Single Particle 2D Electron Crystallography for Membrane Protein Structure Determination

Inauguraldissertation

Zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch – Naturwissenschaftlichen Fakultät
der Universität Basel

von

Ricardo Diogo Righetto

aus Campinas, Brasilien

Basel, 2019

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Henning Stahlberg, Fakultätsverantwortlicher
Prof. Dr. Volker Roth, Korreferent

Basel, 21.05.2019

Prof. Dr. Martin Spiess,
The Dean of the Faculty

Summary

Proteins embedded into or attached to the cellular membrane perform crucial biological functions. Despite such importance, they remain among the most challenging targets of structural biology. Dedicated methods for membrane protein structure determination have been devised since decades, however with only partial success if compared to soluble proteins.

One of these methods is 2D electron crystallography, in which the proteins are periodically arranged into a lipid bilayer. Using transmission electron microscopy to acquire projection images of samples containing such 2D crystals, which are embedded into a thin vitreous ice layer for radiation protection (cryo-EM), computer algorithms can be used to generate a 3D reconstruction of the protein. Unfortunately, in nearly every case, the 2D crystals are not flat and ordered enough to yield high-resolution reconstructions. Single particle analysis, on the other hand, is a technique that aligns projections of proteins isolated in solution in order to obtain a 3D reconstruction with a high success rate in terms of high resolution structures.

In this thesis, we couple 2D crystal data processing with single particle analysis algorithms in order to perform a local correction of crystal distortions. We show that this approach not only allows reconstructions of much higher resolution than expected from the diffraction patterns obtained, but also reveals the existence of conformational heterogeneity within the 2D crystals. This structural variability can be linked to protein function, providing novel mechanistic insights and an explanation for why 2D crystals do not diffract to high resolution, in general. We present the computational methods that enable this hybrid approach, as well as other tools that aid several steps of cryo-EM data processing, from storage to postprocessing.

Contents

1. Introduction

In this chapter, an introduction to techniques for determining the 3D structures of proteins is provided, with an emphasis on membrane proteins. The principles of transmission electron microscopy and its different modalities are also covered. Finally, we explain the challenges of the 2D electron crystallography technique and other computational bottlenecks addressed in this thesis.

Contents

1.1. Protein structure

The cell is the basic unit of life. All living organisms consist of one or more cells. *Eukaryotic* cells feature a nucleus that contains the genetic information of the organism, while *prokaryotic* cells do not have such a nucleus, and carry the genome directly in their cytoplasm. While prokaryotic cells are always unicellular organisms, i.e. bacteria and archaea, eukaryotic cells form more complex organisms that can be uni- or multicellular, such as protozoa, algae, fungi, plants and animals. Functions and processes that are vital to a cell such as replication, catalysis of chemical reactions, biochemical signaling, transporting of molecules and defense are carried out by specialized macromolecules called *proteins*.

The genetic code of an organism is stored in a deoxyribonucleic acid (DNA) molecule, which is formed by a sequence of four nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). Segments of the DNA are used as templates for synthesizing another type of polymer, the ribonucleic acid (RNA), which is a sequence of the following four nucleotides: adenine (A), uracil (U), cytosine (C) and guanine (G). There is a direct relationship between DNA and RNA sequences, hence the name of this process is *transcription* [Alberts 2015].

In a subsequent process, triplets of RNA nucleotides, termed *codons*, encode for an aminoacid, which are the building blocks of proteins. The conversion from a sequence of codons into a sequence of aminoacids is called *translation*. Aminoacids are composed by an *amine* (-NH2) and a *carboxylic acid* (-COOH) group, plus a *side-chain*. There are twenty standard aminoacids with unique side-chains: alanine (Ala/A), arginine (Arg/R), asparagine (Asn/N), aspartic acid (Asp/D), cysteine (Cys/C), glutamic acid (Glu/E), glutamine (Gln/Q), glycine (Gly/G), histidine (His/H), isoleucine (Ile/I), leucine (Leu/L), lysine (Lys/K), methionine (Met/M), phenylalanine (Phe/F), proline (Pro/P), serine (Ser/S), threonine (Thr/T), tryptophan (Trp/W), tyrosine (Tyr/Y) and valine (Val/V).

A portion of DNA that encodes for a specific protein sequence is known as a *gene*. While the transcription from DNA to RNA is a 1:1 mapping between nucleotides and therefore a reversible process, the translation from RNA to protein sequence necessarily implies a loss of information, meaning that it is not possible to unambiguously determine the RNA sequence that originated a given protein sequence. This directional flow of information is known as the *"central dogma of molecular biology"* [Crick 1970].

While the three-dimensional (3D) shape of a protein is in principle dictated by its aminoacid sequence, it is not yet possible to accurately predict the folding of a protein based on sequence information alone[1]. Furthermore, the folding of a

[1] For an account on recent advances in protein structure prediction using machine learning

protein is also affected by multiple environmental factors and by the interaction with other molecules, such that specific claims on their structures ultimately require experimental evidence. The scientific field of *structural biology* is concerned with determining the 3D structures of proteins and nucleic acids as accurately as possible, and, more importantly, how these structures explain protein function at the atomic and molecular level [Branden & Tooze 1999]. This knowledge allows for a precise description of biological processes on physical and chemical grounds, enabling us to make predictions and interventions, such as the design of drugs targeting diseases caused by a malfunctioning protein, or by an external threat to the cells, like viruses, for example.

The dimension of a protein and its components is usually measured in Å units ($1\,\text{Å} = 1 \times 10^{-10}\,\text{m}$)[2], and its mass in Da units ($1\,\text{Da} = 1.660\,538\,92 \times 10^{-27}\,\text{kg}$). Protein structures can be broken down into the following hierarchical levels, in increasing order of complexity:

1. The **primary structure** of a protein is simply the sequence in which the aminoacids appear in the *polypeptide chain*. Protein sequences are read from the N- terminal (positively charged) of the first aminoacid to the C- terminal (negatively charged) of the last aminoacid. In this context, the aminoacids can also be referred to as *residues*.

2. The **secondary structure** consists of how the protein chain folds into specific motifs in 3D. The two most common secondary structure elements found in proteins are α-helices and β-sheets. These structural patterns arise from energetically favorable interactions between adjacent aminoacids in 3D. While the position of α-helices can be observed at a resolution of $\sim$7 Å, its helical *pitch* can only be observed at $\sim$4 Å resolution or better, and β-sheet observation requires a resolution of at least $\sim$5 Å. Secondary structure elements are usually connected by *loops* of residues.

3. The **tertiary structure** is the specific conformation assumed by a folded protein chain in 3D. It is described by the coordinates of all atoms of all residues in the protein. Side-chains for all aminoacids can be unambiguously assigned at a resolution of $\sim$3 Å, but truly resolving separated atomic positions requires a resolution of at least 1.2 Å. Structures resolved at a resolution of $\sim$2 Å are said to be of “near-atomic” resolution, although these terms have been used somewhat liberally across the structural biology field [Wlodawer & Dauter 2017]. The tertiary structure is further subdivided into protein *domains*, which are

techniques, please see the blog post by Mohammed AlQuraishi [WWW Document], n.d. URL https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/ (accessed 22.2.2019)

[2] The radius of a hydrogen atom is approximately 0.5 Å.

independent and functional parts of the aminoacid sequence that can fold and exist on their own [Alberts 2015]. Domains are connected to each other by loops and/or secondary structure elements.

4. The **quaternary structure** is the assembly of multiple protein chains (i.e. tertiary structures) into a *multimer*, or a macromolecular complex. This process is also called *oligomerization*. Multimers formed by identical polypeptide chains are *homomers*, while distinct chains assembled together form an *heteromer*. As the name suggests, multimers are formed by two or more *monomers*, which in functional terms are also referred to as *subunits*. The spatial arrangement between subunits in a macromolecular complex will depend on residue hydrophobicity, electrostatic charges and mechanical forces, among other effects [Branden & Tooze 1999]. In many cases, the quaternary structure of a protein is symmetric in 3D space.

Because genetic codes are evolutionarily related [Branden & Tooze 1999], similar protein sequences may be found across widely different organisms. The high correlation between sequence identity and 3D structure similarity [Chothia & Lesk 1986] implies that determining the structure of a protein from one organism may provide insights for the mechanism of the same or similar protein in other organisms as well. This is why model organisms such as *E. Coli* (prokaryote) or *C. Elegans* and *S. cerevisiae* (eukaryotes) greatly simplify the endeavour of molecular and structural biology [Alberts 2015].

1.1.1. Membrane proteins

The contents of a cell are separated from the surrounding environment by a *membrane*. This membrane is mainly composed of phospholipids. The amphiphilic nature of phospholipids makes them spontaneously assemble into a *lipid bilayer*. That is, the water-insoluble end of the phospholipids (the “tails”) face the inner side of the bilayer, while the water-soluble end (the “head”) face the aqueous environments on both sides of the membrane. Such lipid bilayers act as a container and protector to the cell, and are impermeable to water, ions and other molecules. A cell needs, however, to be able to fetch nutrients from its environment, excrete the wastes of energy production, and communicate with other cells. Such tasks are performed by *membrane proteins*. They are different from *soluble proteins* in that they have large portions of exposed surface that is insoluble. Proteins that have domains fully or partially embedded into the lipid bilayer are called *integral* membrane proteins, while proteins that are only attached or bound to the lipid bilayer are denominated *peripheral* membrane proteins. Cellular organelles and vesicles are also spatially delimited by a lipid mono- or bilayer [Alberts 2015].

Some of the most important membrane proteins are "gatekeepers" like ion and water channels, substrate importers and exporters like ATP-binding cassette (ABC) transporters, signal detectors like G protein-coupled receptors (GPCR), and energy converters such as ATP synthases [Branden & Tooze 1999, Alberts 2015]. As membrane proteins act as mediators between the *internal* and *external* environments of the cell in many ways, they are important targets for drug development [Goldie *et al.* 2014]. Surprisingly, though, very little is known about the structure of membrane proteins in comparison to soluble proteins. Out of the 151,364 protein structures deposited at the Protein Data Bank (PDB)[3], only 2,741 correspond to structures of membrane proteins (as of 28.4.2019), despite them accounting for ∼25% of all genes [Stansfeld *et al.* 2015]. The number narrows further down when considering that, out of those, only 890 correspond to unique structures, i.e., not accounting for different conformations or substrate-binding states of the same protein. The difficulty in handling membrane proteins arises from the hydrophobic nature of their domains, meaning they are not stable in solution [Abeyrathne *et al.* 2012]. Nevertheless, technological advances in structural determination methods, introduced in the next section, are enabling a steady growth in the number of membrane protein structures available [White 2009], as shown in Fig. 1.1.

1.2. Techniques for structure determination

The shortest wavelength of visible light is of about 0.2 µm, meaning that cells, which range in size from ∼1 µm–100 µm ($1\,\mu\text{m} = 1 \times 10^{-6}\,\text{m}$), can be visualized by conventional light microscopy [Alberts 2015]. Protein dimensions are much smaller, however, in the range of ∼1 nm–1000 nm ($1\,\text{nm} = 1 \times 10^{-9}\,\text{m}$). Furthermore, the structural details of proteins describing the spatial arrangement of atoms is one order of magnitude smaller ($1\,\text{Å} = 1 \times 10^{-10}\,\text{m}$). Therefore, special physical techniques are required to retrieve high-resolution information on the structure of proteins. We will introduce in the next sections the three main techniques for high-resolution protein structure determination: X-ray crystallography, nuclear magnetic resonance, and transmission electron microscopy. In all cases, it is assumed that the protein of interest has been expressed and purified in suitable amounts [Branden & Tooze 1999, Alberts 2015].

1.2.1. X-ray crystallography

X-ray crystallography, or X-ray diffraction (XRD), is the oldest and most widely used structural biology technique. A special type of particle accelerator, called a *synchrotron light source*, consists of an electron beam accelerated and injected into

[3] https://www.ebi.ac.uk/pdbe

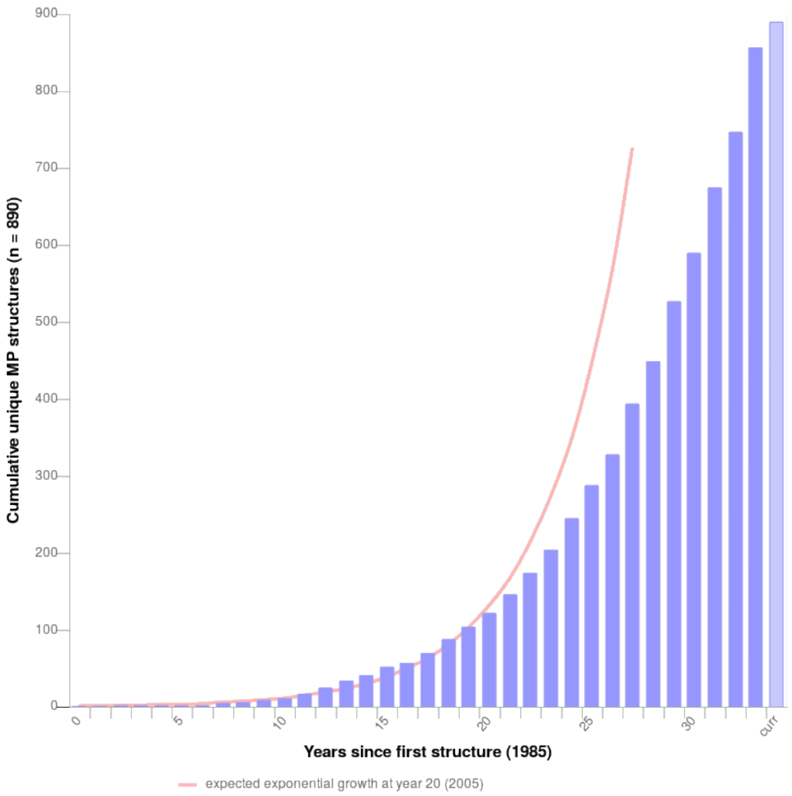


Figure 1.1: **Number of membrane protein structures determined per year.** As of 28.4.2019, there were 2,741 membrane protein structures deposited in the PDB, with only 890 of those corresponding to unique structures. Source: Stephen White Lab at UC Irvine, `https://blanco.biomol.uci.edu/mpstruc/`. Please see [White 2009] for more details.

a storage ring, where it is kept circulating close to the speed of light in high vacuum. The deflection of the electron beam by electromagnetic coils produces X-rays, which are photons with a wavelength range of ~0.01 nm–10 nm. Specifically, for biological macromolecular crystallography, wavelengths in the range of ~0.5 Å–2 Å are typically used. Tangent to the storage ring are *beamlines* where the X-rays are harnessed for imaging and diffraction experiments. In the case of macromolecular crystallography, *diffraction patterns* of protein crystals are recorded at multiple orientations and merged in 3D. Because the crystals contain multiple copies of the protein of interest arranged in a highly periodical manner, the diffraction patterns detected consist of regularly spaced peaks or *diffraction spots*. The repetitive unit in a crystal is called the *unit cell*, which in turn contains one or more copies of the *asymmetric unit* related by symmetry operations. These peaks correspond to the intensities of the Fourier transform (FT) of the unit cell. The nature of X-ray diffraction implies that the phase information of the Fourier transform is not recorded, and has to be estimated by other methods such as molecular replacement (when the structure of a similar protein is already known) or anomalous scattering experiments in order to obtain an electron density map of the structure in real space [Branden & Tooze 1999].

It was from an X-ray diffraction pattern that James Watson and Francis Crick, based on the work of Rosalind Franklin, inferred the double-stranded helical structure of the DNA [Watson & Crick 1953]. A few years later, the first protein structures, myoglobin [Kendrew *et al.* 1960] and hemoglobin [Perutz *et al.* 1960], were determined. However only in 1985 the first high-resolution structure of a membrane protein was obtained [Deisenhofer *et al.* 1985]. There is no limitation to XRD in terms of molecular size or mass, in principle. The main bottleneck of the technique, however, is the very nature of the sample: growing 3D crystals of proteins is difficult in most cases [Branden & Tooze 1999], and even more so for membrane proteins, although dedicated crystallization methods have been devised for them [Landau & Rosenbusch 1996]. Many biochemical conditions have to be screened for crystallization, and often special constructs of the protein chains have to be made, involving mutation, insertion or deletion of residues, such that resulting structures are often unlikely to represent physiological conditions found in nature. Despite these limitations, 83.7% of all structures deposited in the PDB were determined by XRD (126,694 structures as of 29.4.2019)[4].

1.2.2. Nuclear magnetic resonance

A technique that does not require crystals for structure determination is nuclear magnetic resonance (NMR) spectroscopy. Briefly, this method exploits the fact that certain atomic nuclei, such as that of hydrogen, have a non-zero magnetic *spin*. Therefore, in the presence of a strong constant magnetic field, the spins of these atoms will all be aligned. Upon perturbation of the field by radiofrequency (RF) pulses of electromagnetic radiation, the nuclei of these atoms will be excited. When shifting back to the aligned state, they emit radiation that can be detected and quantified. The intensity and frequency of the radiation will depend on the environment around each hydrogen nucleus [Alberts 2015]. Using prior knowledge about the protein sequence and biophysical parameters, computational methods can analyze the NMR spectra and derive pairwise distance relationships between atomic nuclei, and consequently reveal an *ensemble* of structures that agree with the experimental data [Branden & Tooze 1999].

While crystallization is not a requirement, the fact that NMR is usually performed

[4] https://www.rcsb.org/stats/summary

in solution also poses a challenge for studying membrane proteins. Nevertheless, it has been successfully used to solve the structure of membrane proteins giving the membrane environment can be approximated by the use of lipid micelles or nanodiscs [Liang & Tamm 2016]. The main limitation of NMR as a structural determination tool is that the method is restricted to small proteins only, typically smaller than 20 kDa [Alberts 2015]. In the range where it is applicable, it is a powerful tool for investigating protein structure and dynamics, accounting for 7.3% of all the structures currently deposited in the PDB (11,050 structures as of 29.4.2019).

1.2.3. Transmission electron microscopy

The transmission electron microscope (TEM) was invented by Ernst Ruska and Max Knoll in 1933. Electrons accelerated at 300 kV have a wavelength of ∼0.02 Å, much shorter than that of visible light (∼0.2 µm) or X-rays (∼0.01 nm). The TEM exploits this fact to generate images of a thin sample at very high magnification, allowing even atoms to be visible under special conditions [Williams & Carter 2009]. In his visionary lecture *There's plenty of room at the bottom*, Richard Feynman recognized that being able to *look* at things at such a small scale with the TEM would revolutionize fields as diverse as nanotechnology, computer science, and molecular biology [Feynman 1960].

Briefly, electrons emitted by an *electron gun*, such as a tungsten or LaB_6 filament in the simplest machines, are focused by electromagnetic lenses (coils) into a coherent beam that is transmitted through a thin sample in vacuum. Upon interaction with the sample, a *phase shift* of the electron *wave front* occurs. Assuming the sample is thin and light enough to be considered a *phase contrast object* and ignoring dynamic scattering effects, the difference in phase between incident electrons generates contrast in the image plane. This image corresponds to the integrated Coulomb potential across the sample. Details about TEM construction, operation and theory can be found in the excellent books by [Reimer & Kohl 2008] and [Williams & Carter 2009]. In this thesis we will cover only the most important aspects from the point of view of image processing for structural biology. More extensive information about 3D electron microscopy of biological specimens can also be found in the book by Joachim Frank [Frank 2006].

Contrast Transfer Function

Perhaps the most striking and non-intuitive characteristics of TEM images is the modulation by a *contrast transfer function* (CTF) in Fourier space, and its equivalent *point spread function* (PSF) in real space. The CTF arises from constructive and destructive interference patterns between scattered and unscattered electrons incident on the image plane. The precise form of this modulation depends mainly on the defocus at which the image is recorded and the acceleration voltage [Wade 1992, Mindell & Grigorieff 2003]. If defocus varies in different directions, the image is said to be *astigmatic*. The implication is that different spatial frequencies are *transferred* to the image with a resolution-dependent amplitude modulation and phase reversal. At the exact frequencies where the phase reversal occurs, the amplitude modulation is zero, meaning the information at these resolution bands is lost. In Fourier space, this behavior is clearly visualized as a pattern of concentric rings, or ellipses in the presence of astigmatism [Thon 1966]. On top of the oscillatory behavior of the CTF, imperfections in the optical system of the microscope, vibrations and other effects cause an amplitude decay towards the higher resolutions, i.e. an *envelope function*. An example of the CTF and PSF effects is illustrated in Fig. 1.2 for a synthetic image. In contrast to diffraction methods, TEM imaging records both amplitude and phase information. Because of the CTF, however, a direct interpretation of the image is not possible. In order to obtain a meaningful image or reconstruction, it is first necessary to *flip the phases* of the frequencies where contrast is reversed, and ideally also correct for the amplitude modulation by a Wiener-like filter [Penczek 2010c]. Still, the information at the zero-crossings of the CTF is permanently lost, and can only be restored by averaging multiple images taken at different defoci values (i.e. where the zero-crossings appear at different frequencies).

Cryo-Electron Microscopy

Soft matter such as proteins and nucleic acids are highly sensitive to radiation damage by the electron beam [Baker & Rubinstein 2010]. Furthermore, the vacuum inside the electron microscope column is a harsh environment to the sample, which must be kept hydrated. Simply freezing the sample destroys the structural details due to the formation of ice crystals. A breakthrough in overcoming these challenges for high-resolution structural analysis was obtained by the group of Jacques Dubochet, which found a way to freeze the samples so rapidly that the formation of ice crystals is avoided, embedding the sample into a thin layer of vitreous water instead [Adrian *et al.* 1984]. The sample is flash-frozen by plunging it into liquid ethane, and subsequently kept cooled by liquid nitrogen (colder than −160 °C) to maintain its vitreous state. This method significantly decreases beam-induced radiation damage to the sample. Because the sample is kept at cryogenic temperatures, a range of electron microscopy techniques that make use of this type of sample conditioning are commonly referred to as *cryo-electron microscopy*, or simply *cryo-EM*.

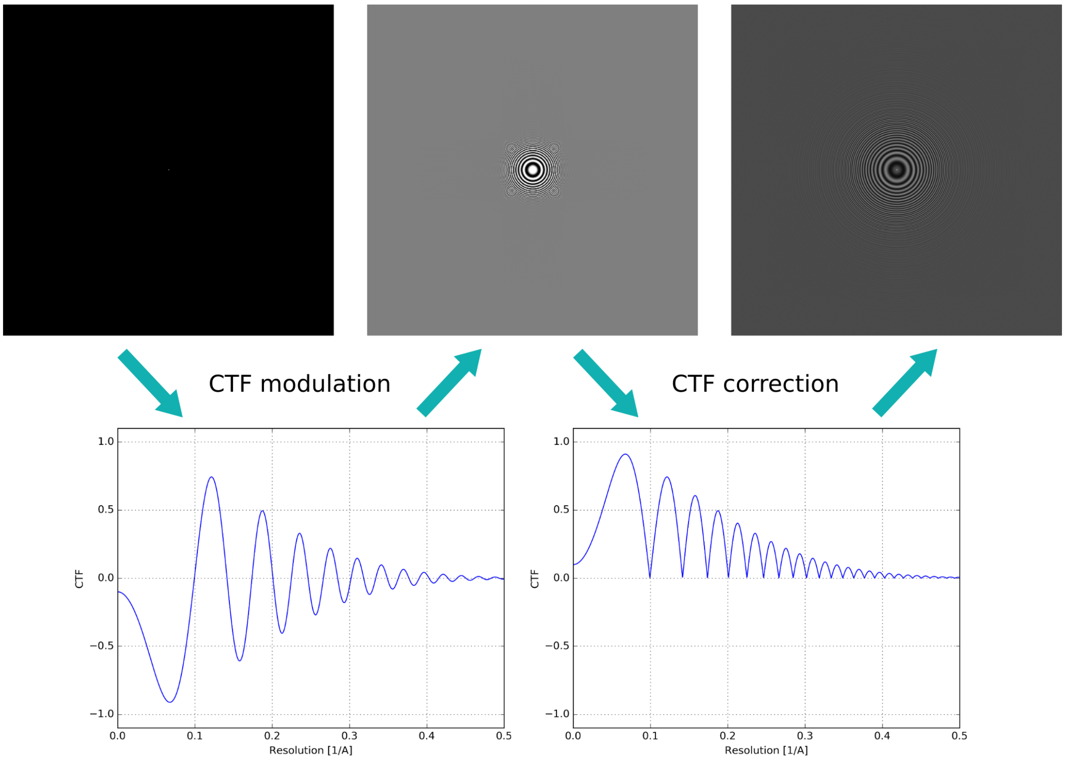


Figure 1.2: **The contrast transfer function and its correction.** The top-left panel shows a simulated image of 512×512 pixels that is zero everywhere, except the central pixel which has a value of 1.0. Following the arrows, the FT of the image is multiplied by a CTF modulation curve (here shown in 1D for convenience), in this example corresponding to a defocus of -0.5 μm, spherical aberration $C_s = 2.7$ mm, amplitude contrast of 0.07 and a B-factor of 80 Å^2 representing an envelope function [Mindell & Grigorieff 2003]. The resulting image in real space (top middle panel) displays reversed contrast and signal delocalisation represented by circular fringes. The modulated image is then corrected in Fourier space by phase-flipping, with the resulting real space image shown in the top-right panel. While the central peak is partially restored, the image still contain fringes related to the lack of amplitude correction, the limited signal sampling, and the irreversible signal loss at the zeros of the CTF.

Direct Electron Detectors

Even with the cryogenic protection described above, biological samples still suffer considerable radiation damage in the TEM. In order to keep this damage to a minimum, regions of the sample are exposed to very low electron doses for image recording. The low dose implies that cryo-TEM images have a very low signal-to-noise (SNR) ratio. For decades, photographic film was used to record electron micrographs. This meant that micrographs had to be developed and digitized (scanned) in order to be analyzed by computer programs, in a very time consuming process. Digital detectors such as charge-coupled devices (CCD) which came into use at the turn of the century [Downing & Hendrickson 1999] offered easier data logistics and the possibility to automate data collection. The drawback of such detectors was that the need to convert electrons into photons by means of a scintillator introduced distortions and reduced the detective quantum efficiency (DQE) of the detector. For this reason, photographic film was still preferred for high-resolution structural analysis for many years after [Zhang *et al.* 2010], despite the convenience of digital data acquisition offered by CCD cameras. The scenario changed with the introduction of *direct electron detectors* (DED) [Milazzo *et al.* 2011]. As the name indicates, this type of camera is capable of detecting when an electron hits a pixel of the sensor without the need of converting them to photons, which significantly increases the DQE [McMullan *et al.* 2016].

Not only the DEDs offer images with a better SNR: as they are capable of acquiring images at a high frame rate, they opened up the possibility to record *movies* instead of static images. Based on this hardware feature, computational methods were developed that correct for *beam-induced motion*, which was one of the main resolution-limiting factors for the technique [Brilot *et al.* 2012]. In short, these algorithms align (globally or locally) the frames of a movie, in order to generate an aligned average that contains much more high-resolution information than the simple, unaligned average [Brilot *et al.* 2012, Li *et al.* 2013a, Zheng *et al.* 2017]. Besides that, as radiation damage increases with the accumulated exposure, recording movies allows the dose to be fractionated along the consecutive frames. Based on this, *dose-weighting* filters were developed, which essentially allow only the initial frames, when the protein is still relatively undamaged, to contribute to the high-resolution frequencies [Grant & Grigorieff 2015, Scheres 2014, Zivanov *et al.* 2019].

3D Reconstruction

In the weak phase object approximation, TEM images are 2D *projections* of the 3D object [Williams & Carter 2009, Frank 2006]. The *central section theorem* states that the Fourier transform of a 2D projection is a central slice through the FT of the 3D object [De Rosier & Klug 1968]. This fact is exploited to obtain 3D reconstructions of proteins and other objects of interest. Given the relative orientations of the projections are known, they can be merged in 3D space and the object can be reconstructed by Fourier inversion [Penczek 2010a]. This is typically performed in an expectation-maximization iterative procedure, where an improved knowledge about the alignment parameters of the individual images leads to a better 3D reconstruction and vice-versa, until convergence.

Cryo-EM is actually an umbrella term encompassing different techniques that differ mainly in the nature of the sample conditioning and the geometry of the data collection, while sharing many common concepts and approaches. Currently, electron microscopy accounts for 1.5% of all the structures currently deposited in the PDB (2,264 structures as of 29.4.2019) and the number is rapidly increasing. The main "modalities" of cryo-EM are briefly explained ahead:

- **Single particle analysis (SPA)** is the mainstream cryo-EM method. It consists of imaging a sample containing multiple copies of the protein of interest, which are randomly oriented in the vitreous ice layer [Frank 2006]. *Picking* is the process of detecting and extracting candidate particles from the micrographs. The goal is then to identify the relative orientation of the multiple projections acquired in order to obtain a 3D reconstruction [Cheng *et al.* 2015b].

  In order to analyze large sets of very noisy images, computational approaches based on statistical pattern recognition methods have been developed. It is particularly important to avoid *overfitting*, i.e. mistakenly assuming random noise for actual features of the structure [Henderson *et al.* 2012, Henderson 2013]. In principle, the particles can be aligned and their orientations assigned by maximizing their similarity to projections of the current 3D volume [Frank 2006]. This procedure is called *projection matching*, and a similarity measure commonly employed is the *cross-correlation* (CC). This approach is, however, prone to overfitting if naively implemented.

  Multivariate statistical analysis (MSA) methods such as principal component analysis (PCA) and related approaches have been introduced early in the field [van Heel & Frank 1981, van Heel *et al.* 2000]. This type of dimensionality reduction provides for computational speedup and prevents overfitting by considering only the most statistically relevant components of the dataset. In other words, random noise tends to be described by principal components of small associated eigenvalues which therefore can be left out of the analysis [van Heel *et al.* 2016].

  More recently, maximum-likelihood [Sigworth 1998, Sigworth *et al.* 2010] and

Bayesian [Scheres 2012a] methods were introduced. These methods help preventing overfitting by accounting for alignment errors in a probabilistic manner, and weighting the contribution of each projection at each spatial frequency according to estimates of their spectral SNR. These methods are, however, computationally expensive, although their implementation in graphics processing unit (GPU) cards have greatly favored their adoption [Kimanius *et al.* 2016].

Other important measures for preventing overfitting, used in conjunction with the methods above, consist in splitting the dataset into random halves that are refined independently [Scheres & Chen 2012], at least beyond an imposed resolution limit [Grigorieff 2016, Grant *et al.* 2018]. In SPA and other cryo-EM techniques, resolution is typically assessed by the Fourier shell correlation (FSC) curve, which indicates how well two reconstructions obtained from random halves of the particle set agree with each other as a function of spatial resolution [Harauz & van Heel 1986]. There is not a consensus, however, on the threshold at which the FSC should be assessed to derive a single number describing the resolution of the reconstruction [Rosenthal & Henderson 2003, van Heel & Schatz 2005], with the best practice being to assess the curve as a whole [Penczek 2010b].

Proteins are molecular machines, meaning they often undergo conformational changes in order to perform their functions. In stark contrast to X-ray crystallography, SPA offers the possibility of disentangling multiple structures from heterogeneous datasets. This is done by unsupervised or semi-supervised classification approaches, that can be based on MSA [van Heel *et al.* 2012], maximum-likelihood [Scheres *et al.* 2005, Lyumkis *et al.* 2013], Bayesian [Scheres 2012a] or stochastic gradient descent (SGD) [Punjani *et al.* 2017] algorithms. These methods allow snapshots of protein conformational changes to be resolved from the same sample, which can arise due to intrinsic structural flexibility or be related to the interaction with substrates and ligands [Nogales & Scheres 2015].

The introduction of DEDs, together with improvements in instrumentation, sample preparation and data processing methods, has greatly increased the popularity of cryo-EM and of SPA in particular [Kühlbrandt 2014], as it opens up the possibility of determining structures that would not crystallize and were too large for NMR. Membrane proteins can be studied by SPA in the presence of detergents [Liao *et al.* 2013] or inserted into a lipid nanodisc [Gao *et al.* 2016]. That is, provided they can be embedded into an environment resembling the cellular membrane, as in NMR. The main limitation of SPA is in the molecular weight: too small proteins (typically smaller than 100 kDa) don't yield enough signal for alignments, although advances in instrumentation are relaxing this constraint [Khoshouei *et al.* 2017]. On the other hand, too big structures, such as large viruses, are computationally demanding to resolve and violate the assumptions commonly made for 3D reconstruction in SPA [Zhang & Hong Zhou 2011].

- **Electron tomography (ET or cryo-ET)** is a method for obtaining a 3D reconstruction of a whole cell, or parts of a cell or tissue. In this technique, the same area of the sample is exposed to the electron beam while rotating along an axis at a constant angular step [Baumeister *et al.* 1999]. As the relative orientation of the micrographs (or movies) recorded varies only by the tilt angle, a 3D reconstruction of the area of interest can be readily obtained, although the actual orientation and alignment should be refined for optimal results [Mastronarde & Held 2017]. If multiple copies of a structure can be detected in reconstructed tomograms, they can then be aligned and classified to generate a higher-resolution map. This process is called *subtomogram averaging* (STA) and is analogous to SPA, only with 3D maps instead of 2D projections as input data [Castaño-Díez *et al.* 2012]. The main limitation of cryo-ET at present consist in the sample holders, that cannot be tilted beyond ∼60°, generating the so-called *missing wedge* of information in Fourier space. Furthermore, tilted images are of much lower quality than the non-tilted images, because of larger drifts and lower contrast due to the increased speciment thickness. Still, in cases of exceptionally well-behaved samples and careful data acquisition, high resolution structures can be obtained by STA [Schur *et al.* 2016].

- **2D Electron Crystallography** is the oldest of the methods based on TEM for structural determination, along with the closely related helical reconstruction method [De Rosier & Klug 1968, Crowther *et al.* 1970]. This method is especially suited for solving membrane protein structures, although it can also be used to study soluble proteins under certain conditions [Schmidt-Krey & Cheng 2013]. In this technique, many thousands of copies of the protein of interest is arranged in a bidimensional periodical array, i.e. the 2D crystal. In the case of membrane proteins, they are typically embedded in a lipid bilayer. These 2D crystals can either be grown by artificially inducing a self-assembling process, or naturally occur in the cell membrane [Abeyrathne *et al.* 2012].

  As in XRD, the periodical nature of the 2D crystals is represented by Bragg spots in Fourier space. By collecting information from many 2D crystals at different orientations, including different tilt angles, and merging their diffraction spots in the 3D reciprocal space, a reconstruction can be obtained by Fourier inversion [Stahlberg *et al.* 2015]. If the TEM is operated in diffraction mode, only diffraction intensities are recorded, and the phase information has to be retrieved from somewhere else, as in XRD [Rossmann & Henderson 1982]. If working in imaging mode, the recorded projections of 2D crystals contain both amplitudes and phase information, although at a much lower SNR if compared to the diffraction mode, and modulated by the CTF. This is the case of interest discussed in the rest of this work.

This technique was used to reveal the first ever 3D structure of a membrane protein, bacteriorhodopsin, although initially only at low resolution [Henderson & Unwin 1975]. The adoption of cryo-EM to prevent radiation damage then enabled bacteriorhodopsin and other structures to be solved at high resolution [Henderson *et al.* 1990, Kühlbrandt *et al.* 1994, Nogales *et al.* 1998, Gonen *et al.* 2005].

As in cryo-ET, the mechanical limit in the tilting angle generates a *missing cone* of information in reciprocal space, whose effect is to make the structures appear elongated along the vertical axis, perpendicular to the membrane plane. This information can be restored to some extent by computational methods using prior knowledge about the sample [Gipson *et al.* 2011, Biyani *et al.* 2018]. The major limitation that has prevented a widespread adoption of the technique is, however, the intrinsic disorder of 2D crystals. The structures mentioned above that have been determined at high resolution represent exceptional cases where the 2D crystals are flat and ordered nearly to perfection. The vast majority of 2D crystals are not that well behaved, displaying curvatures and other imperfections. A key step in ameliorating this problem was the invention of the *unbending* algorithm [Henderson *et al.* 1986]. Briefly, this algorithm compares the predicted unit cell positions from the detected lattice based on the diffraction spots, with their real-space positions as detected by cross-correlation with a reference. This reference may come either from a central area of the 2D crystal itself or from a projection of the current 3D volume. The difference in the positions obtained by the two methods indicate the displacement of the unit cells, that can be computationally corrected by shifting small patches of the 2D crystal in real space. After convergence, the iterative procedure yields a “perfect”, i.e. non-distorted lattice, that is verified by sharper diffraction spots in reciprocal space. This approach, combined with weighting of the indexed spots according to their SNR and iterative refinement of the tilt geometry, allows higher resolution maps to be obtained from 2D electron crystallography. Details of the data processing steps employed in 2D electron crystallography can be obtained in [Schenk *et al.* 2010] and [Arheit *et al.* 2013]. A major limitation remains, however, in that out-of-plane distortions of the 2D crystal are not addressed by conventional unbending.

1.3. Structure and aim of this thesis

The goal of this thesis is to demonstrate whether and how single particle analysis algorithms can be used to process 2D crystal data. Specifically, we are interested in showing that aligning particles extracted from 2D crystals against a 3D reference is equivalent to locally unbending the crystal in 3D, thus addressing out-of-plane distortions and improving the resolution of the reconstructions.

