is required to limit the cytotoxic effects and to induce the fusion by swelling of the cells due to the restored osmotic pressure (Herrmann et al., 1985; Hui et al., 1985; Yang and Shen, 2006). When PEG4000 is applied as a drop to the center of the well, the stock concentration of PEG4000 used was negligible in this study, implying that either PEG is not required for cell fusion, or that diffusion of the more viscous drop is required to reach a certain PEG concentration in some parts of the well that allows syncytia formation.

Uninfected High Five cells and rbv-infected Sf9 cells did not fuse at the conditions established for rbv-infected High Five cells, supporting the impact of another factor to increase the efficiency of cell fusion, next to PEG4000. A candidate for such a factor is the virus-associated gp64. This membrane protein presented on the plasma membrane of rbv-infected cells (Rohrmann, 2013) is absent in uninfected cells, and the viral genes show lower expression levels in Sf9 cells than in High Five cells (Schönherr, 2021; Wilde et al., 2014), supporting these observations. A gp64 impact is also in accordance with an improved cell fusion, resulting in larger syncytia, the later cell fusion is induced. The amount of gp64 increases during the infection cycle and thus fusion can be induced with a higher efficiency the more gp64 is present. It was shown for the HIV glycoprotein gp41 that PEG-mediated vesicle fusion was enhanced (Haque and Lentz, 2002), while assays by Blissard and Wenz proofed that a pH shift below 5.5 for at least 5 s was sufficient to induce syncytia formation of uninfected insect cells transiently expressing baculoviral gp64 (Blissard and Wenz, 1992). A similar approach, called V-fusion, was applied by Gottesmann and coworkers to reproducibly induce cell fusion. For that, the vesicular stomatitis virus G protein (VSV-G) was produced by recombinant expression and a pH shift was induced by applying PBS (pH 6) for 30 s (Gottesman et al., 2010).

During storage, especially under light exposure and at room temperature, PEG solutions decrease in pH and increase in their ionic strength (Votano et al., 2004). During 17 months of room temperature storage, a decrease in pH from 8.8 to 4.5 was observed, while storage at 4°C or -20°C improved its lifetime (Votano et al., 2004). This is highly problematic for the reproducibility of experiments. The PEG solution used here was stored at 4°C for 6 years and the fact that these results were not reproducible with a freshly prepared PEG4000 solution indicated that PEG aging has already affected the results. An examination of the pH of the PEG solution used here revealed an already aged solution with a pH of 3.

These findings suggest that the observed cell fusion is due to a pH-induced conformational change in the viral fusion protein gp64 rather than PEG-induced. If fusion is indeed induced by a pH shift, the procedure described here could be easily adapted by changing the culture medium of rbv-infected cells to PBS (pH 5) for at least 5 s at 39 hpi and a cell density of 0.8 million cells/6-well. Subsequently, PBS needs to be again replaced by the usual cell culture medium. This is highly comparable to the approach described by Gottesmann et al. (2010) and Blissard and Wenz (1992). Moreover, the short pH induction has the potential to extend the fusion zone from a small fusion ring to the entire well, while avoiding the negative effects of PEG on the crystal integrity, and thus resulting in a higher homogeneity of crystal sizes used for diffraction experiments. A quick exchange from cell culture medium to PEG solution and back is prevented by the high viscosity of the PEG solution.

6.3 Structural homologs exhibit similar crystallization behavior

The research on intracellular protein crystallization in Lars Redecke's group demonstrated, that many proteins showed signs of crystal growth in living insect cells, however, either the size or the frequency of the observed structures were too low, to enable the collection of diffraction data. Consequently, only a fraction of the proteins, which induce the formation of ordered structures in rbv-infected insect cells, are structurally investigated. To further improve the methodology and to deepen the understanding of the process of *in cellulo* crystallization, more *in cellulo* crystallizing proteins need to be identified and their intracellular crystallization to be investigated.

The concept of studying homologous proteins for their crystallization propensity is almost as old as protein X-ray crystallography itself. As early as 1958 and 1969, Kendrew and Perutz already tested myoglobin and hemoglobin from different species until they were able to solve the structures from sperm whale myoglobin and horse hemoglobin (Kendrew et al., 1958; Perutz et al., 1960; Rupp, 2010). In the context of the intracellular crystallization approach, this strategy is also successful, as shown in this study: of the HEX-1 orthologs tested, nine out of ten crystallized into X-ray diffracting protein crystals in insect cells, enabling structural studies; the paralog of *Trypanosoma brucei* IMPDH, the *Tb*GMPR, which also crystallizes *in cellulo*, forming needle-shaped structures.

Despite the low sequence conservation (32.4 %), the crystal morphologies of *Tb*IMPDH ori and *Tb*GMPR cyto v2 are highly similar and the protein structures are in large parts identical, whereas the crystal morphologies of the highly conserved HEX-1 proteins (sequence identities between 42.9 and 53.0 %) display a large variation from cuboidal or cubic structures with hexagonal cross-section to rod-shaped forms. The crystallization efficiency of the HEX-1 proteins was comparable, while *Tb*GMPR cyto v2 crystallized with higher efficiency and first crystals were detected earlier than after rbv *Tb*IMPDH ori infection. The high structural similarity of *Tb*IMPDH ori and *Tb*GMPR cyto v2 also continues in their biologically active unit, which assembles into an octameric state. Contradictory, the crystal lattices of *Tb*IMPDH ori and *Tb*GMPR cyto v2 differ: the octamers of *Tb*IMPDH ori are stacked in a linear way, while the *Tb*GMPR cyto v2 octamers are densely staggered.

Despite the high sequence conservation, the observed HEX-1 crystal morphologies cover a large range of possibilities. The addition of one (*Nc*HEX-1 cyto) up to four amino acids (*Nc*HEX-1 cyto v2) due to the gene cloning strategy is able to change the observed crystal shape, from cubic with hexagonal cross-section to spindle shaped or bipyramidal (Schönherr, 2021). Considering that no crystals were detected for *Tb*GMPR ori, while *Tb*GMPR cyto v2 that contains two additional amino acids at the N- and at the C-terminus crystallizes well (Schönherr, 2021), the selected cloning strategy selected obviously influences the outcome of the *in cellulo* crystallization experiment. Other than the crystal morphologies, the 3D structures of the NcHEX-1 variants show only small differences, indicating a minor impact of the additional residues on the fold and structure of HEX-1 proteins. Since all proteins have to be considered as individual in terms of characteristics and crystallization behavior, no general guideline for a cloning strategy that applies to all proteins can be proposed. Depending on whether the changes in the protein sequence are at the solvent accessible surface of the protein or at a protein-protein interface or even at or close to the active site of an enzyme, a severe impact on the outcome of the crystallization trial has to be considered, and thus all cloning strategies should be tested.

The presence of several structures of the same protein obtained under various crystallization conditions is a key for understanding the dynamical behavior of the protein of interest and its function (Marino-Buslje et al., 2019). The structural models obtained from crystals grown under different conditions differ. This gives an impression of the dynamic diversity of a protein in the form of snap shots. The dynamic diversity of the protein cannot be purely derived from homologous proteins and homology modelling, since even small differences can highly impact the protein function and dynamics (Marino-Buslje et al., 2019), as it can be seen for *Tb*GMPR and *Tb*IMPDH. Nevertheless, the crystallization of a protein homologous in sequence or structure can be a valid starting point for crystallization trials. If a structural model of the target protein is already known, *in cellulo* crystallization and diffraction data collection of protein crystals in cells and the following structure elucidation can validate the protein structure and possibly its cofactors in the quasi-native environment of the cell.

6.3.1 The natively crystallizing protein HEX-1 has a conserved fold and an intrinsic crystallization tendency

The Woronin body major protein, HEX-1, is a common feature in the genus Ascomycota (Tang et al., 2020a). It forms a specialized organelle, derived from peroxisomes, the Woronin body. Its major component is a hexagonal protein crystal formed by the HEX-1 protein (Yuan et al., 2003). In this study, the variability of HEX-1 proteins in different fungal species was investigated, similar to the approach of Tang et al., (2020a), who investigated the length deviations of HEX-1 proteins in various fungal species. HEX-1 proteins exhibit a wide distribution of amino acid sequence lengths, ranging from less than 100 to more than 500 amino acids, with two more frequent protein lengths around 200 and 420 amino acids. HEX-1 genes from 10 different species were subjected to closer inspection and were recombinantly expressed in High Five insect cells, revealing spontaneous intracellular crystallization, as expected, for six of them, namely *Ao*HEX-1, *An*HEX-1, *Cg*HEX-1, *Mg*HEX-1, *Vd*HEX-1, and *Vg*HEX-1. Sequence alignment of the 10 selected protein sequences revealed a high similarity to *Nc*HEX-1 for the six HEX-1 proteins that form *in cellulo* crystals, especially in the C-terminal region, as well as a conserved N-terminal heptapeptide in five of these proteins (sequence alignment in Appendix p.145). The four non-crystallizing proteins, namely *Af*HEX-1, *Mr*HEX-1, *Pl*HEX-

1and *Pt*HEX-1, include N-terminal extensions in front of the conserved heptapeptide, as annotated in the gene data bank. Deletion of these extension sequences (designated delN) resulted in spontaneous crystallization of *Af*HEX-1delN, *Mr*HEX-1delN and *Pl*HEX-1delN inside living insect cells, indicating the recovery of their native function, the intracellular crystal formation.

The functional role of HEX-1 is to crystallize in order to form the Woronin body, a rigid plug that seals the septal pores in case of cellular damage (Tang et al., 2020a; Yuan et al., 2003). Since the N-terminal sequence extension prevents the crystallization *in cellulo*, an annotation error in the data base is indicated, which is in accordance with observations of Beck and Ebel (2013) and Vangalis et al. (2020). The results of this study confirm the native crystallization tendency of HEX-1, independent of the cellular environment, the specific cell type, or the organism. This self-assembly tendency is suggested to be an intrinsic characteristic of the three-dimensional structure, depending on the conserved β-strand-rich fold of all HEX-1 molecules. β-sheets can provide a large interaction surface for protein assembly (Doye et al., 2004). Also, other *in cellulo* or *in vivo* crystallizing proteins, like lili-mip or Xpa, as well as *in vivo* aggregating proteins, such as the amyloid-β peptide associated with Alzheimer's disease, fold into β-strand-rich structures, facilitating the self-assembly (Banerjee et al., 2016; Doye et al., 2004; Soto et al., 1995; Tsutsui et al., 2015). β-sheet edges are known to enhance protein interaction (Doye et al., 2004).

Protein crystallization is thought to function through the presence of 'sticky-patches' within the 3D structure, leading to the spontaneous self-assembly of proteins that can result in ordered protein crystals (Derewenda and Godzik, 2017). The formation of β-sheets promotes the stickiness of certain areas of the protein by providing an interaction surface, which would increase the self-assembly capacity of these proteins. In contrast, potential interaction sites can be protected, e.g., by the formation of β-barrels, which inhibits the self-assembly of protein crystals as a regulatory mechanism (Doye et al., 2004). Consequently, the intrinsic fold of natively assembling proteins may then be sufficient to favor crystal growth, which is self-assembly in an ordered fashion, by incorporating evolutionary unprotected 'sticky-patches'.

The conserved heptapeptide is missing in the amino acid sequence of *Verticillium dahliae* HEX-1. Upstream of the annotation of the *Vd*HEX-1 gene sequence, a heptapeptide is encoded, suggesting an incorrect annotation also in this case (Vangalis et al., 2020). However, as proven here by the lack of electron density in *Nc*HEX-1, *Af*-HEX-1delN and *Mg*HEX-1, it seems to be flexible and thus not to be involved in the contacts required for the formation of the crystal. In *Pl*HEX-1delN and *Mr*HEX-1delN the heptapeptide is partly resolved, forming two β strands, which form a β-barrel with the N-terminus of the adjacent HEX-1 molecule, indicating that this region can support the stabilization of protein-protein interactions. To conclude, the heptapeptide seems to be a conserved starting signal of the HEX-1 proteins and supports the formation of crystals.

However, in some species e.g., *Aspergillus fumigatus*, *Aspergillus oryzae*, *Verticillium dahliae* and *Magnaporthe grisea*, alternative splice variants of HEX-1 exist, one encoding a histidine-rich exon and one without this exon (Beck and Ebel, 2013; Curach et al., 2004; Juvvadi et al., 2007; Soundararajan et al., 2004; Vangalis et al., 2020). Five of the sequences tested here for intracellular crystallization encode a histidine-rich sequence, namely *Af*HEX-1delN, *An*HEX-1, *Mr*HEX-1delN, *Pl*HEX-1delN, *Vg*HEX-1, while the remaining five HEX-1 proteins, namely *Nc*HEX-1, *Vd*HEX-1, *Cg*HEX-1, *Mg*HEX-1, and *Ao*HEX-1, did not encode this region. In *Aspergillus fumigatus* and *Metarhizium robertsii* the histidine-rich patch is involved in the localization of the Woronin bodies at the septal pores by interacting with the tethering protein Leashin (Lah) (Beck et al., 2013; Tang et al., 2020a). Both proteins were produced in this study including the histidine-rich region, which seems to be highly flexible, since no electron density was observed. This may indicate that a stabilization of this region occurs upon binding to Lah.

The structural investigations of four HEX-1 proteins did not reveal any evidence for phosphorylation-assisted multimerization in insect cells. Juvvadi et al. (2007) proposed that the multimerization and self-assembly of *Ao*HEX-1 is stabilized by phosphorylation. When protein kinase C (PKC) activity was inhibited, a decrease in multimer stability was observed, which is in line with a suggested involvement of Phosphorylated Ser151 in the multimer assembly. At least for the crystal formation in insect cells, this phosphorylation does not appear to be required for multimer stabilization. The proposed Ser151 is involved in crystal contacts in the structural models of *Mr*HEX-1delN and *Pl*HEX-1delN, but no electron density was observed that indicates phosphorylation. Phosphate groups are rich in electrons and are usually easily identified in electron density maps. No interpretable electron density was obtained for the respective Ser residue in *Af*HEX-1delN and *Mg*HEX-1, indicating that this residue is not involved in the formation of crystal contacts and is not stabilized by protein-protein interactions in crystals grown in the cytosol of living insect cells. However, natively the protein forms crystals in the Woronin body of fungi, derived from peroxisomes, yielding different environmental conditions.

The only HEX-1 protein that did not form intracellular crystals in this study, originating from *Pyrenophora tritici-repentis* (*Pt*HEX-1), includes an extended sequence between the conserved N-terminal heptapeptide and the C-terminal region conserved in all HEX-1 proteins, which not only shows low similarity to the other HEX-1 sequences, but also does not encode a histidine-rich region. This protein sequence seems to have a lower or no intrinsic crystallization tendency compared to the other HEX-1 proteins, suggesting an incorrect annotation of the splice sites, since the biological function of this protein is to form crystals.

To clarify whether this gene may be wrongly annotated in the gene data base, its native function of forming crystals needs to be investigated. An approach combining cell fusion with cross-seeding by a crystallizing HEX-1 homolog may drive the protein into a crystalline state. For cross-seeding of crystals a similar crystal packing is required (Bergfors, 2003), which may be a choice to overcome the rate limiting step of nucleation in proteins that itself did not crystallize in living cells so far, but have a homologous protein that did. Cross-seeding in living cells may be performed by fusing cells infected with one rbv encoding a crystallizing HEX-1 protein to cells infected with the rbv of the non-crystallizing protein *Pt*HEX-1delN. The nucleation of the crystallizing HEX-1 homolog could then serve as a seed to induce the crystal growth of *Pt*HEX-1delN if the annotation of its primary sequence is correct.

6.3.2 Native cofactors bound to *Tb*GMPR cyto v2

Structural investigation of *Tb*GMPR cyto v2 crystals grown in fused, rbv-infected High Five cells revealed the structure of the GMP reductase from *Trypanosoma brucei*. The overall fold and the conformation of the octameric enzyme is comparable to the *Tb*GMPR structure obtained by conventional protein crystallization, which was recently published by Imamura and coworkers (Imamura et al., 2020). Crystallization in its quasi-native environment, the living cell, revealed GTP bound to the canonical binding site II in the Bateman domain. This cofactor is known to induce a twisted oligomeric state, representing the enzyme conformation of the active state, which was also observed here (Imamura et al., 2020). Imamura et al. previously showed, that GTP can occupy the canonical binding site II and the active center. In this study, the active site was also occupied, but identification of the base was not possible due to lack of electron density, indicating a high variability in the conformations of the substrate and the catalyzing residues observed in the crystal of the functional enzyme. During enzyme-assisted catalysis, the involved residues and the substrate change their conformation. Thus, they may not be visible in the electron density obtained after X-ray diffraction data collection, because the electron density is the result of averaging over all intermediate states of these atoms exhibited in the crystal. The monophosphate as well as the ribose moiety are well resolved in the electron density.

The GMPR of Trypanosomatids are known to be potential drug targets for the treatment of African Trypanosomiasis, because these pathogens do not have a *de novo* pathway to produce purine nucleotides (Sarwono et al., 2017). Trypanosomatids and Leishmania encode the only GMPRs that are known to contain the CBS domain, also denoted as Bateman domain, which allosterically regulates the enzyme activity, particularly in IMPDHs (Bessho et al., 2016; Smith et al., 2016). The regulation of both enzymes, IMPDH and GMPR, functions via allosteric binding of nucleotide cofactors to the Bateman domain and exhibit the same intermediate state during catalysis, the thioimidate intermediate to the catalytic cysteine (Patton et al., 2011; Smith et al., 2016).

The structure of *Tb*IMPDH ori, which was obtained by diffraction of intracellular crystals, revealed binding of ATP and GDP in the canonical binding sites I and II.

Comparisons with other crystallized IMPDH and the observed conformation induced by the bound cofactors indicate that the observed octameric assembly is in its inactive, closed state. In contrast to the *Tb*IMPDH N-His structure obtained by Nass and coworkers (2020) from *in cellulo* grown, but for diffraction data collection isolated crystals, no cofactor hydrolysis was observed when crystals were diffracted inside the cells. The *Tb*IMPDH structure obtained by Nass et al. (2020) revealed GMP bound to the canonical binding site II, which is the suggested product of GDP hydrolyzation that occurred during crystal storage after isolation from the living cells. The active center located in the Cys-loop is not visible in the IMPDH structures of intracellular grown crystals, though the enzymes are in their inactive state. Lack of electron density is a sign of high variability of atomic positions in the crystal and non-uniform conformation of the Cys-loop between the individual unit cells. Whereas for *Tb*GMPR the active center is visible in the active conformation of the enzyme with activating cofactors bound. Different to what was observed for the GMP bound *Tb*GMPR including a Cys318 to Ala mutation (Imamura et al., 2020), no electron density for a potassium ion was detected in the active site. At the position of the potassium ion, the C-terminus of the adjacent protein chain was located in the structural *Tb*GMPR model calculated in this study, forming a crystal contact.

The correct crystal symmetry of *Tb*GMPR cyto v2 could not be unambiguously determined in this work. Although, *Zanuda* still suggested P422 as the correct space group, the better choice according to the R-factors was P121, thus a monoclinic lattice. The dense packing of the octamers together with their twisted state complicate the identification of the symmetry, thus the data were processed and modeled in space group P1 without using any symmetry operators. The use of a higher symmetry would have increased the obtained resolution by requiring less data for the same signal-to-noise ratio. Also, Friedel pairs were not treated as the same reflection. In order to compensate the low symmetry, more reflections need to be collected. Due to the narrow spacing of the reflections resulting from the large unit cell parameters, the reflection lunes observed during oscillation of the crystal during data collection are not likely to overlap. Thus, a larger spacing of, e.g., 1° during data collection would increase the amount of lattice points collected and leads to an increased redundancy and completeness of the data set.

The identification of bound cofactors is not always easy from its electron density. However, this problem is closely related to the resolution of the obtained data set and the quality of the electron density map. The higher the resolution of the obtained diffraction data, the more detail is visible in the electron density map. In case of low-resolution diffraction, the consultation of previously published work is required to interpret the visible electron density. Likewise, in case of *Tb*GMPR cyto v2, the substrate bound in the active site cannot clearly be identified as GMP or IMP. Only the determination of the putative binding pattern allows the identification of a ribose moiety with a phosphorylation at the 5’ OH. The same is true for the identification of the GTP bound to the canonical binding site II which regulates the relative orientation of the domains and thus the GMPR activity. According to the Polder omit maps (Appendix p. 149) the two ligands bound to *Tb*GMPR structure are clearly defined and the interpretation of the electron density seems correct. The uncertainty in the ligand identification in a structural model obtained from low-resolution data is a challenge of the intracellular crystallization approach, since the environment of the cell offers all possible cofactors that may bind to the investigated protein. However, in combination with other techniques it brings new insights into the functionality of enzymes. Fortunately, most *in cellulo* grown protein crystals, that lead to structure elucidation so far (HEX-1 homologs, *Tb*IMPDH and *Tb*CatB), diffracted X-rays to high resolution (Gati et al., 2014; Lahey-Rudolph et al., 2021; Nass et al., 2020; Norton-Baker et al., 2021; Redecke et al., 2013). And for examples like *Tb*GMPR, cell fusion has been established in this study to improve the resolution of the collected X-ray diffraction data.

6.4 Conclusion

This work implemented a general workflow for cryo-crystallography in intact insect cells into the *InCellCryst* pipeline, covering all steps from data collection over processing to model building. It also includes a strategy to overcome resolution limits by enlarging the crystal volume inside the cell and validates the opportunity to identify native cofactors in enzyme structures, a benefit provided by the quasi-native environment during protein crystallization in the living cell.

Data processing in general was already established for serially collected diffraction data. However, the salt-rich environment of the cells surrounding the crystal in combination with ice formation under cryo-conditions required the filtering for the resulting reflections before merging and scaling the intensities when data is to be processed with the software suite *CrystFEL*. An alternative approach to process diffraction data with the software suite *XDS* revealed comparable results. *XDS* is designed to process single-crystal data and scripts were developed to enable multiple crystal processing with it. *XDS* harbors the advantage that less crystals are required to obtain a data set, due to the information on the crystal orientation, which is obtained by the rotation series. *CrystFEL* requires more frames but can process all frames individually. For proteins, which do not crystallize efficiently inside the cells, *XDS* is a good choice. It is able to process small rotation diffraction data sets from individual crystals and thus requires less detector frames by using the rotation information to generate a complete data set by merging, instead of collecting 1,000s of empty detector frames by applying helical line scans. For protein crystals that diffract only to low resolution a highly redundant data set processed with *CrystFEL* may get the most out of the data.

In all structures elucidated in this study unusually high B-factors were observed, which seems to be algorithm-induced and could be cured by map-sharpening following real-space refinement. This suggestion needs to be further confirmed and investigated by comparing the processing results of different software suites and by investigating a larger cohort of different data sets.

The cell protects the crystals from environmental stress but also limits manipulation of the crystal growth. However, crystal enlargement by cell fusion enabled the extension of resolution limits. Further investigations regarding the causative mechanism needs to be performed in order to optimize the efficiency of cell fusion from a few cells to the entire cell culture well to increase the effects of the crystal enlargement. The establishment of this comparatively simple method to increase the resolution of a crystal will broaden the use of *in cellulo* grown crystals for structural studies.

The aging of PEG solutions which were not only used for cell fusion here but also as a cryo-protectant in this study should urgently be considered. Freshly prepared PEG solutions enable the reproducibility and possibly reduce the formation of ice during sample preparation. This will further increase the data quality.

7. Outlook

This work implemented the basic techniques required for structure elucidation using *in cellulo* crystals under cryo-conditions. Further improvements in data processing like comparing additional processing software, e.g., *cctbx.xfel*, or including a hierarchical cluster analysis to decide which crystals to merge, as well as improvements in the sample quality, e.g., reduction of ice formation, will improve the quality of the obtained structures and may ease the identification of biological information including native cofactors, particularly in critical cases.

The cell fusion established here has to be considered as a proof of principle that an increase in the crystal volume positively impacts the resolution limit of *in cellulo* grown crystals, leading to structure elucidation that was prevented by resolution problems so far. An implementation of a purely pH-dependent cell fusion using the baculoviral gp64 protein presented on the surface of infected insect cells might enlarge the extend of cell fusion from a specific fusion ring to the entire cell culture well. This would further shift the resolution limits, by shifting the proportion of enlarged to normal sized crystals in favor of the enlarged crystals, resulting in increased signal-to-noise ratios and completeness in higher resolution shells of the resulting data set.

The usage of cell fusion in combination with the use of structural homologs enable the implementation of a cross-seeding approach. Cross-seeding can be performed by fusing cells expressing a protein that does not crystallize efficiently, with cells expressing a homologous protein, that crystallizes to a comparable symmetry and unit cell size. Such an approach harbors the potential to elucidate even more structures applying the *InCellCryst* approach. This could be obtained by the combination of different promoters, which incorporate different expression capacities, in different cells of the same cell culture, that were mixed after infection. Martínez-Solís et al., (2016) observed that the use of the orf46 promoter results in higher protein yields than the polyhedrin promoter that control gene expression in this study. The usage of the polyhedrin promoter or even a weaker promoter for the seed protein does not interfere with the protein production of the protein of interest using the stronger promoter. If cells are fused at the right time point, the formed crystals may function as a seed to drive the protein of interest into crystallization.

In the future, the screening approach developed by Schönherr (2021) may be extended by the implementation of experimental phasing approaches either relying on fusion tags as already described by Duve (2022) and Freise (2020) or by using different crystal morphologies of the same protein to derive the phases required for the electron density calculation, as described by Metz et al., (2021). A *de novo* phasing strategy strengthens the significance of *in cellulo* crystallography to crystallize and model new, unknown proteins together with their cofactors without bias from previous information.

8. References

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science (80-.). *181*, 223–230.

Asiegbu, F.O., Choi, W., Jeong, J.S., and Dean, R.A. (2004). Cloning, sequencing and functional analysis of Magnaporthe grisea MVP1 gene, a hex-1 homolog encoding a putative "woronin body" protein. FEMS Microbiol. Lett. *230*, 85–90.

Assmann, G., Brehm, W., and Diederichs, K. (2016). Identification of rogue datasets in serial crystallography. J. Appl. Crystallogr. *49*, 1021.

Axford, D., Ji, X., Stuart, D.I., and Sutton, G. (2014). In cellulo structure determination of a novel cypovirus polyhedrin. Acta Crystallogr. Sect. D Biol. Crystallogr. *70*, 1435.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876.

Banerjee, S., Coussens, N.P., Gallat, F.X., Sathyanarayanan, N., Srikanth, J., Yagi, K.J., Gray, J.S.S., Tobe, S.S., Stay, B., Chavas, L.M.G., et al. (2016). Structure of a heterogeneous, glycosylated, lipid-bound, in vivo-grown protein crystal at atomic resolution from the viviparous cockroach Diploptera punctata. IUCrJ *3*, 282–293.

Barty, A., Kirian, R.A., Maia, F.R.N.C., Hantke, M., Yoon, C.H., White, T.A., and Chapman, H. (2014). Cheetah: Software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. J. Appl. Crystallogr. *47*, 1118–1131.

Battye, T.G.G., Kontogiannis, L., Johnson, O., Powell, H.R., and Leslie, A.G.W. (2011). iMOSFLM: A new graphical interface for diffraction-image processing with MOSFLM. Acta Crystallogr. Sect. D Biol. Crystallogr. *67*, 271–281.

Beck, J., and Ebel, F. (2013). Characterization of the major Woronin body protein HexA of the human pathogenic mold Aspergillus fumigatus. Int. J. Med. Microbiol. *303*, 90–97.

Beck, J., Echtenacher, B., and Ebel, F. (2013). Woronin bodies, their impact on stress resistance and virulence of the pathogenic mould Aspergillus fumigatus and their anchoring at the septal pore of filamentous Ascomycota. Mol. Microbiol. *89*, 857–871.

Bergfors, T. (2003). Seeds to crystals. J. Struct. Biol. *142*, 66–76.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235.

Bernal, J.D., and Crowfoot, D. (1934). X-ray photographs of crystalline pepsin [3]. Nature *133*, 794–795.

Bessho, T., Okada, T., Kimura, C., Shinohara, T., Tomiyama, A., Imamura, A., Kuwamura, M., Nishimura, K., Fujimori, K., Shuto, S., et al. (2016). Novel Characteristics of Trypanosoma brucei Guanosine 5'-monophosphate Reductase Distinct from Host Animals. PLoS Negl. Trop. Dis. *10*, e0004339.

Blissard, G.W., and Wenz, J.R. (1992). Baculovirus gp64 envelope glycoprotein is sufficient to mediate pH-dependent membrane fusion. J. Virol. *66*, 6829–6835.

Boudes, M., Garriga, D., Fryga, A., Caradoc-Davies, T., and Coulibaly, F. (2016). A pipeline for structure determination of in vivo-grown crystals using in cellulo diffraction. Acta Crystallogr. Sect. D, Struct. Biol. *72*, 576–585.

Boutet, S., Lomb, L., Williams, G.J., Barends, T.R.M.M., Aquila, A., Doak, R.B., Weierstall, U., DePonte, D.P., Steinbrener, J., Shoeman, R.L., et al. (2012). High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography. Science (80-.). *337*, 362.

Bragg, W.L. (1913). The structure of some crystals as indicated by their diffraction of X-rays. Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. Character *89*, 248–277.

Brewster, A.S., Waterman, D.G., Parkhurst, J.M., Gildea, R.J., Michels-Clark, T.M., Young, I.D., Bernstein, H.J., Winter, G., Evans, G., and Sauter, N.K. (2016). Processing XFEL data with cctbx.xfel and DIALS. Comput. Crystallogr. Newsl. *7*, 32–53.