The introduction of DEDs along with microscope automation tools provoked an explosion in the amount of data generated and processed by cryo-EM laboratories. Data compression techniques are therefore desired to minimize the burden of data transfer and storage. Chapter 2 presents an extension to the MRC file format, the most widely used in the field of cryo-EM for storing images and volumes, that addresses this problem. Our extension, MRCZ, natively supports codecs from the *blosc* library, leveraging the CPU cache for high performance I/O. While this chapter is not specifically related to 2D electron crystallography, the Python module created for MRCZ is used as basis for the computational tools developed in the later chapters.

Along the course of this work, tools for the automation of cryo-EM data processing and curation were developed, along with computational methods for improving 2D electron crystallography reconstructions. These tools have been integrated in the software package *FOCUS* and are described in previous publications [Biyani *et al.* 2017, Biyani *et al.* 2018] and doctoral theses [Scherer 2015, Biyani 2017]. Chapter 3 builds on previous studies on the correction of 2D crystal distortions [Scherer *et al.* 2014] to present a new module in the FOCUS package, which allows the user to export a 2D crystallography project for processing with state-of-the-art SPA programs. We show that, by using the new FOCUS module to export patches of the 2D crystals to a modified version of the FREALIGN package [Grigorieff 2016], it is possible to obtain map of the MloK1 potassium channel at higher resolution than by conventional electron crystallography, from the same dataset [Kowal *et al.* 2018]. More interestingly, though, is the finding of heterogeneous conformations of MloK1 within the 2D crystals, something that can only be achieved by 3D classification in SPA.

Chapter 4 further expands on the concepts of single particle 2D electron crystallography, providing more methodological details and practical advice for users. The data processing approach devised should be applicable independent of the specific software package adopted, although with some differences in how they are used and the results interpreted.

In Chapter 5 we present an application of the methods presented in the previous chapters on a challenging biological case. We used the hybrid single particle/2D crystal approach to solve the structure of *focal adhesion kinase* (FAK), a small membrane-attached protein. As with MloK1, the use of SPA algorithms significantly improves resolution compared to the crystallographic case, and reveals distinct crystalline arrangements within the highly mosaic 2D crystals.

Finally, Chapter 6 concludes with a summary of the work presented in each chapter, with a discussion on their advances, limitations, implications and possible directions of future work.

1.4. Publication list

Below is a list of journal articles and conference proceedings derived directly or indirectly from the work presented in this thesis:

Published or submitted articles

Righetto, R. & Stahlberg, H. **Single Particle Analysis for High Resolution 2D Electron Crystallography**. Methods Mol. Biol. (2019). (invited submission for book chapter publication)
This is Chapter 4 of the thesis.

Righetto, R., Biyani, N., Kowal, J., Chami, M. & Stahlberg, H. **Retrieving High-Resolution Information from Disordered 2D Crystals by Single Particle Cryo-EM**. Nat. Commun. 10, 1722 (2019). `http://doi.org/10.1038/s41467-019-09661-5`
This is Chapter 3 of the thesis.

Schmidli C., Albiez S., Rima L., Righetto, R., Mohammed I., Oliva P., Kovacik L., Stahlberg, H. & Braun T. **Microfluidic protein isolation and sample preparation for high resolution cryo-EM**. bioRxiv 556068 (2019). `http://doi.org/10.1101/488403` (preprint; submitted for peer-reviewed publication)

Biyani, N., Scherer, S., Righetto, R., Kowal, J., Chami, & Stahlberg, H. **Image processing techniques for high-resolution structure determination from badly ordered 2D crystals**. J. Struct. Biol. 203, 120–134 (2018). `http://doi.org/10.1016/j.jsb.2018.03.013`

Biyani, N., Righetto, R., McLeod, R., Caujolle-Bert, D., Castaño-Diez, D., Goldie, K. & Stahlberg, H. **Focus: The interface between data collection and data processing in cryo-EM**. J. Struct. Biol. 198, 124–133 (2017). `http://doi.org/10.1016/j.jsb.2017.03.007`

McLeod, R. A., Righetto, R., Stewart, A. & Stahlberg, H. **MRCZ - A file format for cryo-TEM data with fast compression**. J. Struct. Biol. 0–1 (2017). `http://doi.org/10.1016/j.jsb.2017.11.012`
This is Chapter 2 of the thesis.

Conference proceedings

Daday, C., Acebrón, I., Simon, M., Righetto, R., Lietha, D. & Gräter, F. **When an Enzyme Self-Assembles on a Membrane: Focal Adhesion Kinase**. Biophys. J. 114-3, 61 (2018). http://doi.org/10.1016/j.bpj.2017.11.382

In preparation

Righetto, R., Adaixo, R., Anton, L., Schwede, T., Maier, T. & Stahlberg, H. **Revisiting the structure of urease by high-resolution cryo-EM**.

Adaixo, R., Righetto, R., Ni, D., Taylor, N., & Stahlberg, H. **The structure of human thyroglobulin**.

Acebrón, I., Righetto, R., Culley, J., Daday, C., Biyani, N., Redondo, P., Chami, M., Boskovic, J., Gräter, F., Frame, M., Stahlberg, H. & Lietha, D. **Membrane binding induces domain rearrangements and oligomerization that prime FAK for activation**.
This is Chapter 5 of the thesis.

2. MRCZ – A file format for cryo-TEM data with fast compression

We present in this chapter an extension to the MRC file format, called MRCZ, which natively supports fast compression algorithms. MRCZ addresses the rapid increase in the amount of data generated by direct electron detectors in the field of cryo-EM. The C and Python I/O libraries developed for MRCZ also support the computational tools presented in the later chapters.

Contribution: implementation of features for MRCZ file handling in Python and benchmarking scripts.

The following section has been published in:

Journal of Structural Biology
Volume 201, Issue 3, March 2018, Pages 252-257
http://doi.org/10.1016/j.jsb.2017.11.012

MRCZ – A file format for cryo-TEM data with fast compression

Robert A. McLeod[1], Ricardo Diogo Righetto[1], Andy Stewart[2], Henning Stahlberg[1]

[1] Center for Cellular Imaging and NanoAnalytics, Biozentrum, University of Basel, Mattenstrasse 26, CH-4058 Basel, Switzerland

[2] Department of Physics, University of Limerick, Limerick, Ireland

Contents

Abstract

The introduction of fast CMOS detectors is moving the field of transmission electron microscopy into the computer science field of big data. Automated data pipelines control the instrument and initial processing steps which imposes more onerous data transfer and archiving requirements. Here we conduct a technical demonstration whereby storage and read/write times are improved 10× at a dose rate of 1 e^{-}/pix/frame for data from a Gatan K2 direct-detection device by combination of integer decimation and lossless compression. The example project is hosted at github.com/em-MRCZ and released under the BSD license.

2.1. Introduction

The introduction of CMOS-based direct electron detectors for transmission electron microscopy greatly improved the duty cycle to nearly 100% compared to traditional slow-scan CCD detectors. The high duty-cycle allows for nearly continuous readout, such that dose fractionation has become ubiquitous as a means to record many-frame micrograph stacks in-place of traditional 2D images. The addition of a time-dimension, plus the large pixel counts of CMOS detectors, greatly increases both archival and data transfer requirements and associated costs to a laboratory. Many laboratories have a 1 Gbit/s Ethernet connection from their microscope to their computing center, which implies a data transfer rate of around 60–90 MB/s under typical conditions. If the microscope is run with automated data collection, such as SerialEM [Mastronarde 2005], then the so-called ‘movie’ may be 5–20 GB and may be saved every few minutes, or even faster. In such a case, it may not be possible to transfer the data fast enough to keep up with collection. Costs for storing data on spinning (hard disk) storage, for example through the use of Google Cloud[5], is typically US$100–200/TB/year. A cryo-TEM laboratory producing 200 TB of data per year is potentially faced with an annual data storage cost on the same order of magnitude as a post-doctoral fellow salary.

One approach whereby considerable archival savings may be realized is by decimation of the data from floating-point format to integer-format. Nominally, the analog-to-digital converted signal from the detector is typically output as an integer. Due to data processing requirements, it is often necessary to convert the integer data to 32-bit floating point format. The most common initial step that results in decimal data is the application of a gain reference, where the bias of the detector white values is removed. In-addition, conversion to floating-point is often inevitable due to operations such as sub-pixel shifting in drift correction [Li *et al.* 2013a, Li

[5] Google Cloud Storage Pricing | Cloud Storage Documentation [WWW Document], n.d. Google Cloud Platform. URL `https://cloud.google.com/storage/pricing` (accessed 3.11.17)

et al. 2013b, Grant & Grigorieff 2015, McLeod *et al.* 2017b, Zheng *et al.* 2017], or image filtration. If instead the micrographs are stored as 8-bit integers, with the gain reference (and potentially other operations) stored in meta-data, then a 4× reduction in storage and transfer requirements is realized. In this case, the gain reference and other bias corrections must be performed at the computing center, rather than using the software provided by the direct electron detector vendor. Since vendor gain normalization techniques are often proprietary and secret, there is a need for open-source equivalent solutions [Afanasyev *et al.* 2015].

Further improvements in data reduction can be realized by modern high-speed lossless compression codes. Lossless compression methods operate on the basis of repeated patterns in the data. Nominally, purely-random numbers are incompressible. However counting electron data is Poisson distributed, such that its range of pixel histogram covers on only a limited range of values. In such a regime substantial compression ratios may be achieved. Therefore due to the repetition of intensity values, integer-format data can be compressed much more efficiently than gain-normalized floating-point data. Generally when comparing compression algorithms one is interested in the compression rate (in units of megabytes/s) and the compression ratio (in percent). Modern compression codecs such as Z-standard[6] or LZ4[7] are designed for efficient multi-threaded operation on modern, parallel CPUs and can compress on the order of 1–2 GB/s/core, such that the time for read/write/transfer plus compression operations is greatly faster than when operating on uncompressed data.

We demonstrate here combining decimation to 8-bit integer with lossless compression. We utilize an extension of the venerable MRC format[8] [Cheng *et al.* 2015a], where meta-compression is implemented which implies the combination of lossless compression and lossless filtering to improve compressibility as well as execution with efficient blocked and multi-threaded processing. We propose using common serialization tools to embed metadata in the MRC2014 extended header, and compare JSON[9] and Message Pack[10].

[6] facebook/zstd [WWW Document], n.d. GitHub. URL `https://github.com/facebook/zstd` (accessed 3.13.17).

[7] lz4/lz4 [WWW Document], n.d. GitHub. URL `https://github.com/lz4/lz4` (accessed 3.13.17).

[8] MRC/CCP4 file format for images and volume [WWW Document]. n.d. CCP-EM, URL `http://www.ccpem.ac.uk/mrc_format/mrc_format.php` (accessed 7.27.17).

[9] Standard ECMA-404 [WWW Document], n.d. URL `http://www.ecma-international.org/publications/standards/Ecma-404.htm` (accessed 3.11.17).

[10] MessagePack: It's like JSON. but fast and small. [WWW Document], n.d. URL `http://msgpack.org/index.html` (accessed 3.11.17).

2.2. The MRCZ format

The MRC format was introduced by [Crowther *et al.* 1996] as an extension of the CCP4 format. It features a 1024-byte metadata header, followed by binary image data with provisions for 3-dimensions. The supported data types are byte (int8), short (int16), or single-precision floating-point (float32). The simplicity of the MRC format, and its ease of implementation, is a likely reason contributing to its popularity. However, the MRC format suffers from some drawbacks. There is no one standard format for MRC, in-spite of many efforts to define one [Cheng *et al.* 2015a]. Furthermore, it cannot compress the data, so it is inefficient from an archival and transmission/distributed computing perspective.

An alternative public domain archival format for electron microscopy is HDF5. However, HDF5 is a "heavyweight" library consisting of ∼350,000 lines of code and 150-pages of specification[11], which makes integration in existing projects difficult. HDF5 has previously demonstrated compression filters including *blosc* and an additional LZ4-based filter funded by Dectris (Baden, CH)[12].

Here we introduce an evolution of the MRC format, MRCZ, with additional functionalities that have become needed in the era of 'Big Data' in electron microscopy. We provide sample libraries for MRCZ in C/99 and also Python 2.7/3.5, as well as a command-line utility that may be used to compress/decompress MRC files so that legacy software can read the output. To facilitate the introduction of MRCZ into other software packages, we have kept the implementations as small as possible (currently *c-mrcz* is <1000 lines of code).

The MRCZ library package leverages an open-source, meta-compression library, *blosc* (blocking, shuffle, compression), principally written by Francesc Alted[13] and Valentin Haenel [Haenel 2014]. *Blosc* combines multi-threaded compression (currently six different codecs are available) with blocking, such that each operation fits in CPU cache (typically optimized to level 2 cache), and filter operations (namely *shuffle* and *bitshuffle*). In testing on cryo-TEM data *blosc* achieved >10 GB/s compression rates on a modern CPU, and furthermore achieves superior compression ratios to codecs such as LZW [Welch 1984] implemented in TIFF. The performance gain is sufficient such that loading a compressed image stack from disk and applying post-processing gain normalization and outlier pixel filtering to it is faster than loading an uncompressed but pre-processed floating-point result. The Python version

[11] HDF5 File Format Specification Version 3.0 [WWW Document], n.d. URL `https://support.hdfgroup.org/HDF5/doc/H5.format.html` (accessed 1.3.17).

[12] Nexus Format/HDF5-External-Filter-Plugins [WWW Document], n.d. URL `https://github.com/nexusformat/HDF5-External-Filter-Plugins` (accessed 13.9.17).

[13] Alted, K. 2014. *Blosc*, an extremely fast, multi-threaded, meta-compressor library [WWW Document]. *Blosc* Main Page. URL `http://www.blosc.org/index.html` (accessed 3.13.17).

of the library also supports asynchronous file writing and reading, where the file is read or written in a background thread, freeing the interpreter for other tasks.

Here results for three compressors are compared for operation on cryo-EM data. Other codecs were tested, including the new Lizard codec (released in March 2017), but not found to have performance advantages for the test data:

1. lz4 is the fastest compressor, with the worst compression ratio, making it ideal for live situations where distributing the data from a master computer is the priority.

2. zStandard (zstd): achieves the highest compression ratio and has the fastest decompression, making it the best choice for archiving. On the lowest compression level it maintains good compression rates.

3. Zlib is a very common library that has been accelerated by blosc. Zlib provides a valuable baseline for comparison, although blosc compression rate with zlib exceeds that of tools such as pigz.

2.2.1. Blocked compression

Roughly around 2005, further increases in CPU clock-frequencies were slowed due to heat generation limitations. Further performance improvements where then realized by packing parallel arithmetic and logic cores per chip. Most common compression algorithms were designed before the era of parallel processing.

Operations in image processing are often relatively simple and executed on the full-frame consisting of many million elements. With the larger number of cores available on modern CPU, often program execution rate is limited not by processing power but the amount of memory bandwidth available to feed data to the cores. Typically fetching data from random-access memory (RAM) is an order of magnitude slower than the cache found on the CPU die. Therefore if the data can be cut into blocks that fit into the lower-level caches large speed improvements are often observed. Parallel algorithms can be made to work efficiently in the case where a computational task can be cut into blocks, and each block can be dispatched to an individual core, and run through an algorithm to completion, as illustrated in Fig. 2.1a. Parallel algorithms should also avoid branching instructions (e.g. conditional if statements), as modern processors request instructions from memory in-advance, and a wrong guess can leave the process idle waiting for memory. For example, *zStandard* also makes use an of a faster and more effective method for evaluating entropy, known as Asymmetric Numeral System (ANS), than classic compression algorithms. ANS significantly improves compression ratio in data with large degrees of randomness [Duda 2013, Duda *et al.* 2015].

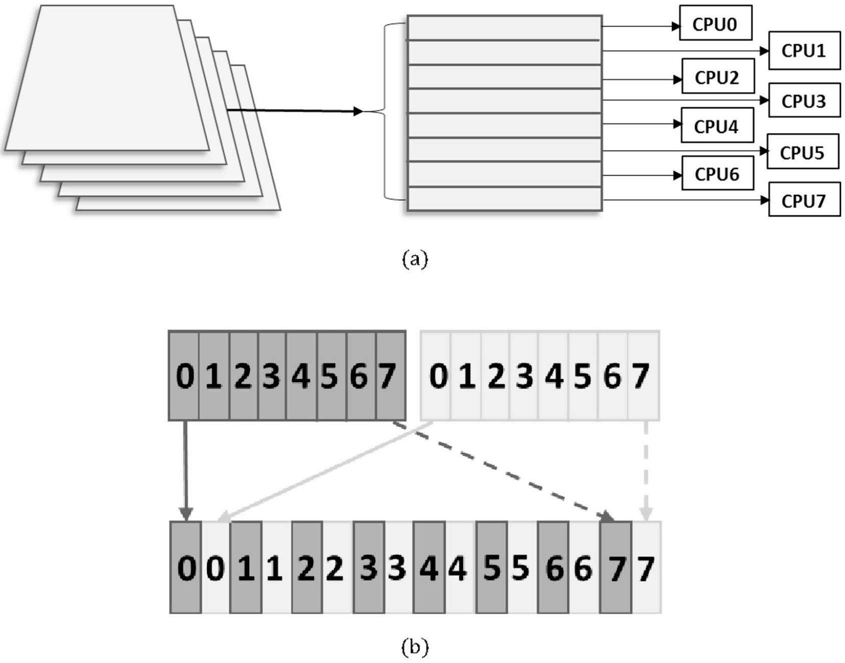


Figure 2.1: **MRC compression and decompression with *blosc*.** (a) Each MRC volume is chunked, such that each z-axis slice is compressed separately. Then in *blosc* each chunk is further sliced into blocks, which are then dispatched to individual CPU cores for compression. Decompression works in reverse. (b) Normally pixel values are stored in memory contiguously (top row). With bit-shuffling (on little endian systems) the most significant bits (7 index) are stored adjacently, and similarly for the least-significant bits (0 index). This improves compressibility and as a result both compression ratio and compression rate are improved.

In blocking strategies, data is conceptually separated into chunks and blocks, with chunks being senior to blocks. In the MRCZ format, each chunk is a single image frame (∼16 million pixels for 4 k detectors, or ∼64 million for 8 k), and each chunk is broken into numerous blocks, with a default block size of 1 MB. Such a block size provides a balanced trade-off between compression rate and the ratio between compressed and uncompressed data. For image stacks, the highest compression ratio would likely be in the time-axis, but this is the least convenient axis for chunking, as it would make retrieving individual frames or slices of frames impossible.

2.2.2. Bit-decimation by shuffling

Direct electron detectors may be operated in counting mode whereby the ratio of dose rate to detector cycle rate is low enough that two electrons landing in the same pixel on a rapidly cycling detector is statistically rare. The order of magnitude of the usable dose rate before mis-counting is on the order of magnitude of $1e^{-}/pix/frame$ per 100 Hz cycle rate of the detector. Typically drift correction is performed on the time scale of a second, so for the K2 Summit (Gatan, Pleasanton, CA) operating at 400 Hz the expected dose per pixel in an integrated frame is ∼1–8. This implies that even a single byte (*uint8*) to store each pixel is too large of a data container, as it can hold data values up to 255. David Mastronarde implemented in SerialEM and IMOD [Mastronarde 2005] a new data type for MRC that incorporates a decimation step where each pixel is packed into 4-bits, leading to maximum per-pixel values of 16 before clipping occurs, thereby providing an effective compression ratio of 2.0 compared to *uint8*. The disadvantage of 4-bit packed data is that it is not a hardware data type, such that two pixels are actually packed into an 8-bit integer. Whenever the data is loaded into memory for processing, it must be unpacked with bit-shifting operations, which is computationally not free. There is also the risk of intensity-value clipping.

Blosc optionally makes use of a filter step, of which there are two currently implemented, *shuffle* and *bitshuffle*. *Shuffle* re-arranges each pixel by its most significant byte to least, whereas *bitshuffle* performs the same task on a bit-level, illustrated in Fig. 2.1b. The shuffle-style filters are highly efficient when the underlying data has a narrow histogram, such that the most-significant digit in a pixel has more commonality with other pixels' most significant digit than its own least-significant digit. For example, if an image saved as *uint8* type contains mostly zeros in its most significant digits, they will be bit-shuffled into a long-series of zeros, which is trivially compressible. As such, *bitshuffle* effectively performs optimized data decimation without any risk of clipping values. Shuffling is also effective for floating-point compression, as the sign bit and the exponent are compressible whereas the mantissa usually does not contain repeated values and therefore it is not especially compressible. The mantissa can be made more compressible by rounding to some significant bits, for example the nearest 0.001 of an electron, but this generates round-off error.

2.2.3. Benchmarks

Benchmarks for synthetic random Poisson data were conducted for images covering a range of electron dose levels consisting of $[0.1, 0.25, 0.5, 1.0, 1.5, 2.0, 4.0]$ electron counts/pixel. The free parameters examined consist of: compression codec, block size, threads, and compression level were all evaluated. Here the term 'compression level' refers to the degree of computing effort the algorithm will use to achieve higher compression ratios. The machine specification for benchmark results is as follows:

Two Intel® Xeon® E5-2680 v3 CPUs operating with Hyperthreading® and Turboboost®:

- No. of physical cores: 2 × 12
- Average clock rate: 2.9 GHz (spec: 2.4 GHz)
- L1 cache size: 32 kB per core
- L2 cache size: 256 kB per core
- L3 cache size: 30,720 kB per processor

The size of the L2 cache generally has a large impact on the compression rate as a function of the blocksize used by *blosc*. For file I/O the RAID0 hard drive used was benchmarked to have a read/write rate of ×300 MB/s, which is comparable to parallel-file systems in general use in cluster environments.

Example benchmarks on cryo-TEM image stacks are shown in Table 2.1 for a variety of *blosc* libraries as well as external compression tools. *Uint4* refers to the SerialEM practice of interlaced packing of two pixels into a single-byte. JPEG2000 and *uint4* were not multi-threaded; all other operations used 48 threads. *Pigz*, *lbzip2* and *pxz* are command-line utilities and hence include a read and write. Indicated times are averages over 20 read/writes. To achieve repeatable result, the disk was flushed between each operation, with the Linux command:

```
echo 3 | sudo tee /proc/sys/vm/drop_caches
```

The gains in compression ratio by using more expensive algorithm such as Burrows-Wheeler (bzip2) or LZMA2 (xz) are quite minimal with cryo-TEM data, likely due to the high degree of underlying randomness (or entropy). *Lbzip2* is the clear winner among command-line compression tools, as it still is faster than reading or writing uncompressed data and achieves the second-best compression ratio. *Blosc* accelerates the read/write by a factor of 3–6x over that of the uncompressed data.

Figures for benchmark results are shown in Fig. 2.2. Best compression ratio as a function of dose rate is shown in Fig. 2.2a. An important consequence of compressing Poisson-like data is that compression ratios increase substantially with sparseness. I.e. compressed size scales sub-linearly with decreasing dose fractions. For example, a cryo-tomography projection of 10 frames of 4 k × 4 k data recorded at a dose rate of would have a compressed size of 13 MB, compared to 670 MB for its uncompressed, gain-normalized image stack. Such compression therefore enables finer-dose fractionation for advanced drift correction algorithms without imposing

Table 2.1: Comparison of read/write times for 60 × 3838 × 3710 cryo-TEM image stacks.

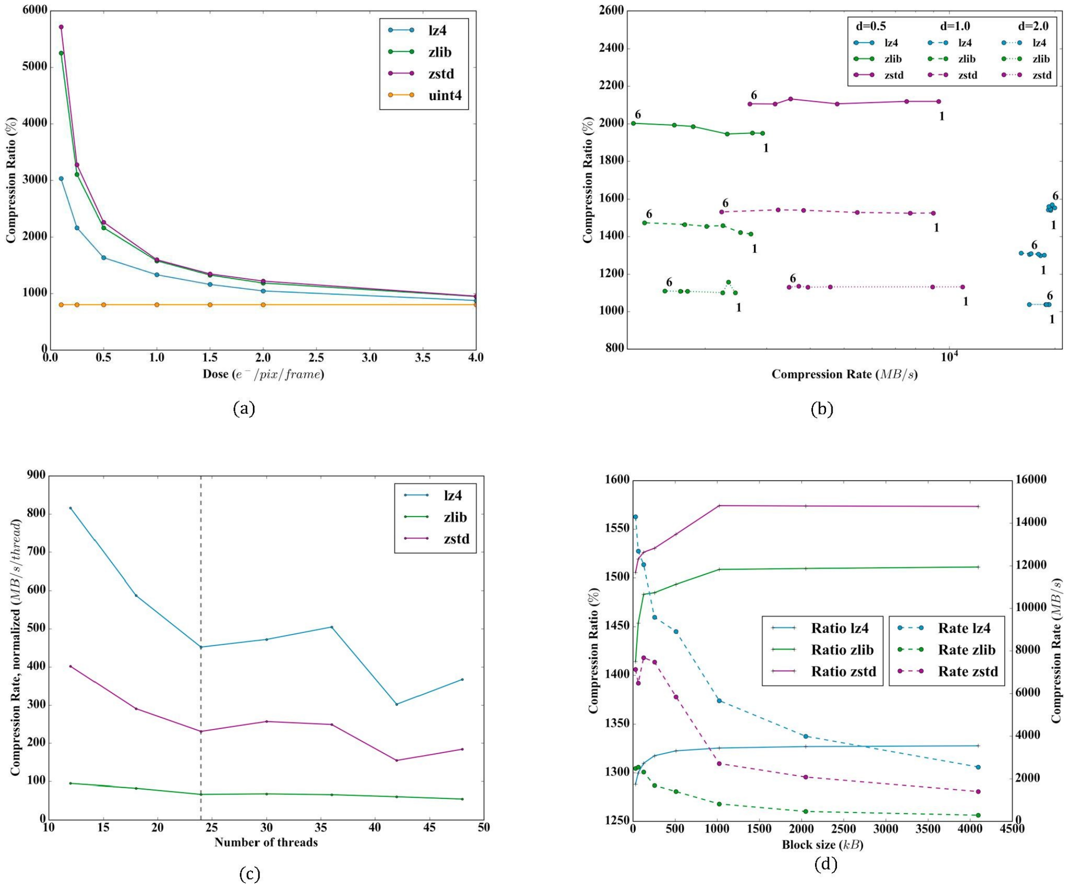
Codec/data type/compression level	Compressed Size (MB)	Compressed Ratio	Compression-Write Time (s)	Decompression-Read Time (s)
None/int8	854	1.00	3.40	3.21
uint4	427	0.50	2.14	6.05
blosc-lz4/int8/9	340	0.40	0.50	0.96
blosc-zstd/int8/1	320	0.37	0.76	1.10
blosc-zstd//int8/5	319	0.37	0.86	1.09
JPEG2000/uint8	317	0.37	106.8	N/A
pigz/int8/1	367	0.43	0.86	4.74
lbzip2/int8/9	314	0.37	3.17	2.23
pxz/int8/6	305	0.36	47.5	31.8

onerous storage requirements. (See Fig. 2.3)

With regards to compression level, shown in Fig. 2.2b, which is a reflection on the effort level of the compressor, generally *zlib* saturates at 4–5, whereas *zstd* saturates at 2–3, and *lz4* sees little disadvantage to running at its highest compression level. Good compromises for processing are compression level 1 for *zstd* and *zlib* and for archiving 3 for *zstd* and 5 for *zlib*. *Lz4* can operate with compression levels of 9 for real-time applications but it is not as suitable for archiving due to the lower compression ratios. The *bitshuffle* filter is important in this situation and contributes heavily to the quickest compression level of 1 being the best compromise between rate and ratio for *zstd*, in that it uses *a priori* knowledge about the structure of the pixel values to pre-align the data into its most compressible order.

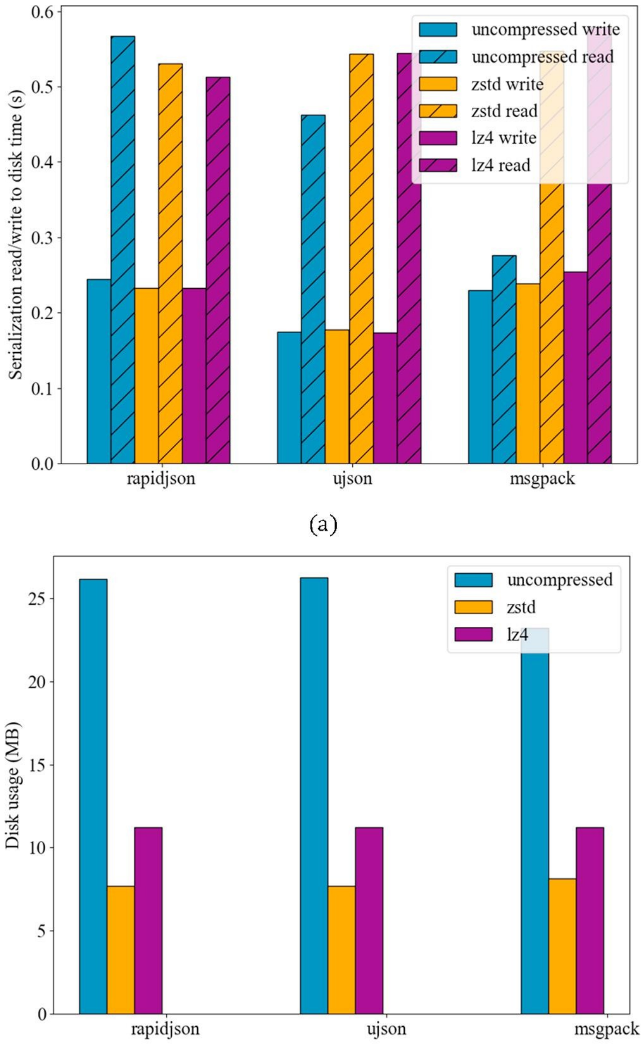
In *blosc* the scaling with threads is roughly $1/0.7N_{threads}$ for $N_{threads} \geq 2$, up to the number of physical cores. When hyper-threading is enabled an oversubscription of approximately $N_{threads} \approx 1.5N_{cores}$ gives the highest absolute compression rate.

Cache sizes are important in that they impose thresholds on data sizes, shown in Fig. 2.1d. *Blosc* chops the data into blocks, and *MRCZ* cuts a volume into single z-axis slices called chunks. For example a 4 k × 4 k image chunk may be cut into 64 separate 256 kB blocks. If the block size fits into the L2 cache (≤256 kB) then compression rate advantage is expected, and this is evident. However, testing on simulated Poisson data shows that larger blocks (which result in larger dictionaries in the compression algorithm) achieve a higher compression ratio. Similarly for chunking, if the chunk size is less than the L3 cache (≤30 MB) then only one memory call is sufficient for the entire chunk.



(a) (b) (c) (d)

Figure 2.2: **Performance for various compression codecs found in blosc.** (a) The dependence of compression ratio varies strongly with the dose. Here *zstd* has the best compression ratio. (b) The dependence of the compression level on the compression ratio is mild, such that for *zstd* and *zlib* typically 1 is used. (c) Scaling on the compression rate with the number of parallel computing threads employed. The machine used has 2×12 physical cores, indicated with the dashed line. The area to the right of the dashed line indicates the region in which Intel Hyperthreading® is active. (d) Dependence of the compression ratio and rate on the blocksize used, which is the most critical parameter examined. Typically a blocksize scaled to fit into L2 cache (256 kB) is optimal for speed, but the compression ratio benefits from a larger blocksize (≥512 kB).



(a)

(b)

Figure 2.3: **Performance evaluation of different serialization methods for meta-data paired with compression.** (a) Read and write times for the profiled serialization methods on a sample of 25 MB of JSON-like text metadata data, when used with and without blosc compression. (b) Size on disk of the metadata.

The optimal block size for compression ratio is expected to be when each block holds one significant bit each. So for 4 k × 4 k × 8-bit images the ideal block size would be $16/8 = 2$ MB, whereas the saturation in compression ratio actually appears at 1 MB.

2.2.4. Enabling electron counting in remote computers with image compression

Direct electron detectors can generate data at very high rates that are difficult to continuously write to storage. Some versions of the K2, such as the In-Situ (IS) model, permit access to the full data rate of at 3838 × 3710 × 400 Hz × uint8, or 5.3 GB/s. Such systems require a large quantity of random-access memory to record the data in bursts. The K2 Summit commonly used in cryo-TEM counts at 400 Hz through the use of Field-Gate Programmable Arrays (FPGA) but the maximum rate available to the user is 40 Hz, in large part due to the data rate. This is unfortunate as it does not permit experimentation with alternative subpixel detection algorithms that might better localize the impact of the primary electron in the detector layer.

However the multi-threaded meta-compression discussed above may be able to alleviate the data flow problem. A master node could, using *zstd* (or failing that, *lz4*) as a compression codec, compress the raw data from a K2 IS on-the-fly and dispatch it to worker computers for counting, thus enabling counting without a FPGA or similar hardware counting solution. The compression ratio achievable would depend heavily on thresholding of low intensity values but could be in the range of 50:1.

2.3. Extended metadata in MRCZ

File formats require complex, nested metadata to be encoded into a stream of bytes. The conversion of metadata to bytes is called serialization. In order to achieve a high level of portability in the future for the metadata, we advise use of a well-established serialization standard. Here two serialization standards are compared, JSON (JavaScript Object Notation), which is the most ubiquitous serialization method in the world, and Message Pack (msgpack.org and pypi.python.org/pypi/msgpack-python), a binary serialization tool with a similar language structure to JSON. Libraries are available for both for many different programming languages, with the exception of Matlab and Fortran for Message Pack. Here two high-speed JSON encoders available for C and Python are profiled, RapidJSON (rapidjson.org and pypi.python.org/pypi/pyrapidjson) and UltraJSON (github.com/esnme/ujson4c and pypi.python.org/pypi/ujson).

The three serialization methods were tested on a sample of 25 MB of complicated

JSON data. All three methods produce more-or-less similar results in terms of read/write times to disk, as shown in Fig. 2.2a. Compression does not speed nor slow read/write times, except for Message Pack read rates. *Lz4* reduces the data size on disk to roughly one-half, and *zstd* to roughly one-third, of the uncompressed size. Message Pack was tested with Unicode-encoding enabled to make it equivalent to the JSON encoders, which slows its read/write time by ~20%.

2.4. Conclusion

The introduction of fast, large-pixel count direct electron detectors has moved the field of electron microscopy inside the domain of "Big Data" in terms of data processing and storage requirements. Here an extension of the MRC file format is demonstrated that permits on-the-fly data compression to lessen both transmission and storage requirements, using the meta-compression library *blosc*. *Blosc* is well suited as a library for Big Data purposes because it does not explicitly endorse any particular algorithm and intends to support new compression methods as they are developed. Development of a *blosc2* standard, with additional features, is currently underway. Major new expected features include principally: first, a super-chunk header, that records the position of individual image chunks in the file, and second, buffered output so the file can be written to disk as it is compressed, lessening memory consumption. Also *blosc* is especially targeted towards high-speed compression codecs. With high-speed compression using the *zStandard* codec, file input-output rates are accelerated and archival storage requirements are reduced.

With the use of compression, sparse data can be compressed to very high ratios. Hence, data can be recorded in smaller dose fractions with a less-than-linear increase in data size. For very small dose fraction applications, such as cryo-electron tomography, electron crystallography, or software electron counting schemes, compression can reduce data transmission and storage requirements by 10–50x.

Acknowledgements

The authors would like to thank the authors of the *blosc* library, Francesc Alted and Valentin Haenel, for their work on the library and feedback during our testing of MRCZ. This work was supported by the Swiss National Science Foundation (grant 205320_166164), and the NCCR TransCure program.

Note

An extended version of this article is available as a preprint on bioRxiv [McLeod *et al.* 2017a].

3. Retrieving High-Resolution Information from Disordered 2D Crystals by Single Particle Cryo-EM

In this publication, we demonstrate that single particle analysis algorithms can be used to process data from disordered 2D crystals. This allows achieving resolutions much higher than would be expected from the diffraction spots observed in the power spectra of the 2D crystal images. Furthermore, the approach reveals that conformational heterogeneity is a factor of disorder in 2D crystals, and can be sorted out by means of 3D classification.

Contribution: implementation of a module in FOCUS for exporting 2D crystal data to SPA software packages, modification of FREALIGN to include alignment restraints and auto-refinement, data processing, model fitting and results analysis.

The following section has been published in:

Nature Communications
10, 1722 (2019)
http://doi.org/10.1038/s41467-019-09661-5

Retrieving High-Resolution Information from Disordered 2D Crystals by Single Particle Cryo-EM

Ricardo D. Righetto[1], Nikhil Biyani[1], Julia Kowal[1,2], Mohamed Chami[1] and Henning Stahlberg[1*]

[1] Center for Cellular Imaging and NanoAnalytics, Biozentrum, University of Basel, Mattenstrasse 26, CH-4058 Basel, Switzerland

[2] Institute for Molecular Biology and Biophysics, ETH, Otto-Stern-Weg 5, CH-8093 Zürich, Switzerland

[*] Corresponding Author: henning.stahlberg@unibas.ch

Contents

Abstract

Electron crystallography can reveal the structure of membrane proteins within 2D crystals under close-to-native conditions. High-resolution structural information can only be reached if crystals are perfectly flat and highly ordered. In practice, such crystals are difficult to obtain. Available image unbending algorithms correct for disorder, but only perform well on images of non-tilted, flat crystals, while out-of-plane distortions are not addressed. Here, we present an approach that employs single-particle refinement procedures to locally unbend crystals in 3D. With this method, density maps of the MloK1 potassium channel with a resolution of 4 Å were obtained from images of 2D crystals that do not diffract beyond 10 Å. Furthermore, 3D classification allowed multiple structures to be resolved, revealing a series of MloK1 conformations within a single 2D crystal. This conformational heterogeneity explains the poor diffraction observed and is related to channel function. The approach is implemented in the FOCUS package.

3.1. Introduction

Electron crystallography of native two-dimensional (2D) crystals of bacteriorhodopsin allowed the determination of the first 3D model of a membrane protein in 1975 [Henderson & Unwin 1975]. Since then, considerable effort has been invested in growing 2D crystals of purified membrane proteins from protein-lipid-detergent mixtures, leading to several high-resolution structures [Henderson *et al.* 1990, Gonen *et al.* 2005, Nogales *et al.* 1998, Kühlbrandt *et al.* 1994]. However, in most cases, grown 2D crystals only diffracted to lower resolution [Abeyrathne *et al.* 2012]. Electron crystallography also requires collecting image data from tilted samples, which is technically difficult and limited in the reachable tilt angle, causing the so-called "missing cone" problem in Fourier space [Stahlberg *et al.* 2015]. Furthermore, images acquired from tilted samples are of lower quality when compared to images of non-tilted samples, because the increased effective specimen thickness also increases the number of inelastic electron scattering events and reduces the image contrast [Biyani *et al.* 2018]. These effects combined limit the resolution along the vertical z-direction, which may make reconstructions appear vertically smeared-out in real space.

Recently, the resolution revolution [Kühlbrandt 2014] in cryo-electron microscopy (cryo-EM), triggered by the development of direct electron detectors (DED) [McMullan *et al.* 2016] and better image processing software, allowed determining the atomic structures of isolated membrane protein particles in detergent or amphipols [Liao *et al.* 2013], or lipid nanodiscs [Gao *et al.* 2016]. In particular, DEDs deliver images at much improved signal-to-noise ratios (SNR) and allow series of dose-fractionated images (movies) to be recorded from the same region, which can be computationally corrected for image drift and merged. Single Particle Analysis (SPA) is now a widespread method capable of determining high-resolution protein structures routinely.

Nevertheless, the capability to analyze the structure of membrane proteins in 2D crystals is important, when (i) such crystals occur naturally in the cell membrane, (ii) the lipid bilayer influences membrane protein function, or (iii) the protein of interest is too small for conventional SPA and the addition of tags to increase the particle size would disturb its function. 2D crystals can also help elucidating conformational changes triggered by ligands [Abeyrathne *et al.* 2012, Stahlberg *et al.* 2015]. In addition, recent advances toward the rational design of scaffolds may provide a more systematic way to present arbitrary proteins as 2D crystals for structural studies [Gonen *et al.* 2015, Suzuki *et al.* 2016]. The capability to reach highest-resolution structural data from badly ordered 2D crystals is important.

In electron crystallography, distortions of the 2D crystal lattice in the image plane can be computationally corrected via an interpolation scheme [van Heel & Hollenberg 1980], correlation averaging [Saxton & Baumeister 1982], or the so-called lattice "unbending" algorithm [Henderson *et al.* 1986]. However, this unbending is performed in the 2D projection images only. Three-dimensional (3D) out-of-plane distortions in the crystals, i.e., curvature or "bumps" in the membrane plane, could not be corrected. SPA, on the other hand, aligns projections of randomly oriented isolated particles in 3D space to reconstruct the density map of the underlying protein structure [Crowther *et al.* 1970, Frank 2006]. Thus, if the unit cells or patches of the 2D crystals are treated as "single particles", SPA offers the framework required to correct for out-of-plane crystal distortions. This rationale is similar to that of processing segments extracted from helical filaments [Egelman 2007, He & Scheres 2016, Sachse *et al.* 2007]. Previous attempts of "3D unbending" 2D crystal datasets did not reach higher resolutions than the conventional 2D crystallographic approach, despite exploiting natural constraints on the orientation of particles extracted from 2D crystals [Scherer *et al.* 2014, Kuang *et al.* 2015]. The lack of DED data at the time prohibited correction for specimen drift, and the algorithms employed were suboptimal compared to the modern, probability-based methodology now used in SPA to account for noisy data while avoiding reference bias [Stewart & Grigorieff 2004, Scheres 2012b, Scheres 2012a, Lyumkis *et al.* 2013].

Here, we present a high-resolution application of SPA to electron crystallography data, using cryo-EM movies of the prokaryotic, cyclic-nucleotide modulated potassium channel MloK1. Previous studies by crystallographic processing on this dataset led to the structure of MloK1 at 4.5 Å [Kowal *et al.* 2018]. Each of the four MloK1 monomers forming the pore has a transmembrane domain (TMD), a voltage sensor domain (VSD), and a soluble cyclic-nucleotide binding domain (CNBD), which lies in the intracellular side. The total molecular weight of the tetramer is 160 kDa. Although the 2D crystal images processed did not diffract beyond 10 Å, we improved the resolution of the MloK1 3D map to 4.0 Å. Furthermore, we identified different conformations of MloK1 tetramers within the disordered 2D crystals by means of single- particle classification.

3.2. Results

3.2.1. Software implementation

As an extension to recently developed movie-mode electron crystallography algorithms [Biyani *et al.* 2018], we implemented a 2D crystal single- particle module into the FOCUS software package [Biyani *et al.* 2017] (Fig. 3.1). Electron dose-fractionated movies of 2D crystals are first corrected for specimen drift and averaged within FOCUS, using external tools[Grant & Grigorieff 2015, Zheng *et al.* 2017]. Subsequently, microscope defocus, sample tilt geometry, crystal lattice vectors, and unit cell positions are determined for all movies [Biyani *et al.* 2018, Gipson *et al.* 2007]. This can be done in an automated and parallelized manner in FOCUS. Next, the graphical user interface (GUI) features a new tab called “Particles” in which the user can perform particle picking, i.e., window patches from the 2D crystal images (Suppl. Fig. A.1), which are boxed, assembled into a particle stack, and submitted towards implemented SPA workflows, using RELION [Scheres 2012a] or FREALIGN [Lyumkis *et al.* 2013] within FOCUS. If more than one lattice is present in one image, only the strongest lattice is considered for picking. The center of each windowed patch, here termed “particle”, corresponds to the center of a crystal unit cell located by the classical unbending algorithm [Henderson *et al.* 1986, Crowther *et al.* 1996], optionally with an additional phase shift applied to translate the center of a protein to the center of the window (Suppl. Fig. A.2). Only particle positions are considered, for which the cross-correlation (CC) peaks found by the unbending algorithm are stronger than a user-definable threshold. Overlap between boxes containing neighboring particles ensures smooth transitions of the local alignment parameters across the distorted 2D crystal lattice.

Alongside the particles, a meta-data file is generated containing information for every particle, such as the micrograph it came from, its Euler angles converted from the 2D crystal tilt geometry [Biyani *et al.* 2018], the picking x,y coordinates, and the calculated defocus and astigmatism values at the center of the box following the CTFTILT conventions [Mindell & Grigorieff 2003]. Optionally, an additional particle stack can be created simultaneously during picking by correcting each particle box for the local contrast transfer function (CTF) of the electron microscope. Available CTF correction methods are phase-flipping, CTF multiplication, or an ad

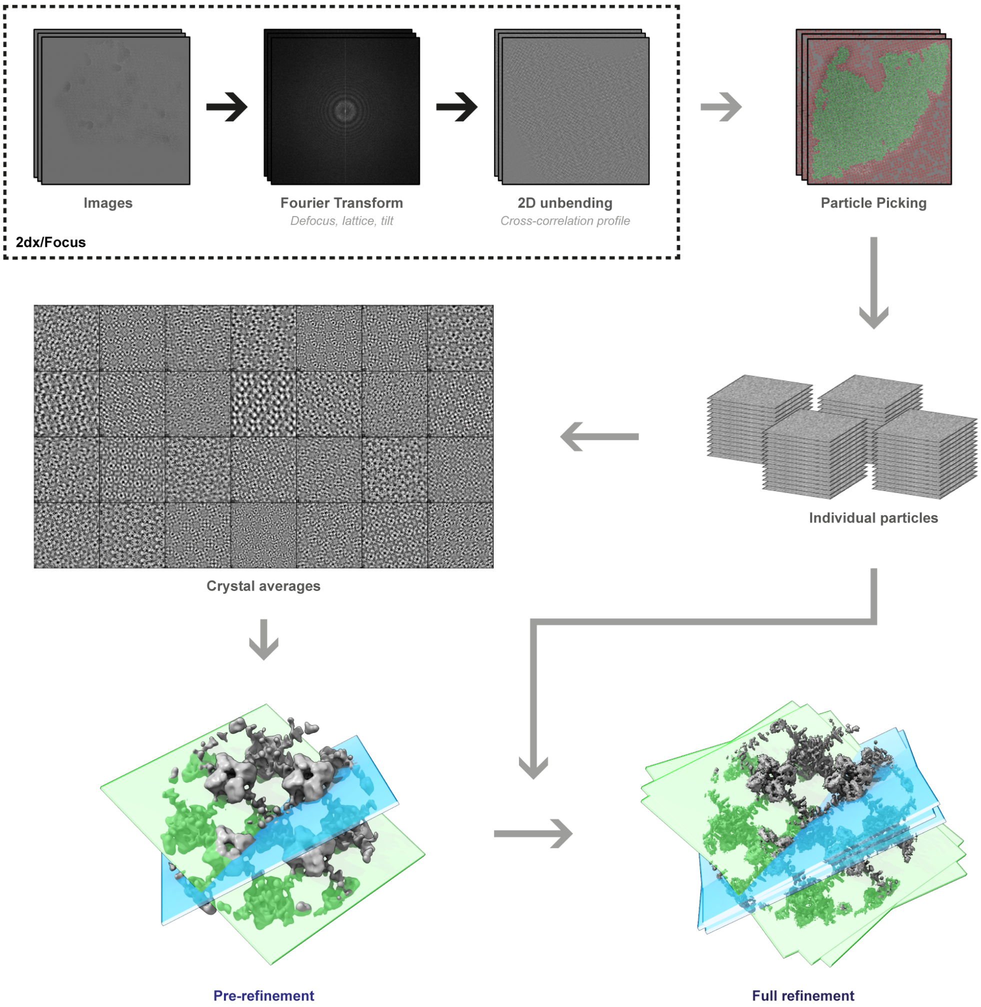


Figure 3.1: **Workflow employed to process 2D crystal data by single-particle analysis programs.** Steps depicted within the dashed box were previously available in the 2D crystallography mode of FOCUS. A new GUI (Suppl. Fig. A.1) was implemented to pick particles from the 2D crystals based on the unit cell positions obtained from classical unbending. Stacks of individual particles are then exported together with the associated metadata for processing with established SPA software. The particles extracted from the 2D crystal in each image are averaged in FOCUS for quality assessment, and rapid pre-refinement of the map is done with FREALIGN using these crystal averages as input. For simplicity, slices representing two crystal averages in different orientations are shown (light blue and light green). Then, the individual particles are initialized with the pre-refined alignment parameters and provided to FREALIGN to perform a full refinement to high resolution. These individual particles might end up having slightly different orientations than the crystal average, due to distortions in the 2D crystal lattice

hoc Wiener filter. When generated, the CTF-corrected stack can be used to perform correlation averaging [Saxton & Baumeister 1982] of the 2D crystals, providing immediate feedback on the dataset quality in similar fashion to the 2D class averages generated in SPA (Suppl. Fig. A.3). Since 2D crystal images present the proteins densely packed in the images, only a comparatively small number (100–1000) of 2D crystal images is usually needed to produce large particle numbers (100,000 to 1,000,000). Windowed particles for each 2D crystal image can be averaged, and these fewer image averages can already be used to determine initial 3D models and pre-refine the alignment parameters that will be propagated to the larger particle dataset, thereby speeding up the structure determination process (Fig. 3.1).

3.2.2. Processing MloK1 2D crystals

In order to test our approach, we utilized a 2D crystal dataset of the potassium channel MloK1 in the presence of cAMP that yielded a 4.5 Å resolution map when processed by classical 2D crystallography [Kowal *et al.* 2018]. The Fourier transforms of the images showed Bragg reflections to 10 Å resolution at best, in most cases worse than 14 Å. Drift-correction of the 346 movies from the previous study was performed with MotionCor2 [Zheng *et al.* 2017] in FOCUS [Biyani *et al.* 2017] and the aligned averages were used for further processing. The defocus at the center of the image, tilt geometry and lattice determined previously were retained. Out of the 346 crystal images, 76 were discarded, as they were associated with a second lattice in the same image. The image data from secondary lattices can be treated independently in classical 2D crystallography, but in SPA processing these would lead to exaggerated resolution estimates due to overlap between particles picked from different lattices in the same image, i.e., it would effectively introduce duplicated particles in the dataset. This can occur for example if two 2D crystals happen to be on top of each other on the support film, or if the periodical structure under study has a 3D component, such as a microtubule or vesicle, which becomes flattened during grid preparation [Stahlberg *et al.* 2015]. Using the new GUI, a total of 231,688 unique particles with a box size of 320 × 320 pixels were windowed from the remaining 270 unique images, which had a pixel size of 1.3 Å on the sample level (Suppl. Fig. A.2). In the non-tilted views, each "particle" was roughly comprised of nine MloK1 tetramers. Because the unit cell of the processed MloK1 2D crystals had $p42_12$ space-group symmetry that contains a screw axis [Kowal *et al.* 2018], we applied a phase origin shift of half a unit cell (180°) in the direction of the first lattice vector to the picking coordinates in order to have one tetramer at the center of each particle box. Phase-flipped copies of the particle projections were calculated and stored at the same time as the non-CTF corrected particle projections for the generation of crystal averages (Suppl. Fig. A.3).

3.2.3. Consensus refinement

Using a modified version of FREALIGN v9.11 [Lyumkis *et al.* 2013] (Suppl. Note 1) and the initial tilt geometry obtained in FOCUS, we calculated a 3D reconstruction at 6.5 Å resolution with C4 symmetry imposed. After pre-refinement using the 270 crystal averages, the global resolution improved to 4.8 Å. Finally, the updated alignment parameters were propagated from the crystal averages to the 231,688 individual raw particles, and refined in a single class using a custom auto-refinement script written for FREALIGN (Suppl. Note 2). The refined map contains nine full MloK1 tetramers (Fig. 3.2a). For further analysis, the central tetramer was cropped out of the full reconstruction and postprocessed (Fig. 3.2b) leading to the "consensus" refinement map at a global resolution of 4.0 Å based on the Fourier shell correlation (FSC) curve [Harauz & van Heel 1986, Rosenthal & Henderson 2003, van Heel & Schatz 2005] (Fig. 3.2c and Suppl. Fig. A.4). To avoid inflation of the FSC curve due to the large overlap between neighboring particle boxes, particles extracted from the same 2D crystal were always assigned to the same half-set throughout the refinement [He & Scheres 2016]. At this resolution level we could identify densities for many side chains in the transmembrane domains (TMD), especially the larger ones, such as phenylalanine and tryptophan and those at the S4–S5 linker (Suppl. Fig. A.5). This 4.0 Å resolution estimate based on the FSC corresponds to an isotropically averaged measurement. The map's resolution is better in the xy membrane plane than it is along the z axis. By calculating the Fourier ring correlation between the central xy slices ($z = 0$) of the unmasked half-maps we estimate resolution to be slightly better than 4.0 Å in this plane, whereas in the orthogonal xz plane ($y = 0$), which is identical to the yz plane due to the imposed symmetry, the estimated resolution is about 4.7 Å considering the FSC 0.143 threshold [Rosenthal & Henderson 2003] (Suppl. Fig. A.4f). These two orthogonal resolution estimates are consistent with the observed features of the consensus map. Rigid-body fitting of our previously published model for MloK1 [Kowal *et al.* 2018] into the higher-resolution consensus map revealed a sequence register shift of 1 residue at the S4–S5 linker, while the X-ray crystallographic model for the TMD [Clayton *et al.* 2008] (PDB: 3BEH) was in good agreement with the experimental densities of our map. This new map allowed refining a full-length atomic model of MloK1, improving both its geometry and fit-to-density indicators compared to the previous model (PDB: 6EO1) obtained by classical 2D electron crystallographic processing [Kowal *et al.* 2018] (Suppl. Table A.1).

3.2.4. 3D Classification

The local resolution estimates and model B-factors for the consensus map (Suppl. Fig. A.4a, b) suggested high conformational variability of the CNBDs of the molecules,

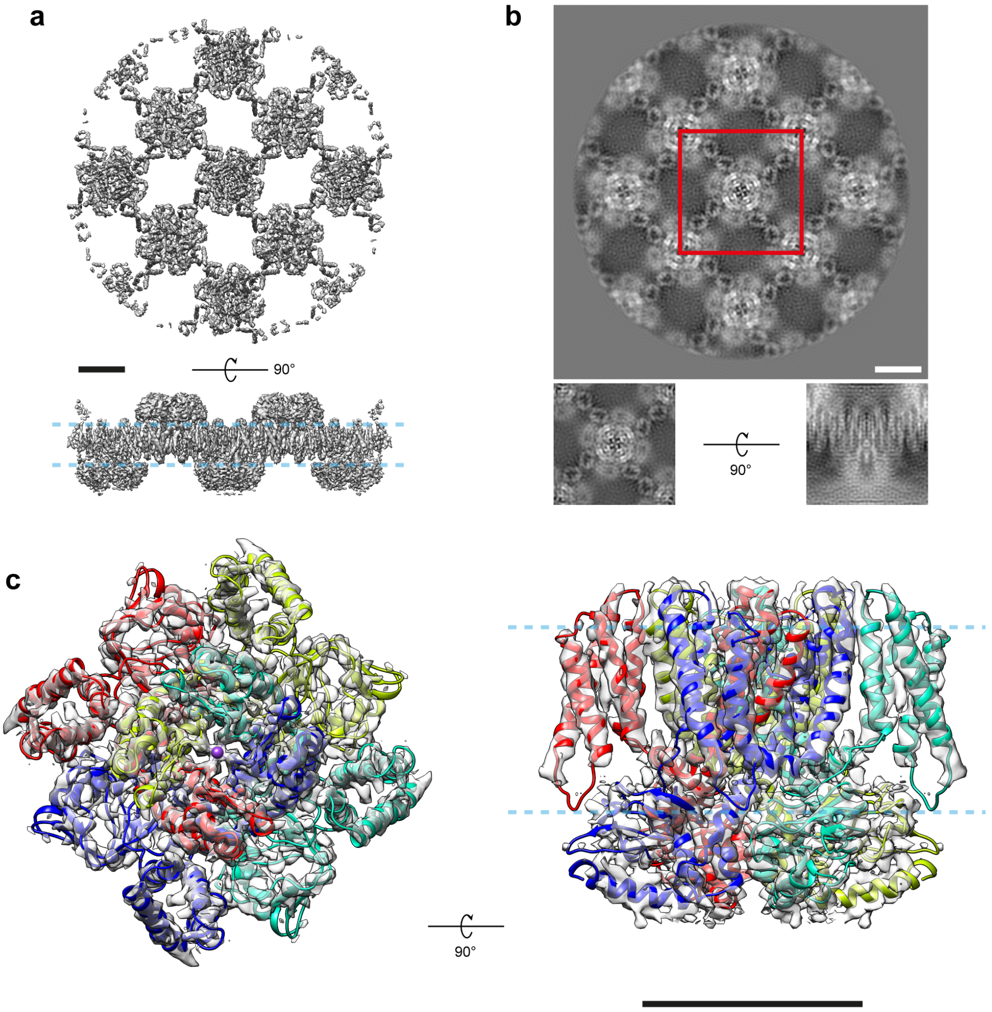


Figure 3.2: **The cryo-EM map of full-length MloK1 at 4 Å.** The map was obtained from the consensus single-particle refinement of 2D crystal data in FREALIGN. a) The full refined map containing approximately nine MloK1 tetramers; total molecular weight over 1 MDa. b) Projection of the full map along the z axis. The sub-volume used for postprocessing is indicated (red box). Insets: projections of the central sub-volume orthogonal to the xy and xz planes. c) The postprocessed map of the central tetramer with the refined atomic model fit into the density, viewed from the extracellular side and parallel to the membrane plane. The pale blue dashed lines indicate the approximate position of the lipid bilayer. Scale bars: 50 Å

inviting for 3D classification of the particles. Because particle images contained more than one MloK1 tetramer, signal subtraction [Bai *et al.* 2015] for the densities corresponding to the neighboring tetramers was applied (Suppl. Fig. A.6), in order to be able to perform a 3D classification on the central tetramer alone. This was especially important for images of tilted samples, where the projections of neighboring tetramers partially overlap with the projection of the central tetramer, which is our classification target. We then applied maximum-likelihood 3D classification in FREALIGN [Lyumkis *et al.* 2013] to determine conformational variations within the 2D crystals, while imposing C4 symmetry and keeping the alignment parameters for the particles constant. Surprisingly, we found that the dataset was quite heterogeneous. 3D classification produced eight 3D classes at resolutions from 4.4 to 5.6 Å (Fig. 3.3 and Suppl. Fig. A.7) that differed conformationally to various degrees (see below for more details). The 3D classification did not correlate with particle positions in the 2D crystals, indicating that no bias from defocus or tilt angle affected the classification (Suppl. Fig. A.8). When employing a localized reconstruction method [Ilca *et al.* 2015, Grigorieff 2016] to look for non-symmetric conformations, no significant deviations from four-fold symmetry were detected.

To obtain more insight, the consensus atomic model was flexibly fit into each map of the 3D classes using Normal Mode Analysis [Lopéz-Blanco & Chacón 2013] and then refined in real space [Afonine *et al.* 2018b]. The resulting models were globally aligned against the consensus map (Suppl. Fig. A.9), showing root mean square deviations (RMSD) ranging from 0.379 Å (class 7) to 0.720 Å (class 1) toward the consensus map. Alignment of the models against the consensus model by the selectivity filter region only (residues 174–180) produced an ensemble depicting a continuum of conformations of the CNBDs and the S2–S3 loops (Fig. 3.4), within which model #1 is the most different from all other models (Suppl. Fig. A.9b) and in an “extended” conformation, with the CNBD farthest away from the membrane plane and the TMD. Conversely, model #4 is the most different from model #1 (RMSD of 1.021 Å after global superposition) and is in a “compact” conformation having the CNBD closest to the membrane plane and the TMD (Fig. 3.4, Suppl. Fig. A.9). The other five models can be understood as intermediate snapshots along the trajectory between models #1 and #4.

Ranking the models in our ensemble in descending order of their pairwise RMSD values (Suppl. Table A.2 and Suppl. Fig. A.9a) allowed us to inspect the largest conformational changes of MloK1. Along the trajectory from the most compact state (model #4), to the most extended state (model #1) depicted in Suppl. Movie 1, the CNBD moves toward the inner pore axis of the channel and simultaneously away from membrane plane. The C-terminal helix of the CNBDs tilts by about 3 degrees further from the membrane, while shifting away from the pore axis. The S2–S3 loop of one monomer closely follows the CNBD β-rolls of the adjacent monomer, which move toward the symmetry axis. At the same time, helix S4 of the VSD extends

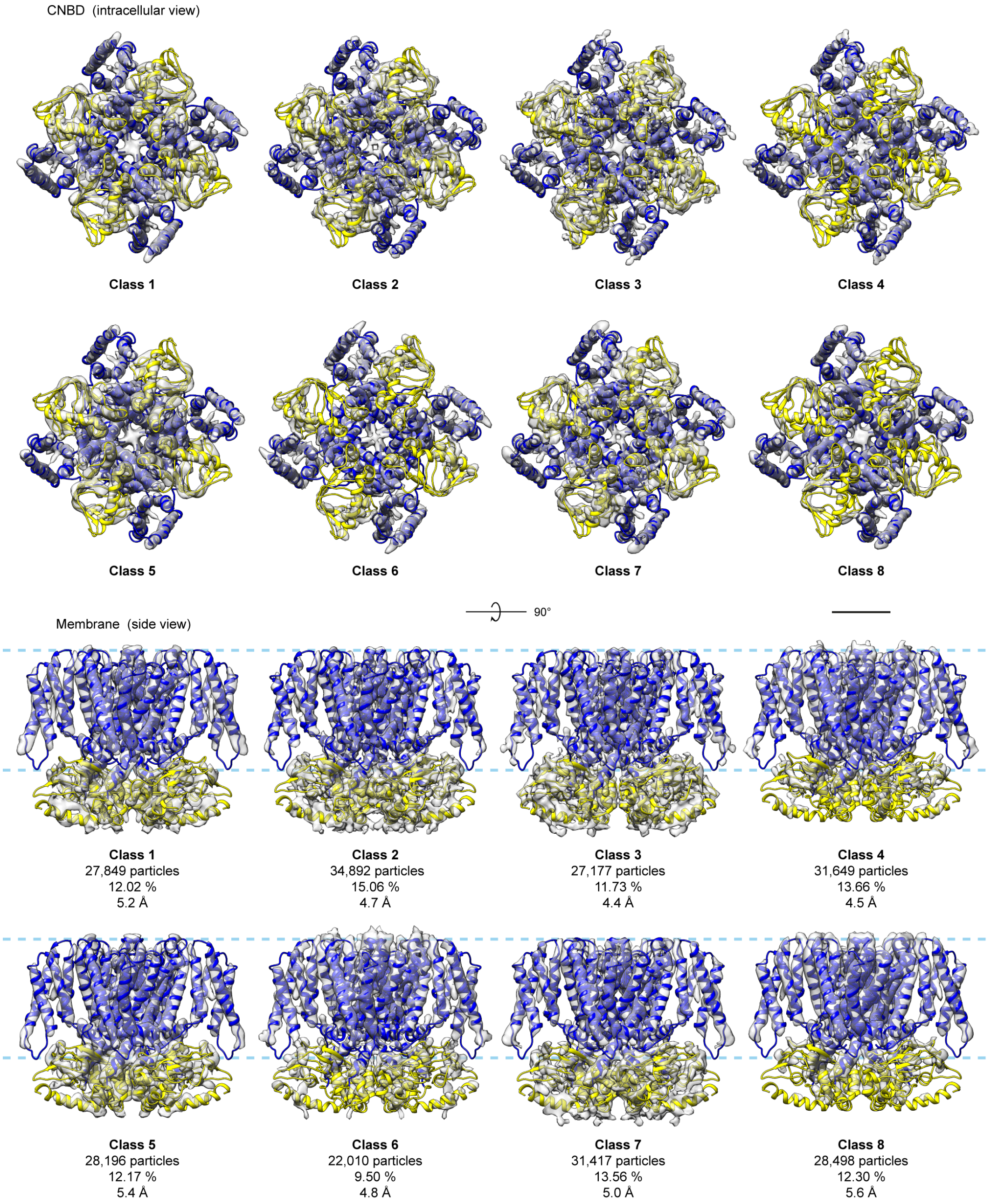


Figure 3.3: **Conformational continuum of MloK1 from 2D crystals.** All the maps were obtained from one 2D crystal dataset using signal subtraction and 3D classification after consensus refinement. The consensus model was flexibly fitted and refined into each 3D class. The CNBD is colored yellow and the TMD is colored blue, while the maps are shown as transparent gray surfaces. The map threshold for each class was calculated such that all iso-surfaces enclose the same volume. The pale blue dashed line indicates the approximate position of the lipid bilayer. Conformational variations among the classes primarily change the height of the CNBDs in relation to the membrane slab and their tilt in relation to the TMD and the VSD. C4 symmetry was applied. Scale bar: 25 Å

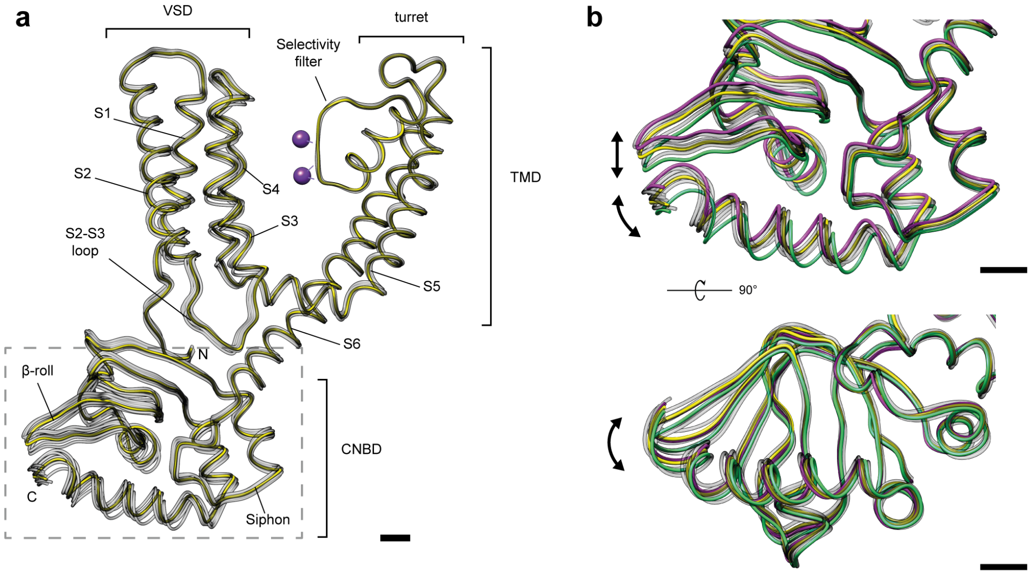


Figure 3.4: **Conformational ensemble of MloK1 within the 2D crystals.** The consensus model (yellow) and the eight atomic models derived from the 3D classes (gray) are shown superposed. For this visualization, the eight models in the ensemble were aligned to the selectivity filter of the consensus model (residues 174–180). a One full chain of the MloK1 tetramer is shown with the different domains indicated; the dashed light-gray box indicates the area highlighted for panels b and c; b close-up view of the CNBD part where most conformational changes occur. Models #1 and #4 are the most different from each other (RMSD 1.021 Å) and are depicted in green and magenta, respectively. The arrows indicate the principal directions of conformational variability. Scale bars: 5 Å

toward the intracellular membrane boundary. The next largest conformational difference appears between the "extended" model #6 and the "compact" model #1 (RMSD: 0.927 Å, Suppl. Table A.2). This trajectory is very similar to that between models #4 and #1, but in addition the helices S1–S4 of the VSD tilt by about 1 degree in such a way that the S1–S2 loop comes closer to the periplasmic pore turrets (Suppl. Movie 2). Among the largest conformational changes observed is a 6-degree helical rotation of the CNBD around the symmetry axis, perpendicular to the membrane plane, between intermediate models #3 and #5 (RMSD: 0.845 Å, Suppl. Table A.2), as shown in Suppl. Movie 3. Compared to model #1, the C-terminal helix in "extended"-like model #3 is rotated by about 4 degrees clockwise when seen from the CNBD side, while in the "compact" model #5 it is rotated by about 3 degrees counter-clockwise. A rotation of the selectivity filter by about the same magnitude also occurs in this trajectory. Finally, the most extended conformers of our ensemble (models #1 and #3) also have the S6 helix crossing bundle and the selectivity filter slightly more constricted (<1 Å) than the most compact ensemble members (models #4, #5, and #6).

3.3. Discussion

Electron crystallography applied to well-ordered 2D crystals has, in the past, delivered structures of membrane proteins at far better resolutions than achievable by SPA. However, the resolution revolution in cryo-EM reversed this, so that today the precision of the physical arrangement of proteins in the 2D crystals is lower than the resolution achievable by SPA. With the addition of detergents or reconstitution in lipid nanodiscs, ion channels similar to MloK1 in mass have been consistently resolved in the resolution range of ∼3.5 Å by SPA recently [Gao *et al.* 2016, Dang *et al.* 2017, Lee & MacKinnon 2017, Li *et al.* 2017] . Nevertheless, cases exist where the structure of a membrane protein in the lipid environment provided by a 2D crystal is important, such as when the crystals occur naturally in the cell membrane, or when a 2D periodic array is artificially designed [Gonen *et al.* 2015, Suzuki *et al.* 2016].

Here, we present a hybrid approach for the analysis of 2D crystal images, combining 2D crystallographic and single particle image processing methods. Even though the missing cone effect is still present, and images of tilted samples are generally of lower quality than of non-tilted samples, the application of SPA algorithms to the MloK1 dataset ameliorated the overall resolution by accounting for local variations in the tilt geometries along the distorted 2D crystals, thus offering a higher diversity of views for the 3D reconstruction. This documents that the lack of crystal planarity can become an advantage in SPA. Also, masking the reference map along the iterative refinement in SPA is akin to projective constrained optimization [Biyani *et al.* 2018, Gipson *et al.* 2011], both methods contributing to reduce the impact of the missing cone.

Our highest resolution structure, derived from the consensus map (Fig. 3.2), is in general agreement with our previously published model [Kowal *et al.* 2018]. The overall resolution of the map improved from 4.5 to 4.0 Å using our single- particle approach. The main bottlenecks to resolution by this method are presumably the lack of side views due to the tilting limitations, and the inherent lower quality of the high tilt images. Also, in this particular MloK1 dataset the pixel size at the sample level was relatively large (1.3 Å), which means the data was collected at a range of the K2 Summit's camera detective quantum efficiency that is suboptimal from the perspective of the SNR [McMullan *et al.* 2016]. However, we here show that the previously determined structure was an average of several coexisting conformations.

Consequently, SPA processing enabled the particles within the 2D crystals to be classified, revealing significant conformational variations. While 2D classification of unit cells has been explored in the past [Frank *et al.* 1988, Sherman *et al.* 1998], our approach enables the retrieval of distinct 3D classes from within the same 2D crystal dataset. The eight obtained 3D classes (Fig. 3.3) have a slightly lower resolution than the consensus map, likely due to the limited number of particles in each class. The conformational differences observed between the individual 3D classes explain the low resolution to which the imaged 2D crystals showed diffraction. The analysis suggests that the CNBDs were only partially occupied with ligands in the cAMP-saturated crystals. Following Rangl et al. [Rangl *et al.* 2016], a possible interpretation is that the extended "up-state" orientation of the CNBDs corresponds to the cAMP-free orientation, while the compact "down-state" orientation with the CNBDs approaching the membrane, is the cAMP-bound conformation [Kowal *et al.* 2018]. Such insights have been obtained previously from lower resolution structures [Kowal *et al.* 2014] and high-speed AFM experiments [Rangl *et al.* 2016]. Our data and the processing workflow presented here, now provides higher-resolution data that invite for a more detailed analysis of this hypothesis in 3D models of MloK1.

Variations in CNBD positions are correlated with changes in the transmembrane regions including the VSD of MloK1, yielding insights into the mechanistic links between CNBDs and channel modulation. The observed interaction between the S2–S3 loop of the VSD and the CNBD β-rolls of the adjacent monomer corroborates the coupling between VSD and CNBD previously hypothesized [Kowal *et al.* 2018]. In this proposed mechanism, the channel is open and in a compact conformation when cAMP is bound, and, upon ligand unbinding, the CNBD extends away from the membrane and closes the channel, accompanied by a tilting of the VSD towards the extracellular side (Suppl. Movie 2). This suggests that a change in the binding state of the CNBD is transduced to the selectivity filter via the VSD, possibly also with help of the N-terminal loop and the S4–S5 linker (Suppl. Movie 1 and Suppl. Movie 2), rather than directly via the C-linker of S6. Furthermore, we also observe a helical rotation of the CNBD around the pore axis during the CNBD extension, which is orchestrated with a rotation of the selectivity filter in the opposite direction (Suppl. Movie 3). Combined with a constriction of the pore, we suggest this mechanism to then prevent the passage of K+ ions at the selectivity filter [Kowal *et al.* 2018, Tombola *et al.* 2006, Liu *et al.* 2015]. While the S6 helix crossing bundle also constricts slightly during this conformational change, it remains too wide to block ion conduction.

In summary, we have shown that the SPA approach greatly reduces the need for perfectly planar, well-ordered 2D crystals. The procedure allows retrieving high-resolution information from disordered and non-flat 2D crystals, as illustrated by the 4 Å consensus model presented. Importantly, SPA also allows detecting conformational variations of proteins in the 2D crystals. Classification of the low-resolution

2D crystal images of the potassium channel MloK1 by SPA resulted in a series of 3D maps with resolutions between 4.4 and 5.6 Å, giving insight into ligand binding and channel gating. The data processing workflow is available from a GUI in the FOCUS package.

3.4. Methods

3.4.1. Protein expression, purification, and 2D crystallization

The expression and purification of MloK1, and the growth of 2D crystals in the presence of cAMP is described in Kowal et al. [Kowal *et al.* 2014]. For convenience, the protocol is summarized here: *E. coli* BL21(DE3) cells with the 6-His-tagged MloK1 construct were grown at 37 °C. Protein expression was then induced with 0.2 mg/ml anhydrotetracycline for 2 h. Cells were sonicated and membrane proteins solubilized with 1.2% *n*-decyl-β-D-maltopyranosite (Anatrace) in the buffer containing 295 mM NaCl, 5 mM KCl, 20 mM Tris-HCl pH 8.0, 10% glycerol, 1 mM phenylmethylsulphonyl, and 0.2 mM cAMP, and incubated for 2.5 h at 4 °C. The MloK1 extract was purified at a Co^{2+} affinity chromatography column, in the same buffer as before but with the addition of 0.2% *n*-decyl-β-D-maltopyranoside and 40/500 mM imidazole (wash/elution). Throughout the purification, 0.2 mM cAMP (Fluka) was present to maintain the integrity of MloK1. For 2D crystallization, detergent-solubilized MloK1 sample was subsequently mixed with E. coli polar lipid extract (Avanti Polar Lipids) at a lipid-to-protein ratio of 0.8–1.0 (w/w) and dialyzed against detergent-free buffer (20 mM KCl, 20 mM Tris-HCl pH 7.6, 1 mM BaCl2, 1 mM EDTA, 0.2 mM cAMP) for 5–10 days.