Carugo, O. (2018). How large B-factors can be in protein crystal structures. BMC Bioinformatics *19*.

Chapman, H.N., Fromme, P., Barty, A., White, T.A., Kirian, R.A., Aquila, A., Hunter, M.S., Schulz, J., Deponte, D.P., Weierstall, U., et al. (2011). Femtosecond X-ray protein nanocrystallography. Nature *470*, 73–77.

Cipriani, F., Röwer, M., Landret, C., Zander, U., Felisaz, F., and Márquez, J.A. (2012). CrystalDirect: A new method for automated crystal harvesting based on laser-induced photoablation of thin films. Acta Crystallogr. Sect. D Biol. Crystallogr. *68*, 1393–1399.

Coulibaly, F., Chiu, E., Ikeda, K., Gutmann, S., Haebel, P.W., Schulze-Briese, C., Mori, H., and Metcalf, P. (2007). The molecular organization of cypovirus polyhedra. Nature *446*, 97–101.

Coulibaly, F., Chiu, E., Gutmann, S., Rajendran, C., Haebel, P.W., Ikeda, K., Mori, H., Ward, V.K., Schulze-Briese, C., and Metcalf, P. (2009). The atomic structure of baculovirus polyhedra reveals the independent emergence of infectious crystals in DNA and RNA viruses. Proc. Natl. Acad. Sci. U. S. A. *106*, 22205.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: A sequence logo generator. Genome Res. *14*, 1188–1190.

Curach, N.C., Te'o, V.S.J., Gibbs, M.D., Bergquist, P.L., and Nevalainen, K.M.H. (2004). Isolation, characterization and expression of the hex1 gene from Trichoderma reesei. Gene *331*, 133–140.

Darwin, C.G. (1914). XXXIV. The theory of X-ray reflexion . London, Edinburgh, Dublin Philos. Mag. J. Sci. *27*, 315–333.

Dauter, Z. (1999). Data-collection strategies. Acta Crystallogr. Sect. D Biol. Crystallogr. *55*, 1703–1717.

Derewenda, Z.S., and Godzik, A. (2017). The "sticky patch" model of crystallization and modification of proteins for enhanced crystallizability. In Methods in Molecular Biology, (NIH Public Access), pp. 77–115.

Dodson, G., and Steiner, D. (1998). The role of assembly in insulin's biosynthesis. Curr. Opin. Struct. Biol. *8*, 189–194.

Doerr, A. (2011). Diffraction before destruction. Nat. Methods *8*, 283.

Doye, J.P.K., and Poon, W.C.K. (2006). Protein crystallization in vivo. Curr. Opin. Colloid Interface Sci. *11*, 40–46.

Doye, J.P.K., Louis, A.A., and Vendruscolo, M. (2004). Inhibition of protein crystallization by evolutionary negative design. *1*, P9.

Duszenko, M., Redecke, L., Mudogo, C.N., Sommer, B.P., Mogk, S., Oberthuer, D., and Betzel, C. (2015). In vivo protein crystallization in combination with highly brilliant radiation sources offers novel opportunities for the structural analysis of post-translationally modified eukaryotic proteins. Acta Crystallogr. Sect. Struct. Biol. Commun. *71*, 929–937.

Duve, T. (2022). Intrazelluläre Kristallisation von IMPDH mit Zink-Finger-Domänen und Prozessierung von Diffraktionsdatensätzen Intracellular crystallization of IMPDH with zinc finger domains and. University of Lübeck.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. Sect. D Biol. Crystallogr. *66*, 486–501.

Ewald, P.P. (1913). Zur Theorie der Interferenzen der Röntgenstrahlen in Kristallen. Phys. Zeitschrift *14*, 465–472.

Flot, D., Mairs, T., Giraud, T., Guijarro, M., Lesourd, M., Rey, V., Van Brussel, D., Morawe, C., Borel, C., Hignette, O., et al. (2010). The ID23-2 structural biology microfocus beamline at the ESRF. J. Synchrotron Radiat. *17*, 107–118.

Franke, J. (2021). Screening von HEX-1 Homologen auf Kristallisation in lebenden Insektenzellen Screening of HEX-1 homologs for crystallization in living insect cells. University of Lübeck.

Freise, L. (2020). Etablieren von Zinkfinger-Sequenzen zum Lösen des Phasenproblems bei intrazellulärer Proteinkristallisation. University of Lübeck.

Fujiwara, T., Oda, K., Yokota, S., Takatsuki, A., and Ikehara, Y. (1988). Brefeldin A causes disassembly of the Golgi complex and accumulation of secretory proteins in the endoplasmic reticulum. J. Biol. Chem. *263*, 18545–18552.

Gati, C., Bourenkov, G., Klinge, M., Rehders, D., Stellato, F., Oberthür, D., Yefanov, O., Sommer, B.P., Mogk, S., Duszenko, M., et al. (2014). Serial crystallography on in vivo grown microcrystals using synchrotron radiation. IUCrJ *1*, 87–94.

Gevorkov, Y., Yefanov, O., Barty, A., White, T.A., Mariani, V., Brehm, W., Tolstikova, A., Grigat, R.R., and Chapman, H.N. (2019). XGANDALF - Extended gradient descent algorithm for lattice finding. Acta Crystallogr. Sect. A Found. Adv. *75*, 694–704.

Gottesman, A., Milazzo, J., and Lazebnik, Y. (2010). V-fusion: a convenient, nontoxic method for cell fusion. Biotechniques *49*, 747–750.

Gould, S.J., Keller, G.A., Hosken, N., Wilkinson, J., and Subramani, S. (1989). A conserved tripeptide sorts proteins to peroxisomes. J. Cell Biol. *108*, 1657–1664.

Haque, E., and Lentz, B.R. (2002). Influence of gp41 fusion peptide on the kinetics of poly(ethylene glycol)-mediated model membrane fusion. Biochemistry *41*, 10866–10876.

Harms, M. (2019). Optimierung der in cellulo- Kristallisation von Proteinen in Insektenzellen. University of Luebeck.

Harrison, S.C. (2015). Viral membrane fusion. Virology *479–480*, 498–507.

Hasegawa, H., Wendling, J., He, F., Trilisky, E., Stevenson, R., Franey, H., Kinderman, F., Li, G., Piedmonte, D.M., Osslund, T., et al. (2011). In vivo crystallization of human IgG in the endoplasmic reticulum of engineered chinese hamster ovary (CHO) cells. J. Biol. Chem. *286*, 19917–19931.

Hasegawa, H., Woods, C.E., Kinderman, F., He, F., and Lim, A.C. (2014). Russell body phenotype is preferentially induced by IgG mAb clones with high intrinsic condensation propensity: Relations between the biosynthetic events in the ER and solution behaviors in vitro. MAbs *6*, 1518–1532.

Hattne, J., Echols, N., Tran, R., Kern, J., Gildea, R.J., Brewster, A.S., Alonso-Mori, R., Glöckner, C., Hellmich, J., Laksmono, H., et al. (2014). Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers. Nat. Methods *11*, 545–548.

Havighurst, R.J. (1926). The Effect of Crystal Size upon the Intensity of X-Ray Reflection. Proc. Natl. Acad. Sci. *12*, 375–380.

Henry, C.J., and Atchison, R.W. (1971). Paracrystal Formation in Cell Cultures Infected with Adenovirus Type 2. J. Virol. *8*, 842–849.

Herrmann, A., Arnold, K., and Pratsch, L. (1985). The effect of osmotic pressure of aqueous PEG solutions on red blood cells. Biosci. Rep. *5*, 689–696.

Holton, J.M., and Frankel, K.A. (2010). The minimum crystal size needed for a complete diffraction data set. Acta Crystallogr. D. Biol. Crystallogr. *66*, 393–408.

Hruban, Z., and Swift, H. (1964). Uricase: Localization in hepatic microbodies. Science (80-.). *146*, 1316–1318.

Hui, S.W., Isac, T., Boni, L.T., and Sen, A. (1985). Action of polyethylene glycol on the fusion of human erythrocyte membranes. J. Membr. Biol. *84*, 137–146.

Hwu, Y., and Margaritondo, G. (2021). Synchrotron radiation and X-ray free-electron lasers (X-FELs) explained to all users, active and potential. *28*, 1014–1029.

Imamura, A., Okada, T., Mase, H., Otani, T., Kobayashi, T., Tamura, M., Kilunga Kubata, B., Inoue, K., Rambo, R.P., Uchiyama, S., et al. (2020). Allosteric regulation accompanied by oligomeric state changes of Trypanosoma brucei GMP reductase through cystathionine-ß-synthase domain. Nat. Commun. *11*.

Jakobi, A.J., Passon, D.M., Knoops, K., Stellato, F., Liang, M., White, T.A., Seine, T., Messerschmidt, M., Chapman, H.N., and Wilmanns, M. (2016). In cellulo serial crystallography of alcohol oxidase crystals inside yeast cells. IUCrJ *3*, 88–95.

Jayaram, H.N., Grusch, M., Cooney, D.A., and Krupitza, G. (2022). Consequences of IMP Dehydrogenase Inhibition, and its Relationship to Cancer and Apoptosis. Curr. Med. Chem. *6*, 561–574.

Ji, X., Sutton, G., Evans, G., Axford, D., Owen, R., and Stuart, D.I. (2010). How baculovirus polyhedra fit square pegs into round holes to robustly package viruses. EMBO J. *29*, 505–514.

Johansson, L.C., Stauch, B., Ishchenko, A., and Cherezov, V. (2017). A Bright Future for Serial Femtosecond Crystallography with XFELs. Trends Biochem. Sci. *42*, 749–762.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. *36*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

Juvvadi, P.R., Maruyama, J.I., and Kitamoto, K. (2007). Phosphorylation of the Aspergillus oryzae Woronin body protein, AoHex1, by protein kinase C: evidence for its role in the multimerization and proper localization of the Woronin body protein. Biochem. J. *405*, 533.

Kabsch, W. (2010). Xds. 20124692. Acta Crystallogr. D. Biol. Crystallogr. *66*, 125–132.

Kabsch, W. (2014). Processing of X-ray snapshots from crystals in random orientations. Acta Crystallogr. Sect. D Biol. Crystallogr. *70*, 2204–2216.

Kardoost, A., Schönherr, R., Deiter, C., Redecke, L., Lorenzen, K., Schulz, J., and Diego, I. de (2023). Convolutional neural network approach for the automated identification of in cellulo crystals. BioRxiv 2023.03.28.533948.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., and Phillips, D.C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature *181*, 662–666.

Kennedy, P.G.E. (2019). Update on human African trypanosomiasis (sleeping sickness). J. Neurol. *266*, 2334–2337.

Koiwai, K., Tsukimoto, J., Higashi, T., Mafuné, F., Miyajima, K., Nakane, T., Matsugaki, N., Kato, R., Sirigu, S., Jakobi, A., et al. (2019). Improvement of Production and Isolation of Human Neuraminidase-1 in Cellulo Crystals. ACS Appl. Bio Mater. *2*, 4941–4952.

Krissinel, E., and Henrick, K. (2005). Multiple alignment of protein structures in three dimensions. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) *3695 LNBI*, 67–78.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. *372*, 774–797.

Kubo, Y., Fujihara, N., Harata, K., Neumann, U., Robin, G.P., and O'Connell, R. (2015). Colletotrichum orbiculare FAM1 encodes a novel woronin body-associated pex22 peroxin required for appressorium-mediated plant infection. MBio *6*.

Lahey-Rudolph, J.M., Schönherr, R., Jeffries, C.M., Blanchet, C.E., Boger, J., Ramos, A.S.F., Riekehr, W.M., Triandafillidis, D.P., Valmas, A., Margiolaki, I., et al. (2020). Rapid screening of in cellulo grown protein crystals via a small-angle X-ray scattering/X-ray powder diffraction synergistic approach. J. Appl. Crystallogr. *53*, 1169–1180.

Lahey-Rudolph, J.M., Schönherr, R., Barthelmess, M., Fischer, P., Seuring, C., Wagner, A., Meents, A., and Redecke, L. (2021). Fixed-target serial femtosecond crystallography using in cellulo grown microcrystals . IUCrJ *8*, 665–677.

Lebedev, A.A., and Isupov, M.N. (2014). Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program Zanuda. Acta Crystallogr. Sect. D Biol. Crystallogr. *70*, 2430–2443.

Lentz, B.R. (2007). PEG as a tool to gain insight into membrane fusion. Eur. Biophys. J. *36*, 315–326.

Lentz, B.R., and Lee, J.K. (1999a). Poly(ethylene glycol) (PEG)-mediated fusion between pure lipid bilayers: A mechanism in common with viral fusion and secretory vesicle release? Mol. Membr. Biol. *16*, 279–296.

Lentz, B.R., and Lee, J.K. (1999b). Poly(ethylene glycol) (PEG)-mediated fusion between pure lipid bilayers: A mechanism in common with viral fusion and secretory vesicle release? Mol. Membr. Biol. *16*, 279–296.

Liang, L., Gao, H., Li, J., Liu, L., Liu, Z., and Zhang, K.Q. (2017). The Woronin body in the nematophagous fungus Arthrobotrys oligospora is essential for trap formation and efficient pathogenesis. Fungal Biol. *121*, 11–20.

Liebschner, D., Afonine, P. V., Baker, M.L., Bunkoczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.W., Jain, S., McCoy, A.J., et al. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix.

Acta Crystallogr. Sect. D Struct. Biol. *75*, 861–877.

Lindley, P.F. (1999). Macromolecular crystallography with a third-generation synchrotron source. Acta Crystallogr. Sect. D Biol. Crystallogr. *55*, 1654–1662.

Liu, C., and Xiong, Y. (2014). Electron Density Sharpening as a General Technique in Crystallographic Studies.

Lohkamp, B., Emsley, P., and Cowtan, K.D. (2005). Coot News. Web Doc. 1–5.

Marino-Buslje, C., Monzon, A.M., Zea, D.J., Fornasari, M.S., and Parisi, G. (2019). On the dynamical incompleteness of the Protein Data Bank. Brief. Bioinform. *20*, 356–359.

Martiel, I., Müller-Werkmeister, H.M., and Cohen, A.E. (2019). Strategies for sample delivery for femtosecond crystallography. Acta Crystallogr. Sect. D Struct. Biol. *75*, 160–177.

Martínez-Ripoll, M. (2022). Crystallography. Scattering and diffraction. The Bragg's Law.

Martínez-Solís, M., Gómez-Sebastián, S., Escribano, J.M., Jakubowska, A.K., and Herrero, S. (2016). A novel baculovirus-derived promoter with high activity in the baculovirus expression system. PeerJ *2016*.

Maruyama, J.I., Juvvadi, P.R., Ishi, K., and Kitamoto, K. (2005). Three-dimensional image analysis of plugging at the septal pore by Woronin body during hypotonic shock inducing hyphal tip bursting in the filamentous fungus Aspergillus oryzae. Biochem. Biophys. Res. Commun. *331*, 1081–1088.