3.4.2. Sample preparation

Grids for transmission electron microscopy were prepared as described in Kowal et al. [Kowal *et al.* 2018]: 4 μl of the MloK1 2D crystal solution (0.8 mg/ml) were applied to glow-discharged Quantifoil holey carbon grids (R3.5/1, Cu 400 mesh), which had been coated with an additional <3-nm-thin amorphous carbon layer. Using an FEI Vitrobot IV with the environmental chamber set at 90% humidity and 20 °C, excess solution on the grid was blotted for 3.5 s, and the grids were then flash-frozen into liquid ethane.

3.4.3. Cryo-EM imaging

The dataset of 346 movies recorded and processed for Kowal et al. [Kowal *et al.* 2018] was employed. As reported, the data were collected on an FEI Titan Krios TEM equipped with a Gatan K2 DED. Total dose: 40 $e^{-}/Å^{2}$ distributed over 40 movie frames. Pixel size: 1.3 Å on the sample level (counting mode). Nominal tilt range: $-55°$ to $+55°$.

3.4.4. Data processing

Movies were drift-corrected using MotionCor2 [Zheng *et al.* 2017] via FOCUS [Biyani *et al.* 2017]. Micrographs were processed using the FOCUS package until the 3D merging step, following standard 2D electron crystallography procedures as previously implemented in the 2dx package [Gipson *et al.* 2007]. This yielded the defocus, tilt geometry, lattice, and phase origin information for each micrograph. Using our newly implemented GUI, particles, i.e., patches of the 2D crystal image centered on crystal unit cells, were extracted from positions indicated by the cross-correlation profile of the classical unbending algorithm [Henderson *et al.* 1986]. Only one lattice per image was considered, resulting in the exclusion of 76 duplicated images due to the presence of a second lattice. Unit cells with a cross-correlation (CC) peak above each micrograph's average CC peak value were picked, an approach closely coinciding with the auto-masking procedure for 2D crystals. As the unit cell of the MloK1 2D crystals had $p42_12$ symmetry [Kowal *et al.* 2018, Kowal *et al.* 2014, Chiu *et al.* 2007], a shift of 180 degrees along the first lattice vector was applied to the phase origin (the crystallographic unit cell) to position a MloK1 tetramer at the center of the particle box and allow the imposition of C4 symmetry in the single-particle refinement steps. The box size was of 320 square pixels roughly comprising nine MloK1 tetramers in the non-tilted views. Individual particle images were CTF-corrected by phase flipping and averaged on a per-crystal basis using new scripts in FOCUS. The particle export script also ensures that particles picked from the same 2D crystal stay in the same half-set in order to prevent inflated resolution estimations based on the FSC because of the overlap between neighboring particle boxes [He & Scheres 2016].

3.4.5. Consensus refinement

Pre-refinement using the crystal averages was performed using a custom auto-refinement script based on FREALIGN version 9.11 [Lyumkis *et al.* 2013] (see Suppl. Note A.5 for details). Subsequent refinement after convergence was performed using the same auto-refinement procedure, but now using all particles and inheriting the alignments determined in the pre-refinement and defocus values as estimated at the center of the particle window according to the initial tilt geometry, with CTF correction performed internally by FREALIGN using Wiener filtering [Stewart & Grigorieff 2004, Sindelar & Grigorieff 2012]. Both, in the pre-refinement and the refinement steps, a spherical mask was initially applied to the reference 3D reconstruction leaving a region comprised of about nine MloK1 tetramers for processing. To prevent reference bias, the highest resolution limit used for particle alignment was 7.52 Å.

3.4.6. 3D classification

A focused spherical mask on the central MloK1 tetramer was applied to subtract the signal from the neighboring tetramers using RELION [Scheres 2012b, Bai *et al.* 2015]. Afterwards, the signal-subtracted particle stack was subjected to 90 cycles of maximum-likelihood 3D classification in FREALIGN [Lyumkis *et al.* 2013] using eight classes and a resolution limit of 7.0 Å. No alignments were performed at this stage. To decrease the processing time, for this classification the particle stack was downsampled to a pixel size of 2.6 Å by Fourier cropping. Asymmetric classification using both C1 symmetry and the localized reconstruction [Ilca *et al.* 2015, Grigorieff 2016] of a single suitably masked monomer to search for deviations from C4 symmetry were also attempted.

3.4.7. Map postprocessing

For post-processing and analysis of every map, a box of 104 cubic voxels containing only the central MloK1 tetramer was extracted from the larger half-maps. The box center was translated by 12 voxels in the Z direction to coincide with the center of the MloK1 tetramer before cropping out the smaller volume. FSC curves between half-maps were calculated using a spherical mask of 42 voxels radius and a soft cosine-edge of 6 voxels width, and corrected for the relative volumes of the particle and the mask within the box [Grigorieff 2016]. The maps were sharpened by deconvolving the MTF curve of the Gatan K2 detector at 300 kV and then using the phenix.auto_sharpen program [Terwilliger *et al.* 2018b], and low pass-filtered at the respective resolution cutoffs defined by the 0.143 threshold criterion [Rosenthal & Henderson 2003] using a soft cosine-edge filter. Resolution was also assessed using the 1/2-bit criterion [van Heel & Schatz 2005] with an estimated particle diameter of 100 Å and four-fold symmetry. A Python script called *focus.postprocess* was written based on the MRCZ module [McLeod *et al.* 2018] and included as a command-line tool in FOCUS. Local resolution maps were calculated using Blocres [Cardone *et al.* 2013].

3.4.8. Model building

A new model was assembled by taking the N-terminal from PDB 6EO1 [Kowal *et al.* 2018] (residues 1–6), the TMD from PDB 3BEH [Clayton *et al.* 2008] (residues 7–219), the CNBD from PDB 3CL1 [Altieri *et al.* 2008] (residues 220–349) and the C-terminus also from PDB 6EO1 [Kowal *et al.* 2018] (residues 350–355). These domains were individually rigid-body fitted into our consensus map using UCSF Chimera [Pettersen *et al.* 2004] and then saved together as a single chain in a new PDB file. Atoms of incomplete residues were filled in using Coot [Emsley & Cowtan 2004]. The model was then flexibly fit into the consensus map using Normal Mode Analysis with iMODFIT [Lopéz-Blanco & Chacón 2013] at a resolution limit of 4.0 Å. Riding hydrogens were added to prevent steric clashes in the subsequent refinement [Word *et al.* 1999]. Secondary structure annotation was calculated using *ksdssp* [Kabsch & Sander 1983] and manually adjusted in UCSF Chimera. This single chain was then refined into the map using *phenix.real_space_refine* [Afonine *et al.* 2018b] with electron scattering form factors, global minimization and B-factor refinement. Modelling issues, such as Ramachandran outliers, rotamer outliers, and steric clashes, were monitored using Molprobity [Davis *et al.* 2007] and manually corrected in Coot, always followed by real-space refinement rounds in PHENIX [Adams *et al.* 2002]. Upon convergence, three more copies of the chain were generated and rigid-body fitted into the map in UCSF Chimera to account for the tetrameric channel. This was followed again by iterations of real-space refinement in PHENIX and manual tweaking in Coot whenever necessary, which were always followed by refinement rounds in PHENIX. Finally, potassium ions and associated restraints were added to the model at putative positions to optimize the geometry of the selectivity filter and real-space refined again in PHENIX. Model quality metrics were assessed throughout refinement using Molprobity, EMRinger [Barad *et al.* 2015], and per-residue plots in Coot.

Based on the consensus model and the eight maps obtained after convergence of 3D classification in FREALIGN, we generated an ensemble of models representing the conformational variability of MloK1. The consensus model was flexibly fit into each map using Normal Mode Analysis by iMODFIT [Lopéz-Blanco & Chacón 2013] followed by five cycles of real-space refinement in PHENIX with electron scattering form factors, global minimization and B-factor refinement, always using the global resolution determined for each 3D class, which ranged from 4.4 Å (class 3) to 5.6 Å (class 8). For comparison, every model in the ensemble was superposed against each other and against the consensus model using *phenix.superpose_pdbs*, and the RMSD between the C-α atoms was computed as a similarity measure. Based on the pairwise RMSD matrix resulting from the eight models, hierarchical agglomerative clustering [Kelley *et al.* 1996] was performed using the single linkage criterion with the *scikit-learn* Python module [Pedregosa *et al.* 2011]. For visual analysis, all members of the ensemble were superposed onto the consensus model based on the selectivity filter only (residues 174:180). Distances and angles were calculated using the Axes/Planes/Centroids tool [Meng *et al.* 2006] in UCSF Chimera.

3.4.9. Data analysis

Results were analyzed with the aid of Python scripts based on the MRCZ [McLeod *et al.* 2018], NumPy (`http://www.numpy.org`), scikit-learn [Pedregosa *et al.* 2011], and SciPy (`http://www.scipy.org`) modules.

3.4.10. Figures and animations

Plots were generated using the matplotlib Python module (`http://www.matplotlib.org`). Figures and movies were made with the aid of UCSF Chimera [Pettersen *et al.* 2004].

Acknowledgements

We thank N. Grigorieff for software support, D. Herbst and T. Maier for computational resources and discussions, C. Nimigean and R. Adaixo for discussions, and S. Müller for support in manuscript preparation. Calculations were performed at sciCORE of the University of Basel. R.D.R. acknowledges funding from the Fellowships for Excellence program sponsored by the Werner-Siemens Foundation and the University of Basel. This work was supported by the Swiss National Science Foundation, grant 205320_166164 and the NCCR TransCure.

Author contributions

R.D.R. conceived the computational experiments, processed the data and analyzed the results. N.B. and R.D.R. wrote or modified computer programs and scripts. J.K. expressed and purified MloK1 and prepared the 2D crystal sample. M.C. and J.K. collected cryo-EM data. H.S. initiated and supervised the project. R.D.R. and H.S. wrote the manuscript, with the support from all authors.

Data availability

The raw data are deposited in the EMPIAR database, accession code EMPIAR-10233. The full consensus map is deposited at the EMDB, accession code EMD-4439. The cropped consensus map used for resolution estimation is deposited at the EMDB under accession code EMD-4432 and the fitted model is deposited at the PDB, accession code 6I9D. The ensemble of maps and models derived from 3D classification have been deposited to the EMDB and the PDB under the following accession codes, respectively: class 1, EMD-4441, PDB 6IAX; class 2, EMD-4513, PDB 6QCY; class 3, EMD-4514, PDB 6QCZ; class 4, EMD-4515, PDB 6QD0; class 5, EMD-4516, PDB-6QD1; class 6, EMD-4517, PDB 6QD2; class 7, EMD-4518, PDB 6QD3; class 8, EMD-4519, PDB 6QD4. Other data are available from the corresponding author upon request.

Code availability

The source code for exporting a 2D crystal project to SPA programs is available within the FOCUS project at `http://github.com/C-CINA/focus`. The source code for the modified FREALIGN version is available from the repository at `http://github.com/C-CINA/frealign-2dx`.

Supplementary Information

Supplementary information associated with this article can be found in Appendix A.

4. Single Particle Analysis for High Resolution 2D Electron Crystallography

In this chapter, we expand on the single particle analysis methodology applied to 2D crystals, presented previously. Here we give special emphasis on the methodological details, and provide practical guidelines to users.

Contribution: implementation of modifications in the FOCUS and FREALIGN packages to enable single particle 2D electron crystallography, testing of other SPA packages on 2D crystal data, addressing of 3D classification and resolution estimation issues.

The following section will be submitted as
an invited chapter in the following book:

Methods in Molecular Biology

Edited by:
Brent Nannenga and Tamir Gonen
`http://www.springer.com/series/7651`

Single Particle Analysis for High Resolution 2D Electron Crystallography

Ricardo D. Righetto[1] and Henning Stahlberg[1]*

[1] Center for Cellular Imaging and NanoAnalytics, Biozentrum, University of Basel, Mattenstrasse 26, CH-4058 Basel, Switzerland

* Corresponding Author: henning.stahlberg@unibas.ch

Contents

Abstract

Electron crystallography has been used for decades to determine three-dimensional structures of membrane proteins embedded in a lipid bilayer. However, high resolution information could only be retrieved from samples, where the 2D crystals were well ordered and perfectly flat. This is rarely the case in practice. We implemented in the FOCUS package a module to export transmission electron microscopy images of 2D crystals for 3D reconstruction by single particle algorithms. This approach allows for correcting local distortions of the 2D crystals, yielding much higher resolution reconstructions than otherwise expected from the observable diffraction spots. In addition, the single particle framework enables classification of heterogeneous structures coexisting within the 2D crystals. We provide here a detailed guide on single particle analysis of 2D crystal data based on the FOCUS and FREALIGN packages.

4.1. Introduction

Historically, the structures of membrane proteins have been much more difficult to determine than those of their soluble counterparts. This is mainly due to the hydrophobic nature of the transmembrane domains, which render these proteins challenging both for crystallization and solubilization. However, a special type of crystallography, in which the proteins form ordered two-dimensional (2D) periodical arrays in a lipid membrane, is suitable for gathering information on their three-dimensional (3D) structures. It was from a naturally occurring 2D crystal, the purple membrane of *Halobacterium salobium*, that the first 3D structure of a membrane protein was obtained [Henderson & Unwin 1975]. This breakthrough made the transmission electron microscope (TEM) a promising instrument for studying membrane proteins, as long as 2D crystals could be obtained. In fact, after it was demonstrated that embedding the sample in vitreous ice eliminated drying and fixation artifacts and reduced the impact of radiation damage [Adrian *et al.* 1984], electron crystallography became the method of choice for solving membrane protein structures, reaching near-atomic resolution in a number of cases, such as bacteriorhodopsin [Henderson *et al.* 1990], the plant light-harvesting complex [Kühlbrandt *et al.* 1994], $\alpha\beta$-tubulin [Nogales *et al.* 1998] and aquaporin-0 [Gonen *et al.* 2005]. Unfortunately, it was also realized that satisfying the sample quality requirements for achieving high resolution diffraction, namely the perfect ordering and flatness of the 2D crystals, in most cases was very difficult or even impossible.

While it is possible in some cases to employ X-ray crystallography [Deisenhofer & Michel 1989, Landau & Rosenbusch 1996] or nuclear magnetic resonance (NMR) [Liang & Tamm 2016] to solve structures of membrane proteins, these techniques also have their own shortcomings, namely the need for suitably large 3D crystals, or being constrained to small molecular weights, respectively. Furthermore, in both techniques, the membrane protein environment is often not comparable to the native cellular membrane. In single particle cryo-electron microscopy (cryo-EM), however, most of these technical challenges are overcome by performing 3D reconstructions from proteins randomly oriented in solution [Frank 2006, Kühlbrandt 2014]. Membrane proteins, specifically, can be solubilized if surrounded by detergent micelles [Liao *et al.* 2013] or a lipid nanodisc [Gao *et al.* 2016].

Nevertheless, there are specific cases where working with 2D crystals is necessary or even advantageous compared to these other techniques. One example is when such periodical arrangements occur natively in the cell membrane [Henderson & Unwin 1975, Henderson *et al.* 1990], or when their natural formation is associated with a biological specific function [Goñi *et al.* 2014]. Another example is when the 2D array is designed artificially, enabling studies of proteins and other molecules that would be difficult otherwise [Gonen *et al.* 2015, Subramanian *et al.* 2018]. Other advantages are that 2D crystallography offers the highest possible efficiency in terms of number of particles in the field of view, and allows reconstructions from arbitrarily small proteins, which can be challenging for conventional single particle analysis (SPA). Those are specially interesting aspects if combined with the single particle approach outlined in this chapter.

2D electron crystallography can be performed both in the diffraction or the imaging mode of the TEM [Schenk *et al.* 2010]. In this chapter, we will cover only the imaging mode, in which several projection images of 2D crystals at different orientations are recorded. As in single particle analysis, the principle of 3D reconstruction is based on the central section theorem [Frank 2006, De Rosier & Klug 1968]. The periodical nature of the two-dimensional array in real space appears as diffraction spots (Bragg peaks) in reciprocal space. These spots can then be indexed, measured, and corrected for the contrast transfer function (CTF) of the microscope. After merging the detected spots in 3D reciprocal space, a real-space map of the protein can be obtained by Fourier inversion [Schenk *et al.* 2010, Stahlberg *et al.* 2015]. Algorithms for performing these operations have been implemented in the MRC package [Crowther *et al.* 1996], and subsequently in the *2dx* [Gipson *et al.* 2007] and FOCUS [Biyani *et al.* 2017] packages, among others [van Heel *et al.* 1996]. A technical limitation of 2D electron crystallography is that samples in the microscope can typically only be tilted up to about 60 degrees for geometric reasons, therefore leaving a cone of missing information in 3D Fourier space. Furthermore, only if the 2D crystals are perfectly flat and well ordered, high resolution diffraction spots will be observed. Because this is rarely the case in practice, image processing algorithms were developed to correct for the crystal distortions *in silico* [van Heel & Hollenberg 1980, Saxton & Baumeister 1982]. One such algorithm, the so-called *image unbending* [Henderson *et al.* 1986], has been particularly successful [Henderson *et al.* 1990]. This method works by moving small patches of the 2D crystal image by iteratively comparing cross-correlation peaks indicating the locations of unit cells in real space with their predicted positions. However, this approach is intrinsically limited to distortions in the image plane, and cannot account for azimuth angle variations that are present when 2D crystals with larger in-plane distortions are imaged at higher tilts. In order to be able to correct for out-of-plane distortions, such as "bumps" in the 2D crystal, the patches need to be compared (e.g., cross-correlated) against a reference in 3D space, which is essentially what single particle refinement and reconstruction algorithms do [Frank 2006]. First attempts in this direction showed promise [Scherer *et al.* 2014, Kuang *et al.* 2015], but were still limited to low resolution reconstructions. This was in part because the datasets analyzed in these works had not yet been collected on direct electron detectors (DED) [McMullan *et al.* 2016], and partially because the algorithms implemented did not account for the very low signal-to-noise (SNR) ratios of cryo-EM images in a probabilistic manner [Sigworth *et al.* 2010].

We have since then implemented a module in the FOCUS package that allows the user to export a 2D crystallography project for processing with standard single-particle analysis software. This not only brings to electron crystallography the maturity that these packages have achieved in terms of robustness and performance [Punjani *et al.* 2017, Grant *et al.* 2018, Zivanov *et al.* 2018], but, more importantly, it opens the possibility of classifying heterogeneous structures [Scheres *et al.* 2007, Lyumkis *et al.* 2013] coexisting within the 2D crystals. In the **??** section, we will describe in detail the workflow for processing 2D crystal data within the single particle framework, with an emphasis on practical tips and tricks. The approach presented here is based on the one we used to process poorly diffracting 2D crystals of MloK1, a 160 kDa prokaryotic potassium channel [Kowal *et al.* 2014, Kowal *et al.* 2018]. Using single particle refinements, we were able to obtain a map of MloK1 at 4 Å resolution and to observe distinct conformations of its cyclic-nucleotide binding domain (CNBD), helping to elucidate the mechanism of gating for this channel [Righetto *et al.* 2019]. The method is generally applicable to other 2D crystal samples as well.

4.2. Materials

4.2.1. Software

In this work, we refer to the software package FOCUS (`http://www.focus-em.org`) [Biyani *et al.* 2017] and a version of the FREALIGN package [Grigorieff 2016] extended with additional features useful for the processing of 2D crystal data, hereby called frealign-2dx to avoid confusion with the original implementation (`http://github.com/C-CINA/frealign-2dx`) [Righetto *et al.* 2019]. Both, FOCUS and frealign-2dx are freely available as open-source software and run on Linux-based operating systems.

4.2.2. Hardware

A typical computing workstation to carry out the data processing steps described ahead will have the following hardware components (see Note 1):

- 2× 12-core CPU
- 256 GB RAM
- $>$ 50 TB HD storage
- 1 TB SSD storage (used as a fast "scratch" disk)
- 2x NVIDIA GTX 1080 GPU card (see Note 2)

With this setup, it is possible to carry out real-time data preprocessing (*e.g.*, drift-correction and CTF estimation, see Section 4.3.1), classical 2D crystal processing (see Section 4.3.2) and single particle refinements (Section 5 onwards). However, for the processing of large datasets, in particular with large box sizes, a high-performance computing (HPC) cluster may be required.

4.3. Methods

We here describe the computational steps required to process 2D crystal data with single particle software. For details and protocols on the growth of 2D crystals and their sample preparation for TEM imaging, please refer to [Abeyrathne *et al.* 2010, Abeyrathne *et al.* 2012, Goldie *et al.* 2014, Schmidt-Krey & Cheng 2013]. After exporting the data from FOCUS [Biyani *et al.* 2017], we will use frealign-2dx, an extended version of the FREALIGN package [Grigorieff 2016] (see Supplementary Notes 1 and 2 of [Righetto *et al.* 2019]), although any other single particle analysis package can also be used, in principle (see Note 3).

4.3.1. Data acquisition and initial processing of movies

To ensure a good spectral SNR (SSNR) in the high-resolution range, we recommend the acquisition of movies at a microscope magnification corresponding to a pixel size <1 Å at the sample level, to benefit from the improved detective quantum efficiency (DQE) of DEDs at lower detector resolutions [McMullan *et al.* 2016]. Total exposures in the range of 40-50 $e^-/\text{Å}^2$ per movie are known to work well. Multiple 2D crystals should be imaged, with tilt angles ranging from 0 to ±60 degrees. A typical dataset will contain ~100 to ~1000 movies, which should be spread accross the tilt range.

Movies of 2D crystals should be corrected for beam-induced drift [Brilot *et al.* 2012] and for beam-induced resolution loss by applying dose-dependent temperature factors to each frame before computing frequency-weighted averages from all frames [Grant & Grigorieff 2015], generating a drift-corrected average image. Also, defocus and astigmatism have to be estimated for each image [Rohou & Grigorieff 2015, Zhang 2016]. Using FOCUS, these steps can be carried out in real time during the microscopy session and the dataset can be conveniently pruned based on parameters such as total drift, measured defocus and estimated CTF resolution, among others [Biyani *et al.* 2017].

4.3.2. Conventional 2D crystallographic reconstruction

Before starting with the single particle approach, it is necessary to process the data in conventional 2D electron crystallography fashion. For each 2D crystal image in the dataset, the crucial parameters to be obtained at this stage are:

- The defocus and astigmatism at the center of the image
- The tilt geometry of the specimen plane, defined by the angles TLTANG and TLTAXIS
- The orientation of the 2D crystal on the specimen plane, defined by the angle TAXA
- The phase origin (translational shifts)
- The unit cell positions provided by the unbending algorithm [Henderson *et al.* 1986]

Ideally, the best possible 3D map should be obtained in this way by iterative 3D merging and refinement before proceeding, although an initial, low-resolution map may suffice (see Note 4). For more information and protocols for the processing of 2D crystal data, please refer to [Schenk *et al.* 2010, Stahlberg *et al.* 2015, Gipson *et al.* 2007, Schmidt-Krey & Cheng 2013, Biyani *et al.* 2018].

4.3.3. Exporting a 2D crystal project for single particle analysis

Having performed conventional 2D crystallography data processing, we can proceed to the single particle analysis of 2D crystals. The single particle scripts are available under the `Particles` tab on the top bar of the FOCUS GUI. The first script, `Pick & export particles`, is the core of our approach, allowing the user to extract the particles and export these data in a way that is understood by single particle programs. The script will operate on the images that have been selected via the FOCUS library GUI (see Note 5). The script presents a number of parameters to the user, the most important of which will be explained below.

Particle Picking

The classical unbending algorithm [Henderson *et al.* 1986] generates a cross-correlation (CC) profile (Fig. 4.1a). The peaks in this CC profile indicate the position of the unit cells, available after running the `Unbend II` script during the 2D crystal processing in Section 4.3.2 (Fig. 4.1b). This information, stored in the file `<image_name>_profile.dat` for each micrograph, will be used to pick the particles from the 2D crystals (Fig. 4.1c).

Optionally, a `Phase Origin Shift` can be applied to the CC peaks. This is a translational shift, here presented in degrees in relation to the unit cell dimensions. A phase origin shift of 180° in one direction means shifting by half the length of the unit cell. Such a shift may be required when the center of the protein does not coincide with the center of the crystal unit cell, as shown in Fig. 4.2. The shift is calculated accordingly for non-tilted crystals, based on the current estimate of the tilt geometry.

The `Box size` parameter defines the side length of the (squared) box into which the particles will be windowed, in pixels. The particles, in this context, are patches of the 2D crystals, and this parameters defines how large such patches should be. Differently from conventional SPA, it is not possible here to have a single protein unit isolated in the box. Also, in the tilted views the projection of the neighboring proteins overlaps with that of the central protein, which has to be taken into account during the particle alignments (see Fig. 4.4). For this reason, the box size should be large enough to contain the protein of interest plus at least one neighbor protein on each side, implying an overlap between the boxes of adjacent particles, as shown by the blue boxes in Fig. 4.1c. This also favors the smoothness of the alignment parameters across the 2D crystal, because of the natural correlations existing among adjacent particles (proteins located close to each other in the 2D crystal tend to appear in similar orientations). Furthermore, larger boxes contain more signal,

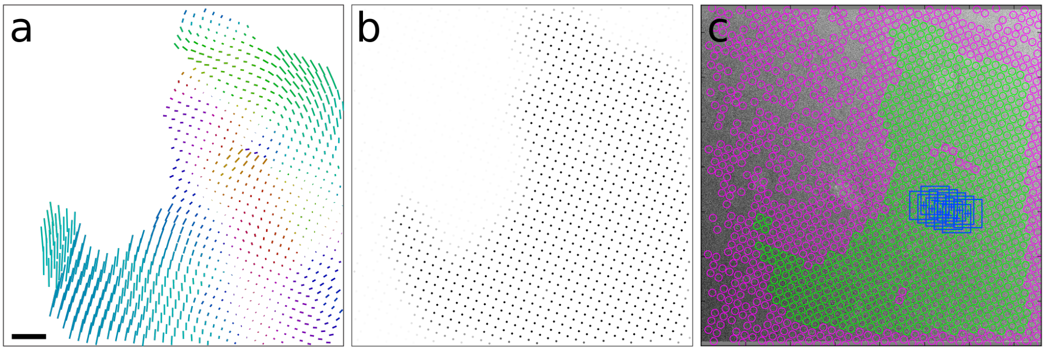


Figure 4.1: **Particle picking from unbending profile.** a) The unbending algorithm compares the observed and predicted unit cell positions. Displacement vectors are 10× exaggerated for visualization, with colors indicating the displacement direction. b) CC peaks indicating unit cell positions after unbending. Darker dots indicate stronger peaks. c) Particle picking based on the information from b). Green circles are picked particles (*i.e.*, above user-selected threshold) and magenta circles are ignored particles. The blue squares indicate the size (416 Å in this case) of the overlapping patches to be extracted, with the crosses indicating the box centers. Scale bar: 500 Å.

rendering the alignments more reliable. On the other hand, a box too large detracts from the goal of correcting local crystal distortions as accurately as possible. The optimal box size will depend mainly on the actual size and molecular weight of the protein. Smaller proteins will certainly require larger patches comprising more neighboring units, and vice-versa.

The `Threshold to include particles` parameter allows to define how strict the picking should be:

$$CC_{thr} \geq \mu_{CC} + z\sigma_{CC} \tag{4.1}$$

where CC_{thr} is the threshold above which candidate particles defined by the CC peaks in a given 2D crystal will be picked, μ_{CC} is the average of all CC peaks in that crystal and σ_{CC} their standard deviation. z is a multiplier of the standard deviation to make the threshold more or less selective. Higher (positive) values of z make the picking more stringent and in theory selects "better" particles. Conversely, lower (negative) values will pick more, potentially also "worse", particles. The default for `Threshold to include particles` is $z = 0$, meaning only CC peaks above average will be boxed, which in most cases is a good compromise (see green and magenta circles in Fig. 4.1c).

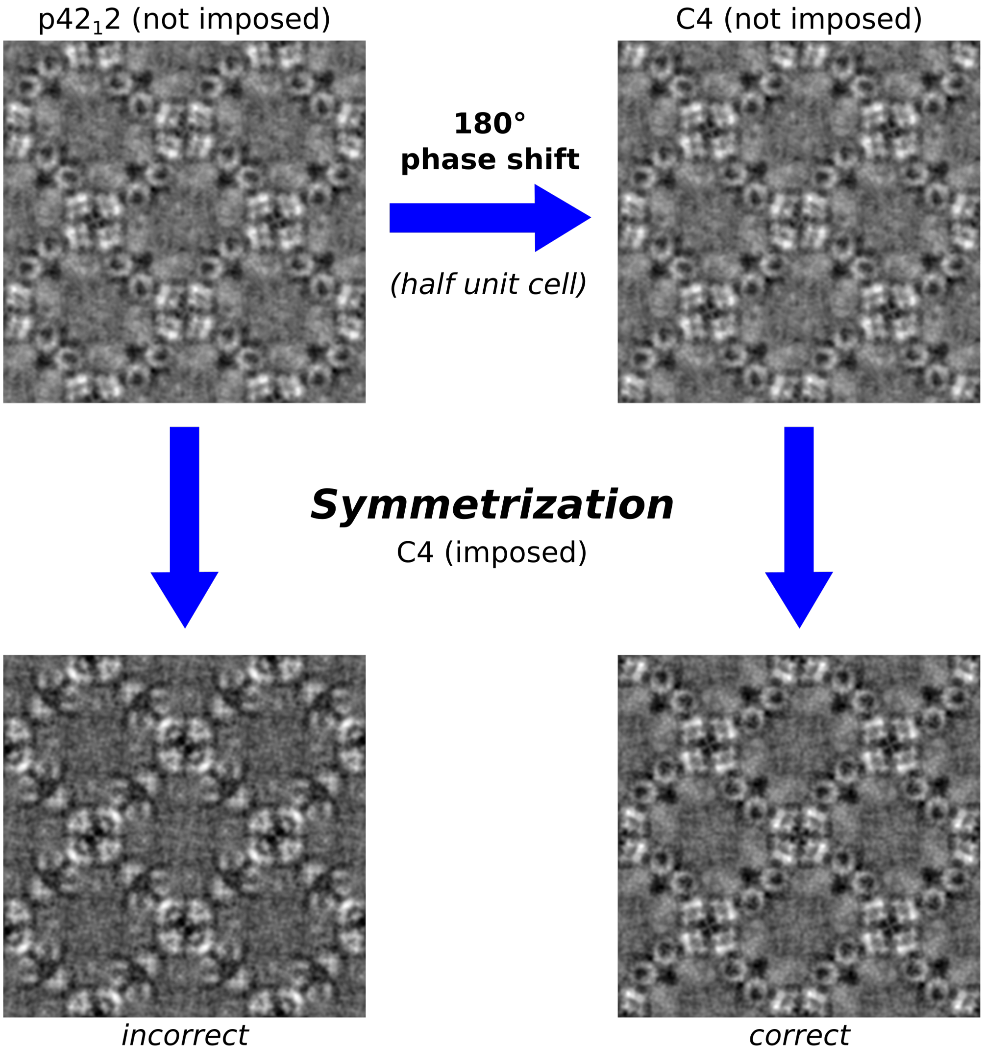


Figure 4.2: **Shifting the phase origin.** In single particle analysis it is desired to have the protein of interest at the center of the reconstruction box. This is especially important when the object is symmetrical in 3D. Crystallographic reconstructions may have *point-group* symmetry, while single particle reconstructions may have *space-group* symmetry. In the case of MloK1, the phase origin of non-tilted views has to be shifted by half a unit cell (180°), to bring one tetramer to the center of the particle box and then be able to apply C4 symmetry. The procedure is illustrated for the projection map of a non-tilted crystal. The tilted views are also shifted accordingly by the cosine of the tilt angle.

CTF Correction

Each particle will be assigned a defocus value. For perfectly non-tilted images, the defocus will be the same for all particles. However, for tilted images there will be a defocus gradient. The defocus at the center of each particle box is computed according to the CTFTILT equation [Mindell & Grigorieff 2003]:

$$\Delta f = \Delta f_0 + [(x - x_0)sin\phi - (y - y_0)cos\phi]tan\gamma \tag{4.2}$$

where Δf is the defocus at the center of the particle box, Δf_0 is the defocus at the center of the image, (x, y) and (x_0, y_0) are the coordinates of the box and the image centers, respectively, ϕ is the angle between the tilt axis and the X-axis (*i.e.*, TLTAXIS) and γ is the tilt angle (TLTANG). Astigmatism is assumed to be constant across the whole image. The particles are boxed without CTF correction, which will be performed later by frealign-2dx. However, in addition to boxing uncorrected particles, it is possible and recommended to generate also a CTF corrected particle stack. There are three options for this CTF correction in FOCUS: *phase flipping*, *CTF multiplication*, or *Wiener filtering*. The latter requires a user-defined *ad hoc* constant proportional to the inverse of the dataset's SNR. Any of the three options should yield similar results for the purposes of calculating crystal averages (Section 4.3.4).

Exporting the metadata

When generating the particle metadata files, the orientation for each particle will be inherited from the 2D crystal it was picked from. As frealign-2dx uses the SPIDER angle convention [Frank *et al.* 1996], the crystallographic tilt geometry is converted to Euler angles as follows [Biyani *et al.* 2018]:

$$\begin{aligned} \psi &= 270^\circ - TLTAXIS \\ \theta &= TANGL \\ \phi &= 90^\circ - TAXA \end{aligned} \tag{4.3}$$

where ψ is an in-plane rotation (2D), θ is the tilt angle, and ϕ is a 3D rotation. These values, along with defocus information and other metadata, are stored in a text file called `particles.par`. The particle stack is stored in the file `particles.mrcs`. If CTF-corrected stacks are generated, additional similar files will also be created. All these files are located in the directory `stacks/` inside the location specified by the `General Single-Particle directory` parameter. If an absolute path is not specified, this location will be relative to the `merge/` directory within the FOCUS project directory. Optionally, figures indicating the included and ignored picking coordinates overlayed on the micrographs can also be saved (inside the `picking/` directory), and the metadata can also be saved in `.star` format for RELION (see Note 6). The underlying Python script that performs particle picking and metadata creation can be efficiently run in parallel, with a user-defined number of threads.

It is important to notice that while creating the particle stacks and respective metadata, the script will by default ensure that particles extracted from the same 2D crystal stay assigned to the same "half-set". This is required to prevent inflated resolution estimates, because of the large overlap between adjacent particles (Fig. 4.1c) [He & Scheres 2016]. In frealign-2dx this is accomplished by interleaving the 2D crystals in the particle stack and in the `.par` file. In RELION, this is controlled by the `rlnRandomSubset` and `rlnHelicalTubeID` labels in the `.star` file (see Note 7).

4.3.4. Pre-Refinement

After picking and exporting the particles, it should be possible to start a single particle refinement straight away (Section 4.3.5). It is fast and useful, however, to run a "pre-refinement" (see below), using crystal averages prior to the full refinement with individual particles.

Crystal averages

The next script in the pipeline, `Generate crystal averages`, will calculate *crystal averages*, using the CTF-corrected particle stacks (Section 5). These averages are analogous to 2D class averages in SPA. They provide a straightforward way of visually assessing the picking parameters chosen (emphe.g., phase origin shift, box size, etc.). This procedure is also known as the *correlation averaging* method [Saxton & Baumeister 1982]. The Fourier ring correlation (FRC) for each crystal is also computed, its plots can be consulted under the `FRC/` directory. As a validation measure, the crystal averages should be highly similar to their corresponding projection maps, such as those shown in Fig. 4.2 (except for the phase origin, if a shift was applied as described in Section 5). The difference is that the crystal averages are computed by averaging the boxed unit cells directly, while the projection maps are computed from the information in the diffraction spots alone (which is also a form of averaging). Examples of crystal averages are shown in Supplementary Fig. 3 of [Righetto *et al.* 2019].

Running the pre-refinement

The script `Prepare pre-refinement` offers a graphical interface to generate an `mparameters` file for frealign-2dx [Grigorieff 2016]. It will also put the files required by frealign-2dx into the `pre-refine` directory, by default. The default values should be reasonable, but the user is encouraged to fine-tune the parameters to her or his needs. We draw the attention here to the `Auto-refinement` parameters available in our modified version of frealign-2dx:

- `FSC threshold for resolution limit? (thresh_fsc_ref)`: the resolution limit used in refinement (for preventing overfitting) will be taken as the resolution where the Fourier shell correlation (FSC) curve crosses this threshold (default: 0.800)
- `FSC threshold for map improvement evaluation? (thresh_fsc_eval)`: any map improvement between refinement cycles will be assessed by looking at the resolution where the FSC curve crosses this threshold (default: 0.143)
- `Minimum resolution limit? (res_min)`: this parameter prevents the auto-refiner from using a resolution too low, which may cause the refinement to diverge (default: 40 Å)
- `Stay away from current map resolution? (ref_stay_away)`: this parameter prevents the resolution limit from getting too close to the current map resolution, which would potentially introduce bias in the refinement (default: 2 Å)
- `Try different combinations of parameter mask? (change_pmask)`: if enabled, this parameter makes the auto-refiner try different combinations among the active refinement parameters $(\phi, \theta, \psi, x, y)$, before declaring convergence, which helps avoiding local minima (default: *yes*)
- `Keep the tilt angle fixed? (no_theta)`: if active, the tilt angle (θ) will be kept fixed. Sometimes this is helpful for 2D crystal data (default: *no*)

For more information, please see Supplementary Note 2 of [Righetto *et al.* 2019]. After preparing the pre-refinement, it can be launched via the `Run pre-refinement` script. Alternatively, it can be launched from the command line by running the `frealign_run_refine_auto` program inside the pre-refinement directory. The crystal averages are only a few hundred for a typical dataset and they have a high SNR. The pre-refinement should therefore complete in only a few minutes and noticeably improve the tilt geometry estimation (now converted to Euler angles), as shown in Fig. 4.3. Plots of the angular changes can also be obtained by running the `Analyze pre-refinement results` script.

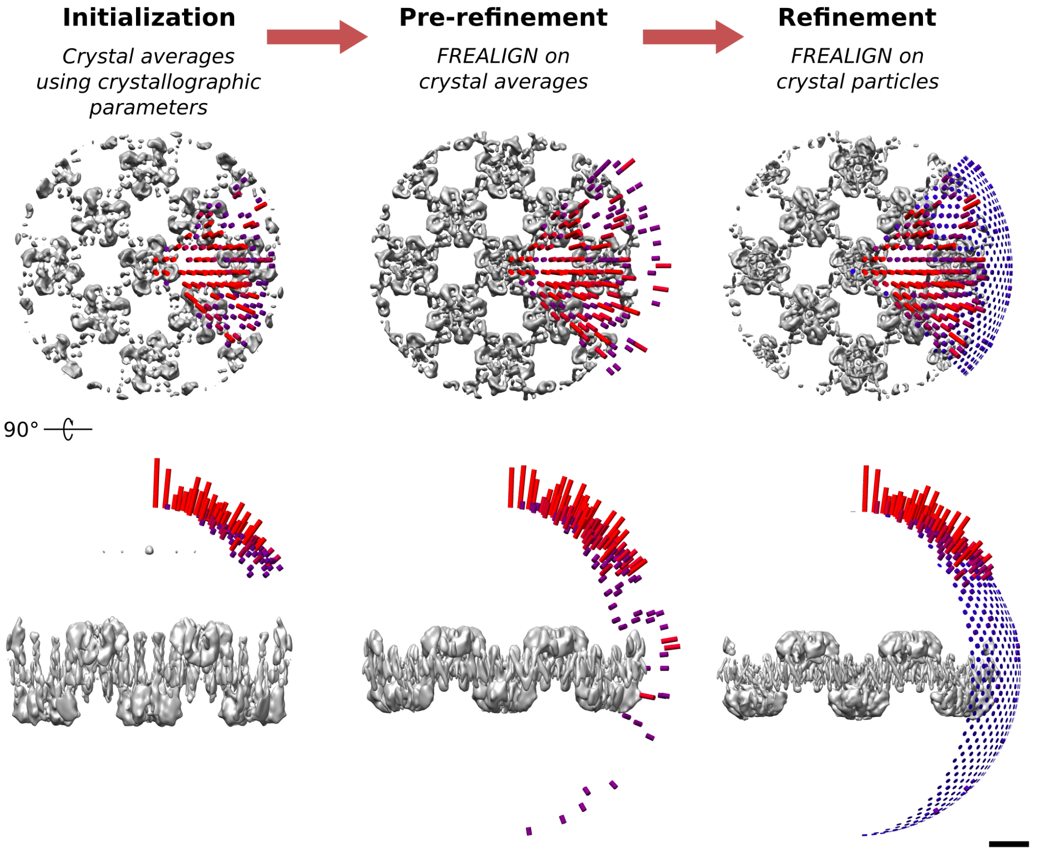

Initialization
Crystal averages using crystallographic parameters
Pre-refinement
FREALIGN on crystal averages
Refinement
FREALIGN on crystal particles
90°

Figure 4.3: **Single particle refinement from 2D crystals.** Left: The initial reconstruction uses parameters from 2D crystallographic data processing. Center: The map is then improved by doing a “pre-refinement” with frealign-2dx using crystal averages. Right: Using particles (local patches) extracted from the 2D crystals starting with parameters obtained in the pre-refinement, the map is refined to high resolution with frealign-2dx. Both the height and the color of the cylinders (here shown for the asymmetric unit only) indicate the proportion of particles or crystal averages in the corresponding orientation, relative to each dataset (red: high particle density, blue: low particle density; a.u.). Maps shown are not sharpened. Scale bar: 50 Å.

4.3.5. Consensus Refinement

After the pre-refinement is complete, the full refinement using the extracted particles can take place. This refinement finally corrects for the local distortions of the 2D crystals. As in the pre-refinement, the FOCUS script `Prepare refinement` will generate the `mparameters` file and generate the proper file structure for the frealign-2dx refinement. At this stage, the extracted particles will inherit the pre-refined alignment parameters from their "parent" 2D crystals (Section 4.3.4). This is done internally by the Python script `SPR_FrealignParameterInheritance`. The extracted particles have a much lower SNR than the crystal averages, and their alignment parameters are expected to not deviate too much from the crystal average. Therefore, besides the auto-refinement parameters (Section 4.3.4), it is now recommended to activate the alignment restraints available in frealign-2dx (see Note 8):

- `Restraint for Euler angles, in degrees (sigma_angles)`: standard deviation for a Gaussian restraint on the Euler angles
- `Restraint for x,y shifts, in pixels (sigma_shifts)`: standard deviation for a Gaussian restraint on the x, y shifts

More information about these restraints can be found in the Supplementary Note 1 of [Righetto *et al.* 2019]. Upon convergence, a higher quality map should have been obtained, and a more diverse set of views should be observed, as shown in Fig. 4.3. This evidences that particles from the same 2D crystal are not exactly always in the same orientation. This map, commonly referred to as a "consensus map" because no 3D classification has been performed (yet), can now be postprocessed as described in Section 4.3.7. The user is also encouraged to experiment with more advanced features, such as defocus refinement [Grant *et al.* 2018, Grigorieff 2016].

4.3.6. 3D Classification

Perhaps the most interesting feature of the single particle method is its ability to classify heterogeneous data into conformationally homogeneous classes [Lyumkis *et al.* 2013, Scheres 2012a]. Structural variability is also a source of disorder in 2D crystals, thus limiting the achievable resolution by conventional crystallographic methods. We have used 3D classification to detect distinct conformations of the MloK1 CNBD in relation to the transmembrane domain (TMD) [Righetto *et al.* 2019].

Differently from conventional SPA, however, in 2D crystal particles the protein of interest is always surrounded by neighboring proteins. This means that, in tilted views, the projection of the neighbors overlaps with the projection of the classification target. This problem is illustrated in Fig. 4.4.

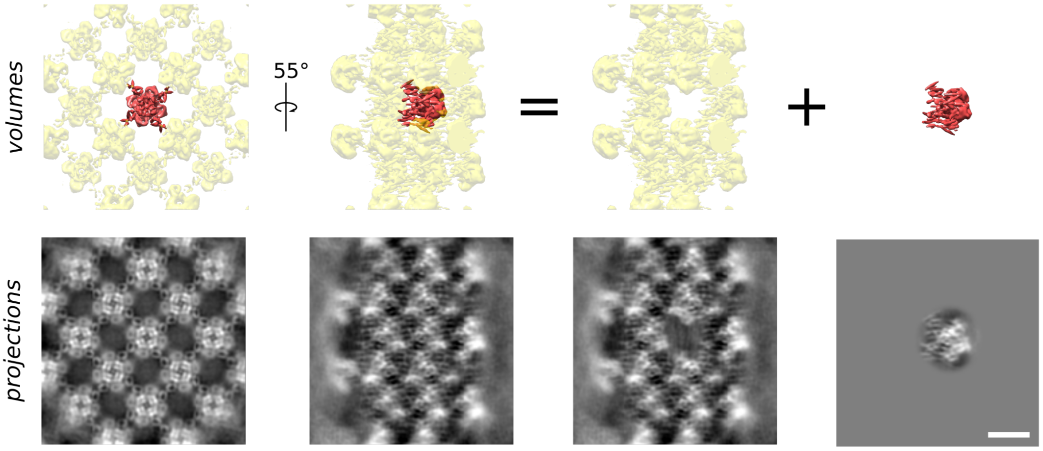


Figure 4.4: **Signal subtraction for 3D classification.** For 2D crystal data, the projection of the neighboring proteins (transparent yellow densities) overlap with the projection of the classification target (red density) in the tilted views. In order to perform 3D classification, it is necessary to subtract the signal of the neighboring proteins from the particles. Shown here is an example from the MloK1 dataset. Scale bar: 50 Å.

Therefore, 3D classification will only work properly if the signal from the neighbor proteins is subtracted from the particle images. For more accurate results in signal subtraction of 2D crystal data, the user should calculate a completely unmasked reconstruction in frealign-2dx (see Fig. 4.4). This can be done by setting a reconstruction radius larger than the particle box in the `mparameters` file and then running the command `frealign_calc_reconstructions` from the refinement directory (see Note 9). Based on this reconstruction, the user should then define a soft mask focussed on the protein of interest (emphe.g., the central MloK1 tetramer), and invert this mask to select everything else that should be subtracted from the experimental projections (*i.e.*, the particles). The `focus.postprocess` tool can be used to create such masks (see Section 4.3.7). The signal subtraction itself can be accomplished from the command line using RELION [Bai *et al.* 2015]. An example command line for this would be:

```
relion_project --i particles_9_r1-subtract.mrc --subtract_exp --angpix 2.68
--ctf --ang particles_9_r1.star --o subtracted_particles
```

The `mparameters` file can then be edited to point to the new `subtracted_particles.mrcs` stack. If required, the new particle stack and the subtracted 3D reference can be downsampled to a coarser pixel size for computational speedup. Also, the user

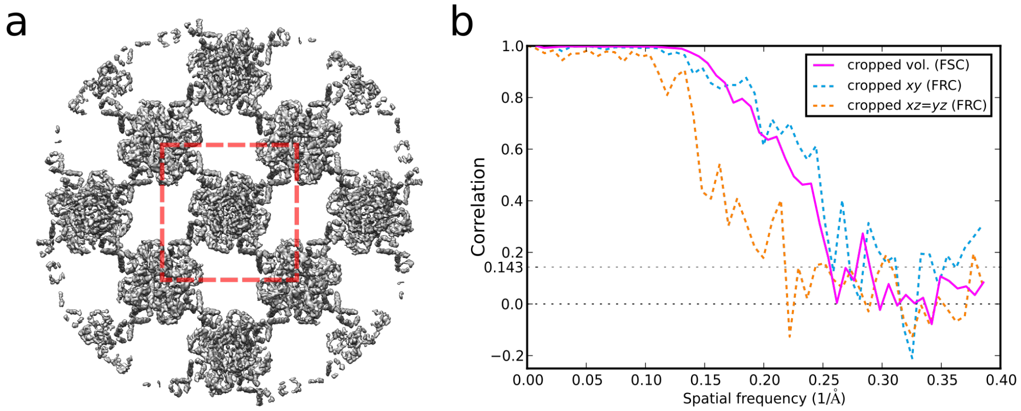


Figure 4.5: **Resolution estimation.** a) The full, sharpened consensus map of MloK1, with the central tetramer, which was cropped for resolution estimation, highlighted. b) FSC curve for the central tetramer, after adjusting for the protein volume, and FRC curves for the xy and $xz = yz$ planes.

should set the `nclasses` parameter to the number of classes desired. In order to run the 3D classification, the user should start immediately after the refinement cycle in which the consensus refinement converged (emphe.g., 10, if the auto-refinement converged in cycle 9). Because the signal-subtracted particles may contain too little signal for reliable alignments, it is recommended to disable the optimization of any alignment parameters at this stage and run the classification for a pre-defined number of cycles (emphe.g., 40). In this case, the auto-refiner will not be used and the 3D classification can be launched with the command `frealign_run_refine`. The occupancies of the particles in each class will be randomized to initialize the classification. Convergence of the 3D classification can be assessed by the user with the `frealign_calc_stats`.

4.3.7. Assessing Resolution and Post-Processing

Resolution is typically assessed by computing the FSC curves [Harauz & van Heel 1986, Rosenthal & Henderson 2003] between independently (beyond a certain resolution) refined half-maps. As explained in Section 5, FOCUS ensures that particles from the same 2D crystal are assigned to the same half-sets to prevent an inflation of the FSC curve. Normally, only a sub-volume is of interest, for example, the central tetramer in the case of MloK1, as shown in Fig. 4.5.

We developed a program called `focus.postprocess` that conveniently offers a number of volume cropping, masking, resolution estimation and sharpening operations. A typical command line is:

```
focus.postprocess map1.mrc map2.mrc --angpix 1.3 --crop_center 0,0,-12
--crop_size 104,104,104 --mtf data_mtf_k2_300kv.star --mw 160.0
--mask_radius 42 --out consensus
```

The `focus.postprocess` tool has features that allow the user to assess the resolution anisotropy of the map due to the missing cone (options `--cone_aperture`, `--xy_only`, `--xz_only` and `--yz_only`). All options can be found by typing `focus.postprocess --help` on the command line.

If maps from multiple 3D classes have been obtained and refined, the same procedures should be applied on them. Finally, the postprocessed maps can be used for modelling the atomic structure of the protein [Brown *et al.* 2015].

4.4. Conclusions

Single particle refinement algorithms can obtain 3D reconstructions from disordered 2D crystals at resolutions well beyond the range of observable diffraction diffraction spots. A dedicated module in the FOCUS package and the frealign-2dx program, a modified version of FREALIGN, enable this method. The approach is not limited to these specific implementations, though. A likely reason for so many 2D crystals being disordered is conformational heterogeneity. The single particle approach to 2D crystals allows signal subtraction and 3D classification, which can be used to disentangle different conformations as demonstrated for the MloK1 potassium channel dataset. In summary, this method presents a viable alternative to conventional 2D crystallography when the 2D crystal form is related to the protein function or intentionally designed.

4.5. Notes

1. For other possible hardware examples, see the official online FOCUS documentation: `http://focus.c-cina.unibas.ch/wiki/doku.php?id=1_0:hardware_linux`

2. Depending on which packages you will use, GPU cards may not be required. Please consult the specific documentation of the software packages present in your data processing pipeline.

3. We have also successfully used *cis*TEM [Grant *et al.* 2018], RELION [Scheres 2012a] and cryoSPARC [Punjani *et al.* 2017] with 2D crystal data, although these packages have not been as extensively tested and customized as frealign-2dx in this context.

4. Actually, just merging in 2D may be enough in some cases, as long as phase origins are sufficiently

accurate. In this case, however, an initial 3D map needs to be obtained by other means. For example, we have successfully used stochastic gradient descent (SGD) [Punjani *et al.* 2017] to this end.

5. It is assumed the dataset has been pruned accordingly before and/or during the conventional 2D crystal reconstruction workflow.

6. The program `par2star.py` can also be called from the command line as a standalone tool.

7. If using cryoSPARC, please note that, at present, it ignores these fields and always randomizes the order of the particles when assigning its half-sets. In this case it is up to the user to ensure that the resolution estimation is unbiased by, for example, re-splitting the data based on the 2D crystals after the refinement and calculating half-maps using FREALIGN or RELION.

8. If using RELION, similar behavior would be accomplished by using local searches, *i.e.*, setting the initial angular sampling equal to the sampling from which to perform local searches.

9. For the signal subtraction and 3D classification tasks, the user should work from the command line, as these options are not available from the FOCUS GUI.

Acknowledgements

Calculations were performed at the sciCORE cluster of the University of Basel. RDR acknowledges funding from the Fellowships for Excellence program sponsored by the Werner-Siemens Foundation and the University of Basel. This work was supported by the Swiss National Science Foundation through grant 205320_166164 and the NCCR TransCure.

5. Application: Membrane binding induces domain rearrangements and oligomerization that prime FAK for activation

This chapter presents a biologically significant application of the methods described in Chapters 3 and 4. Focal Adhesion Kinase (FAK) is an enzyme that self-assembles attached to a lipid membrane. It is involved in the processes of cell adhesion and migration, and is known to be overexpressed in cancer cells. Because the spontaneously formed arrays are largely disordered, our single particle approach to 2D crystals is required to determine the structure of membrane-assembled FAK.

Contribution: processing of FAK 2D crystal data with SPA algorithms, model fitting and refinement.

The following section is part of a publication in preparation:

Membrane binding induces domain rearrangements and oligomerization that prime FAK for activation

Acebrón, I.[1], Righetto, R.[2], Culley, J.[3], Daday, C.[4], Biyani, N.[2], Redondo, P.[1], Chami, M.[2], Boskovic, J.[1], Gräter, F.[4], Frame, M.[3], Stahlberg, H.[2] & Lietha, D.[1,5*]

[1] Cell Signalling and Adhesion Group, Structural Biology Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain

[2] Center for Cellular Imaging and NanoAnalytics, Biozentrum, University of Basel, Mattenstrasse 26, CH-4058 Basel, Switzerland

[3] Edinburgh Cancer Research UK Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XR, United Kingdom

[4] Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany

[5] Cell Signalling and Adhesion Group, Structural and Chemical Biology, Centro de Investigaciones Biológicas (CIB), Spanish National Research Council (CSIC), 28040 Madrid, Spain

* Corresponding Author: daniel.lietha@cib.csic.es

Contents

5.1. Introduction

Focal adhesions are complex macromolecular structures that allow cells to mechanically sense the external environment, i.e., the extracellular matrix (ECM). Located on the cytoplasmic side of the cell membrane, they transduce mechanical forces into signalling between the ECM and internal, long (> 10 μm) actin microfilaments. The enzyme focal adhesion kinase (FAK) is a key component of focal adhesion sites. Because of its role in regulating events such as cell migration, adhesion and survival, FAK is an anticancer drug target [Lietha *et al.* 2007, Bauer *et al.* 2019].

Although atomic structures for individual domains have been determined, little is known about the quaternary structure of FAK and its assembling with respect to the membrane and other proteins [Lietha *et al.* 2007]. It has been previously observed that a particular phospholipid of the cell membrane, PIP_2, is involved in FAK activation [Goñi *et al.* 2014] and allosteric regulation [Zhou *et al.* 2015]. In one of these studies, FAK was shown to assemble in clusters [Goñi *et al.* 2014]. Further *in vitro* experiments revealed that FAK tends to spontaneously assemble into relatively ordered 2D arrays in the presence of a PIP_2 monolayer. Such a layer is expected to closely resemble the native environment of the eukaryotic cell membrane where the protein is found, and therefore prompted the use of 2D electron crystallography and our single-particle approach for solving the FAK structure. The chicken FAK studied here preserves ~95% sequence identity to human FAK. The construct comprises the catalyic kinase domain (32 kDa) [Nowakowski *et al.* 2002] and its autoinhibiting FERM domain (42 kDa) [Ceccarelli *et al.* 2006], with a total molecular weight of approximately 74 kDa. Despite advances in the development of phase plates for enhancing contrast in transmission electron microscopy (TEM) [Khoshouei *et al.* 2017], membrane proteins of this size still pose great challenge for single particle cryo-EM. This is not a constraint, however, for 2D electron crystallography.

5.2. Results

We first collected cryo-EM data from 2D crystals of FAK alone, hereby denoted the “FAK apo” dataset, and subsequently of FAK in the presence of the AMP-PNP ligand, hereby denoted the “FAK AMP-PNP” dataset. Fig. 5.1 shows a typical micrograph imaged from these samples and its Fourier transform (FT). It is readily seen that these 2D crystals diffract only to ~25 Å or worse.

Nevertheless, we proceeded with standard 2D crystallographic image processing, resulting in a map where only the FERM domains were visible. In the 2D crystals, the FAK monomers tend to assemble in p2 symmetrical dimers, which interface other dimers via the FERM domain. The low resolution of this map invited the appli-

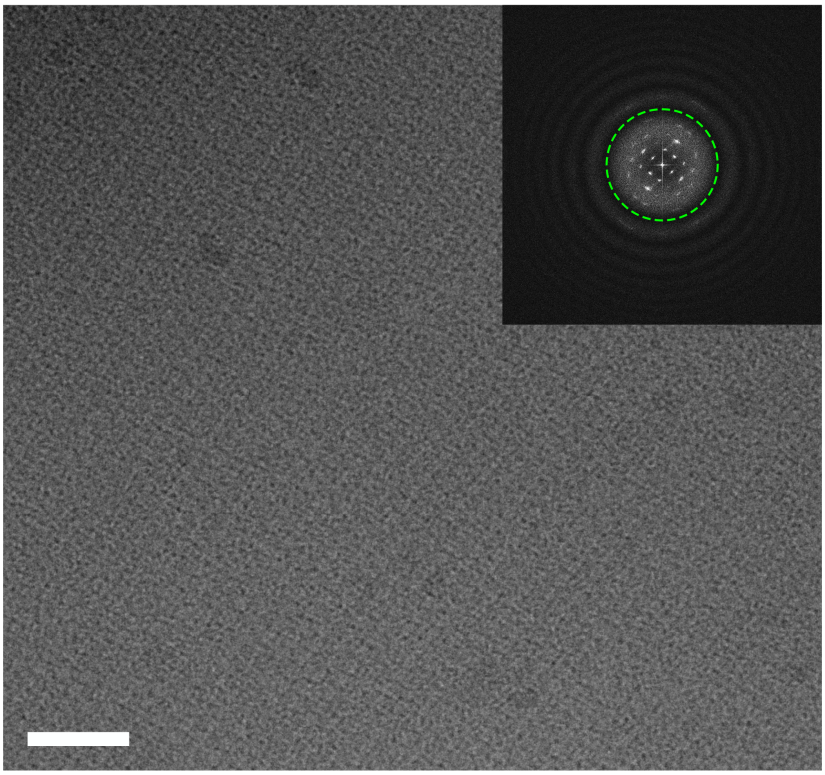


Figure 5.1: **Typical micrograph of a FAK 2D crystal.** A typical non-tilted micrograph from the FAK apo dataset is shown, taken at $-2.3\ \mu$m defocus. The inset shows the central part of the micrograph FT amplitudes. The green dashed ring corresponds to a resolution of 1/22 Å^{-1}. The few diffraction spots visible are well under this limit, indicating that the 2D crystals are poorly ordered. Scale bar: 500 Å

cation of our single particle approach to 2D crystals. Using our newly developed interface in the FOCUS package for picking and exporting 2D crystal patches as single particles, a set of 361,976 particles from 595 micrographs was obtained. The map was then reconstructed and refined with our modified FREALIGN implementation [Grigorieff 2016, Righetto *et al.* 2019] starting from the previously determined tilt geometry and C2 symmetry imposition. While this indeed led to a better map, the kinase domains were still poorly resolved. Surprisingly, an *ab initio* map calculation without symmetry in cryoSPARC [Punjani *et al.* 2017] revealed an even better map where the kinase domains could be identified, which was further improved by imposition of C2 symmetry and refinement of the particle alignments. The evolution of the FAK apo map is shown in Fig. 5.2, with 106,783 particles in the final reconstruction.

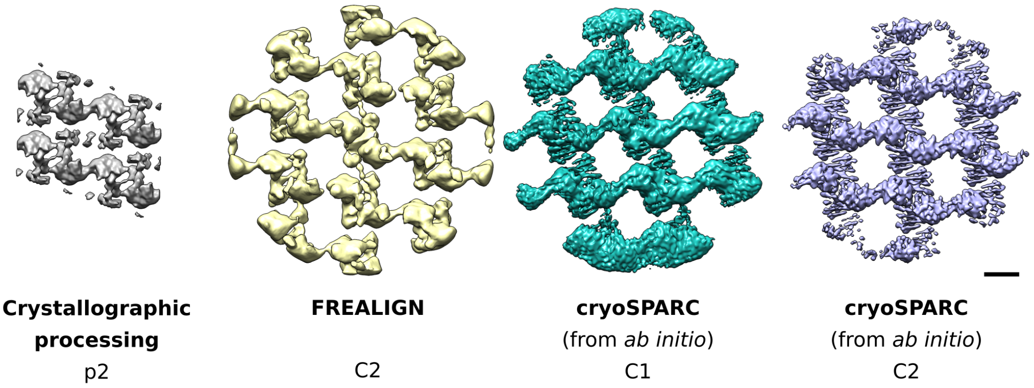


Figure 5.2: **Maps obtained from the FAK apo dataset with different algorithms.** The first map (left to right) was obtained by conventional 2D crystallographic processing imposing p2 symmetry. The second map was refined with FREALIGN imposing C2 symmetry, using the crystallographic tilt geometry as initialization for the alignment parameters. The third and fourth maps were refined in cryoSPARC from intial maps obtained from the SGD-based *ab initio* algorithm, without symmetry and with C2 symmetry imposed, respectively. Scale bar: 50 Å.

Based on the experience with the "apo" dataset, the FAK AMP-PNP data were directly exported into cryoSPARC after initial 2D crystallographic processing and particle picking in FOCUS. From this dataset we were able to extract nearly 2.7 million particles. Interestingly, already at an early stage we could identify the existence of two different types of "crystal packings", or spatial arrangements of the FAK dimers, as shown in Fig. 5.3. After 2D and 3D classification rounds, only one of these yielded a map of acceptably high resolution, with the same type of packing observed in the "apo" dataset (Fig. 5.2). The best reconstruction comprised 548,394 particles.

As the initial tilt geometry for both datasets was not very accurate, we further refined the particle orientations and defocus values in cisTEM [Grant *et al.* 2018]. An issue in resolution estimation arised, however, as cryoSPARC shuffled the order of the particles assigned to the half-sets, hence inflating the Fourier shell correlation (FSC) curves [Harauz & van Heel 1986] (Fig. B.1). For a more accurate (and more conservative) resolution assessment, we calculated new half-maps ensuring that particles from the same 2D crystal were always assigned to the same half-set, using the final set of alignment parameters. After masking, the overall resolution of the FAK dimer at the center of the reconstructed maps was of ∼6 Å for both datasets (Fig. 5.4a). Local resolution estimation indicated the FERM domains as more stable, while the kinase domain, that contains the catalytic site, is likely more flexible (Fig. 5.4b).

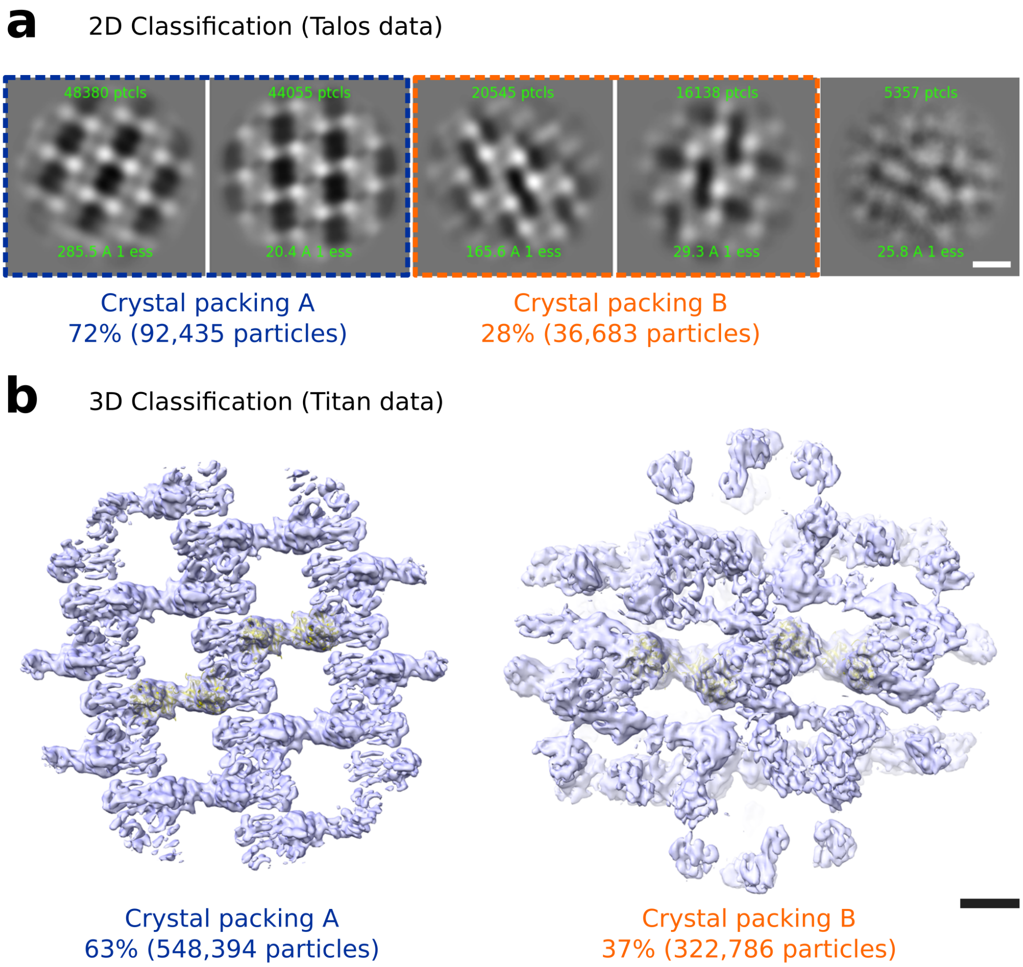


Figure 5.3: **Two types of 2D crystal packings in the FAK-AMPPNP sample.** a) The 2D class averages from cryoSPARC reveal two different types of 2D crystal packing in the FAK-AMPPNP dataset. They were identified already at the stage of sample screening at the Talos TEM. b) The maps obtained with 3D classification in cryoSPARC on the full dataset collected at the Titan confirmed the two types of crystal packing. Scale bars: 50 Å.

At this resolution level, it was possible to observe α-helices and confidently fit the FERM and kinase domains, as shown in Fig. 5.5.

5.3. Discussion

Using 2D electron crystallography combined with single particle analysis, we obtained reconstructions of FAK with and without the AMP-PNP ligand from spontaneously self-assembled 2D crystals on a PIP_2 lipid monolayer. This condition is believed to closely resemble the cellular membrane where focal adhesion sites are natively formed. Despite the 2D crystals being very disordered, not showing diffraction beyond $\sim$25 Å (Fig. 5.1), we were able to obtain maps at $\sim$6 Å resolution (Fig. 5.4a).

Furthermore, the power of 2D and 3D classification in single particle analysis allowed better maps to be resolved, including the detection of two different types of FAK assembling in 2D crystals in the AMP-PNP dataset (Fig. 5.3). It is not clear if and how these two types of crystal packings are functionally related. However, the assembly that led to the highest resolution map in the presence of AMP-PNP was of the same type as in the *apo* state. Interestingly, the FERM-FERM interfaces between FAK dimers also seem to be very flexible according to our local resolution estimations (Fig. 5.4b). Altogether, these are likely reasons why the FAK 2D crystals show high mosaicity and diffract only to low resolution.

With the single particle approach, however, different types of (pseudo-)crystalline arrangements and disorder can be identified and discarded where appropriate. Coupled with the localised refinements, this leads to maps of much higher resolutions than would be expected by conventional crystallographic methods, as demonstrated here for FAK. Furthermore, we have shown that virtually any single-particle software can be used to process 2D crystal data, with three different packages being used in this project (FREALIGN, cryoSPARC, and cisTEM). However, special care has to be taken to ensure that resolution estimates are not exaggerated by the large overlap between adjacent particles, as discussed in Section 5.4.5 (Figs. 5.4 and B.1).

Using flexible fitting methods, we fitted the known structures of the FERM and kinase domains into our maps, revealing some conformational differences between the FAK apo and AMP-PNP-bound states (Fig. 5.5). These models, which are still under interpretation, will provide new insights on mechanism and function, in light of recent studies on FAK regulation by interaction with lipids [Goñi *et al.* 2014, Zhou *et al.* 2015] and mechanical forces [Bauer *et al.* 2019].

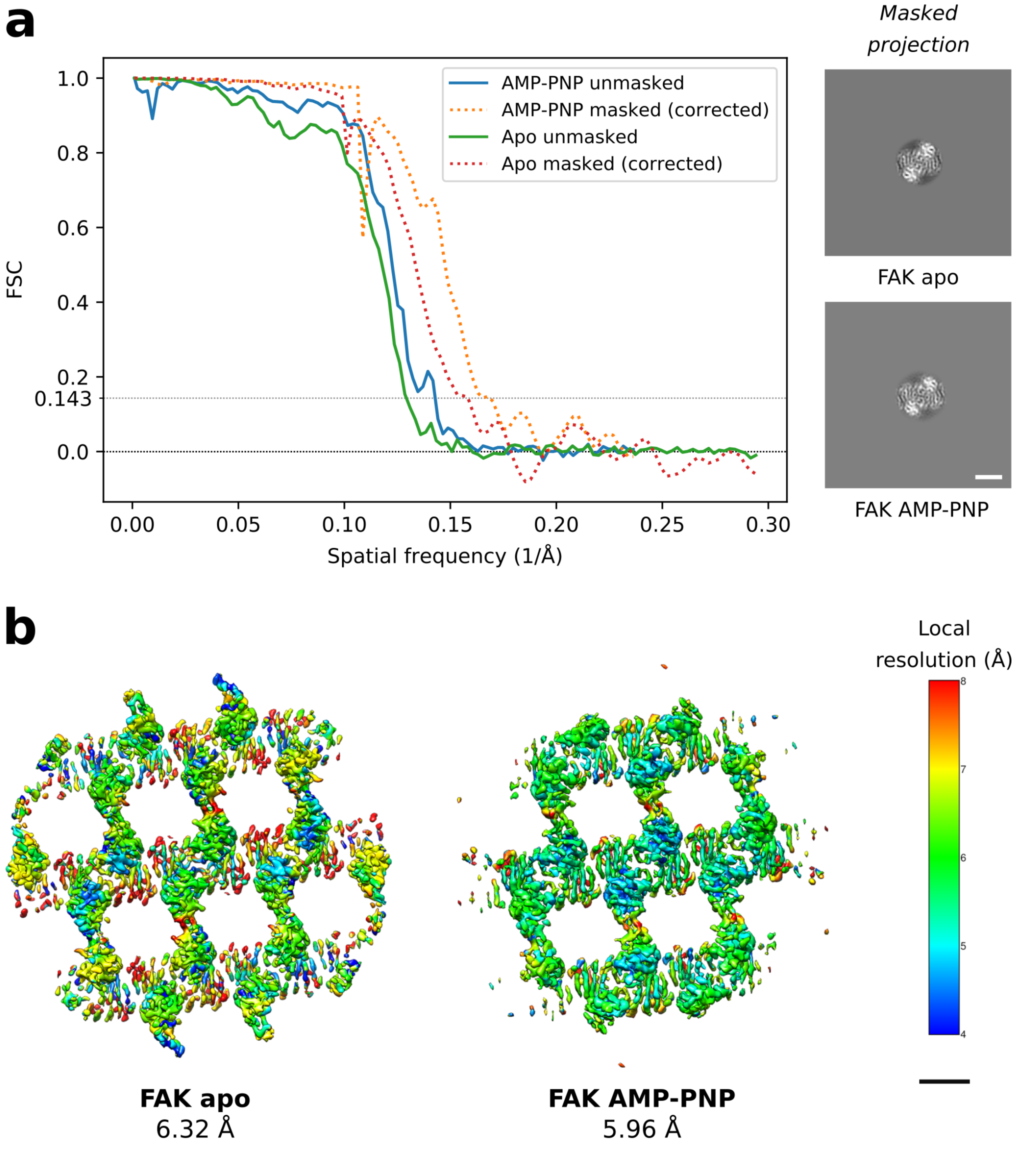


Figure 5.4: **Resolution estimation of the FAK maps.** a) FSC curves for the apo and AMP-PNP maps, with and without masking. The curves were calculated after re-splitting the dataset into random half-sets based on the micrographs instead of the particles (see Suppl. Fig. 1). Masked FSC curves were corrected for artificial correlations introduced by the mask [Chen *et al.* 2013]. The insets show projections of the maps after application of a spherical mask to select the central dimer only. b) Local resolution maps estimated with Blocres [Cardone *et al.* 2013], with the global resolution indicated. Scale bars: 50 Å.