McGill, R., Tukey, J.W., and Larsen, W.A. (1978). Variations of box plots. Am. Stat. *32*, 12–16.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., Cowley, A.P., and Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. *41*, W597–W600.

Metz, M., Arnal, R.D., Brehm, W., Chapman, H.N., Morgan, A.J., and Millane, R.P. (2021). Macromolecular phasing using diffraction from multiple crystal forms. Acta Crystallogr. Sect. A, Found. Adv. *77*, 19–35.

Moreau, D.W., Atakisi, H., and Thorne, R.E. (2021). Ice in biomolecular cryocrystallography. Acta Crystallogr. Sect. D Struct. Biol. *77*, 540–554.

Mudogo, C.N., Falke, S., Brognaro, H., Duszenko, M., and Betzel, C. (2020). Protein phase separation and determinants of in cell crystallization. *21*, 220–230.

Nagaratnam, N., Tang, Y., Botha, S., Saul, J., Li, C., Hu, H., Zaare, S., Hunter, M., Lowry, D., Weierstall, U., et al. (2020). Enhanced X-ray diffraction of in vivo-grown μnS crystals by viscous jets at XFELs. Acta Crystallogr. Sect. F Struct. Biol. Commun. *76*, 278–289.

Nass, K., Meinhart, A., Barends, T.R.M., Foucar, L., Gorel, A., Aquila, A., Botha, S., Doak, R.B., Koglin, J., Liang, M., et al. (2016). Protein structure determination by single-wavelength anomalous diffraction phasing of X-ray free-electron laser data. IUCrJ *3*, 180–191.

Nass, K., Redecke, L., Perbandt, M., Yefanov, O., Klinge, M., Koopmann, R., Stellato, F., Gabdulkhakov, A., Schönherr, R., Rehders, D., et al. (2020). In cellulo crystallization of Trypanosoma brucei IMP dehydrogenase enables the identification of genuine co-factors. Nat. Commun. *11*, 1–13.

Neutze, R., Wouts, R., van der Spoel, D., Weckert, E., and Hajdu, J. (2000). Potential for biomolecular imaging with femtosecond X-ray pulses. Nature *406*, 752–757.

Norton-Baker, B., Mehrabi, P., Boger, J., Schönherr, R., Von Stetten, D., Schikora, H., Kwok, A.O., Martin, R.W., Miller, R.J.D., Redecked, L., et al. (2021). A simple vapor-diffusion method enables protein crystallization inside the HARE serial crystallography chip. Acta Crystallogr. Sect. D Struct. Biol. *77*, 820–834.

Oeda, K., Inouye, K., Ibuchi, Y., Oshie, K., Shimizu, M., Nakamura, K., Nishioka, R., Takada, Y., and Ohkawa, H. (1989). Formation of crystals of the insecticidal proteins of Bacillus thuringiensis subsp. aizawai IPL7 in Escherichia coli. J. Bacteriol. *171*, 3568–3571.

Oscarsson, M., Beteva, A., Flot, D., Gordon, E., Guijarro, M., Leonard, G., McSweeney, S., Monaco, S., Mueller-Dieckmann, C., Nanao, M., et al. (2019). MXCuBE2: the dawn of MXCuBE Collaboration. J. Synchrotron Radiat. *26*, 393–405.

Patton, G.C., Stenmark, P., Gollapalli, D.R., Sevastik, R., Kursula, P., Flodin, S., Schuler, H., Swales, C.T., Eklund, H., Himo, F., et al. (2011). Cofactor mobility determines reaction outcome in the IMPDH and GMPR (2- ±) 8 barrel enzymes. Nat. Chem. Biol. *7*, 950–958.

Perrakis, A., and Sixma, T.K. (2021). AI revolutions in biology. EMBO Rep. *22*, e54046.

Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., and North, A.C. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature *185*, 416–422.

Pflugrath, J.W. (1999). The finer things in X-ray diffraction data collection. Acta Crystallogr. Sect. D Biol. Crystallogr. *55*, 1718–1725.

Phillips, J.C., Wlodawer, A., Yevitz, M.M., and Hodgson, K.O. (1976). Applications of synchrotron radiation to protein crystallography: Preliminary results. Proc. Natl. Acad. Sci. U. S. A. *73*, 128–132.

Redecke, L., Nass, K., DePonte, D.P., White, T.A., Rehders, D., Barty, A., Stellato, F., Liang, M., Barends, T.R.M.M., Boutet, S., et al. (2013). Natively inhibited Trypanosoma brucei cathepsin B structure determined by using an X-ray laser. Science *339*, 227–230.

Reed, L.J., and Muench, H. (1938). A simple method of estimating fifty per cent endpoints,. Am. J. Epidemiol. *27*, 493–497.

Rohrmann, G.F. (2013). Baculovirus Molecular Biology.

Röntgen, W.C. (1896). On a new kind of rays. Nature *53*, 274–276.

Rørvig-Lund, A., Bahadori, A., Semsey, S., Bendix, P.M., and Oddershede, L.B. (2015). Vesicle Fusion Triggered by Optically Heated Gold Nanoparticles. Nano Lett. *15*, 4183–4188.

Rupp, B. (2010). Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology (New York: Garland Science).

Sarwono, A.E.Y., Suganuma, K., Mitsuhashi, S., Okada, T., Musinguzi, S.P., Shigetomi, K., Inoue, N., and Ubukata, M. (2017). Identification and characterization of guanosine 5'-monophosphate reductase of Trypanosoma congolense as a drug target. Parasitol. Int. *66*, 537–544.

Sawaya, M.R., Cascio, D., Gingery, M., Rodriguez, J., Goldschmidt, L., Colletier, J.P., Messerschmidt, M.M., Boutet, S., Koglin, J.E., Williams, G.J., et al. (2014). Protein crystal structure obtained at 2.9 Å resolution from injecting bacterial cells into an X-ray free-electron laser beam. Proc. Natl. Acad. Sci. U. S. A. *111*, 12769–12774.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: An open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

Schlichting, I. (2015). Serial femtosecond crystallography: The first five years. IUCrJ *2*, 246–255.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: A new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Schoenherr, R., Boger, J., Lahey-Rudolph, J.M., Harms, M., Kaiser, J., Nachtschatt, S., Wobbe, M., Duden, R., Koenig, P., Bourenkov, G.P., et al. (2023). InCellCryst - A streamlined approach to structure elucidation using in cellulo crystallized recombinant proteins. BioRxiv 2023.06.16.545255.

Schönherr, R. (2021). Entwicklung einer Pipeline für die Kristallisation von Proteinen in lebenden Insektenzellen. Unitversity of Luebeck.

Schönherr, R., Klinge, M., Rudolph, J.M., Fita, K., Rehders, D., Lübber, F., Schneegans, S., Majoul, I. V., Duszenko, M., Betzel, C., et al. (2015). Real-time investigation of dynamic protein crystallization in living cells. Struct. Dyn. *2*.

Schönherr, R., Rudolph, J.M., and Redecke, L. (2018). Protein crystallization in living cells. Biol. Chem. *399*, 751–772.

Smith, S., Boitz, J., Chidambaram, E.S., Chatterjee, A., Ait-Tihyaty, M., Ullman, B., and Jardim, A. (2016). The cystathionine- β -synthase domains on the guanosine 5′′-monophosphate reductase and inosine 5′-monophosphate dehydrogenase enzymes from Leishmania regulate enzymatic activity in response to guanylate and adenylate nucleotide levels. Mol. Microbiol. *100*, 824–840.

Son, M., Lee, K.-M., Yu, J., Kang, M., Park, J.M., Kwon, S.-J., and Kim, K.-H.

(2013). The HEX1 gene of Fusarium graminearum is required for fungal asexual reproduction and pathogenesis and for efficient viral RNA accumulation of Fusarium graminearum virus 1. J. Virol. *87*, 10356–10367.

Soto, C., Castano, E.M., Frangione, B., and Inestrosa, N.C. (1995). The α-helical to ß-strand transition in the amino-terminal fragment of the amyloid ß-peptide modulates amyloid formation. J. Biol. Chem. *270*, 3063–3067.

Soundararajan, S., Jedd, G., Li, X., Ramos-Pamploña, M., Chua, N.H., and Naqvi, N.I. (2004). Woronin body function in Magnaporthe grisea is essential for efficient pathogenesis and for survival during nitrogen starvation stress. Plant Cell *16*, 1564–1574.

Steinberg, G., Harmer, N.J., Schuster, M., and Kilaru, S. (2017). Woronin body-based sealing of septal pores. *109*.

Stevens, R.C. (2003). The cost and value of three-dimensional protein structures. Drug Discov. World *4*, 35–48.

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques *28*.

Takamori, S., Cicuta, P., Takeuchi, S., and Di Michele, L. (2022). DNA-assisted selective electrofusion (DASE) of Escherichia coli and giant lipid vesicles. Nanoscale *14*, 14255–14267.

Tang, G., Shang, Y., Li, S., and Wang, C. (2020a). MrHex1 is Required for Woronin Body Formation, Fungal Development and Virulence in Metarhizium robertsii. J. Fungi *6*, 1–14.

Tang, Y., Saul, J., Nagaratnam, N., Martin-Garcia, J.M., Fromme, P., Qiu, J., and LaBaer, J. (2020b). Construction of gateway-compatible baculovirus expression vectors for high-throughput protein expression and in vivo microcrystal screening. Sci. Rep. *10*, 13323.

Tenney, K., Hunt, I., Sweigard, J., Pounder, J.I., McClain, C., Bowman, E.J., and Bowman, B.J. (2000). hex-1, a gene unique to filamentous fungi, encodes the major protein of the Woronin body and functions as a plug for septal pores. Fungal Genet. Biol. *31*, 205–217.

Tsutsui, H., Jinno, Y., Shoda, K., Tomita, A., Matsuda, M., Yamashita, E., Katayama, H., Nakagawa, A., and Miyawaki, A. (2015). A diffraction-quality protein crystal processed as an autophagic cargo. Mol. Cell *58*, 186–193.

Vangalis, V., Papaioannou, I.A., Markakis, E.A., Knop, M., and Typas, M.A. (2020). Hex1, the Major Component of Woronin Bodies, Is Required for Normal Development, Pathogenicity, and Stress Response in the Plant Pathogenic Fungus Verticillium dahliae. J. Fungi *6*, 1–23.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444.

Votano, J., Parham, M., and Hall, L. (2004). PEG Stability: A Look at pH and Conductivity Changes over Time in Polyethylene Glycols.

Wallace, A.C., Laskowski, R.A., and Thornton, J.M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng. *8*, 127–134.

Wang, W., and Hedstrom, L. (1997). Kinetic mechanism of human inosine 5'-monophosphate dehydrogenase type II: Random addition of substrates and ordered release of products. Biochemistry *36*, 8479–8483.

Weik, M., Ravelli, R.B.G., Kryger, G., McSweeney, S., Raves, M.L., Harel, M., Gros, P., Silman, I., Kroon, J., and Sussman, J.L. (2000). Specific chemical and structural damage to proteins produced by synchrotron radiation. Proc. Natl. Acad. Sci. *97*, 623–628.

White, T.A. (2019). Processing serial crystallography data with crystFEL: A step-by-step guide. Acta Crystallogr. Sect. D Struct. Biol. *75*, 219–233.

White, T.A., Kirian, R.A., Martin, A. V., Aquila, A., Nass, K., Barty, A., and Chapman, H.N. (2012). CrystFEL: A software suite for snapshot serial crystallography. J. Appl. Crystallogr. *45*, 335–341.

White, T.A., Barty, A., Stellato, F., Holton, J.M., Kirian, R.A., Zatsepin, N.A., and Chapman, H.N. (2013). Crystallographic data processing for free-electron laser sources. Acta Crystallogr. Sect. D Biol. Crystallogr. *69*, 1231–1240.

White, T.A., Mariani, V., Brehm, W., Yefanov, O., Barty, A., Beyerlein, K.R., Chervinskii, F., Galli, L., Gati, C., Nakane, T., et al. (2016). Recent developments in CrystFEL. J. Appl. Crystallogr. *49*, 680–689.

Wilde, M., Klausberger, M., Palmberger, D., Ernst, W., and Grabherr, R. (2014). Tnao38, high five and Sf9-evaluation of host-virus interactions in three different insect cell lines: Baculovirus production and recombinant protein expression. Biotechnol. Lett. *36*, 743–749.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., et al. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. Sect. D Biol. Crystallogr. *67*, 235–242.

Woronin, M. (1864). Zur Entwicklungsgeschichte des Ascobolus pulcherrimus Cr. und Pezizen. Abh. Senkenb. Naturforsch. *5*, 333–334.

Yang, J., and Shen, M.H. (2006). Polyethylene glycol-mediated cell fusion. Methods Mol. Biol. *325*, 59–66.

Yuan, J., Li, D., Qin, L., Shen, J., Guo, X., Tumukunde, E., Li, M., and Wang, S. (2019). HexA is required for growth, aflatoxin biosynthesis and virulence in Aspergillus flavus. BMC Mol. Biol. *20*.

Yuan, P., Jedd, G., Kumaran, D., Swaminathan, S., Shio, H., Hewitt, D., Chua, N.H., and Swaminathan, K. (2003). A HEX-1 crystal lattice required for Woronin body function in Neurospora crassa. Nat. Struct. Biol. *10*, 264–270.

Zander, U., Bourenkov, G., Popov, A.N., De Sanctis, D., Svensson, O., Mccarthy, A.A., Round, E., Gordeliy, V., Mueller-Dieckmann, C., and Leonard, G.A. (2015). MeshAndCollect: An automated multi-crystal data-collection workflow for synchrotron macromolecular crystallography beamlines. Acta Crystallogr. Sect. D Biol. Crystallogr. *71*, 2328–2343.

Zielinski, K.A., Prester, A., Andaleeb, H., Bui, S., Yefanov, O., Catapano, L., Henkel, A., Wiedorn, M.O., Lorbeer, O., Crosas, E., et al. (2022). Rapid and efficient room-temperature serial synchrotron crystallography using the CFEL TapeDrive. IUCrJ *9*, 778–791.

Zimmermann, U. (1982). Electric field-mediated fusion and related electrical phenomena. BBA - Rev. Biomembr. *694*, 227–277.