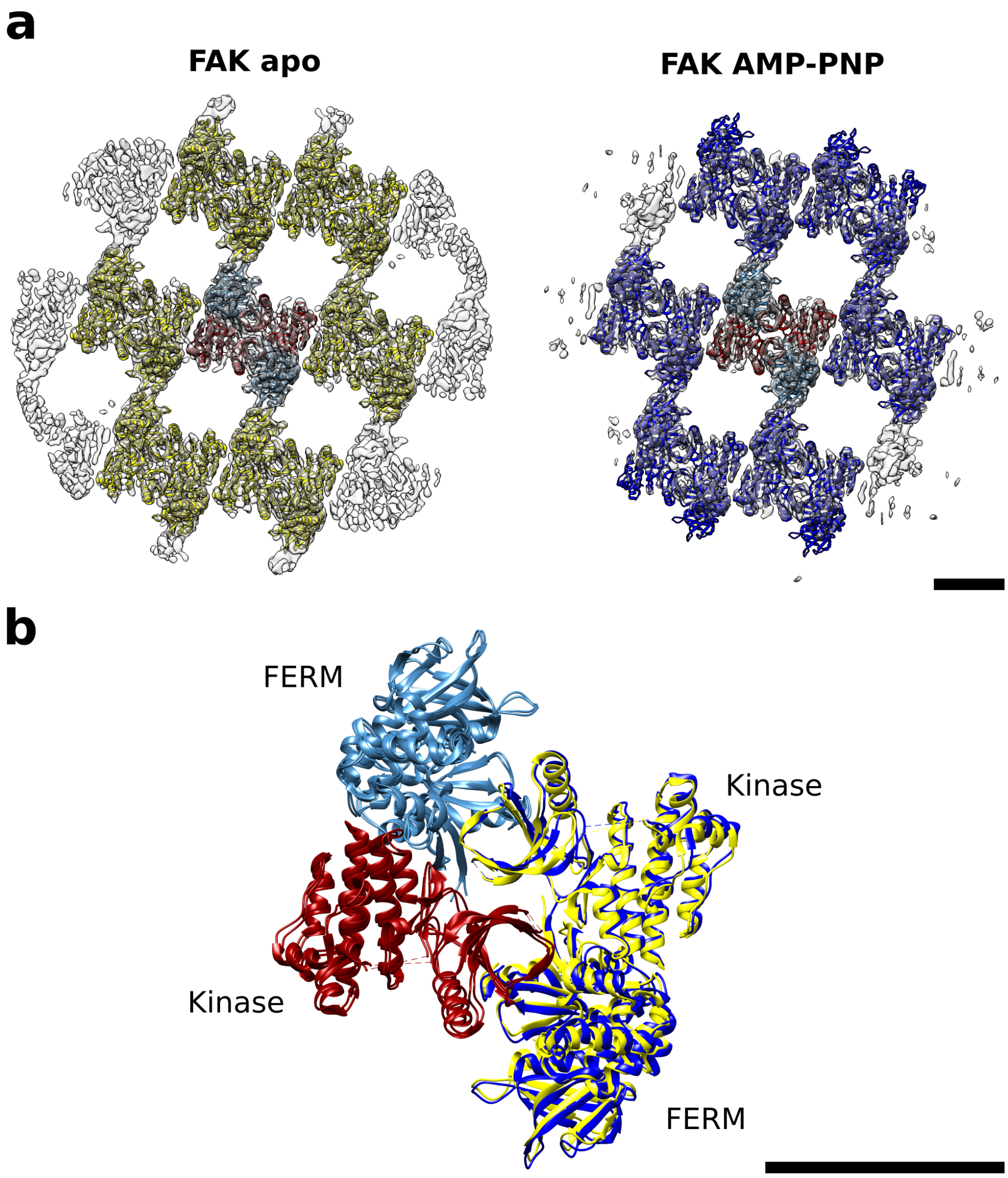


Figure 5.5: **Flexible fitting of FAK models into the maps.** a) Atomic models of FAK were flexibly fit into the apo (yellow ribbons) and AMP-PNP (blue ribbons) maps. The kinase and the FERM domains are shown in dark red and steel blue ribbons, respectively. b) Close-up on the central dimer only, with the models for the apo and AMP-PNP states superposed. One monomer (left) is colored according to the domains (kinase: dark red; FERM: steel blue), while the other monomer (right) is colored according to the ligand binding state (apo: yellow; AMP-PNP: blue). Scale bars: 50 Å.

5.4. Methods

5.4.1. Protein expression and purification

Purified FAK was generated as reported in [Goñi *et al.* 2014]. In brief, chicken FAK31-686 containing an N-terminal 6xHis tag was expressed by transient expression in HEK293GnT1 suspension cultures using polyethylenimine (PEI) as transfection agent. Purification was performed by Nickel-NTA affinity chromatography, followed by anion exchange (Source 15Q column) and size exclusion (Superdex200) chromatography. Purified FAK is concentrated in the storage buffer (20 mM Tris, 200 mM NaCl, 5% Glycerol, 2 mM TCEP] to ~5-10 mg/ml and aliquots are flash frozen in LN_2 and stored at −80 °C.

5.4.2. 2D crystallization and sample preparation

PIP_2 monolayers are prepared by dispensing PIP_2 lipid (brain porcine, Avanti Polar Lipids) dissolved in chloroform/methanol onto a FAK protein solution at 0.05 mg/ml. 2D crystals form spontaneously over ~3 days at 4 °C. Initial crystals were polycrystalline, after optimization, larger single-lattice crystals were grown in 50 mM sodium acetate pH 5.5, 300 mM $LiSO_4$. Lipid monolayers with bound 2D FAK crystals were then picked up with a holey non-charged carbon grid (Lacey Carbon 300 mesh), blotted and frozen using an FEI Vitrobot IV plunge freezer.

5.4.3. Cryo-EM imaging

FAK apo

Grids containing the FAK apo sample were imaged at an FEI Tecnai Polara and an FEI Titan Krios TEMs, both equipped with a Gatan K2 DED. A Gatan Imaging Filter (GIF) with a slit aperture of 20 eV was also employed at the Titan Krios. A total of 1,145 dose-fractionated movies (612 from the Titan and 533 from the Polara) each with an accumulated exposure of 40 $e^-/Å^2$ distributed over 80 movie frames were acquired in counting mode. Pixel size: 1.088 Å (Polara), 1.7 Å or 1.058 Å (Titan), on the sample level. Nominal tilt angles: 0°, 15°, 30°, 45°, 60°.

FAK AMP-PNP

Grids containing the FAK AMP-PNP sample were imaged at the same FEI Titan Krios TEM as above. A total of 2,608 dose-fractionated movies each with an accumulated exposure of 80 $e^-/Å^2$ distributed over 40 movie frames were acquired in counting mode. Pixel size: 1.058 Å on the sample level. Nominal tilt angles: 0°, 10°, 20°, 30°, 40°, 50°.

In both cases, the data collection was automated via SerialEM scripts [Schorb *et al.* 2018], with drift correction and dose weighting performed by MotionCor2 [Zheng *et al.* 2017] and CTF estimation by CTFFIND4 [Rohou & Grigorieff 2015] via FOCUS [Biyani *et al.* 2017] during the microscopy session.

5.4.4. Data processing

FAK apo

The 2D crystals displayed high mosaicity and hence diffracted only to low-resolution (typically <25 Å, as shown in Fig. 5.1). Initial processing was carried out in FOCUS [Biyani *et al.* 2017] following conventional 2D electron crystallographic procedures: determination of defocus, tilt geometry, lattice, unbending, and merging [Gipson *et al.* 2007, Schenk *et al.* 2010]. These steps resulted in an initial, low-resolution 3D reconstruction with p2 symmetry imposed, where the FERM domain could be identified but not the kinase domain (Fig. 5.2). A total of 595 micrographs were deemed good enough to be included in the reconstruction.

For this reason, we decided to try the single particle approach to 2D crystals as described in Chapters 3 and 4. We used the FOCUS module [Righetto *et al.* 2019] to pick and export particles extracted from the FAK 2D crystals. Given the three different magnifications present in the dataset (1.7 Å and 1.058 Å from the Titan, and 1.088 Å from the Polara), the particles were downsampled to a pixel size of 1.7 Å when needed. In total, 361,976 particles were extracted at a box size of 238 × 248 pixels. Starting from the converted crystallographic tilt geometry, the map was then refined with the modified version of FREALIGN [Grigorieff 2016, Righetto *et al.* 2019] with C2 symmetry imposed. While the map did improve visibly, resolution was still not good enough to confidently place the kinase domain (Fig. 5.2). We then decided to try cryoSPARC and its stochastic gradient descent (SGD) *ab initio* algorithm on this FAK dataset. After 2D classification for cleaning up the set of particles, a much better initial map was obtained and then refined to comparably higher resolutions, initially in C1 and afterwards with C2 symmetry imposed, as shown in Fig. 5.2. Finally, the C2-symmetric map was exported into cisTEM [Grant *et al.* 2018] for further rounds of alignment and defocus refinement. The final reconstruction comprised 106,783 particles.

FAK AMP-PNP

As before, the 2D crystals displayed high mosaicity, with only low resolution diffraction spots. Initial processing was carried out in FOCUS [Biyani *et al.* 2017] following the standard 2D electron crystallographic procedures as above: determination of defocus, tilt geometry, lattice, unbending, and merging [Gipson *et al.* 2007, Schenk *et al.* 2010]. From the 2,608 movies acquired, 850 were deemed good enough to be included in the reconstruction.

We then used the FOCUS module [Righetto *et al.* 2019] to pick and export particles extracted from the FAK AMP-PNP 2D crystals. In total, 2,666,154 particles were extracted at a pixel size of 2.116 Å and a box size of 200 × 200 pixels. Downsampling was applied for speedup reasons. With this dataset, we went directly for single particle processing with cryoSPARC. After 2D classification rounds, the multi-class *ab initio* map estimation confirmed the existence of two distinct types of 2D crystal packings in this dataset, which had been already observed during the sample screening stage (Fig. 5.3). The best 3D classes representing each crystal packing contained 548,394 (class 1) and 322,786 particles (class 2), respectively. The best 3D class (class 1) was then refined further in cryoSPARC and subsequently in cisTEM [Grant *et al.* 2018], including defocus refinement.

5.4.5. Postprocessing and resolution estimation

During refinement in cryoSPARC and cisTEM it was observed that the FSC curves never dropped below zero (Fig. B.1). This happened because cryoSPARC shuffled the order of the particles for its half-set assignment, violating the crystal-based split performed by the FOCUS as explained in Sections 3.4.4 and 4.3.3. Therefore, in order to obtain more reliable resolution estimates, the half-sets were split again based on the 2D crystals (i.e. particles from the same 2D crystal were always assigned to the same half-set) and reconstructed using the `relion_reconstruct` program [Scheres 2012b].

The FSC curves [Harauz & van Heel 1986] were then calculated using `focus.postprocess` [Righetto *et al.* 2019]. For assessing the resolution at the central dimer only, a spherical mask with a soft edge was applied, and the FSC curve was corrected for the correlations introduced by the mask [Chen *et al.* 2013] (Fig. 5.4a). The 0.143 threshold was used for resolution estimation and the maps were sharpened by deconvolving the MTF of the detector and reversing the contrast loss by a negative B-factor

[Rosenthal & Henderson 2003]. Finally, local resolution was estimated using Bsoft [Cardone *et al.* 2013] (Fig. 5.4b).

5.4.6. Modelling

A model for the FAK monomer was created in UCSF Chimera [Pettersen *et al.* 2004] by taking the FERM domain from PDB 2AEH [Ceccarelli *et al.* 2006] and the kinase domain from PDB 1MP8 [Nowakowski *et al.* 2002]. Flexible fitting into the FAK apo and FAK AMP-PNP maps was performed using normal mode analysis by iMODFIT [Lopéz-Blanco & Chacón 2013], followed by real space refinement in PHENIX [Afonine *et al.* 2018b]. Copies of the monomer were gradually inserted, fitted and refined into the 2D crystal supercell to optimize the geometry at the interfaces.

Acknowledgements

Calculations were performed at the sciCORE cluster of the University of Basel.

6. Conclusions

The adoption of direct electron detectors (DED) has drastically changed the landscape of structural biology [Kühlbrandt 2014]. The higher SNR [McMullan *et al.* 2016] and the ability to correct for beam-induced motion [Brilot *et al.* 2012] offered by these detectors have generally improved the resolutions obtainable by cryo-EM techniques, even to near-atomic resolution in some cases [Zivanov *et al.* 2018]. Combined with tools for automation of the data collection [Schorb *et al.* 2018] and data processing pipelines [Biyani *et al.* 2017, Zivanov *et al.* 2018], the amount of data generated for cryo-EM experiments has also grown rapidly. Aiming at addressing this challenge, in Chapter 2 we introduced MRCZ, an extension to the widely adopted MRC format, that natively supports modern data compression codecs that take advantage of CPU multi-threading and caching capabilities. In particular, we found the *zStandard* codec to generally offer the best tradeoff between compression *ratio* and compression *rate*, although other codec options are available to satisfy specific needs. As *blosc* is an open-source *meta*-compression library, it is actually possible to include any compression codec desired, which would be readily available for MRCZ.

We have also developed a software package, *FOCUS*, that offers a user-friendly graphical front-end to several other software packages employed in cryo-EM [Biyani *et al.* 2017]. By pipelining common processing steps such as drift-correction, CTF estimation and particle picking, FOCUS enables real-time data processing, and the ability to easily prune the dataset according to pre-defined parameters related to data quality. This makes cryo-EM data collection more efficient and less vulnerable to problems that would waste a multi-day microscopy session, as problems can be readily diagnosed and fixed as needed. Since its release, we have been maintaining FOCUS and including new functionality, such as new features to support high-throughput data acquisition using beam-image shift [Cheng *et al.* 2018, Zivanov *et al.* 2018].

Concerning 2D electron crystallography of membrane proteins, we have also developed improved algorithms that, for example, perform more accurate *unbending* based on DED movie frames and estimate the tilt geometry with higher local accuracy by cutting the 2D crystal image into tiles [Biyani *et al.* 2018]. While such advances indeed lead to better reconstructions by 2D crystallography [Kowal *et al.* 2018], they still lag behind the resolutions offered by the single particle analysis (SPA) technique since the recent adoption of new sample preparation methods [Gao *et al.* 2016, Lee & MacKinnon 2017] and improvements in data processing algorithms [Scheres 2012a, Grigorieff 2016, Punjani *et al.* 2017, Grant *et al.* 2018]. In Chapters 3 and 4 we demonstrated that these SPA software packages can also be used to process 2D crystal data, enabled by a new module in FOCUS. In this way, local distortions can be corrected by aligning small patches of the 2D crystals to a reference in 3D. Although conceptually simple, the idea of using SPA algorithms on “particles” extracted from 2D crystals is not straightforward, as special care has to be taken to avoid exaggerated resolution estimations, and the inherent local correlations across the 2D crystal can be exploited by choosing a suitable box size and imposing restraints on the particle alignments. We modified the FREALIGN package [Grigorieff 2016], written in the FORTRAN programming language, to include such restraints and an automated refinement algorithm, reducing the need for user intervention and subjective interpretations [Righetto *et al.* 2019]. Also, we developed a postprocessing tool written in Python, `focus.postprocess`, that conveniently offers a number of map filtration, masking and other operations from a single command-line interface and is not specific to any SPA software package. The development of these scientific computing tools, which are written in a variety of programming languages, required the adoption of modern software engineering practices such as version control, continuous integration and code profiling.

As reported in Chapter 3, using the hybrid single particle 2D electron crystallography approach, we obtained a “consensus” map of the MloK1 potassium at ∼4 Å resolution from 2D crystals that would not show diffraction beyond ∼10 Å in the best cases. Even more interesting was the confirmation that the MloK1 2D crystals contained heterogeneous conformations of the CNBD, in agreement with observations from high-speed AFM experiments [Rangl *et al.* 2016], a movement that is linked to the channel function [Kowal *et al.* 2018]. For the first time, distinct conformations of a protein were resolved in 3D from 2D crystals, thanks to the ability of SPA algorithms to disentangle heterogeneous datasets into more homogeneous classes of particles [Righetto *et al.* 2019]. This finding challenges the very notion of regarding proteins periodically arranged in a lipid bilayer as “crystals”, which are normally thought to be a perfectly ordered arrangement of identical objects in a lattice. In retrospect, it also provides an explanation to why the majority of 2D crystals would not show diffraction at high resolution: they are likely heterogeneous, and such structural variability induces lattice distortions.

It is true, however, that the single particle approach to 2D crystals here devised is still more complicated and less likely to reach high resolutions ≤3.5 Å than conventional SPA with current methods and algorithms. Nevertheless, the question of whether 2D crystals or lipid nanodiscs can better emulate the cellular membrane environment and thus a more physiological condition remains. The answer will likely depend on the protein being studied and its relationship to specific phospholipids and to other membrane proteins. For example, some potassium channels are thought to form clusters in the membrane [Duncan *et al.* 2017], while other proteins can even naturally assemble into 2D crystals [Henderson & Unwin 1975].

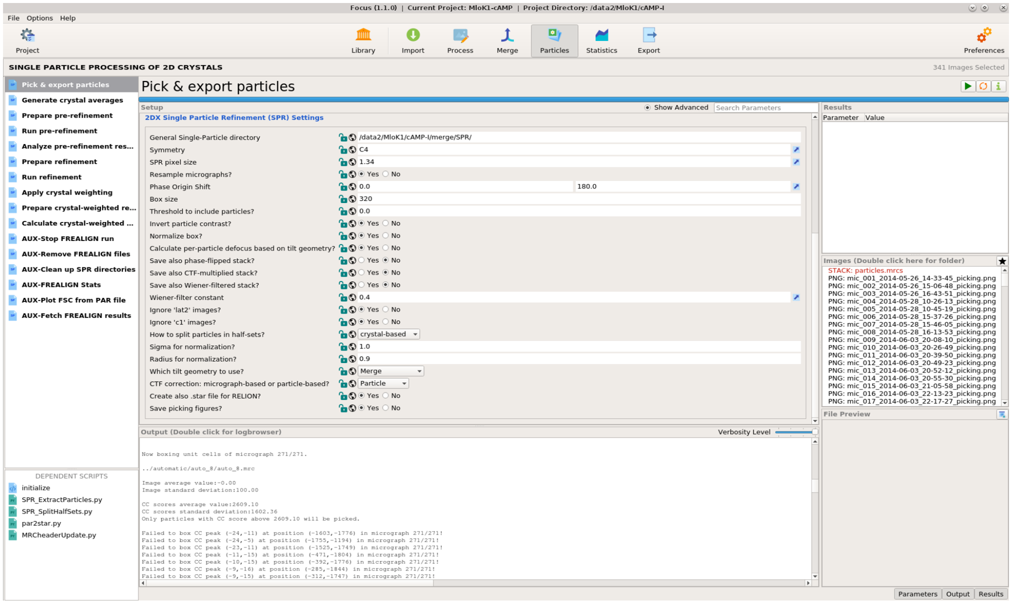
We applied our method on one such case where the protein tends to spontaneously assemble into 2D crystals, *focal adhesion kinase* (FAK), a protein involved in mechanical sensing and cell migration processes. As shown in Chapter 5, we could obtain maps at ~6 Å resolution from highly disordered 2D crystals that did not show any diffraction beyond ~25 Å, and revealed two different types of crystal packings using single particle 2D and 3D classification. It is for cases like this that we believe our method is a suitable tool that allows structure determination in a functionally relevant condition. Furthermore, it can also be valuable for studying 2D arrays of proteins that are artificially designed [Gonen *et al.* 2015, Suzuki *et al.* 2016].

The adoption of phase plates to increase contrast in cryo-EM is already enabling smaller structures to be resolved [Khoshouei *et al.* 2017]. In combination with better detectors and lower-energy machines [Peet *et al.* 2019], the main instrumental limitations are likely to be addressed in the near-future. From the sample preparation perspective, microfluidic protein isolation from nanoliters of cell lysate, a problem on which we have also worked [Schmidli *et al.* 2019], will greatly enhance efficiency and allow the proteome of specific cells to be studied at high resolution. In addition to 2D crystals and lipid nanodiscs, membrane protein structures can now also be solved when embedded in liposomes, which allows the application of an electrical potential difference between each side of the lipid bilayer [Tonggu & Wang 2018]. On the computational side, more automated and robust methods for model building [Terwilliger *et al.* 2018a] and validation [Afonine *et al.* 2018a] are being developed. It is especially interesting the recent introduction of *deep learning* and other modern machine learning methods in the field, in applications such as *ab initio* map estimation [Punjani *et al.* 2017], automatic tomogram segmentation and annotation [Chen *et al.* 2017], particle picking [Tegunov & Cramer 2018] and protein dynamics estimation from heterogeneous datasets [Dashti *et al.* 2019]. In this scenario, not only can we expect a higher throughput in cryo-EM structure determination, but, more importantly, the integration of data from different and complementary experimental sources will allow a better understanding of the complex cell machinery [Russel *et al.* 2012].

Appendices

A. Supplementary Information for "Retrieving High-Resolution Information from Disordered 2D Crystals by Single Particle Cryo-EM"

A.1. Supplementary Figures



2DX Single Particle Refinement (SPR) Settings

General Single-Particle directory: /data2/MloK1/cAMP-I/merge/SPR/
Symmetry: C4
SPR pixel size: 1.34
Resample micrographs? ● Yes ○ No
Phase Origin Shift: 0.0 180.0
Box size: 320
Threshold to include particles? 0.0
Invert particle contrast? ● Yes ○ No
Normalize box? ● Yes ○ No
Calculate per-particle defocus based on tilt geometry? ● Yes ○ No
Save also phase-flipped stack? ○ Yes ● No
Save also CTF-multiplied stack? ○ Yes ● No
Save also Wiener-filtered stack? ○ Yes ● No
Wiener-filter constant: 0.4
Ignore 'lat2' images? ● Yes ○ No
Ignore 'c1' images? ● Yes ○ No
How to split particles in half-sets? crystal-based
Sigma for normalization? 1.0
Radius for normalization? 0.9
Which tilt geometry to use? Merge
CTF correction: micrograph-based or particle-based? Particle
Create also .star file for RELION? ● Yes ○ No
Save picking figures? ● Yes ○ No

```
Now boxing unit cells of micrograph 271/271.

../automatic/auto_8/auto_8.mrc

Image average value:-0.00
Image standard deviation:100.00

CC scores average value:2609.10
CC scores standard deviation:1602.36
Only particles with CC score above 2609.10 will be picked.

Failed to box CC peak (-24,-11) at position (-1603,-1776) in micrograph 271/271!
Failed to box CC peak (-24,-5) at position (-1755,-1194) in micrograph 271/271!
Failed to box CC peak (-23,-11) at position (-1525,-1749) in micrograph 271/271!
Failed to box CC peak (-11,-15) at position (-471,-1804) in micrograph 271/271!
Failed to box CC peak (-10,-15) at position (-392,-1776) in micrograph 271/271!
Failed to box CC peak (-9,-16) at position (-285,-1844) in micrograph 271/271!
Failed to box CC peak (-9,-15) at position (-312,-1747) in micrograph 271/271!
```

STACK: particles.mrcs
PNG: mic_001_2014-05-26_14-33-45_picking.png
PNG: mic_002_2014-05-26_15-06-48_picking.png
PNG: mic_003_2014-05-26_16-43-51_picking.png
PNG: mic_004_2014-05-28_10-26-13_picking.png
PNG: mic_005_2014-05-28_10-45-19_picking.png
PNG: mic_006_2014-05-28_15-37-26_picking.png
PNG: mic_007_2014-05-28_15-46-05_picking.png
PNG: mic_008_2014-05-28_16-13-53_picking.png
PNG: mic_009_2014-06-03_20-08-10_picking.png
PNG: mic_010_2014-06-03_20-26-49_picking.png
PNG: mic_011_2014-06-03_20-39-50_picking.png
PNG: mic_012_2014-06-03_20-49-23_picking.png
PNG: mic_013_2014-06-03_20-52-12_picking.png
PNG: mic_014_2014-06-03_20-55-30_picking.png
PNG: mic_015_2014-06-03_21-05-58_picking.png
PNG: mic_016_2014-06-03_22-13-23_picking.png
PNG: mic_017_2014-06-03_22-17-27_picking.png

Figure A.1: **New graphical user interface in FOCUS created to export a 2D electron crystallography dataset for single particle analysis.** The full pipeline for refinement of a single map based on FREALIGN can be executed following the sequence of scripts shown in the panel on the left. Alternatively, all scripts can be called from the command line.

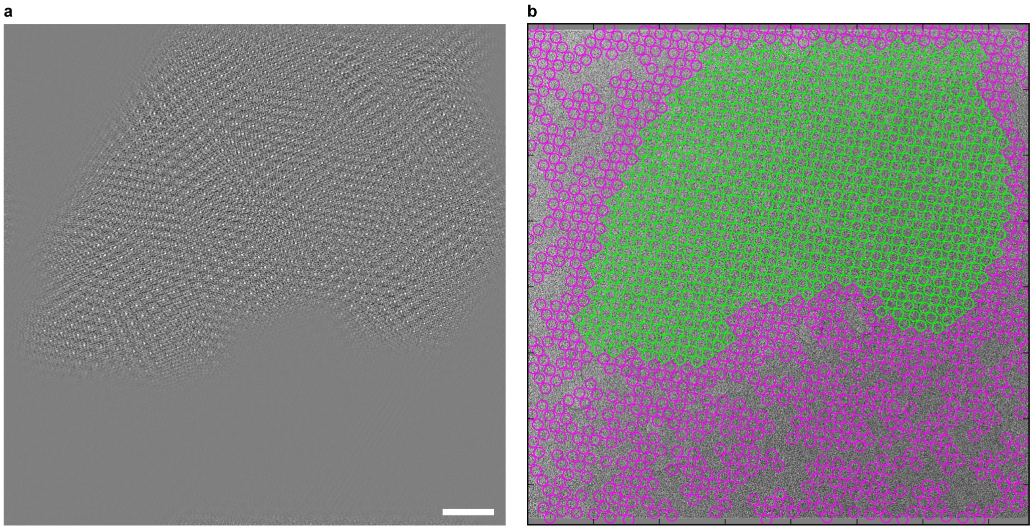


Figure A.2: **Particle picking from 2D crystals.** The center of each particle (a windowed patch of the 2D crystal) corresponds to the center of each unit cell determined by the classical unbending algorithm [Crowther *et al.* 1970, Henderson *et al.* 1986], optionally with an additional phase shift applied to translate the center of a protein to the center of the window. a) The cross-correlation (CC) profile of a 2D crystal of Mlok1 obtained by the classical unbending procedure5. The CC peaks indicate putative positions of unit cells. b) After application of a threshold to the values of the CC peaks to facilitate particle selection: green, particles picked (above threshold); magenta, particles ignored because they are probably bad or false positives (below threshold). Scale bar: 500 Å.

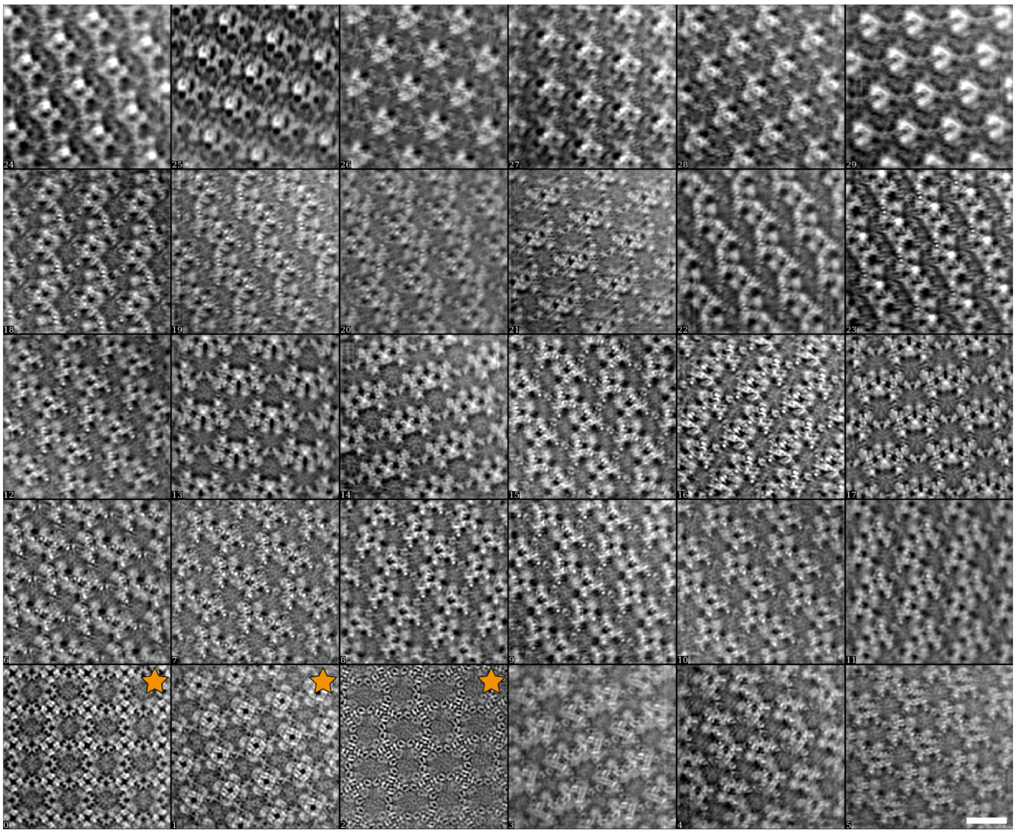


Figure A.3: **Randomly selected crystal averages from the MloK1 dataset.** The particles contain one MloK1 tetramer in the center and at least eight neighboring complete tetramers, depending on the crystal tilt angle. Non-tilted images (tilt angle < 5°) are marked with a star for ease of visualization. Scale bar: 100 Å.

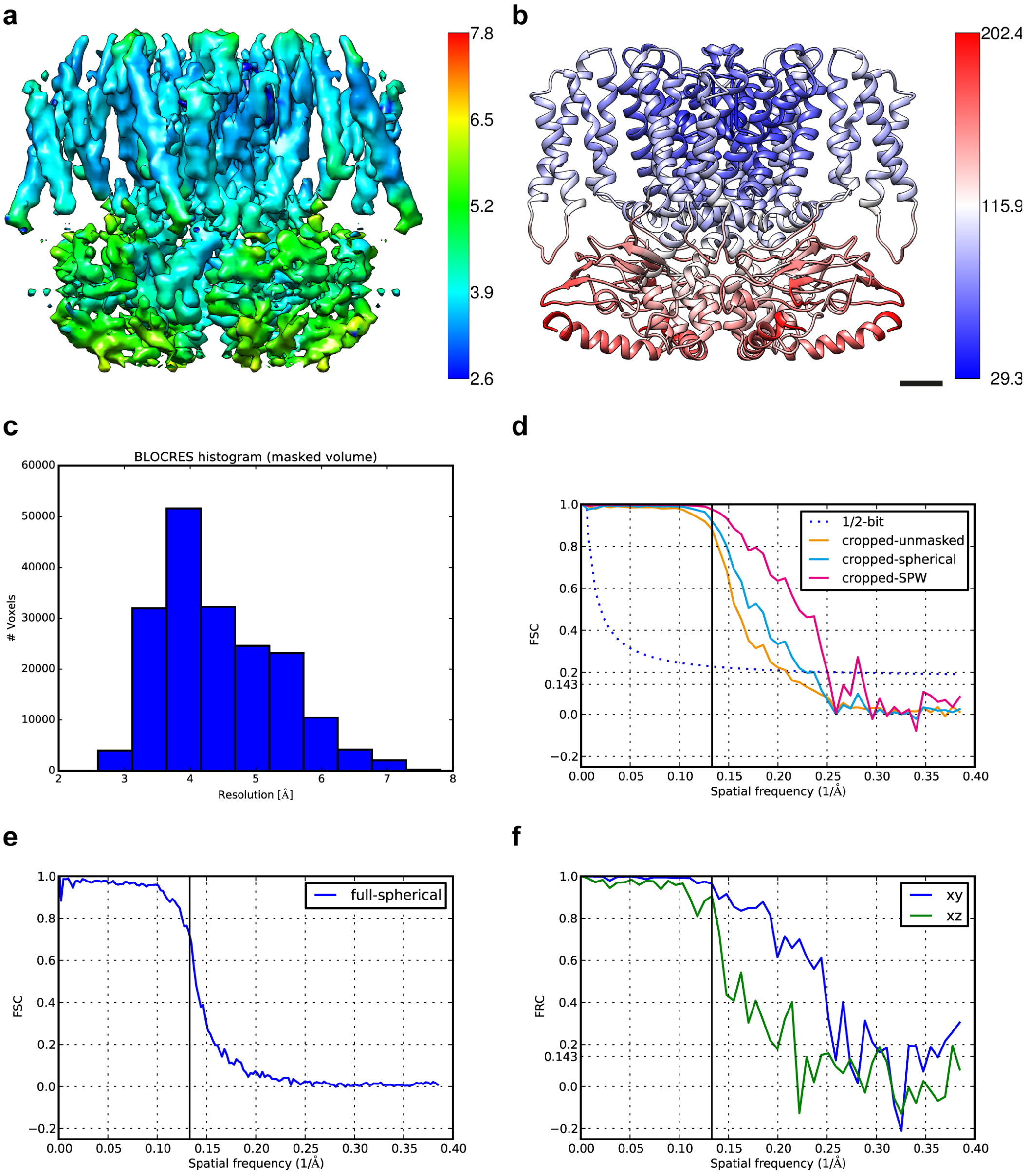


Figure A.4: **Resolution of the consensus map.** *(continued on next page)*

a) Local resolution map according to Blocres using a kernel of 20 cubic voxels and an FSC threshold of 0.143 with colorbar units in Å. b) Average B-factor per residue of the refined MloK1 atomic model with colorbar units in Å^2. c) Histogram of voxel-assigned resolutions according to Blocres as in a). d) FSC of the central area cropped for postprocessing and analysis of the central MloK1 tetramer, with a box size of 104 cubic voxels. The solid lines correspond respectively to: orange, FSC between the unmasked cropped half-volumes; cyan, FSC between the cropped half-volumes after applying a soft-edged spherical mask of radius 54.6 Å; magenta, using the same soft-edged spherical mask, but after adjusting the masked FSC for the relative volumes of the molecule (MW=160 kDa) and the mask according to the Single Particle Wiener (SPW) Filter [Sindelar & Grigorieff 2012] (Fpart/Fmask = 0.288). For resolution assessment, both the 0.143 cutoff and the 1/2-bit criterion curve [van Heel & Schatz 2005] (blue dotted line) are shown (calculated for 4-fold symmetric particle with longest dimension of 100 Å). e) FSC between full half-maps at the end of consensus refinement, corresponding to a box size of 320 cubic voxels (i.e., before cropping) and masked by a soft-edged sphere with radius of 192.96 Å. f) FRC between the unmasked, cropped half-volumes as in d) along orthogonal central slices only: blue, along the xy-plane; green, along the xz-plane (identical to the FRC along the yz-plane due to imposition of C4 symmetry, not shown). The solid vertical black line shown in panels d), e) and f) indicates the highest resolution limit used to align the particles in FREALIGN: 7.52 Å.

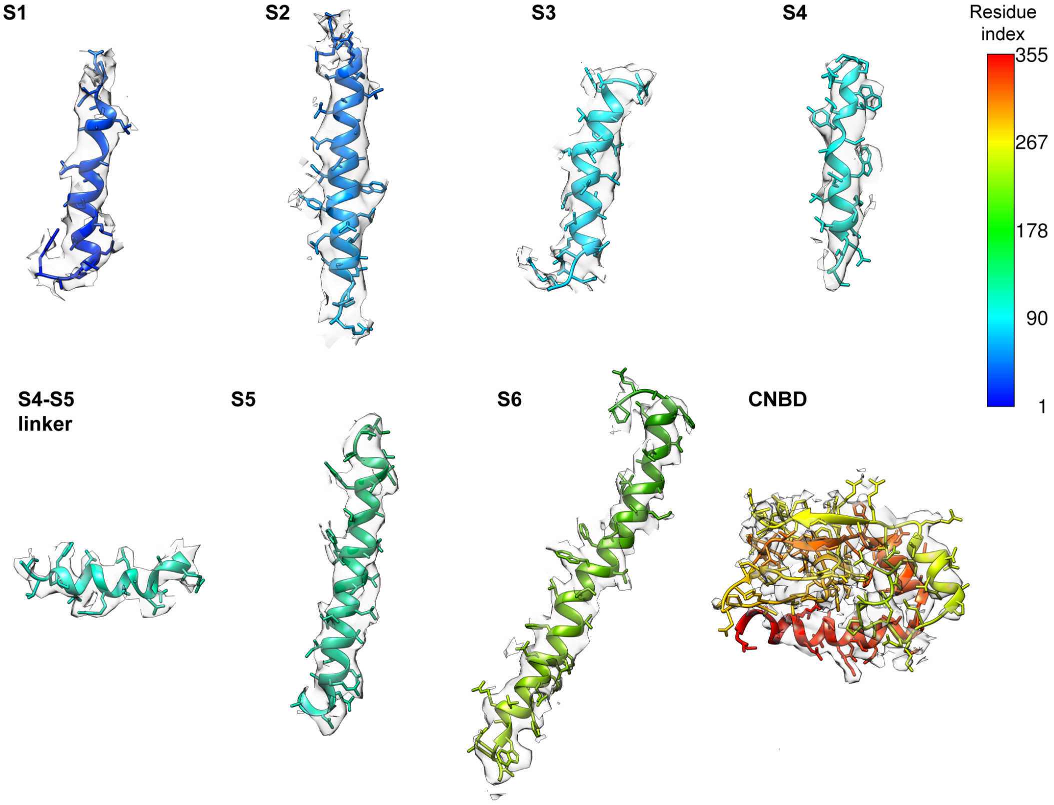


Figure A.5: **MloK1 model fit to map.** Selected fragments along one chain of the refined MloK1 atomic model shown inside the consensus electron density map. The map was globally sharpened using the *phenix.auto_sharpen* program [Terwilliger *et al.* 2018b]. A zone of 2.5 Å around the model was used to mask the map in UCSF Chimera [Pettersen *et al.* 2004]. Color bar indicates the residue index from the N- to the C- terminal.