9. Register

9.1 List of Figures

9.2 List of Tables

9.3 List of Equations

10. Appendix

10.1 Abbreviations

Abbreviation	Meaning
Å	Angstroms
AcMNPV	*Autographa californica* multiple nucleopolyhedrovirus
aa	Amino acids
Af	*Aspergillus fumigatus*
An	*Aspergillus nidulans*
Ao	*Aspergillus oryzae*
ATP	Adenosine triphosphate
bp	Base pairs
CatB	Cathepsin B
CBS	Cystathionine β-synthase
Cg	*Chaetomium globosum*
DIC	Differential interference contrast
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleic triphosphate
dpi	Days post infection
dsDNA	Double strand DNA
EC	Emboss contrast
EDTA	N,N′-(Ethane-1,2-diyl)bis[N-(carboxymethyl)glycine]
EMBL	European Molecular Biology Laboratory
ER	Endoplasmic reticulum
fwd	Forward
GMP	Guanosine monophosphate
GMPR	GMP reductase
GTP	Guanosine triphosphate
hpi	Hours post infection
HIV	Human immunodeficiency virus
IMP	Inosine monophosphate
IMPDH	IMP dehydrogenase
IPTG	Propan-2-yl 1-thio-β-D-galactopyranoside
kb	1,000 bp
kDa	1,000 Dalton
Lah	Tethering protein Leashin
LCLS	Linac Coherent Light Source
Mg	*Magnaporthe grisea*
MOI	Multiplicity of infection
Mr	*Metarhizium robertsii*
MTS	Mitochondrial matrix targeting sequence
Nc	*Neurospora crassa*
N-His	N-terminal Histidine-rich tag
NLS	Nuclear targeting sequence
OD_{600}	Optical density at a wavelength of 600 nm
pdb	Protein database
PEG	Polyethylene glycol
PKC	Protein kinase C

Abbreviation	Meaning
Pl	*Purpureocillium lilacinum*
Pt	*Pyrenophora tritici-repentis*
PTS	Peroxisomal targeting sequence
rbv	Recombinant baculovirus
rev	Reverse
RMSD	Root mean square deviation
rpm	Rounds per minute
SAXS	Small-angle X-ray scattering
SFX	Serial femtosecond crystallography
SDS	Sodium dodecylsulfate
SNR	Signal-to-noise ratio
SSX	Serial synchrotron crystallography
Tb	*Trypanosoma brucei*
Tris	2-Amino-2-(hydroxymethyl)propane-1,3-diol
Vd	*Verticillium dahliae*
Vg	*Verruconis gallopava*
X-Gal	5-Bromo-4-chloro-1H-indol-3-yl β-D-galactopyranoside
XMP	Xanthosine monophosphate

10.2 Source code of scripts and commands

10.2.1 Find crystal wedges

```
BEGIN{

}
$1=="#"{
        if (c>1){
    for (j=1; j<c-10; j=j+8){
        count[data]+=1
        for(i=j; i<=j+9; i++){
        if(maxValue[data][count[data]]<=spotcount[data][i]){
            maxFrame[data][count[data]]=frameNo[data][i]
            maxValue[data][count[data]]=spotcount[data][i]
        }

        }
    }
     }
    data+=1
    count[data]=0

    getline
    c=1
    runNo[data] = $1
    frameNo[data][c] = $2
    spotcount[data][c] = $3
    spotFrame[data][$2] = $3


}
$1~/^[0-9]/{
    c+=1
    frameNo[data][c]=$2
    spotcount[data][c]=$3
    spotFrame[data][$2]=$3
}
END{
        for (j=1; j<c-10; j=j+8){
        count[data]+=1
        for(i=j; i<=j+9; i++){
        if(maxValue[data][count[data]]<=spotcount[data][i]){
            maxFrame[data][count[data]]=frameNo[data][i]
            maxValue[data][count[data]]=spotcount[data][i]
        }
        }
    }

#    printf "%s\t%s\t%s\t%s\n", "# win" ,"run", "frame" , "spots"
    for(d=1; d<=data; d++){ #anzahl datensaetze
    for (k[d]=1; k[d]<=count[d]; k[d]++){ # counter der Windows
        if(maxValue[d][k[d]]!=0){
        crystalCount+=1
#       printf "%i\t%i\t%i\t%i\n",  k[d], runNo[d] , maxFrame[d][k[d]], maxValue[d][k[d]]
        if(maxValue[d][k[d]]>=threshold){
            crystalPosEnd[crystalCount]=maxFrame[d][k[d]]
            crystalPosStart[crystalCount]=maxFrame[d][k[d]]
            for(u=1; u<=crystalLength; u++){
            idxPosU=maxFrame[d][k[d]]+u
            spotPosU=spotFrame[d][idxPosU]
            if(spotPosU>=0.3*maxValue[d][k[d]]&&spotPosU>=threshold)
                crystalPosEnd[crystalCount]=idxPosU
            else
                break
            }
```

```
         negCrystalLength=crystalLength*(-1)
              for(v=-1; v>=negCrystalLength; v--){
              idxPosV=maxFrame[d][k[d]]+v
              spotPosV=spotFrame[d][idxPosV]
              if(spotPosV>=0.3*maxValue[d][k[d]]&&spotPosV>=threshold)
                   crystalPosStart[crystalCount]=idxPosV
              else
                   break

            }
            maxF[crystalCount]=maxFrame[d][k[d]]
        }
        }
    }
    }
    printf "%s\t%s\t%s\n", "#Max", "Start", "End"
    for(a=1; a<=crystalCount; a++){
    if(crystalPosStart[a]!=crystalPosEnd[a]){
        system("mkdir /path_to_processingfolder/
        xds_Protein_ori_"runNo[data]"_"crystalPosStart[a]"-"crystalPosEnd[a]"; cd /
        path_to_processingfolder/xds_Protein_ori_"runNo[data]"_"crystalPosStart[a]"-"crystalPosEnd[a]"; cp
        /path_to_original_XDS_Input/OriginalXDSINPperRun/"runNo[data]"_XDS.INP ./XDS.INP ; echo
        \"DATA_RANGE= "crystalPosStart[a]" "crystalPosEnd[a]" \"  >> XDS.INP ; echo \"STARTING_FRAME=
        "crystalPosStart[a]"\" >> XDS.INP; echo \"SPACE_GROUP_NUMBER= 1\" >> XDS.INP echo
        \"UNIT_CELL_CONSTANTS= 57.24 57.24 198.19 90.0 90.0 120.0 \" >> XDS.INP;  cd ../../.. ")
      #  close(system)
    }
    printf "%i\t%i\t%i\n", maxF[a], crystalPosStart[a], crystalPosEnd[a]
    }




}
```

10.2.2 Filter for individual crystal wedges

```
BEGIN{
 getline
 runNoNew=$1
 startNew=$2
 endNew=$3
 lengthNew=endNew-startNew
}
$1==runNoNew{
  runNoOld=runNoNew
  startOld=startNew
  endOld=endNew
  lengthOld=lengthNew
  runNoNew=$1
  startNew=$2
  endNew=$3
  lengthNew=endNew-startNew
  meanLength=(lengthOld+lengthNew)/2
  if (startOld<startNew&&endOld<=startNew||startNew<startOld&&endNew<=startOld) { #if no overlapp
   print "INPUT_FILE= /path_to_processingfolder/xds_Protein_ori_"runNoOld"_"startOld"-"endOld"/
  XDS_ASCII.HKL" # then add Run to XSCALE
  } else { #if overlapp
    if ((endOld-startNew)>=(2)){ # if overlapp is more than 2 frames
      cmdCO1 = "grep \"UNIT_CELL_C-AXIS\" /path_to_processingfolder/
      xds_Protein_ori_"runNoOld"_"startOld"-"endOld"/XDS_ASCII.HKL | awk '{print $2}' "
      cmdCN1 = "grep \"UNIT_CELL_C-AXIS\" /path_to_processingfolder/
      xds_Protein_ori_"runNoNew"_"startNew"-"endNew"/XDS_ASCII.HKL| awk '{print $2}'"
      cmdCO2 = "grep \"UNIT_CELL_C-AXIS\" /path_to_processingfolder/
      xds_Protein_ori_"runNoOld"_"startOld"-"endOld"/XDS_ASCII.HKL | awk '{print $3}' "
    cmdCN2 = "grep \"UNIT_CELL_C-AXIS\" /path_to_processingfolder/
    xds_Protein_ori_"runNoNew"_"startNew"-"endNew"/XDS_ASCII.HKL| awk '{print $3}'"
    cmdCO3 = "grep \"UNIT_CELL_C-AXIS\" /path_to_processingfolder/
    xds_Protein_ori_"runNoOld"_"startOld"-"endOld"/XDS_ASCII.HKL | awk '{print $4}' "
    cmdCN3 = "grep \"UNIT_CELL_C-AXIS\" /path_to_processingfolder/
    xds_Protein_ori_"runNoNew"_"startNew"-"endNew"/XDS_ASCII.HKL| awk '{print $4}'"

     cmdCO1 | getline X1      #print $0
     cmdCN1 | getline X2
     cmdCO2 | getline Y1
     cmdCN2 | getline Y2
     cmdCO3 | getline Z1
   cmdCN3 | getline Z2
 #
      close(cmdCO1)
      close(cmdCN1)
      close(cmdCO2)
      close(cmdCN2)
      close(cmdCO3)
      close(cmdCN3)

      if (((X1*X2+Y1*Y2+Z1*Z2)/((X1+Y1+Z1)^2))>0.99||((X1*X2+Y1*Y2+Z1*Z2)/((X1+Y1+Z1)^2))<-0.99){ #
      crystals are equal
        if (lengthNew<=lengthOld){ # use the old one else skip the old one and check the new one
        against the next (is that helpful?)
          print "INPUT_FILE= /path_to_processingfolder/
          xds_Protein_ori_"runNoOld"_"startOld"-"endOld"/XDS_ASCII.HKL"
          getline
      runNoNew=$1
      startNew=$2 #startOld
          endNew=$3 #endOld
      lengthNew=$3-$2 #endOld-startOld
         }else if (lengthNew>lengthOld){


    }
```

```
            }else{ #wenn Kristalle ungleich, dann behalte beide.
               print "INPUT_FILE= /path_to_processingfolder/
xds_Protein_ori_"runNoOld"_"startOld"-"endOld"/XDS_ASCII.HKL"
        }}
            else{ # if overlapp is less than 2 frames
                print "INPUT_FILE= /path_to_processingfolder/



xds_Protein_ori_"runNoOld"_"startOld"-"endOld"/XDS_ASCII.HKL"   # then add oldRun to
XSCALE


}
}
$1!=runNoNew{        #if new run, set old variables new
    print "INPUT_FILE= /path_to_processingfolder/xds_Protein_ori_"runNoOld"_"startOld"-"endOld"/
XDS_ASCII.HKL"
    runNoNew=$1
    startNew=$2
    endNew=$3
    lengthNew=endNew-startNew
#    getline
}END{

    print "INPUT_FILE= /path_to_processingfolder/xds_Protein_ori_"runNoOld"_"startOld"-"endOld"/
XDS_ASCII.HKL"
}
```

10.2.3 Command list for *XDS* processing

```
awk -f find_spot_count.awk runlist_HEX1ori.txt # erstelle Datei Spots pro Frame; Input Table RunNo Width
Height of grid
bash make_start_xdsinp.sh # make Start XDS-File from original XDS.INP and saves them in a seperate folder
bash divide_spotcounts_into_rows.sh # puts the spot count in the correct format to find the maximum and
crystal wedges in the next step
awk -f new_find_local_max.awk -v threshold=150 -v crystalLength=3 Prozessierung/spotcounts/2.txt # defines
crystal wedges, some are defined in 2 ways --> error message occurs, ignore this
bash run_xds.sh
 bash optimize.sh
 ls Prozessierung/XdsFiles2/xds_Hex1_ori_*/XDS_ASCII.HKL | sed 's/_/\t/g' | sed 's/\//\t/g' | sed 's/-/\t/
g' | awk '{print $6, $7, $8}' | awk -f check_overlorequal.awk >> Prozessierung/XSCALE/XSCALE-2/XSCALE.INP
```

10.2.4 Filtering for salt and ice reflections

```
## written by Juliane Boger
## usage awk -f filterBadIntensities.awk your.stream > yourFiltered.stream

BEGIN{
    cmd = "awk '$1~/^##/&&$4~/^J/&&$5~/^B/{print 1}' filterBadIntensities.awk"
    cmd | getline flag
    if (flag!=1){
    print "failed"
    exit
    }
}
$1=="panel0/corner_x"{
    beamX=-1*(int($3))
}
$1=="panel0/corner_y"{
    beamY=-1*(int($3))
}
$0 ~ /Reflections measured after indexing/{
    print $0
    getline
    print $0
    getline
    while ($0 !~ /End of reflections/){
    if ($6<2000&&(sqrt((beamX-$8)^2+(beamY-$9)^2)<=1290)){ ## reflections below 2000 adu in a distance of max. 1290 Pixels away from the  beam stop are considered as good
        print $0
    }else if ($6<150&&(sqrt((beamX-$8)^2+(beamY-$9)^2)>1290)&&(sqrt((beamX-$8)^2+(beamY-$9)^2)<=1790)){ ## reflections below 150 adu in a distance of max. 1790 and min. 1290 Pixels away from the  beam stop are considered as good
        print $0
    }else if ($6<55&&(sqrt((beamX-$8)^2+(beamY-$9)^2)>1790)&&(sqrt((beamX-$8)^2+(beamY-$9)^2)<=2140)){## reflections below 55 adu in a distance of max. 2140 and min. 1790 Pixels away from the  beam stop are considered as good
        print $0
        }
    else if ($6<28&&(sqrt((beamX-$8)^2+(beamY-$9)^2)>2140)){ ## reflections below 28 adu in a distance of min. 2140 Pixels away from the  beam stop are considered as good
        print $0
        }
    getline
    }
}
$0 !~ /Reflections measured after indexing/{
    print $0
}
```

10.3 Example input files

10.3.1 *CrystFEL* geometry file

```
clen = 0.42118
photon_energy = 12700

adu_per_photon = 1
res = 13333.3

panel0/min_fs = 0
panel0/min_ss = 0
panel0/max_fs = 4147
panel0/max_ss = 4361
panel0/corner_x = -2052.78
panel0/corner_y = -2155.79
panel0/fs = x
panel0/ss = y


panel0/flag_lessthan = -0.5
panel0/flag_morethan = 65000
badr1/min_x= -50
badr1/max_x= 50
badr1/min_y= -50
badr1/max_y= 2000
```

10.3.2 *CrystFEL* unit cell file

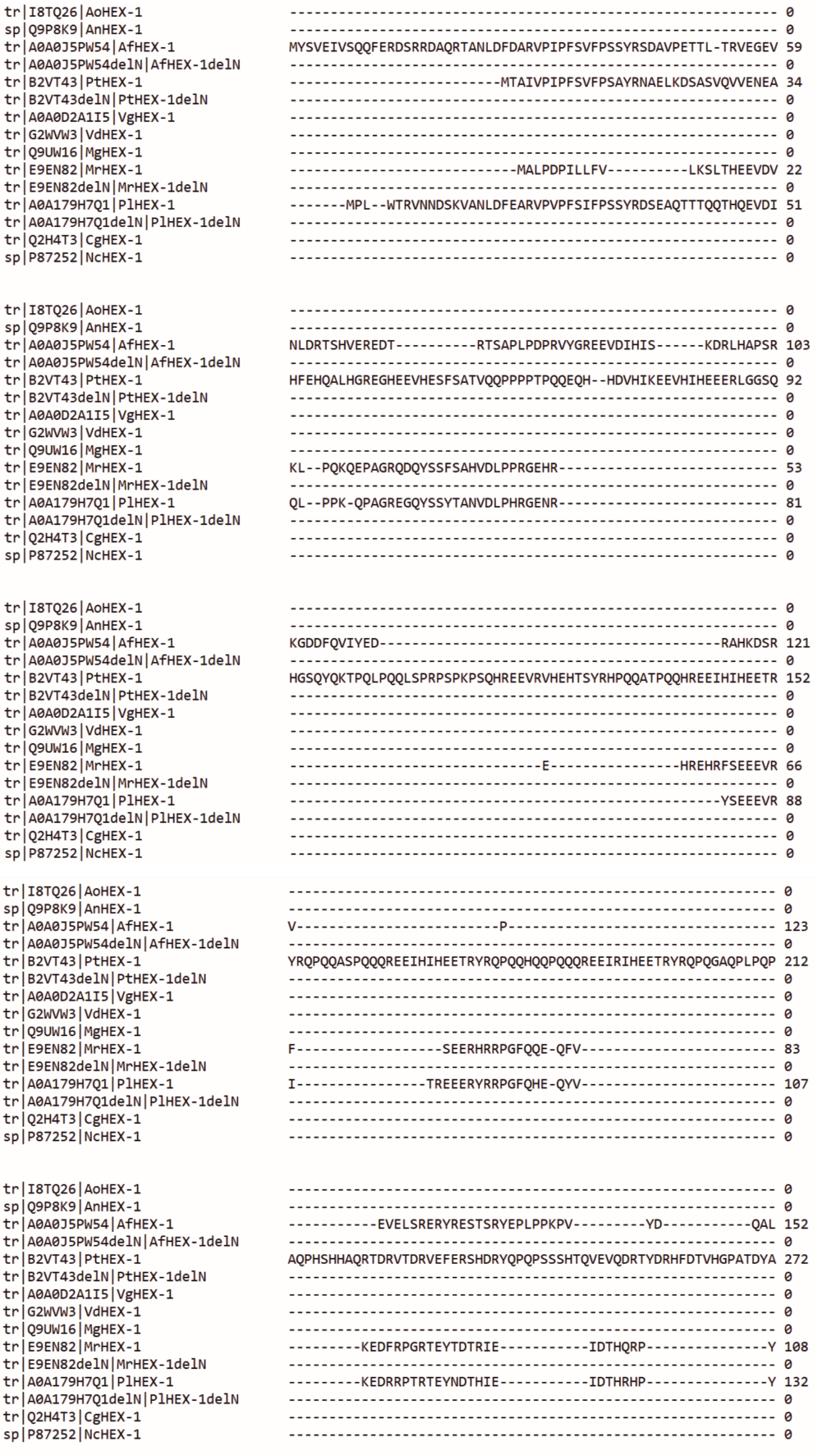
```
CrystFEL unit cell file version 1.0

lattice_type = tetragonal
unique_axis = c
centering = P
a = 168.72 A
b = 168.72 A
c = 280.37 A
al = 90.00 deg
be = 90.00 deg
ga = 90.00 deg
```

10.3.3*XDS* input file

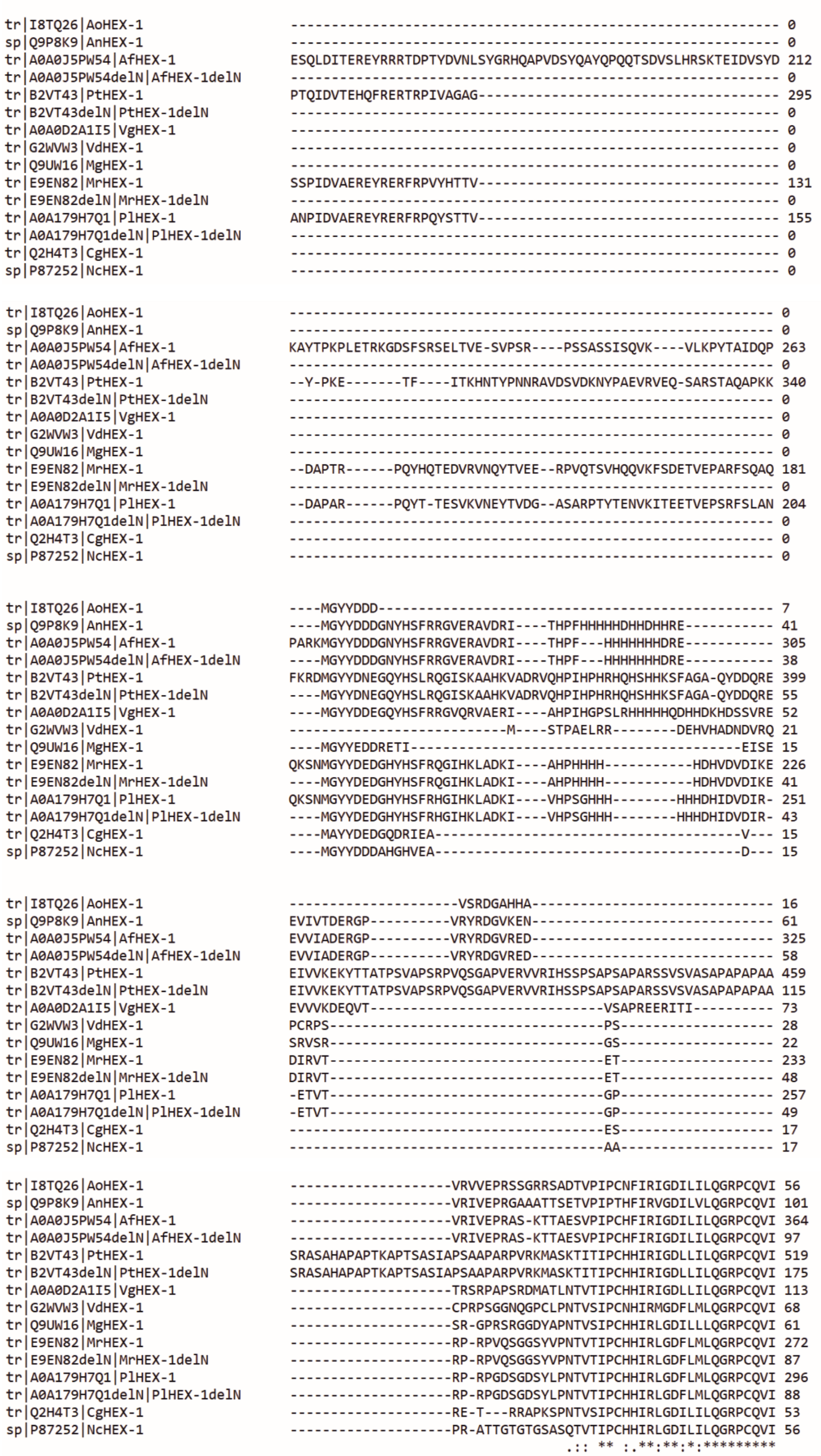
```
DETECTOR=EIGER
MINIMUM_VALID_PIXEL_VALUE=0  OVERLOAD=10048500
SENSOR_THICKNESS= 0.45

DIRECTION_OF_DETECTOR_X-AXIS= 1.0 0.0 0.0
DIRECTION_OF_DETECTOR_Y-AXIS= 0.0 1.0 0.0
ROTATION_AXIS= 0.0 1.0 0.0
INCIDENT_BEAM_DIRECTION=0.0 0.0 1.0
POLARIZATION_PLANE_NORMAL= 0.0 1.0 0.0
INCLUDE_RESOLUTION_RANGE= 999 0.0
FRACTION_OF_POLARIZATION=0.98

NX=4150 NY=4371  QX=0.075 QY=0.075
ORGX=2079.9702 ORGY=2160.9768
DETECTOR_DISTANCE= 374.465
X-RAY_WAVELENGTH= 0.9763       !Angstroem
OSCILLATION_RANGE= 1.0000
STARTING_ANGLE= 191.2600
STARTING_FRAME= 1326
NAME_TEMPLATE_OF_DATA_FRAMES= /mnt/beegfs/P14/2019/p3l-redecke/20191127/RAW_DATA/GMPRcyto_v2/
GMPRcytov2_191_?????.cbf.gz
DATA_RANGE= 1326 1350

JOB= XYCORR INIT COLSPOT IDXREF DEFPIX INTEGRATE CORRECT
FRIEDEL'S_LAW= TRUE
MAXIMUM_NUMBER_OF_PROCESSORS= 8
MAXIMUM_NUMBER_OF_JOBS= 22
MINIMUM_FRACTION_OF_INDEXED_SPOTS= 0.4
```