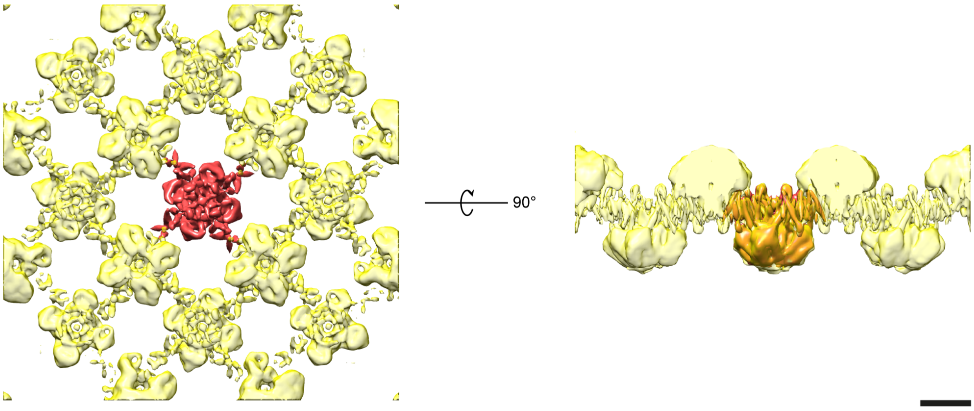


Figure A.6: **Signal subtraction.** The central MloK1 tetramer, shown in red, was masked using a soft spherical mask. This mask was inverted in order to mask all the neighboring tetramers, shown in transparent yellow, from a fully unmasked reconstruction of the consensus map. This masked map without the central tetramer was subtracted from all experimental particle images using the alignments determined in the consensus refinement, prior to the 3D classification. For this operation, the map and the particles were coarsened by a factor of 2 (i.e., a pixel size of 2.6 Å) by Fourier cropping. Scale bar: 50 Å.

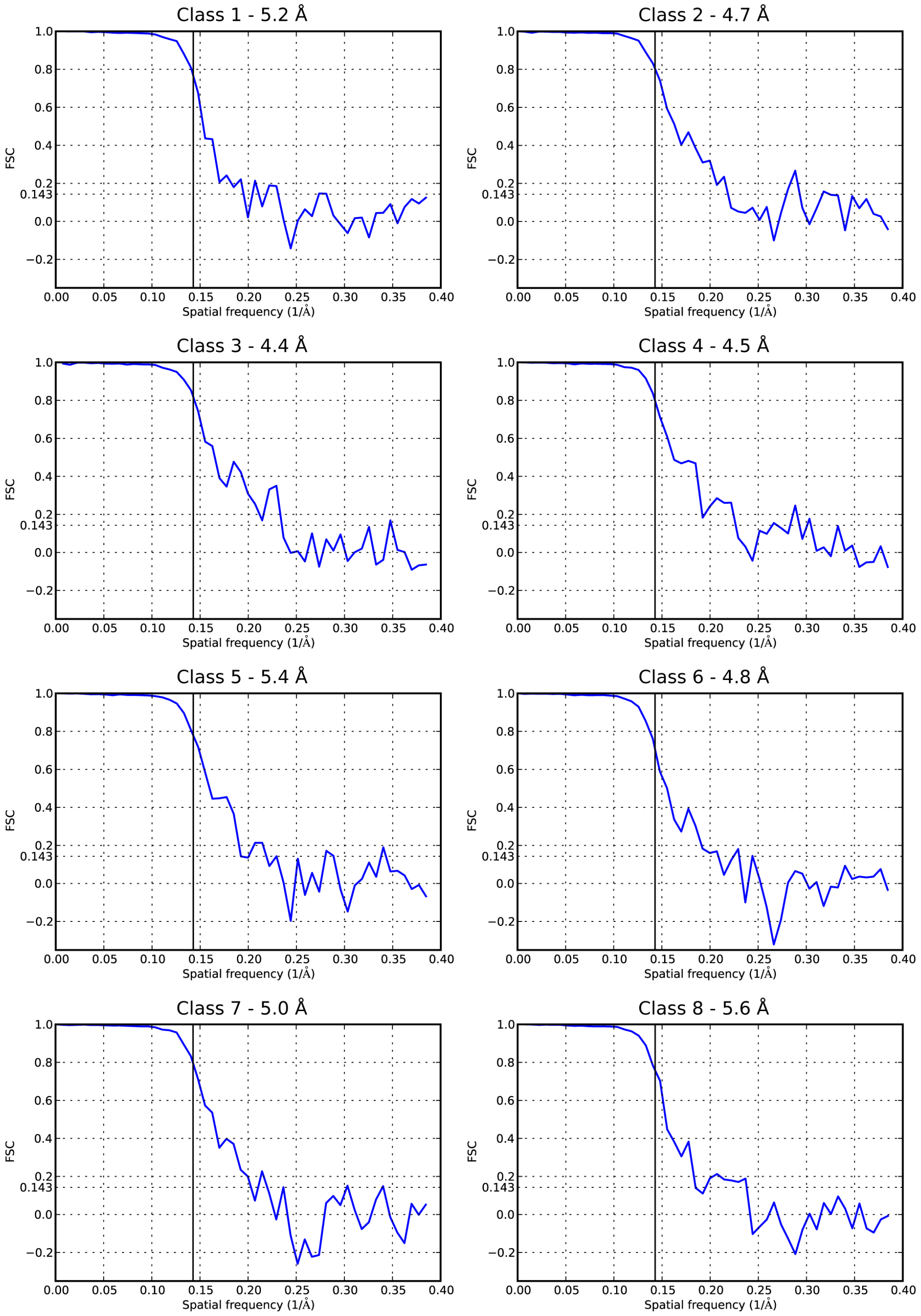


Figure A.7: **Resolution of the 3D classes.** *(continued on next page)*

FSC curves of the central area cropped for postprocessing and analysis of the central MloK1 tetramer, with a box size of 104 cubic voxels, as performed for the consensus map (Supplementary Figure A.4). Likewise, all FSC curves were calculated after masking the half-volumes with a soft-edged spherical mask with a radius of 54.6 Å and adjusted for the relative volumes of the molecules (MW=160 kDa) and the mask according to the Single Particle Wiener filter [Sindelar & Grigorieff 2012]. The resolution stated next to the class number on the title of each plot corresponds to the 0.143 FSC threshold [Rosenthal & Henderson 2003]. The solid vertical black line indicates the resolution limit used to classify the particles in FREALIGN: 7.0 Å. No particle alignment was performed at the stage of 3D classification.

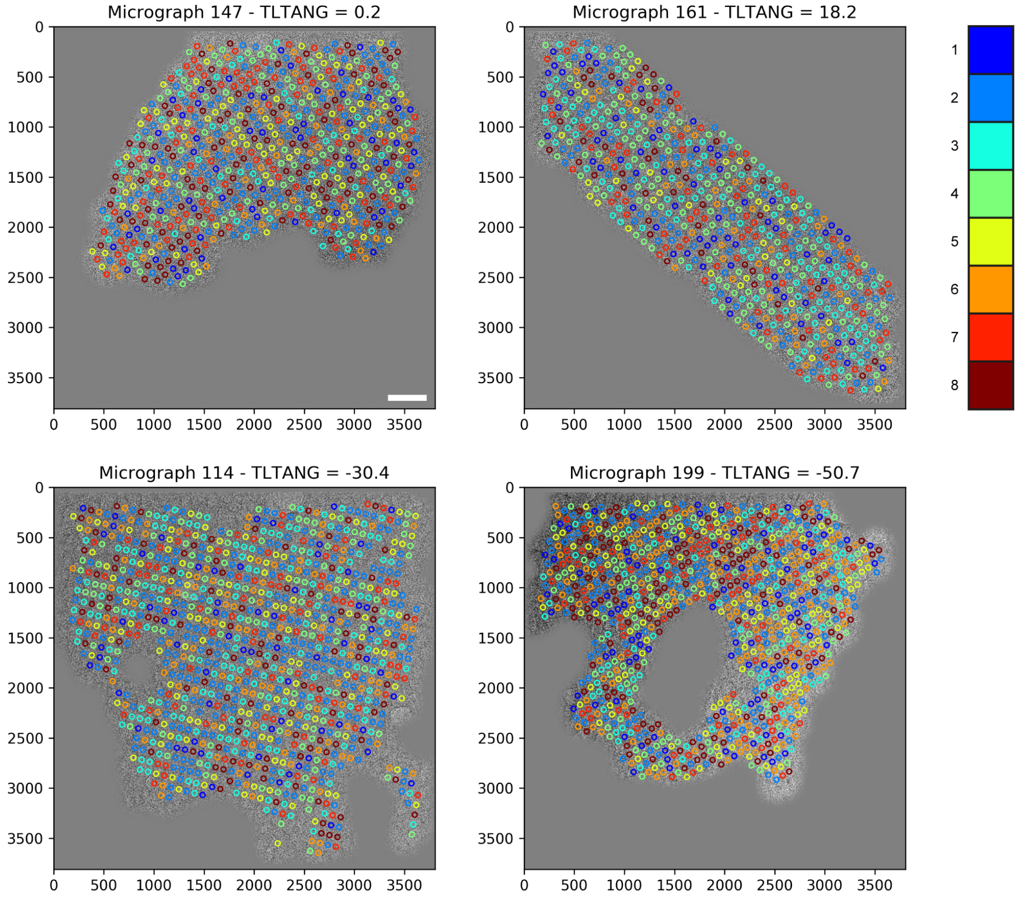

Micrograph 147 - TLTANG = 0.2
Micrograph 161 - TLTANG = 18.2
Micrograph 114 - TLTANG = -30.4
Micrograph 199 - TLTANG = -50.7
0
500
1000
1500
2000
2500
3000
3500
1
2
3
4
5
6
7
8

Figure A.8: **3D classification results and particle positions.** The locations where particles were picked from the 2D crystals are shown on four representative micrographs of the MloK1 dataset. Each circle around the center of the particle is color coded according to the class with the highest occupancy (maximum likelihood) for that particle. Scale bar is 500 Å.

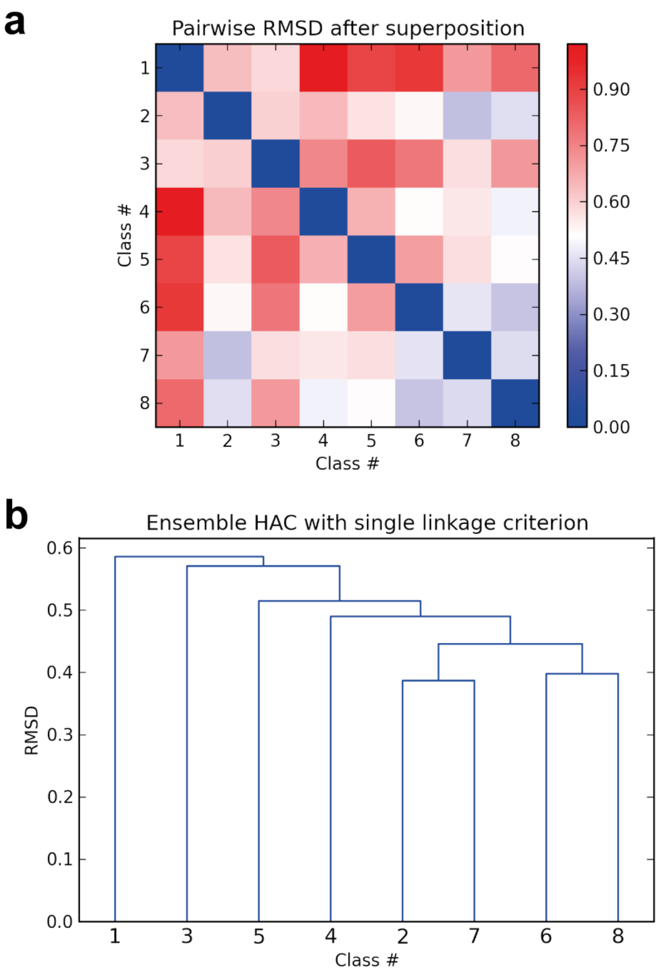


Figure A.9: **Pairwise similarity within ensemble of atomic models.** a) The pairwise RMSD between the eight atomic models derived from the 3D classes was calculated after superposition of the C-α atoms and plotted as a symmetrical matrix for ease of comparison. b) Based on the RMSD matrix from a), a hierarchical agglomerative clustering (HAC) was calculated using the single linkage criterion8 and the dendrogram was plotted to indicate similarity relationships within the ensemble. Model #1 is the most different from all other models in the ensemble, followed by model #3; conversely, models #2 and #7 are the most similar to each other, followed by models #6 and #8. In both panels the RMSD values are given in Angstroms. See also Supplementary Table A.2.

A.2. Supplementary Tables

Table A.1: **Summary of MloK1 single-particle refinement from 2D crystals.** Values in parentheses refer to the previously published structure [Kowal *et al.* 2018].

Symmetry applied	C4 (p42$_1$2)
Global resolution	4.0 Å (4.5 Å)
Number of micrographs	270 (346)
Number of particles	231,688
Pixel size	1.3 Å
Box size	320
Clashscore	6.30
Ramachandran outliers	0.00 %
Ramachandran favored	88.31 %
Rotamer outliers	0
Molprobity score [Davis *et al.* 2007]	1.94
EMRinger score [Barad *et al.* 2015]	0.810
CCmask [Afonine *et al.* 2018b]	0.677

Table A.2: **Pairwise RMSD values among members of the model ensemble.** Values computed between all C-α atoms after global superposition, sorted in descending order. See also Supplementary Figure 9.

Model #	**Model #**	**RMSD [Å]**
1	4	1.021
1	6	0.927
1	5	0.892
3	5	0.845
1	8	0.808
3	6	0.781
3	4	0.742
3	8	0.710
1	7	0.708
5	6	0.701
4	5	0.661
2	4	0.643
1	2	0.638
2	3	0.601
1	3	0.586
3	7	0.571
5	7	0.571
2	5	0.563
4	7	0.558
2	6	0.525
4	6	0.518
5	8	0.515
4	8	0.490
6	7	0.461
2	8	0.448
7	8	0.446
6	8	0.398
2	7	0.387

A.3. Supplementary Movies

Supplementary Movie 1. Blinking of the CNBD. A simple morph from model $4 ("compact" conformation) to model $1 ("extended" conformation) in the ensemble derived from the 3D classes. Each chain is shown with a different ribbon color. Potassium ions are colored purple, and the side chains in the selectivity filter are explicitly shown (residues 175-178). a) Side view; b) CNBD view; c) pore view. Movie generated in UCSF Chimera [Pettersen *et al.* 2004].

Supplementary Movie 2. Blinking of the CNBD and tilting of the VSD. A simple morph from model $6 ("compact" conformation) to model $1 ("extended" conformation) in the ensemble derived from the 3D classes. Each chain is shown with a different ribbon color. Potassium ions are colored purple, and the side chains in the selectivity filter are explicitly shown (residues 175-178). a) Side view; b) CNBD view; c) pore view. Movie generated in UCSF Chimera [Pettersen *et al.* 2004].

Supplementary Movie 3. Rotation of the CNBD and the selectivity filter with respect to the TMD. A simple morph from model $5 (intermediate "compact" conformation) to model $3 (intermediate "extended" conformation) in the ensemble derived from the 3D classes. Each chain is shown with a different ribbon color. Potassium ions are colored purple, and the side chains in the selectivity filter are explicitly shown (residues 175-178). a) Side view; b) CNBD view; c) pore view. Movie generated in UCSF Chimera [Pettersen *et al.* 2004].

A.4. Supplementary Note 1

Alignment restraints in FREALIGN

When processing 2D crystal data with FREALIGN v9.11, it might be necessary to restrain the changes in Euler angles and x, y shifts. We therefore introduced optional restraints to the scoring function being optimized, in the form of Gaussian priors. These restraints have the Gaussian form previously described for x, y shifts and defocus [Chen *et al.* 2009] and for helical parameters [Alushin *et al.* 2010]. Following the notation from Chen et al. [Chen *et al.* 2009], FREALIGN maximizes a weighted similarity measure CC_w between each particle image X and a projection A of the 3D model:

$$S(\phi;\Theta) = CC_w(\phi;\Theta) + \frac{\sigma^2}{|X||A|} \ln f(\phi;\Theta) \tag{A.1}$$

where ϕ is the set of parameters being optimized, Θ is a set of parameters governing the restraints imposed on or on a subset of its parameters, σ is an estimate of the noise standard deviation equal to $|X - A|/\sqrt{N}$, N is the number of pixels in the image and f is the restraint function. For the x, y shifts of an image, the restraint imposed on refinement cycle i is has the same form as the restraint described in Chen et al. [Chen *et al.* 2009]:

$$f_{xy}(\phi;\Theta) = \exp\left[\frac{-(x_i - x_{i-1})^2}{2\sigma_x^2} - \frac{-(y_i - y_{i-1})^2}{2\sigma_y^2}\right] \tag{A.2}$$

where $\sigma_x = \sigma_y$ is a parameter defined by the user controlling how strongly restrained the translational change must be. For an Euler angle ψ, the restraint is then:

$$f_{\psi}(\phi;\Theta) = \exp\left[\frac{-(\psi_i - \psi_{i-1})^2}{2\sigma_{\psi}^2}\right] \tag{A.3}$$

and analogously for the Euler angles θ and Φ. For simplicity, we assume $\sigma_{\psi} = \sigma_{\theta} = \sigma_{\Phi}$.

The translational and angular restraints can be specified by their respective keywords in FREALIGN's `mparameters` file as:

```
# Alignment-restraint parameters
```

```
sigma_angles 0.0 !  When greater than 0:  Restrains the Euler angles to avoid
they change too much in one cycle (STD in degrees).

sigma_shifts 0.0 !  When greater than 0:  Restrains the x,y shifts to avoid they
change too much in one cycle (STD in Angstrom).
```

A.5. Supplementary Note 2

Auto-refinement in FREALIGN

We implemented a single-particle auto-refinement algorithm based on FREALIGN v9.11. The algorithm proceeds by evaluating the FSC between the reconstructed half-maps at the end of each refinement cycle at two different thresholds: a "high" threshold (`thresh_fsc_ref`), for example FSC = 0.5, and a lower threshold (`thresh_fsc_eval`), for example FSC = 0.143 [Rosenthal & Henderson 2003]. For the next refinement cycle, it uses the resolution limit based on `thresh_fsc_ref`. Any FSC improvement beyond this resolution limit, evaluated at `thresh_fsc_eval`, is considered to be unbiased. It is important to remark that these FSC thresholds are arbitrary and not necessarily related to the actual map resolution. If the map does not improve, based on this criterion, then the same refinement cycle can be run again trying different combinations of parameters (PSI, THETA, PHI, SHX, SHY) for refinement (`change_pmask` option), thus changing and/or reducing the dimensionality of the refinement optimization problem. If all parameter combinations have been exhausted and the FSC does not improve from one cycle to the next, refinement is considered to have converged. Even if auto-refinement has converged in the previous step, in some cases, further resolution improvements may still be obtained by modifying other FREALIGN parameters and starting a new auto-refinement procedure from the last cycle of the previous run. We note that a similar auto-refinement strategy was implemented in the cisTEM package [Grant *et al.* 2018].

The auto-refiner can be tuned by the user via the following keywords in FREALIGN's `mparameters` file:

```
# Auto-refinement parameters (only used if calling frealign_run_refine_auto script)

thresh_fsc_ref 0.8 !  Auto-refiner will take the resolution where the FSC crosses
this threshold as the limit for the refinement (typically 0.4 to 0.8).

thresh_fsc_eval 0.143 !  Auto-refiner will evaluate map improvement by looking
at the resolution where the FSC crosses this threshold.

res_min 40.0 !  Auto-refiner won't ever use a resolution lower than this as the
limit for the refinement.

ref_stay_away 2.0 !  Auto-refiner won't ever use a limit for alignment that is
less than this value away from the current map's resolution.

change_pmask T ! T or F. Set to T to allow auto-refiner to try different combinations
of parameter_mask if necessary.
```

```
no_theta F ! T or F. Set to T to keep the tilt angle THETA fixed (always unchanged)
in auto-refinement.  Useful for 2D crystal data.
```

A modified version of FREALIGN v9.11 supporting these features is available at `http://github.com/C-CINA/frealign-2dx`.

B. Supplementary Information for “Membrane binding induces domain rearrangements and oligomerization that prime FAK for activation”

B.1. Supplementary Figures

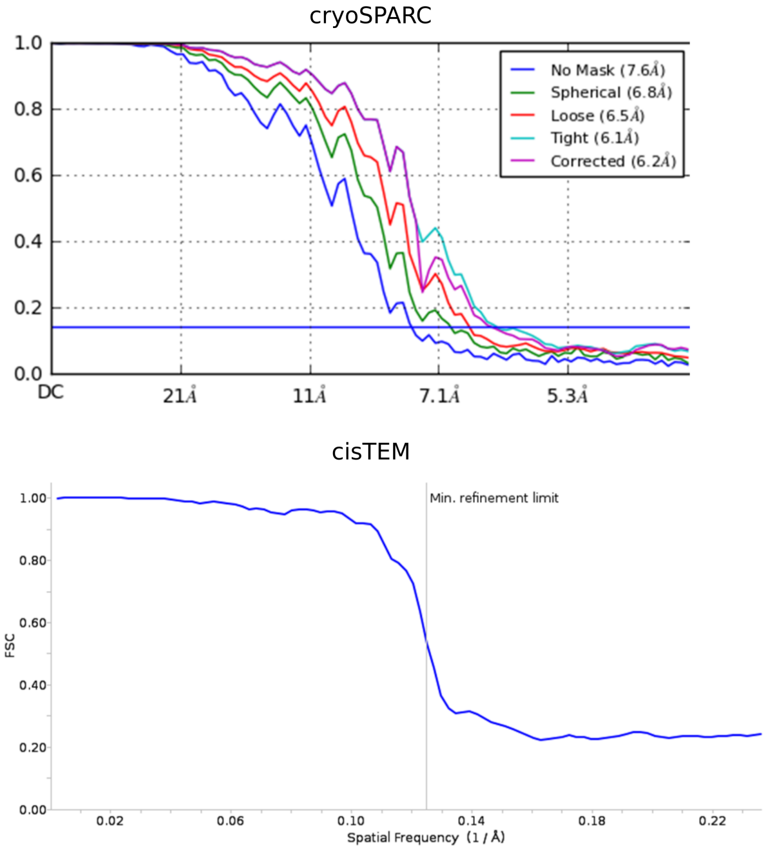


Figure B.1: **Resolution estimation of the AMP-PNP map in cryoSPARC and cisTEM.** The FSC curves at the end of the refinement in cryoSPARC (top) and cisTEM (bottom) are shown. As the half-sets in these cases have been randomly split by cryoSPARC based on the particles, without any regard to which 2D crystal they came from, the FSC curves never fall below zero, due to the artificial correlations introduced by the overlap between adjacent particle boxes.

Acknowledgments

I would like to thank the following people, without whom this doctoral thesis would not have been possible:

First of all, my wife, Ana Paula, who accepted to join me in the adventure of moving from Campinas to Basel and always stands by my side in good and bad times.

My advisor, Henning Stahlberg, for the opportunity of carrying out my thesis research in his group, with great freedom and encouragement.

Volker Roth, for kindly accepting to co-referee this thesis.

My colleagues at C-CINA with whom I had the opportunity to collaborate in exciting projects: Nikhil, Robb, Julia, Mohamed, Ricardo A., Thomas, Claudio and Stefan. Ariane, Kenny, Daniel C.B. and Lubomir for taking good care of our microscopes. Martin Jacquot and the sciCORE team for cluster computing support. Karen Bergmann and Angie Klarer for the administrative support. Ricardo G., Stefano, Anastasya, Paula, Cedric, Raphi, Max, Inay, Kai, and all friends from C-CINA.

Daniel Lietha, Ivan Acebrón and Jaska Boskovic, for the exciting collaboration on FAK.

The researchers in the field of structural biology, in particular the cryo-EM community, with whom I had the opportunity to discuss my ideas and learn a lot from: Jan Pieter Abrahams, Timm Maier, Daniel Castaño-Diez, Misha Kudryashev, Sjors Scheres, Niko Grigorieff, Marin van Heel, among others.

The advisors of my master thesis, Rodrigo Portugal and Fernando von Zuben.

The friends from the Biozentrum: Dominik H. (from whom I first heard about the Fellowships for Excellence program), Luca, Martino, Keith, Anne, Joka, Susi, the organizers from the 2016 and 2017 PhD Retreats, and many others.

The Brazilian friends who I gladly met in Basel: Letícia, Joyce, Raphael and Layara.

My parents, Sérgio and Adriana, who have always supported and encouraged me into science. I also thank the love from Elisa, Fábio, Luciana, Iolita, Gabriel, Eliana and our whole family, and all my friends from Brazil.

List of Abbreviations

2D two-dimensional.

3D three-dimensional.

ABC ATP-binding cassette.

ATP adenosine triphosphate.

blosc blocked, shuffle, compression library.

cAMP cyclic adenosine monophosphate.

CC cross-correlation.

CCD charge-coupled device.

CMOS complementary metal-oxide semiconductor.

CNBD cyclic-nucleotide binding domain.

CPU central processing unit.

cryo-EM cryo-electron microscopy.

cryo-ET cryo-electron tomography.

CTF contrast transfer function.

DED direct electron detector.

DNA deoxyribonucleic acid.

DQE detective quantum efficiency.

ECM extracellular matrix.

ED electron diffraction.

EMDB Electron Microscopy Data Bank.

EMPIAR Electron Microscopy Public Image Archive.

FAK focal adhesion kinase.

FFT fast Fourier transform.

float32 floating-point 32-bit computer data, ∼6 significant figures.

FPGA Field-Gate Programmable Arrays.

FRC Fourier ring correlation.

FREALIGN Fourier REconstruction and ALIGNment.

FSC Fourier shell correlation.

FT Fourier transform.

GB Gigabyte (2^{30} bytes).

Gb Gigabit, network (10^9 bits).

GPCR G protein-coupled receptor.

GPU graphics processing unit.

GUI graphical user interface.

HD hard disk.

HPC high-performance computing.

MB Megabyte (2^{20} bytes).

microED micro-electron diffraction.

MRC Medical Research Council.

MSA multivariate statistical analysis.

MTF mutual transfer function.

NMR nuclear magnetic resonance.

PCA principal component analysis.

PDB Protein Data Bank.

PSF point spread function.

RAM random access memory.

RELION REgularised LIkelihood OptimizatioN.

RNA ribonucleic acid.

SGD stochastic gradient descent.

SNR signal-to-noise ratio.

SPA single particle analysis.

SPA single particle reconstruction.

SSD solid state drive.

SSNR spectral SNR.

STA subtomogram averaging.

TEM transmission electron microscope.

TMD transmembrane domain.

uint8 unsigned 8-bit integer computer data, range 0–255.

VSD voltage-sensing domain.

XRD X-ray diffraction.

List of Figures

List of Tables

References

[Abeyrathne *et al.* 2010] Priyanka D Abeyrathne, Mohamed Chami, Radosav S Pantelic, Kenneth N Goldie and Henning Stahlberg. *Preparation of 2D Crystals of Membrane Proteins for High-Resolution Electron Crystallography Data Collection.* In Grant J Jensen B T Methods in Enzymology, editeur, Cryo-EM Part A Sample Preparation and Data Collection, volume Volume 481, pages 25–43. Academic Press, 2010. (Cited on page 59.)

[Abeyrathne *et al.* 2012] P.D. D. Abeyrathne, M. Arheit, F. Kebbel, D. Castano-Diez, K.N. N. Goldie, M. Chami, H. Stahlberg, L. Renault and W. Kühlbrandt. *Analysis of 2-D Crystals of Membrane Proteins by Electron Microscopy.* In Comprehensive Biophysics, volume 1, pages 277–310. Elsevier, Amsterdam, 2012. (Cited on pages 5, 14, 36, 37 and 59.)

[Adams *et al.* 2002] Paul D Adams, Ralf W Grosse-Kunstleve, Li-Wei Hung, Thomas R Ioerger, Airlie J McCoy, Nigel W Moriarty, Randy J Read, James C Sacchettini, Nicholas K Sauter and Thomas C Terwilliger. *PHENIX: building new software for automated crystallographic structure determination.* Acta Crystallographica Section D, vol. 58, no. 11, pages 1948–1954, nov 2002. (Cited on page 51.)

[Adrian *et al.* 1984] Marc Adrian, Jacques Dubochet, Jean Lepault and Alasdair W. McDowall. *Cryo-electron microscopy of viruses.* Nature, vol. 308, no. 5954, pages 32–36, mar 1984. (Cited on pages 9 and 56.)

[Afanasyev *et al.* 2015] Pavel Afanasyev, Raimond B. G. Ravelli, Rishi Matadeen, Sacha De Carlo, Gijs van Duinen, Bart Alewijnse, Peter J. Peters, Jan-Pieter Abrahams, Rodrigo V. Portugal, Michael Schatz and Marin van Heel. *A posteriori correction of camera characteristics from large image data sets.* Sci. Rep., vol. 5, no. 1, page 10317, jun 2015. (Cited on page 23.)

[Afonine *et al.* 2018a] Pavel V Afonine, Bruno P Klaholz, Nigel W Moriarty, Billy K Poon, Oleg V Sobolev, Thomas C Terwilliger, Paul D Adams and Alexandre Urzhumtsev. *New tools for the analysis and validation of Cryo-EM maps and atomic models.* bioRxiv, jan 2018. (Cited on page 89.)

[Afonine *et al.* 2018b] Pavel V Afonine, Billy K Poon, Randy J Read, Oleg V Sobolev, Thomas C Terwilliger, Alexandre Urzhumtsev and Paul D Adams. *Real-space refinement in PHENIX for cryo-EM and crystallography.* Acta Crystallographica Section D, vol. 74, no. 6, pages 531–544, jun 2018. (Cited on pages 43, 51, 85 and 104.)

[Alberts 2015] Bruce Alberts, editeur. Molecular biology of the cell. Garland Science, New York, N.Y., 6th ed édition, 2015. (Cited on pages 2, 4, 5, 7 and 8.)

[Altieri *et al.* 2008] Stephen L. Altieri, Gina M. Clayton, William R. Silverman, Adrian O. Olivares, Enrique M. De La Cruz, Lise R. Thomas and João H. Morais-Cabral. *Structural and Energetic Analysis of Activation by a Cyclic Nucleotide Binding Domain.* Journal of Molecular Biology, vol. 381, no. 3, pages 655–669, 2008. (Cited on page 51.)

[Alushin *et al.* 2010] Gregory M. Alushin, Vincent H. Ramey, Sebastiano Pasqualato, David A. Ball, Nikolaus Grigorieff, Andrea Musacchio and Eva Nogales. *The Ndc80 kinetochore complex forms oligomeric arrays along microtubules.* Nature, vol. 467, no. 7317, pages 805–10, 2010. (Cited on page 107.)

[Arheit *et al.* 2013] Marcel Arheit, Daniel Castaño-Díez, Raphaël Thierry, Bryant R. Gipson, Xiangyan Zeng and Henning Stahlberg. *Image Processing of 2D Crystal Images.* In Ingeborg Schmidt-Krey and Yifan Cheng, editeurs, Electron Crystallography of Soluble and Membrane Proteins SE - 10, volume 955 of *Methods in Molecular Biology*, pages 171–194. Humana Press, 2013. (Cited on page 15.)

[Bai *et al.* 2015] Xiao Chen Bai, Eeson Rajendra, Guanghui Yang, Yigong Shi and Sjors H.W. Scheres. *Sampling the conformational space of the catalytic subunit of human g-secretase.* eLife, vol. 4, no. December2015, page e11182, dec 2015. (Cited on pages 43, 50 and 69.)

[Baker & Rubinstein 2010] Lindsay A Baker and John L Rubinstein. *Chapter 15 - Radiation Damage in Electron Cryomicroscopy.* Methods in enzymology, vol. 481, pages 371–388, 2010. (Cited on page 9.)

[Barad *et al.* 2015] Benjamin A. Barad, Nathaniel Echols, Ray Yu Ruei Wang, Yifan Cheng, Frank Dimaio, Paul D. Adams and James S. Fraser. *EMRinger: Side chain-directed model and map validation for 3D cryo-electron microscopy.* Nature Methods, vol. 12, no. 10, pages 943–946, 2015. (Cited on pages 51 and 104.)

[Bauer *et al.* 2019] Magnus Sebastian Bauer, Fabian Baumann, Csaba Daday, Pilar Redondo, Ellis Durner, Markus Andreas Jobst, Lukas Frederik Milles, Davide Mercadante, Diana Angela Pippig, Hermann Eduard Gaub, Frauke Gräter and Daniel Lietha. *Structural and mechanistic insights into mechanoactivation of focal adhesion kinase.* Proceedings of the National Academy of Sciences, page 201820567, mar 2019. (Cited on pages 75 and 79.)

[Baumeister *et al.* 1999] Wolfgang Baumeister, Rudo Grimm and Jochen Walz.

Electron tomography of molecules and cells. Trends in Cell Biology, vol. 9, no. 2, pages 81–85, feb 1999. (Cited on page 14.)

[Biyani *et al.* 2017] Nikhil Biyani, Ricardo D. Righetto, Robert McLeod, Daniel Caujolle-Bert, Daniel Castano-Diez, Kenneth N. Goldie and Henning Stahlberg. *Focus: The interface between data collection and data processing in cryo-EM.* Journal of Structural Biology, vol. 198, no. 2, pages 124–133, 2017. (Cited on pages 16, 38, 40, 49, 57, 58, 59, 60, 83, 84 and 87.)

[Biyani *et al.* 2018] Nikhil Biyani, Sebastian Scherer, Ricardo D. Righetto, Julia Kowal, Mohamed Chami and Henning Stahlberg. *Image processing techniques for high-resolution structure determination from badly ordered 2D crystals.* Journal of Structural Biology, vol. 203, no. 2, pages 120–134, 2018. (Cited on pages 15, 16, 36, 38, 46, 60, 64 and 87.)

[Biyani 2017] Nikhil Biyani. *Novel image processing tools and techniques in cryo-electron microscopy.* PhD thesis, Basel, 2017. (Cited on page 16.)

[Branden & Tooze 1999] Carl Branden and John Tooze. Introduction to protein structure. Garland Science, 2 édition, 1999. (Cited on pages 3, 4, 5 and 7.)

[Brilot *et al.* 2012] Axel F Brilot, James Z Chen, Anchi Cheng, Junhua Pan, Stephen C Harrison, Clinton S Potter, Bridget Carragher, Richard Henderson and Nikolaus Grigorieff. *Beam-induced motion of vitrified specimen on holey carbon film.* Journal of Structural Biology, vol. 177, no. 3, pages 630–637, mar 2012. (Cited on pages 11, 60 and 87.)

[Brown *et al.* 2015] Alan Brown, Fei Long, Robert A. Nicholls, Jaan Toots, Paul Emsley and Garib Murshudov. *Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions.* Acta Crystallographica Section D Biological Crystallography, vol. 71, no. 1, pages 136–153, jan 2015. (Cited on page 71.)

[Cardone *et al.* 2013] Giovanni Cardone, J Bernard Heymann and Alasdair C Steven. *One number does not fit all: Mapping local variations in resolution in cryo-EM reconstructions.* Journal of Structural Biology, vol. 184, no. 2, pages 226–236, 2013. (Cited on pages 50, 80 and 85.)

[Castaño-Díez *et al.* 2012] Daniel Castaño-Díez, Mikhail Kudryashev, Marcel Arheit and Henning Stahlberg. *Dynamo: A flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments.* Journal of Structural Biology, vol. 178, no. 2, pages 139–151, may 2012. (Cited on page 14.)

[Ceccarelli *et al.* 2006] Derek F J Ceccarelli, Hyun Kyu Song, Florence Poy, Michael D Schaller and Michael J Eck. *Crystal Structure of the FERM Domain of Focal Adhesion Kinase.* Journal of Biological Chemistry, vol. 281, no. 1, pages 252–259, jan 2006. (Cited on pages 75 and 85.)

[Chen *et al.* 2009] James Z Chen, Ethan C Settembre, Scott T Aoki, Xing Zhang, A Richard Bellamy, Philip R Dormitzer, Stephen C Harrison and Nikolaus Grigorieff. *Molecular interactions in rotavirus assembly and uncoating seen by high-resolution cryo-EM.* Proceedings of the National Academy of Sciences of the United States of America, vol. 106, no. 26, pages 10644–8, 2009. (Cited on page 107.)

[Chen *et al.* 2013] Shaoxia Chen, Greg McMullan, Abdul R. Faruqi, Garib N. Murshudov, Judith M. Short, Sjors H.W. W Scheres and Richard Henderson. *High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy.* Ultramicroscopy, vol. 135, no. 0, pages 24–35, dec 2013. (Cited on pages 80 and 84.)

[Chen *et al.* 2017] Muyuan Chen, Wei Dai, Stella Y Sun, Darius Jonasch, Cynthia Y He, Michael F. Schmid, Wah Chiu and Steven J Ludtke. *Convolutional neural networks for automated annotation of cellular cryo-electron tomograms.* Nature Methods, vol. 14, no. 10, pages 983–985, oct 2017. (Cited on page 89.)