10.4 Amino acid sequences

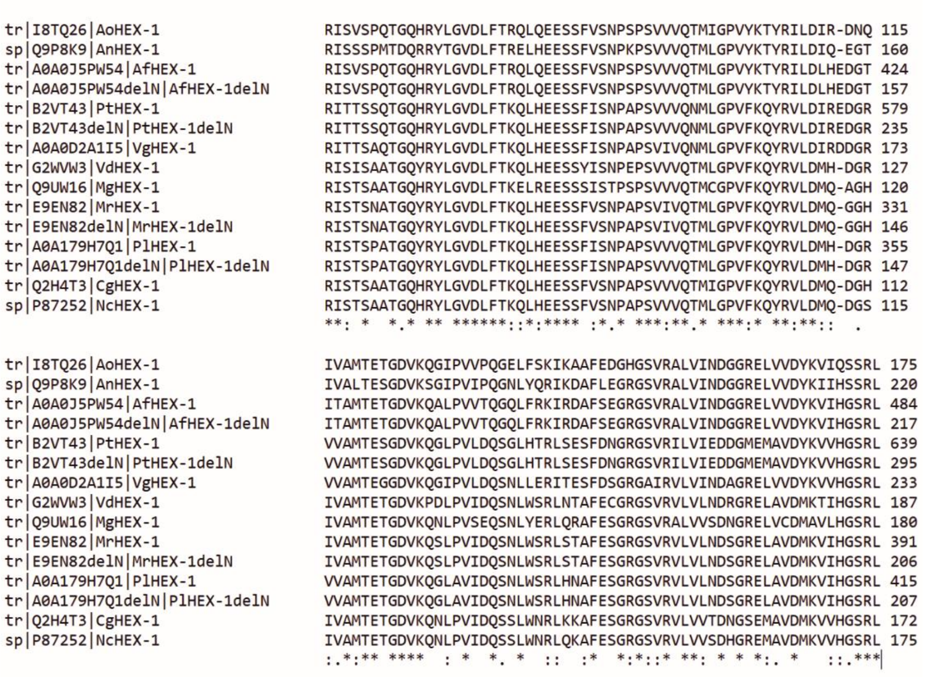
10.4.1 HEX-1 proteins



```
tr|I8TQ26|AoHEX-1                   ------------------------------------------------------------ 0
sp|Q9P8K9|AnHEX-1                   ------------------------------------------------------------ 0
tr|A0A0J5PW54|AfHEX-1               MYSVEIVSQQFERDSRRDAQRTANLDFDARVPIPFSVFPSSYRSDAVPETTL-TRVEGEV 59
tr|A0A0J5PW54delN|AfHEX-1delN       ------------------------------------------------------------ 0
tr|B2VT43|PtHEX-1                   --------------------------MTAIVPIPFSVFPSAYRNAELKDSASVQVVENEA 34
tr|B2VT43delN|PtHEX-1delN           ------------------------------------------------------------ 0
tr|A0A0D2A1I5|VgHEX-1               ------------------------------------------------------------ 0
tr|G2WVW3|VdHEX-1                   ------------------------------------------------------------ 0
tr|Q9UW16|MgHEX-1                   ------------------------------------------------------------ 0
tr|E9EN82|MrHEX-1                   ----------------------------MALPDPILLFV----------LKSLTHEEVDV 22
tr|E9EN82delN|MrHEX-1delN           ------------------------------------------------------------ 0
tr|A0A179H7Q1|PlHEX-1               -------MPL--WTRVNNDSKVANLDFEARVPVPFSIFPSSYRDSEAQTTTQQTHQEVDI 51
tr|A0A179H7Q1delN|PlHEX-1delN       ------------------------------------------------------------ 0
tr|Q2H4T3|CgHEX-1                   ------------------------------------------------------------ 0
sp|P87252|NcHEX-1                   ------------------------------------------------------------ 0


tr|I8TQ26|AoHEX-1                   ------------------------------------------------------------ 0
sp|Q9P8K9|AnHEX-1                   ------------------------------------------------------------ 0
tr|A0A0J5PW54|AfHEX-1               NLDRTSHVEREDT----------RTSAPLPDPRVYGREEVDIHIS------KDRLHAPSR 103
tr|A0A0J5PW54delN|AfHEX-1delN       ------------------------------------------------------------ 0
tr|B2VT43|PtHEX-1                   HFEHQALHGREGHEEVHESFSATVQQPPPPTPQQEQH--HDVHIKEEVHIHEEERLGGSQ 92
tr|B2VT43delN|PtHEX-1delN           ------------------------------------------------------------ 0
tr|A0A0D2A1I5|VgHEX-1               ------------------------------------------------------------ 0
tr|G2WVW3|VdHEX-1                   ------------------------------------------------------------ 0
tr|Q9UW16|MgHEX-1                   ------------------------------------------------------------ 0
tr|E9EN82|MrHEX-1                   KL--PQKQEPAGRQDQYSSFSAHVDLPPRGEHR--------------------------- 53
tr|E9EN82delN|MrHEX-1delN           ------------------------------------------------------------ 0
tr|A0A179H7Q1|PlHEX-1               QL--PPK-QPAGREGQYSSYTANVDLPHRGENR--------------------------- 81
tr|A0A179H7Q1delN|PlHEX-1delN       ------------------------------------------------------------ 0
tr|Q2H4T3|CgHEX-1                   ------------------------------------------------------------ 0
sp|P87252|NcHEX-1                   ------------------------------------------------------------ 0


tr|I8TQ26|AoHEX-1                   ------------------------------------------------------------ 0
sp|Q9P8K9|AnHEX-1                   ------------------------------------------------------------ 0
tr|A0A0J5PW54|AfHEX-1               KGDDFQVIYED------------------------------------------RAHKDSR 121
tr|A0A0J5PW54delN|AfHEX-1delN       ------------------------------------------------------------ 0
tr|B2VT43|PtHEX-1                   HGSQYQKTPQLPQQLSPRPSPKPSQHREEVRVHEHTSYRHPQQATPQQHREEIHIHEETR 152
tr|B2VT43delN|PtHEX-1delN           ------------------------------------------------------------ 0
tr|A0A0D2A1I5|VgHEX-1               ------------------------------------------------------------ 0
tr|G2WVW3|VdHEX-1                   ------------------------------------------------------------ 0
tr|Q9UW16|MgHEX-1                   ------------------------------------------------------------ 0
tr|E9EN82|MrHEX-1                   --------------------------------E---------------HREHRFSEEEVR 66
tr|E9EN82delN|MrHEX-1delN           ------------------------------------------------------------ 0
tr|A0A179H7Q1|PlHEX-1               -----------------------------------------------------YSEEEVR 88
tr|A0A179H7Q1delN|PlHEX-1delN       ------------------------------------------------------------ 0
tr|Q2H4T3|CgHEX-1                   ------------------------------------------------------------ 0
sp|P87252|NcHEX-1                   ------------------------------------------------------------ 0

tr|I8TQ26|AoHEX-1                   ------------------------------------------------------------ 0
sp|Q9P8K9|AnHEX-1                   ------------------------------------------------------------ 0
tr|A0A0J5PW54|AfHEX-1               V-------------------------P--------------------------------- 123
tr|A0A0J5PW54delN|AfHEX-1delN       ------------------------------------------------------------ 0
tr|B2VT43|PtHEX-1                   YRQPQQASPQQQREEIHIHEETRYRQPQQHQQPQQQREEIRIHEETRYRQPQGAQPLPQP 212
tr|B2VT43delN|PtHEX-1delN           ------------------------------------------------------------ 0
tr|A0A0D2A1I5|VgHEX-1               ------------------------------------------------------------ 0
tr|G2WVW3|VdHEX-1                   ------------------------------------------------------------ 0
tr|Q9UW16|MgHEX-1                   ------------------------------------------------------------ 0
tr|E9EN82|MrHEX-1                   F------------------SEERHRRPGFQQE-QFV------------------------ 83
tr|E9EN82delN|MrHEX-1delN           ------------------------------------------------------------ 0
tr|A0A179H7Q1|PlHEX-1               I----------------TREEERYRRPGFQHE-QYV------------------------ 107
tr|A0A179H7Q1delN|PlHEX-1delN       ------------------------------------------------------------ 0
tr|Q2H4T3|CgHEX-1                   ------------------------------------------------------------ 0
sp|P87252|NcHEX-1                   ------------------------------------------------------------ 0


tr|I8TQ26|AoHEX-1                   ------------------------------------------------------------ 0
sp|Q9P8K9|AnHEX-1                   ------------------------------------------------------------ 0
tr|A0A0J5PW54|AfHEX-1               -----------EVELSRERYRESTSRYEPLPPKPV---------YD-----------QAL 152
tr|A0A0J5PW54delN|AfHEX-1delN       ------------------------------------------------------------ 0
tr|B2VT43|PtHEX-1                   AQPHSHHAQRTDRVTDRVEFERSHDRYQPQPSSSHTQVEVQDRTYDRHFDTVHGPATDYA 272
tr|B2VT43delN|PtHEX-1delN           ------------------------------------------------------------ 0
tr|A0A0D2A1I5|VgHEX-1               ------------------------------------------------------------ 0
tr|G2WVW3|VdHEX-1                   ------------------------------------------------------------ 0
tr|Q9UW16|MgHEX-1                   ------------------------------------------------------------ 0
tr|E9EN82|MrHEX-1                   ---------KEDFRPGRTEYTDTRIE-----------IDTHQRP---------------Y 108
tr|E9EN82delN|MrHEX-1delN           ------------------------------------------------------------ 0
tr|A0A179H7Q1|PlHEX-1               ---------KEDRRPTRTEYNDTHIE-----------IDTHRHP---------------Y 132
tr|A0A179H7Q1delN|PlHEX-1delN       ------------------------------------------------------------ 0
tr|Q2H4T3|CgHEX-1                   ------------------------------------------------------------ 0
sp|P87252|NcHEX-1                   ------------------------------------------------------------ 0
```



```
tr|I8TQ26|AoHEX-1                   ------------------------------------------------------------ 0
sp|Q9P8K9|AnHEX-1                   ------------------------------------------------------------ 0
tr|A0A0J5PW54|AfHEX-1               ESQLDITEREYRRRTDPTYDVNLSYGRHQAPVDSYQAYQPQQTSDVSLHRSKTEIDVSYD 212
tr|A0A0J5PW54delN|AfHEX-1delN       ------------------------------------------------------------ 0
tr|B2VT43|PtHEX-1                   PTQIDVTEHQFRERTRPIVAGAG------------------------------------- 295
tr|B2VT43delN|PtHEX-1delN           ------------------------------------------------------------ 0
tr|A0A0D2A1I5|VgHEX-1               ------------------------------------------------------------ 0
tr|G2WVW3|VdHEX-1                   ------------------------------------------------------------ 0
tr|Q9UW16|MgHEX-1                   ------------------------------------------------------------ 0
tr|E9EN82|MrHEX-1                   SSPIDVAEREYRERFRPVYHTTV------------------------------------- 131
tr|E9EN82delN|MrHEX-1delN           ------------------------------------------------------------ 0
tr|A0A179H7Q1|PlHEX-1               ANPIDVAEREYRERFRPQYSTTV------------------------------------- 155
tr|A0A179H7Q1delN|PlHEX-1delN       ------------------------------------------------------------ 0
tr|Q2H4T3|CgHEX-1                   ------------------------------------------------------------ 0
sp|P87252|NcHEX-1                   ------------------------------------------------------------ 0

tr|I8TQ26|AoHEX-1                   ------------------------------------------------------------ 0
sp|Q9P8K9|AnHEX-1                   ------------------------------------------------------------ 0
tr|A0A0J5PW54|AfHEX-1               KAYTPKPLETRKGDSFSRSELTVE-SVPSR----PSSASSISQVK----VLKPYTAIDQP 263
tr|A0A0J5PW54delN|AfHEX-1delN       ------------------------------------------------------------ 0
tr|B2VT43|PtHEX-1                   --Y-PKE-------TF----ITKHNTYPNNRAVDSVDKNYPAEVRVEQ-SARSTAQAPKK 340
tr|B2VT43delN|PtHEX-1delN           ------------------------------------------------------------ 0
tr|A0A0D2A1I5|VgHEX-1               ------------------------------------------------------------ 0
tr|G2WVW3|VdHEX-1                   ------------------------------------------------------------ 0
tr|Q9UW16|MgHEX-1                   ------------------------------------------------------------ 0
tr|E9EN82|MrHEX-1                   --DAPTR------PQYHQTEDVRVNQYTVEE--RPVQTSVHQQVKFSDETVEPARFSQAQ 181
tr|E9EN82delN|MrHEX-1delN           ------------------------------------------------------------ 0
tr|A0A179H7Q1|PlHEX-1               --DAPAR------PQYT-TESVKVNEYTVDG--ASARPTYTENVKITEETVEPSRFSLAN 204
tr|A0A179H7Q1delN|PlHEX-1delN       ------------------------------------------------------------ 0
tr|Q2H4T3|CgHEX-1                   ------------------------------------------------------------ 0
sp|P87252|NcHEX-1                   ------------------------------------------------------------ 0


tr|I8TQ26|AoHEX-1                   ----MGYYDDD------------------------------------------------- 7
sp|Q9P8K9|AnHEX-1                   ----MGYYDDDGNYHSFRRGVERAVDRI----THPFHHHHHDHHDHHRE----------- 41
tr|A0A0J5PW54|AfHEX-1               PARKMGYYDDDGNYHSFRRGVERAVDRI----THPF---HHHHHHHDRE----------- 305
tr|A0A0J5PW54delN|AfHEX-1delN       ----MGYYDDDGNYHSFRRGVERAVDRI----THPF---HHHHHHHDRE----------- 38
tr|B2VT43|PtHEX-1                   FKRDMGYYDNEGQYHSLRQGISKAAHKVADRVQHPIHPHRHQHSHHKSFAGA-QYDDQRE 399
tr|B2VT43delN|PtHEX-1delN           ----MGYYDNEGQYHSLRQGISKAAHKVADRVQHPIHPHRHQHSHHKSFAGA-QYDDQRE 55
tr|A0A0D2A1I5|VgHEX-1               ----MGYYDDEGQYHSFRRGVQRVAERI----AHPIHGPSLRHHHHHQDHHDKHDSSVRE 52
tr|G2WVW3|VdHEX-1                   ---------------------------M----STPAELRR--------DEHVHADNDVRQ 21
tr|Q9UW16|MgHEX-1                   ----MGYYEDDRETI-----------------------------------------EISE 15
tr|E9EN82|MrHEX-1                   QKSNMGYYDEDGHYHSFRQGIHKLADKI----AHPHHHH-----------HDHVDVDIKE 226
tr|E9EN82delN|MrHEX-1delN           ----MGYYDEDGHYHSFRQGIHKLADKI----AHPHHHH-----------HDHVDVDIKE 41
tr|A0A179H7Q1|PlHEX-1               QKSNMGYYDEDGHYHSFRHGIHKLADKI----VHPSGHHH--------HHHDHIDVDIR- 251
tr|A0A179H7Q1delN|PlHEX-1delN       ----MGYYDEDGHYHSFRHGIHKLADKI----VHPSGHHH--------HHHDHIDVDIR- 43
tr|Q2H4T3|CgHEX-1                   ----MAYYDEDGQDRIEA--------------------------------------V--- 15
sp|P87252|NcHEX-1                   ----MGYYDDDAHGHVEA--------------------------------------D--- 15


tr|I8TQ26|AoHEX-1                   --------------------VSRDGAHHA------------------------------- 16
sp|Q9P8K9|AnHEX-1                   EVIVTDERGP----------VRYRDGVKEN------------------------------ 61
tr|A0A0J5PW54|AfHEX-1               EVVIADERGP----------VRYRDGVRED------------------------------ 325
tr|A0A0J5PW54delN|AfHEX-1delN       EVVIADERGP----------VRYRDGVRED------------------------------ 58
tr|B2VT43|PtHEX-1                   EIVVKEKYTTATPSVAPSRPVQSGAPVERVVRIHSSPSAPSAPARSSVSVASAPAPAPAA 459
tr|B2VT43delN|PtHEX-1delN           EIVVKEKYTTATPSVAPSRPVQSGAPVERVVRIHSSPSAPSAPARSSVSVASAPAPAPAA 115
tr|A0A0D2A1I5|VgHEX-1               EVVVKDEQVT-----------------------------VSAPREERITI---------- 73
tr|G2WVW3|VdHEX-1                   PCRPS----------------------------------PS------------------- 28
tr|Q9UW16|MgHEX-1                   SRVSR----------------------------------GS------------------- 22
tr|E9EN82|MrHEX-1                   DIRVT----------------------------------ET------------------- 233
tr|E9EN82delN|MrHEX-1delN           DIRVT----------------------------------ET------------------- 48
tr|A0A179H7Q1|PlHEX-1               -ETVT----------------------------------GP------------------- 257
tr|A0A179H7Q1delN|PlHEX-1delN       -ETVT----------------------------------GP------------------- 49
tr|Q2H4T3|CgHEX-1                   ---------------------------------------ES------------------- 17
sp|P87252|NcHEX-1                   ---------------------------------------AA------------------- 17

tr|I8TQ26|AoHEX-1                   --------------------VRVVEPRSSGRRSADTVPIPCNFIRIGDILILQGRPCQVI 56
sp|Q9P8K9|AnHEX-1                   --------------------VRIVEPRGAAATTSETVPIPTHFIRVGDILVLQGRPCQVI 101
tr|A0A0J5PW54|AfHEX-1               --------------------VRIVEPRAS-KTTAESVPIPCHFIRIGDILILQGRPCQVI 364
tr|A0A0J5PW54delN|AfHEX-1delN       --------------------VRIVEPRAS-KTTAESVPIPCHFIRIGDILILQGRPCQVI 97
tr|B2VT43|PtHEX-1                   SRASAHAPAPTKAPTSASIAPSAAPARPVRKMASKTITIPCHHIRIGDLLILQGRPCQVI 519
tr|B2VT43delN|PtHEX-1delN           SRASAHAPAPTKAPTSASIAPSAAPARPVRKMASKTITIPCHHIRIGDLLILQGRPCQVI 175
tr|A0A0D2A1I5|VgHEX-1               --------------------TRSRPAPSRDMATLNTVTIPCHHIRIGDLLILQGRPCQVI 113
tr|G2WVW3|VdHEX-1                   --------------------CPRPSGGNQGPCLPNTVSIPCNHIRMGDFLMLQGRPCQVI 68
tr|Q9UW16|MgHEX-1                   --------------------SR-GPRSRGGDYAPNTVSIPCHHIRLGDILLLQGRPCQVI 61
tr|E9EN82|MrHEX-1                   --------------------RP-RPVQSGGSYVPNTVTIPCHHIRLGDFLMLQGRPCQVI 272
tr|E9EN82delN|MrHEX-1delN           --------------------RP-RPVQSGGSYVPNTVTIPCHHIRLGDFLMLQGRPCQVI 87
tr|A0A179H7Q1|PlHEX-1               --------------------RP-RPGDSGDSYLPNTVTIPCHHIRLGDFLMLQGRPCQVI 296
tr|A0A179H7Q1delN|PlHEX-1delN       --------------------RP-RPGDSGDSYLPNTVTIPCHHIRLGDFLMLQGRPCQVI 88
tr|Q2H4T3|CgHEX-1                   --------------------RE-T---RRAPKSPNTVSIPCHHIRLGDILILQGRPCQVI 53
sp|P87252|NcHEX-1                   --------------------PR-ATTGTGTGSASQTVTIPCHHIRLGDILILQGRPCQVI 56
                                                                      .:: ** :.**:**:*:*********
```



```
tr|I8TQ26|AoHEX-1                 RISVSPQTGQHRYLGVDLFTRQLQEESSFVSNPSPSVVVQTMIGPVYKTYRILDIR-DNQ 115
sp|Q9P8K9|AnHEX-1                 RISSSPMTDQRRYTGVDLFTRELHEESSFVSNPKPSVVVQTMLGPVYKTYRILDIQ-EGT 160
tr|A0A0J5PW54|AfHEX-1             RISVSPQTGQHRYLGVDLFTRQLQEESSFVSNPSPSVVVQTMLGPVYKTYRILDLHEDGT 424
tr|A0A0J5PW54delN|AfHEX-1delN     RISVSPQTGQHRYLGVDLFTRQLQEESSFVSNPSPSVVVQTMLGPVYKTYRILDLHEDGT 157
tr|B2VT43|PtHEX-1                 RITTSSQTGQHRYLGVDLFTKQLHEESSFISNPAPSVVVQNMLGPVFKQYRVLDIREDGR 579
tr|B2VT43delN|PtHEX-1delN         RITTSSQTGQHRYLGVDLFTKQLHEESSFISNPAPSVVVQNMLGPVFKQYRVLDIREDGR 235
tr|A0A0D2A1I5|VgHEX-1             RITTSAQTGQHRYLGVDLFTKQLHEESSFISNPAPSVIVQNMLGPVFKQYRVLDIRDDGR 173
tr|G2WVW3|VdHEX-1                 RISISAATGQYRYLGVDLFTKQLHEESSYISNPEPSVVVQTMLGPVFKQYRVLDMH-DGR 127
tr|Q9UW16|MgHEX-1                 RISTSAATGQHRYLGVDLFTKELREESSSISTPSPSVVVQTMCGPVFKQYRVLDMQ-AGH 120
tr|E9EN82|MrHEX-1                 RISTSNATGQYRYLGVDLFTKQLHEESSFVSNPAPSVIVQTMLGPVFKQYRVLDMQ-GGH 331
tr|E9EN82delN|MrHEX-1delN         RISTSNATGQYRYLGVDLFTKQLHEESSFVSNPAPSVIVQTMLGPVFKQYRVLDMQ-GGH 146
tr|A0A179H7Q1|PlHEX-1             RISTSPATGQYRYLGVDLFTKQLHEESSFISNPAPSVVVQTMLGPVFKQYRVLDMH-DGR 355
tr|A0A179H7Q1delN|PlHEX-1delN     RISTSPATGQYRYLGVDLFTKQLHEESSFISNPAPSVVVQTMLGPVFKQYRVLDMH-DGR 147
tr|Q2H4T3|CgHEX-1                 RISTSAATGQHRYLGVDLFTKQLHEESSFVSNPAPSVVVQTMIGPVFKQYRVLDMQ-DGH 112
sp|P87252|NcHEX-1                 RISTSAATGQHRYLGVDLFTKQLHEESSFVSNPAPSVVVQTMLGPVFKQYRVLDMQ-DGS 115
                                  **: *  *.* ** ******::*:**** :*.* ***:**.* ***:* **:**::  .

tr|I8TQ26|AoHEX-1                 IVAMTETGDVKQGIPVVPQGELFSKIKAAFEDGHGSVRALVINDGGRELVVDYKVIQSSRL 175
sp|Q9P8K9|AnHEX-1                 IVALTESGDVKSGIPVIPQGNLYQRIKDAFLEGRGSVRALVINDGGRELVVDYKIIHSSRL 220
tr|A0A0J5PW54|AfHEX-1             ITAMTETGDVKQALPVVTQGQLFRKIRDAFSEGRGSVRALVINDGGRELVVDYKVIHGSRL 484
tr|A0A0J5PW54delN|AfHEX-1delN     ITAMTETGDVKQALPVVTQGQLFRKIRDAFSEGRGSVRALVINDGGRELVVDYKVIHGSRL 217
tr|B2VT43|PtHEX-1                 VVAMTESGDVKQGLPVLDQSGLHTRLSESFDNGRGSVRILVIEDDGMEMAVDYKVVHGSRL 639
tr|B2VT43delN|PtHEX-1delN         VVAMTESGDVKQGLPVLDQSGLHTRLSESFDNGRGSVRILVIEDDGMEMAVDYKVVHGSRL 295
tr|A0A0D2A1I5|VgHEX-1             VVAMTEGGDVKQGIPVLDQSNLLERITESFDSGRGAIRVLVINDAGRELVVDYKVVHGSRL 233
tr|G2WVW3|VdHEX-1                 IVAMTETGDVKPDLPVIDQSNLWSRLNTAFECGRGSVRVLVLNDRGRELAVDMKTIHGSRL 187
tr|Q9UW16|MgHEX-1                 IVAMTETGDVKQNLPVSEQSNLYERLQRAFESGRGSVRALVVSDNGRELVCDMAVLHGSRL 180
tr|E9EN82|MrHEX-1                 IVAMTETGDVKQSLPVIDQSNLWSRLSTAFESGRGSVRVLVLNDSGRELAVDMKVIHGSRL 391
tr|E9EN82delN|MrHEX-1delN         IVAMTETGDVKQSLPVIDQSNLWSRLSTAFESGRGSVRVLVLNDSGRELAVDMKVIHGSRL 206
tr|A0A179H7Q1|PlHEX-1             VVAMTETGDVKQGLAVIDQSNLWSRLHNAFESGRGSVRVLVLNDSGRELAVDMKVIHGSRL 415
tr|A0A179H7Q1delN|PlHEX-1delN     VVAMTETGDVKQGLAVIDQSNLWSRLHNAFESGRGSVRVLVLNDSGRELAVDMKVIHGSRL 207
tr|Q2H4T3|CgHEX-1                 IVAMTETGDVKQNLPVIDQSSLWNRLKKAFESGRGSVRVLVVTDNGSEMAVDMKVVHGSRL 172
sp|P87252|NcHEX-1                 IVAMTETGDVKQNLPVIDQSSLWNRLQKAFESGRGSVRVLVVSDHGREMAVDMKVVHGSRL 175
                                  :.*:** ****  : *  *. *  ::  :*  *:*::* **: * * *:. *   ::.***
```

10.4.2 Trypanosoma brucei GMPR and IMPDH

```
IMPDH                      MENTNLRTKTLRDGTTAEELFSQDGLSFNDFIILPGFIDFDS-SKVNVSGQFTKNILLHL  59
GMPR                       -------------MSFNESASIPTGLTYDDVLIIPQHSRVTSRKEVNTTTRLSRNVKLSI  47
                                         :  *.     **:::*.:*:* .  . * .:**.: ::::*: * :

IMPDH                      PLVSSPMDTVTESSMARAMALMGGIGVIHNNCTVEQQARMVRSVKLYRNGFIMKPKSVSP  119
GMPR                       PIVASNMDTVCEQRMAVAMAREGGIGILHRFCSIEEQCAMLREVKRAQSFLIESPRIILP  107
                           *:*:* **** *. ** ***  ****::*. *::*:*. *:*.**  :. :* .*: : *

IMPDH                      DVPVSTI---RNIKSEKGISGILVTEGGKYDGKLLGIVCTKDIDFVKDASAPVSQYMTRR  176
GMPR                       HETAREAWEGLNWKGRVGGVGCLLVVNCKNERKLLGIITRHDLKLADEST-TVESLMTPV  166
                           .  .       * *.. *  * *:. . * : *****:  :*:.:..:::  *.. **

IMPDH                      ENMT-VERYPIKLEEAMDVLNRSRHGYLPVLNDKDEVVCLCSRRDAVRARDYPNSSLDRN  235
GMPR                       DKMVVSTNTSISLEEVTHLMRKGRTANVPIVGQNGQLLYLVTLSDVVKLRKNKQASLDSR  226
                           ::*.   .  *.***. .::.:.* . :*::.::.::: * :  *.*: *.  ::*** .

IMPDH                       GHLLCAAATSTREADKGRVAALSEAGIDVLVLDSSQGNTIYQVSFIRWVKK--TYPHLEV 293
GMPR                       GRLLVGAAVGVKKDDMNRAIRLVEAGADVLVVDIAHGHSDLCINMVKRLKGDPRTASVDI  286
                           *:** .**...:: * .*. * *** ****:* ::*::   :.::: :*       :::

IMPDH                      VAGNVVTQDQAKNLIDAGADSLRIGMGSGSICITQEVLACGRPQATAIYKVARYAASRGV  353
GMPR                       IAGNIASAEAAEALIDAGADGLKIGVGPGSICITRLVAGAGVPQLSAVLACTRVARRRGV  346
                           :***:.: : *: *******.*:**:* ******: * ..* ** :*:   :* *  ***

IMPDH                      PCVADGGLRNVGDVCKALAVGANVAMLGSMIAGTSETPGEYFFKDGMRLKGYRGMGSIDA  413
GMPR                       PCIADGGLRTSGDISKAIGAGADTVMLGNMLAGTDEAPGRVLVKDGQKVKIIRGMAGFGA  406
                           **:******. **:.**:..**:..***.*:***.*:**. :.*** ::*  ***..:.*

IMPDH                      MLQGRESGKRYLSENETLQVAQGVAGAVLDKGSVLKLLAYIHKGLQQSAQDIGEVSFDAI  473
GMPR                       NLSKAERERTQDEDVFSSLVPEGVEGSVACKGPVGPIVRQLVGGLRSGMSYSGAKSIEEM  466
                            *.  *  :   .:  :  * :** *:*  ** *  ::  :  **:.. .  *  *:: :

IMPDH                      REKVYEGQVLFNRRTLTAQSEGAVHSLHHYERKLFASKL       512
GMPR                       QRR-----TRFVRMTGAGLRESGSHGVAKL---------       491
                           :.:     . * * * :.  *.. *.: :
```

10.5 Supplementary figures

10.5.1 Ligand binding sites in *Tb*IMPDH ori processed with *XDS*

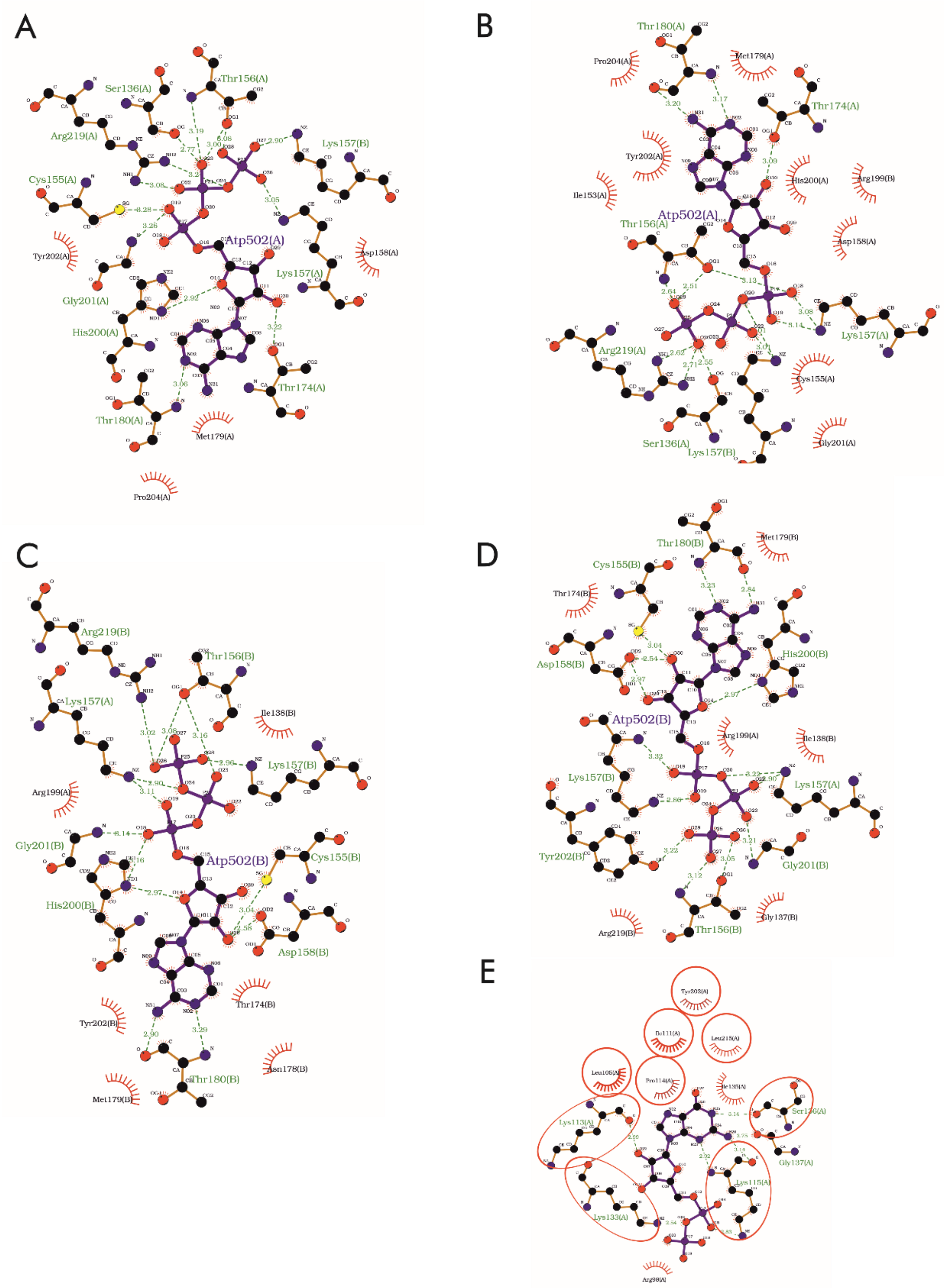


Figure 10.1: **LigPlot** ***representation of interaction sites of ATP bound to the canonical binding site I in the*** **Tb*****IMPDH ori structure which was processed***

with XDS. Interactions of ATP conformer B of chain A (A) and conformer A of chain A (B) as well as conformer B of chain B (C) and conformer A of chain B (D) are schematically displayed. (E): Overlay of the interactions of chain A and chain B GDP with the protein. Interactions occurring in both chains are marked by a red circle. Hydrogen bonds are depicted as dashed, green line while apolar interactions are indicated by red spikes.

10.5.2 Polder maps of cofactors bound to *Tb*GMPR cyto v2

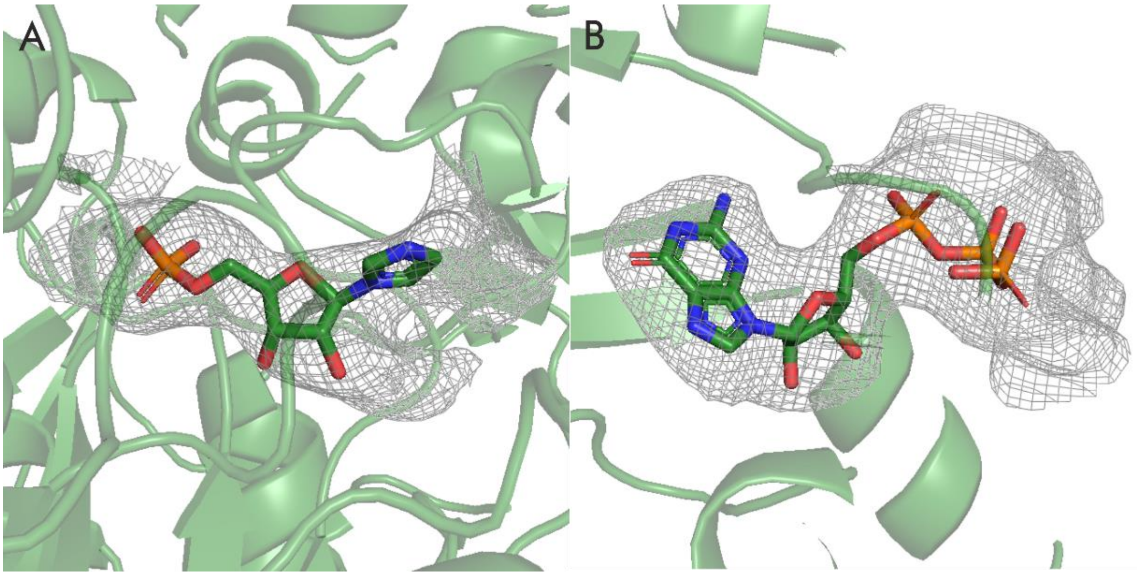


Figure 10.2: Polder omit maps of the monophosphate ribonucleotide (A) and the GTP moiety (B) of chain A bound to the protein structure of* Tb*GMPR cyto v2. The polder omit map is contoured at 3 sigma and carved at 4 Å.