[Cheng *et al.* 2015a] Anchi Cheng, Richard Henderson, David Mastronarde, Steven J Ludtke, Remco H.M. Schoenmakers, Judith Short, Roberto Marabini, Sargis Dallakyan, David Agard and Martyn Winn. *MRC2014: Extensions to the MRC format header for electron cryo-microscopy and tomography.* Journal of Structural Biology, vol. 192, no. 2, pages 146–150, nov 2015. (Cited on pages 23 and 24.)

[Cheng *et al.* 2015b] Yifan Cheng, Nikolaus Grigorieff, Pawel A. Penczek and Thomas Walz. *A Primer to Single-Particle Cryo-Electron Microscopy.* Cell, vol. 161, no. 3, pages 438–449, apr 2015. (Cited on page 12.)

[Cheng *et al.* 2018] Anchi Cheng, Edward T. Eng, Lambertus Alink, William J. Rice, Kelsey D. Jordan, Laura Y. Kim, Clinton S. Potter and Bridget Carragher. *High resolution single particle cryo-electron microscopy using beam-image shift.* Journal of Structural Biology, vol. 204, no. 2, pages 270–275, 2018. (Cited on page 87.)

[Chiu *et al.* 2007] Po-Lin Chiu, Matthew D Pagel, James Evans, Hui-Ting Chou, Xiangyan Zeng, Bryant Gipson, Henning Stahlberg and Crina M Nimigean. *The Structure of the Prokaryotic Cyclic Nucleotide-Modulated Potassium Channel MloK1 at 16 Å Resolution.* Structure, vol. 15, no. 9, pages 1053–1064,

2007. (Cited on page 49.)

[Chothia & Lesk 1986] C Chothia and A M Lesk. *The relation between the divergence of sequence and structure in proteins.* The EMBO journal, vol. 5, no. 4, pages 823–6, apr 1986. (Cited on page 4.)

[Clayton *et al.* 2008] Gina M Clayton, Steve Altieri, Lise Heginbotham, Vinzenz M Unger and João H Morais-Cabral. *Structure of the transmembrane regions of a bacterial cyclic nucleotide-regulated channel.* Proceedings of the National Academy of Sciences, vol. 105, no. 5, pages 1511 LP – 1515, feb 2008. (Cited on pages 41 and 51.)

[Crick 1970] Francis Crick. *Central Dogma of Molecular Biology.* Nature, vol. 227, no. 5258, pages 561–563, aug 1970. (Cited on page 2.)

[Crowther *et al.* 1970] R A Crowther, D J DeRosier and A Klug. *The Reconstruction of a Three-Dimensional Structure from Projections and its Application to Electron Microscopy.* Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 317, no. 1530, pages 319–340, jun 1970. (Cited on pages 14, 37 and 94.)

[Crowther *et al.* 1996] R.A. Crowther, R. Henderson and J.M. Smith. *MRC Image Processing Programs.* Journal of Structural Biology, vol. 116, no. 1, pages 9–16, jan 1996. (Cited on pages 24, 38 and 57.)

[Dang *et al.* 2017] Shangyu Dang, Shengjie Feng, Jason Tien, Christian J. Peters, David Bulkley, Marco Lolicato, Jianhua Zhao, Kathrin Zuberbühler, Wenlei Ye, Lijun Qi, Tingxu Chen, Charles S. Craik, Yuh Nung Jan, Daniel L. Minor, Yifan Cheng and Lily Yeh Jan. *Cryo-EM structures of the TMEM16A calciumactivated chloride channel.* Nature, vol. 552, no. 7685, pages 426–429, 2017. (Cited on page 46.)

[Dashti *et al.* 2019] Ali Dashti, Mrinal Shekhar, Danya Ben Hail, Ghoncheh Mashayekhi, Peter Schwander, Amedee des Georges, Joachim Frank, Abhishek Singharoy and Abbas Ourmazd. *Functional Pathways of Biomolecules Retrieved from Single-particle Snapshots.* bioRxiv, page 291922, jan 2019. (Cited on page 89.)

[Davis *et al.* 2007] Ian W Davis, Andrew Leaver-Fay, Vincent B Chen, Jeremy N Block, Gary J Kapral, Xueyi Wang, Laura W Murray, W Bryan Arendall, Jack Snoeyink, Jane S Richardson and David C Richardson. *MolProbity: all-atom contacts and structure validation for proteins and nucleic acids.* Nucleic Acids Research, vol. 35, no. Web Server issue, pages W375–W383, jul 2007. (Cited on pages 51 and 104.)

[De Rosier & Klug 1968] D. J. De Rosier and A. Klug. *Reconstruction of three dimensional structures from electron micrographs.* Nature, vol. 217, no. 5124, pages 130–134, 1968. (Cited on pages 12, 14 and 57.)

[Deisenhofer & Michel 1989] Johann Deisenhofer and Hartmut Michel. *The Photosynthetic Reaction Center from the Purple Bacterium Rhodopseudomonas viridis.* Science, vol. 245, no. 4925, pages 1463–1473, sep 1989. (Cited on page 56.)

[Deisenhofer *et al.* 1985] J Deisenhofer, O Epp, K Miki, R Huber and H Michel. *Structure of the protein subunits in the photosynthetic reaction centre of Rhodopseudomonas viridis at 3Å resolution.* Nature, vol. 318, no. 6047, pages 618–624, 1985. (Cited on page 7.)

[Downing & Hendrickson 1999] Kenneth H Downing and Felicia M Hendrickson. *Performance of a 2k CCD camera designed for electron crystallography at 400 kV.* Ultramicroscopy, vol. 75, no. 4, pages 215–233, jan 1999. (Cited on page 11.)

[Duda *et al.* 2015] Jarek Duda, Khalid Tahboub, Neeraj J Gadgil and Edward J Delp. *The use of asymmetric numeral systems as an accurate replacement for Huffman coding.* In 2015 Picture Coding Symposium (PCS), pages 65–69. IEEE, 2015. (Cited on page 26.)

[Duda 2013] Jarek Duda. *Asymmetric numeral systems: entropy coding combining speed of Huffman coding with compression rate of arithmetic coding.* arXiv preprint arXiv:1311.2540, 2013. (Cited on page 26.)

[Duncan *et al.* 2017] Anna L. Duncan, Tyler Reddy, Heidi Koldsø, Jean Hélie, Philip W. Fowler, Matthieu Chavent and Mark S. P. Sansom. *Protein crowding and lipid complexity influence the nanoscale dynamic organization of ion channels in cell membranes.* Scientific Reports, vol. 7, no. 1, page 16647, dec 2017. (Cited on page 88.)

[Egelman 2007] Edward H. Egelman. *The iterative helical real space reconstruction method: Surmounting the problems posed by real polymers.* Journal of Structural Biology, vol. 157, no. 1, pages 83–94, 2007. (Cited on page 37.)

[Emsley & Cowtan 2004] Paul Emsley and Kevin Cowtan. *Coot: model-building tools for molecular graphics.* Acta Crystallographica Section D, vol. 60, no. 12, pages 2126–2132, dec 2004. (Cited on page 51.)

[Feynman 1960] Richard P Feynman. *There's plenty of room at the bottom.* Engineering and Science, vol. 23, no. 5, pages 22–36, 1960. (Cited on page 8.)

[Frank *et al.* 1988] Joachim Frank, Wah Chiu and Laura Degn. *The characterization of structural variations within a crystal field*. Ultramicroscopy, vol. 26, no. 4, pages 345–360, 1988. (Cited on page 47.)

[Frank *et al.* 1996] Joachim Frank, Michael Radermacher, Pawel Penczek, Jun Zhu, Yanhong Li, Mahieddine Ladjadj and Ardean Leith. *SPIDER and WEB: Processing and Visualization of Images in 3D Electron Microscopy and Related Fields*. Journal of Structural Biology, vol. 116, no. 1, pages 190–199, jan 1996. (Cited on page 64.)

[Frank 2006] Joachim Frank. Three-Dimensional Electron Microscopy of Macromolecular Assemblies. Oxford University Press, 2nd édition, 2006. (Cited on pages 8, 11, 12, 37, 57 and 58.)

[Gao *et al.* 2016] Yuan Gao, Erhu Cao, David Julius and Yifan Cheng. *TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action*. Nature, vol. advance on, pages 1–17, may 2016. (Cited on pages 13, 36, 46, 57 and 87.)

[Gipson *et al.* 2007] Bryant Gipson, Xiangyan Zeng, Zi Yan Zhang and Henning Stahlberg. *2dx—User-friendly image processing for 2D crystals*. Journal of Structural Biology, vol. 157, no. 1, pages 64–72, jan 2007. (Cited on pages 38, 49, 57, 60, 83 and 84.)

[Gipson *et al.* 2011] Bryant R Gipson, Daniel J Masiel, Nigel D Browning, John Spence, Kaoru Mitsuoka and Henning Stahlberg. *Automatic recovery of missing amplitudes and phases in tilt-limited electron crystallography of two-dimensional crystals*. Physical Review E, vol. 84, no. 1, page 11916, jul 2011. (Cited on pages 15 and 46.)

[Goldie *et al.* 2014] Kenneth N Goldie, Priyanka Abeyrathne, Fabian Kebbel, Mohamed Chami, Philippe Ringler and Henning Stahlberg. *Cryo-electron Microscopy of Membrane Proteins*. In John Kuo, editeur, Electron Microscopy: Methods and Protocols, pages 325–341. Humana Press, Totowa, NJ, 2014. (Cited on pages 5 and 59.)

[Gonen *et al.* 2005] Tamir Gonen, Yifan Cheng, Piotr Sliz, Yoko Hiroaki, Yoshinori Fujiyoshi, Stephen C Harrison and Thomas Walz. *Lipid-protein interactions in double-layered two-dimensional AQP0 crystals*. Nature, vol. 438, no. 7068, pages 633–638, dec 2005. (Cited on pages 15, 36 and 56.)

[Gonen *et al.* 2015] Shane Gonen, Frank DiMaio, Tamir Gonen and David Baker. *Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces*. Science, vol. 348, no. 6241, pages 1365–1368, jun 2015. (Cited on pages 37, 46, 57 and 89.)

[Goñi *et al.* 2014] Guillermina M Goñi, Carolina Epifano, Jasminka Boskovic, Marta Camacho-Artacho, Jing Zhou, Agnieszka Bronowska, M Teresa Martín, Michael J Eck, Leonor Kremer, Frauke Gräter, Francesco Luigi Gervasio, Mirna Perez-Moreno and Daniel Lietha. *Phosphatidylinositol 4,5-bisphosphate triggers activation of focal adhesion kinase by inducing clustering and conformational changes.* Proceedings of the National Academy of Sciences, vol. 111, no. 31, pages E3177 LP – E3186, aug 2014. (Cited on pages 57, 75, 79 and 82.)

[Grant & Grigorieff 2015] Timothy Grant and Nikolaus Grigorieff. *Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6.* eLife, vol. 4, may 2015. (Cited on pages 11, 23, 38 and 60.)

[Grant *et al.* 2018] Timothy Grant, Alexis Rohou and Nikolaus Grigorieff. *cisTEM, user-friendly software for single-particle image processing.* eLife, vol. 7, page e35383, 2018. (Cited on pages 13, 58, 68, 71, 77, 84, 87 and 109.)

[Grigorieff 2016] N Grigorieff. *Frealign: An Exploratory Tool for Single-Particle Cryo-EM.* Academic Press, 2016. (Cited on pages 13, 16, 43, 50, 58, 59, 66, 68, 76, 83, 87 and 88.)

[Haenel 2014] Valentin Haenel. *Bloscpack: a compressed lightweight serialization format for numerical data.* arXiv preprint arXiv:1404.6383, 2014. (Cited on page 24.)

[Harauz & van Heel 1986] George Harauz and Marin van Heel. *Exact filters for general geometry three dimensional reconstruction.* Optik, vol. 78, no. 4, pages 146–156, 1986. (Cited on pages 13, 41, 70, 77 and 84.)

[He & Scheres 2016] Shaoda He and Sjors H W Scheres. *Helical reconstruction in RELION.* Journal of Structural Biology, pages 1–27, 2016. (Cited on pages 37, 41, 49 and 65.)

[Henderson & Unwin 1975] Richard Henderson and P Nigel T Unwin. *Three-dimensional model of purple membrane obtained by electron microscopy.* Nature, vol. 257, no. 5521, pages 28–32, 1975. (Cited on pages 15, 36, 56, 57 and 88.)

[Henderson *et al.* 1986] R. Henderson, J. M. Baldwin, K. H. Downing, J. Lepault and F. Zemlin. *Structure of purple membrane from halobacterium halobium: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution.* Ultramicroscopy, vol. 19, no. 2, pages 147–178, 1986. (Cited on pages 15, 37, 38, 49, 57, 60, 61 and 94.)

[Henderson *et al.* 1990] R. Henderson, J.M. M Baldwin, T.A. A Ceska, F. Zemlin, E. Beckmann and K.H. H Downing. *Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy.* Journal of Molecular Biology, vol. 213, no. 4, pages 899–929, jun 1990. (Cited on pages 15, 36, 56 and 57.)

[Henderson *et al.* 2012] Richard Henderson, Andrej Sali, Matthew L Baker, Bridget Carragher, Batsal Devkota, Kenneth H Downing, Edward H Egelman, Zukang Feng, Joachim Frank, Nikolaus Grigorieff, Wen Jiang, Steven J Ludtke, Ohad Medalia, Pawel a Penczek, Peter B Rosenthal, Michael G Rossmann, Michael F Schmid, Gunnar F Schröder, Alasdair C Steven, David L Stokes, John D Westbrook, Willy Wriggers, Huanwang Yang, Jasmine Young, Helen M Berman, Wah Chiu, Gerard J Kleywegt and Catherine L Lawson. *Outcome of the first electron microscopy validation task force meeting.* Structure, vol. 20, no. 2, pages 205–14, feb 2012. (Cited on page 12.)

[Henderson 2013] Richard Henderson. *Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise.* Proceedings of the National Academy of Sciences of the United States of America, pages 1–5, oct 2013. (Cited on page 12.)

[Ilca *et al.* 2015] Serban L. Ilca, Abhay Kotecha, Xiaoyu Sun, Minna M. Poranen, David I. Stuart and Juha T. Huiskonen. *Localized reconstruction of subunits from electron cryomicroscopy images of macromolecular complexes.* Nature Communications, vol. 6, page 8843, 2015. (Cited on pages 43 and 50.)

[Kabsch & Sander 1983] Wolfgang Kabsch and Christian Sander. *Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, vol. 22, no. 12, pages 2577–2637, dec 1983. (Cited on page 51.)

[Kelley *et al.* 1996] L A Kelley, S P Gardner and M J Sutcliffe. *An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies.* Protein engineering, vol. 9, no. 11, pages 1063–1065, nov 1996. (Cited on page 51.)

[Kendrew *et al.* 1960] John C Kendrew, Richard E Dickerson, Bror E Strandberg, Robert G Hart, D R Davies, D C Phillips and V C Shore. *Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å. resolution.* Nature, vol. 185, no. 4711, page 422, 1960. (Cited on page 7.)

[Khoshouei *et al.* 2017] Maryam Khoshouei, Radostin Danev, Juergen M. Plitzko and Wolfgang Baumeister. *Revisiting the Structure of Hemoglobin and Myoglobin with Cryo-Electron Microscopy.* Journal of Molecular Biology, vol. 429, no. 17, pages 2611–2618, aug 2017. (Cited on pages 13, 75 and 89.)

[Kimanius *et al.* 2016] Dari Kimanius, Bjorn O Forsberg, Sjors Scheres and Erik Lindahl. *Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2.* bioRxiv, jun 2016. (Cited on page 13.)

[Kowal *et al.* 2014] Julia Kowal, Mohamed Chami, Paul Baumgartner, Marcel Arheit, Po-Lin Chiu, Martina Rangl, Simon Scheuring, Gunnar F Schröder, Crina M Nimigean and Henning Stahlberg. *Ligand-induced structural changes in the cyclic nucleotide-modulated potassium channel MloK1.* Nat Commun, vol. 5, jan 2014. (Cited on pages 47, 48, 49 and 58.)

[Kowal *et al.* 2018] Julia Kowal, Nikhil Biyani, Mohamed Chami, Sebastian Scherer, Andrzej J. Rzepiela, Paul Baumgartner, Vikrant Upadhyay, Crina M. Nimigean and Henning Stahlberg. *High-Resolution Cryoelectron Microscopy Structure of the Cyclic Nucleotide-Modulated Potassium Channel MloK1 in a Lipid Bilayer.* Structure, vol. 26, no. 1, pages 20–27.e3, 2018. (Cited on pages 16, 37, 40, 41, 46, 47, 48, 49, 51, 58, 87, 88 and 104.)

[Kuang *et al.* 2015] Qie Kuang, Pasi Purhonen, Thirupathi Pattipaka, Yohannes H Ayele, Hans Hebert and Philip J B Koeck. *A Refined Single-Particle Reconstruction Procedure to Process Two-Dimensional Crystal Images from Transmission Electron Microscopy.* Microscopy and Microanalysis, vol. 21, no. 04, pages 876–885, 2015. (Cited on pages 37 and 58.)

[Kühlbrandt *et al.* 1994] Werner Kühlbrandt, D N Wang and Y Fujiyoshi. *Atomic model of plant light-harvesting complex by electron crystallography.* Nature, vol. 367, no. 6464, pages 614–621, 1994. (Cited on pages 15, 36 and 56.)

[Kühlbrandt 2014] Werner Kühlbrandt. *The Resolution Revolution.* Science, vol. 343, no. March, pages 1443–1444, mar 2014. (Cited on pages 13, 36, 57 and 87.)

[Landau & Rosenbusch 1996] E. M. Landau and J. P. Rosenbusch. *Lipidic cubic phases: A novel concept for the crystallization of membrane proteins.* Proceedings of the National Academy of Sciences of the United States of America, vol. 93, no. 25, pages 14532–14535, 1996. (Cited on pages 7 and 56.)

[Lee & MacKinnon 2017] Chia Hsueh Lee and Roderick MacKinnon. *Structures of the Human HCN1 Hyperpolarization-Activated Channel.* Cell, vol. 168, no. 1-2, pages 111–120.e11, 2017. (Cited on pages 46 and 87.)

[Li *et al.* 2013a] Xueming Li, Paul Mooney, Shawn Zheng, Christopher R Booth, Michael B Braunfeld, Sander Gubbens, David A Agard and Yifan Cheng. *Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM.* Nature Methods, vol. 10, no. 6, pages

584–590, jun 2013. (Cited on pages 11 and 23.)

[Li *et al.* 2013b] Xueming Li, Shawn Q Zheng, Kiyoshi Egami, David A Agard and Yifan Cheng. *Influence of electron dose rate on electron counting images recorded with the K2 camera.* Journal of Structural Biology, vol. 184, no. 2, pages 251–260, nov 2013. (Cited on page 23.)

[Li *et al.* 2017] Minghui Li, Xiaoyuan Zhou, Shu Wang, Ioannis Michailidis, Ye Gong, Deyuan Su, Huan Li, Xueming Li and Jian Yang. *Structure of a eukaryotic cyclic-nucleotide-gated channel.* Nature, vol. 542, no. 7639, pages 60–65, 2017. (Cited on page 46.)

[Liang & Tamm 2016] Binyong Liang and Lukas K Tamm. *NMR as a tool to investigate the structure, dynamics and function of membrane proteins.* Nature Structural &Amp; Molecular Biology, vol. 23, page 468, jun 2016. (Cited on pages 8 and 56.)

[Liao *et al.* 2013] Maofu Liao, Erhu Cao, David Julius and Yifan Cheng. *Structure of the TRPV1 ion channel determined by electron cryo-microscopy.* Nature, vol. 504, no. 7478, pages 107–112, dec 2013. (Cited on pages 13, 36 and 57.)

[Lietha *et al.* 2007] Daniel Lietha, Xinming Cai, Derek F.J. Ceccarelli, Yiqun Li, Michael D. Schaller and Michael J. Eck. *Structural Basis for the Autoinhibition of Focal Adhesion Kinase.* Cell, vol. 129, no. 6, pages 1177–1187, jun 2007. (Cited on page 75.)

[Liu *et al.* 2015] Shian Liu, Paul J. Focke, Kimberly Matulef, Xuelin Bian, Pierre Moënne-Loccoz, Francis I. Valiyaveetil and Steve W. Lockless. *Ion-binding properties of a K <sup>+</sup> channel selectivity filter in different conformations.* Proceedings of the National Academy of Sciences, vol. 112, no. 49, pages 15096–15100, 2015. (Cited on page 47.)

[Lopéz-Blanco & Chacón 2013] José Ramón Lopéz-Blanco and Pablo Chacón. *IMODFIT: Efficient and robust flexible fitting based on vibrational analysis in internal coordinates.* Journal of Structural Biology, vol. 184, no. 2, pages 261–270, 2013. (Cited on pages 43, 51 and 85.)

[Lyumkis *et al.* 2013] Dmitry Lyumkis, Axel F Brilot, Douglas L Theobald and Nikolaus Grigorieff. *Likelihood-based classification of cryo-EM images using FREALIGN.* Journal of structural biology, vol. 183, no. 3, pages 377–388, 2013. (Cited on pages 13, 37, 38, 41, 43, 49, 50, 58 and 68.)

[Mastronarde & Held 2017] David N. Mastronarde and Susannah R. Held. *Automated tilt series alignment and tomographic reconstruction in IMOD.* Journal of Structural Biology, vol. 197, no. 2, pages 102–113, 2017. (Cited on page 14.)

[Mastronarde 2005] David N Mastronarde. *Automated electron microscope tomography using robust prediction of specimen movements.* Journal of Structural Biology, vol. 152, no. 1, pages 36–51, oct 2005. (Cited on pages 22 and 27.)

[McLeod *et al.* 2017a] Robert A McLeod, Ricardo Diogo Righetto, Andy Stewart and Henning Stahlberg. *MRCZ - A proposed fast compressed MRC file format and direct detector normalization strategies.* bioRxiv, mar 2017. (Cited on page 34.)

[McLeod *et al.* 2017b] Robert A. McLeod, Julia Kowal, Philippe Ringler and Henning Stahlberg. *Robust image alignment for cryogenic transmission electron microscopy.* Journal of Structural Biology, vol. 197, no. 3, pages 279–293, mar 2017. (Cited on page 23.)

[McLeod *et al.* 2018] Robert A. McLeod, Ricardo Diogo Righetto, Andy Stewart and Henning Stahlberg. *MRCZ – A file format for cryo-TEM data with fast compression.* Journal of Structural Biology, vol. 201, no. 3, pages 252–257, mar 2018. (Cited on pages 50 and 52.)

[McMullan *et al.* 2016] G McMullan, A R Faruqi and R Henderson. Direct Electron Detectors, volume 579. Elsevier Inc., 1 édition, 2016. (Cited on pages 11, 36, 46, 58, 60 and 87.)

[Meng *et al.* 2006] Elaine C Meng, Eric F Pettersen, Gregory S Couch, Conrad C Huang and Thomas E Ferrin. *Tools for integrated sequence-structure analysis with UCSF Chimera.* BMC Bioinformatics, vol. 7, no. 1, page 339, 2006. (Cited on page 52.)

[Milazzo *et al.* 2011] Anna-Clare Milazzo, Anchi Cheng, Arne Moeller, Dmitry Lyumkis, Erica Jacovetty, James Polukas, Mark H Ellisman, Nguyen-Huu Xuong, Bridget Carragher and Clinton S Potter. *Initial evaluation of a direct detection device detector for single particle cryo-electron microscopy.* Journal of Structural Biology, vol. 176, no. 3, pages 404–408, dec 2011. (Cited on page 11.)

[Mindell & Grigorieff 2003] Joseph a. Mindell and Nikolaus Grigorieff. *Accurate determination of local defocus and specimen tilt in electron microscopy.* Journal of Structural Biology, vol. 142, no. 3, pages 334–347, jun 2003. (Cited on pages 9, 10, 38 and 64.)

[Nogales & Scheres 2015] Eva Nogales and Sjors H.W. H.W. Scheres. *Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity.* Molecular Cell, vol. 58, no. 4, pages 677–689, 2015. (Cited on page 13.)

[Nogales *et al.* 1998] Eva Nogales, Sharon G Wolf and Kenneth H Downing. *Structure of the [alpha][beta] tubulin dimer by electron crystallography.* Nature, vol. 391, no. 6663, pages 199–203, jan 1998. (Cited on pages 15, 36 and 56.)

[Nowakowski *et al.* 2002] Jacek Nowakowski, Ciarán N Cronin, Duncan E McRee, Mark W Knuth, Christian G Nelson, Nikola P Pavletich, Joe Rogers, Bi-Ching Sang, Daniel N Scheibe, Ronald V Swanson and Devon A Thompson. *Structures of the Cancer-Related Aurora-A, FAK, and EphA2 Protein Kinases from Nanovolume Crystallography.* Structure, vol. 10, no. 12, pages 1659–1667, dec 2002. (Cited on pages 75 and 85.)

[Pedregosa *et al.* 2011] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot and E Duchesnay. *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, vol. 12, pages 2825–2830, 2011. (Cited on pages 51 and 52.)

[Peet *et al.* 2019] Mathew J Peet, Richard Henderson and Christopher J Russo. *The energy dependence of contrast and damage in electron cryomicroscopy of biological molecules.* Ultramicroscopy, vol. 203, pages 125–131, aug 2019. (Cited on page 89.)

[Penczek 2010a] Pawel A Penczek. *Chapter One - Fundamentals of Three-Dimensional Reconstruction from Projections.* Methods in enzymology, vol. 482, pages 1–33, 2010. (Cited on page 12.)

[Penczek 2010b] Pawel A Penczek. *Chapter Three - Resolution measures in molecular electron microscopy.* Methods in enzymology, vol. 482, pages 73–100, 2010. (Cited on page 13.)

[Penczek 2010c] Pawel A Penczek. *Chapter Two - Image Restoration in Cryo-Electron Microscopy.* Methods in enzymology, vol. 482, pages 35–72, 2010. (Cited on page 9.)

[Perutz *et al.* 1960] Max F Perutz, Michael G Rossmann, Ann F Cullis, Hilary Muirhead, Georg Will and A C T North. *Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis.* Nature, vol. 185, no. 4711, page 416, 1960. (Cited on page 7.)

[Pettersen *et al.* 2004] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng and Thomas E Ferrin. *UCSF Chimera—A visualization system for exploratory research and analysis.* Journal of Computational Chemistry, vol. 25, no. 13, pages 1605–1612, oct 2004. (Cited on pages 51, 52, 85, 98 and 106.)

[Punjani *et al.* 2017] Ali Punjani, John L. Rubinstein, David J Fleet and Marcus A. Brubaker. *cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination.* Nature Methods, vol. 14, no. 3, pages 290–296, feb 2017. (Cited on pages 13, 58, 71, 72, 76, 87 and 89.)

[Rangl *et al.* 2016] Martina Rangl, Atsushi Miyagi, Julia Kowal, Henning Stahlberg, Crina M Nimigean and Simon Scheuring. *Real-time visualization of conformational changes within single MloK1 cyclic nucleotide-modulated channels.* Nature Communications, vol. 7, page 12789, sep 2016. (Cited on pages 47 and 88.)

[Reimer & Kohl 2008] Ludwig Reimer and Helmut Kohl. Transmission electron microscopy: physics of image formation, volume 36. Springer, 2008. (Cited on page 8.)

[Righetto *et al.* 2019] Ricardo D Righetto, Nikhil Biyani, Julia Kowal, Mohamed Chami and Henning Stahlberg. *Retrieving high-resolution information from disordered 2D crystals by single-particle cryo-EM.* Nature Communications, vol. 10, no. 1, page 1722, dec 2019. (Cited on pages 58, 59, 65, 66, 68, 76, 83, 84 and 88.)

[Rohou & Grigorieff 2015] Alexis Rohou and Nikolaus Grigorieff. *CTFFIND4: Fast and accurate defocus estimation from electron micrographs.* Journal of Structural Biology, vol. 192, no. 2, pages 216–221, nov 2015. (Cited on pages 60 and 83.)

[Rosenthal & Henderson 2003] Peter B Rosenthal and Richard Henderson. *Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy.* Journal of Molecular Biology, vol. 333, no. 4, pages 721–745, oct 2003. (Cited on pages 13, 41, 50, 70, 85, 101 and 109.)

[Rossmann & Henderson 1982] M G Rossmann and R Henderson. *Phasing electron diffraction amplitudes with the molecular replacement method.* Acta Crystallographica Section A, vol. 38, no. 1, pages 13–20, jan 1982. (Cited on page 14.)

[Russel *et al.* 2012] Daniel Russel, Keren Lasker, Ben Webb, Javier Velázquez-Muriel, Elina Tjioe, Dina Schneidman-Duhovny, Bret Peterson and Andrej Sali. *Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies.* PLoS Biol, vol. 10, no. 1, page e1001244, jan 2012. (Cited on page 89.)

[Sachse *et al.* 2007] Carsten Sachse, James Z Chen, Pierre-Damien Coureux, M Elizabeth Stroupe, Marcus Fändrich and Nikolaus Grigorieff. *High-resolution Electron Microscopy of Helical Specimens: A Fresh Look at Tobacco Mosaic Virus.* Journal of Molecular Biology, vol. 371, no. 3, pages 812–835, aug 2007. (Cited on page 37.)

[Saxton & Baumeister 1982] W O Saxton and W Baumeister. *The correlation averaging of a regularly arranged bacterial cell envelope protein.* Journal of Microscopy, vol. 127, no. 2, pages 127–138, aug 1982. (Cited on pages 37, 40, 57 and 65.)

[Schenk *et al.* 2010] Andreas D Schenk, Daniel Castaño-Díez, Bryant Gipson, Marcel Arheit, Xiangyan Zeng and Henning Stahlberg. *Chapter Four - 3D Reconstruction from 2D Crystal Image and Diffraction Data.* In Grant J Jensen B T Methods in Enzymology, editeur, Cryo-EM, Part B: 3-D Reconstruction, volume Volume 482, pages 101–129. Academic Press, 2010. (Cited on pages 15, 57, 60, 83 and 84.)

[Scherer *et al.* 2014] Sebastian Scherer, Marcel Arheit, Julia Kowal, Xiangyan Zeng and Henning Stahlberg. *Single particle 3D reconstruction for 2D crystal images of membrane proteins.* Journal of Structural Biology, vol. 185, no. 3, pages 267–277, mar 2014. (Cited on pages 16, 37 and 58.)

[Scherer 2015] Sebastian Scherer. *High-throughput high-resolution cryo-electron crystallography.* PhD thesis, Diss. Phil.-Nat. Univ. Basel, 2015 .- Ref.: Henning Stahlberg ; Korr.: Volker Roth, Basel, 2015. (Cited on page 16.)

[Scheres & Chen 2012] Sjors H.W. W Scheres and Shaoxia Chen. *Prevention of overfitting in cryo-EM structure determination.* Nature Methods, vol. 9, no. 9, pages 853–854, jul 2012. (Cited on page 13.)

[Scheres *et al.* 2005] Sjors H. W. Scheres, M Valle, R Nunez, C O S Sorzano, R Marabini, G T Herman and J M Carazo. *Maximum-likelihood multi-reference refinement for electron microscopy images.* Journal of Molecular Biology, vol. 348, no. 1, pages 139–149, apr 2005. (Cited on page 13.)

[Scheres *et al.* 2007] Sjors H. W. Scheres, Haixiao Gao, Mikel Valle, Gabor T Herman, Paul P B Eggermont, Joachim Frank and Jose-maria Carazo. *Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization.* Nature methods, vol. 4, no. 1, pages 27–29, 2007. (Cited on page 58.)

[Scheres 2012a] Sjors H. W. Scheres. *A Bayesian view on cryo-EM structure determination.* Journal of Molecular Biology, vol. 415, no. 2, pages 406–18, jan 2012. (Cited on pages 13, 37, 38, 68, 71 and 87.)

[Scheres 2012b] Sjors H. W. Scheres. *RELION: Implementation of a Bayesian approach to cryo-EM structure determination.* Journal of Structural Biology, vol. 180, no. 3, pages 519–530, dec 2012. (Cited on pages 37, 50 and 84.)

[Scheres 2014] Sjors H w Scheres. *Beam-induced motion correction for submegadalton cryo-EM particles.* eLife, vol. 3, page e03665, 2014. (Cited on page 11.)

[Schmidli *et al.* 2019] Claudio Schmidli, Stefan Albiez, Luca Rima, Ricardo Righetto, Inayatulla Mohammed, Paolo Oliva, Lubomir Kovacik, Henning Stahlberg and Thomas Braun. *Microfluidic protein isolation and sample preparation for high resolution cryo-EM.* bioRxiv, page 556068, jan 2019. (Cited on page 89.)

[Schmidt-Krey & Cheng 2013] Ingeborg Schmidt-Krey and Yifan Cheng. Electron Crystallography of Soluble and Membrane Proteins: Methods and Protocols. Humana Press, 2013. (Cited on pages 14, 59 and 60.)

[Schorb *et al.* 2018] Martin Schorb, Isabella Haberbosch, Wim Hagen, Yannick Schwab and David Mastronarde. *Software tools for automated transmission electron microscopy.* bioRxiv, page 389502, 2018. (Cited on pages 83 and 87.)

[Schur *et al.* 2016] Florian K M Schur, Martin Obr, Wim J H Hagen, William Wan, Arjen J Jakobi, Joanna M Kirkpatrick, Carsten Sachse, Hans-Georg Kräusslich and John A G Briggs. *An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation.* Science, vol. 353, no. 6298, pages 506 LP – 508, jul 2016. (Cited on page 14.)

[Sherman *et al.* 1998] Michael B Sherman, Toshinori Soejima, Wah Chiu and Marin van Heel. *Multivariate analysis of single unit cells in electron crystallography.* Ultramicroscopy, vol. 74, no. 4, pages 179–199, sep 1998. (Cited on page 47.)

[Sigworth *et al.* 2010] Fred J Sigworth, Peter C Doerschuk, Jose-Maria Carazo and Sjors H W Scheres. *Chapter Ten - An Introduction to Maximum-Likelihood Methods in Cryo-EM.* In Grant J Jensen B T Methods in Enzymology, editeur, Cryo-EM, Part B: 3-D Reconstruction, volume Volume 482, pages 263–294. Academic Press, 2010. (Cited on pages 12 and 58.)

[Sigworth 1998] F J Sigworth. *A maximum-likelihood approach to single-particle image refinement.* Journal of structural biology, vol. 122, no. 3, pages 328–39, jan 1998. (Cited on page 12.)

[Sindelar & Grigorieff 2012] Charles V Sindelar and Nikolaus Grigorieff. *Optimal noise reduction in 3D reconstructions of single particles using a volume-normalized filter.* Journal of structural biology, may 2012. (Cited on pages 50, 97 and 101.)

[Stahlberg *et al.* 2015] Henning Stahlberg, Nikhil Biyani and Andreas Engel. *3D reconstruction of two-dimensional crystals.* Archives of Biochemistry and Biophysics, vol. 581, pages 68–77, sep 2015. (Cited on pages 14, 36, 37, 40, 57 and 60.)

[Stansfeld *et al.* 2015] Phillip J. Stansfeld, Joseph E. Goose, Martin Caffrey, Elisabeth P. Carpenter, Joanne L. Parker, Simon Newstead and Mark S.P. Sansom. *MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes.* Structure, vol. 23, no. 7, pages 1350–1361, jul 2015. (Cited on page 5.)

[Stewart & Grigorieff 2004] Alex Stewart and Nikolaus Grigorieff. *Noise bias in the refinement of structures derived from single particles.* Ultramicroscopy, vol. 102, no. 1, pages 67–84, dec 2004. (Cited on pages 37 and 50.)

[Subramanian *et al.* 2018] Rohit H. Subramanian, Sarah J. Smith, Robert G. Alberstein, Jake B. Bailey, Ling Zhang, Giovanni Cardone, Lauri Suominen, Mohamed Chami, Henning Stahlberg, Timothy S. Baker and F. Akif Tezcan. *Self-Assembly of a Designed Nucleoprotein Architecture through Multimodal Interactions.* ACS Central Science, vol. 4, pages 1578–1586, 2018. (Cited on page 57.)

[Suzuki *et al.* 2016] Yuta Suzuki, Giovanni Cardone, David Restrepo, Pablo D. Zavattieri, Timothy S. Baker and F. Akif Tezcan. *Self-assembly of coherently dynamic, auxetic, two-dimensional protein crystals.* Nature, vol. 533, no. 7603, pages 369–373, 2016. (Cited on pages 37, 46 and 89.)

[Tegunov & Cramer 2018] Dimitry Tegunov and Patrick Cramer. *Real-time cryo-EM data pre-processing with Warp.* bioRxiv, jan 2018. (Cited on page 89.)

[Terwilliger *et al.* 2018a] Thomas C. Terwilliger, Paul D. Adams, Pavel V. Afonine and Oleg V. Sobolev. *A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps.* Nature Methods, vol. 15, no. 11, pages 905–908, nov 2018. (Cited on page 89.)

[Terwilliger *et al.* 2018b] Thomas C Terwilliger, Oleg V Sobolev, Pavel V Afonine and Paul D Adams. *Automated map sharpening by maximization of detail and connectivity.* Acta Crystallographica Section D, vol. 74, no. 6, jun 2018. (Cited on pages 50 and 98.)

[Thon 1966] F. Thon. *Zur Defokussierungsabhängigkeit des Phasenkontrastes bei der elektronenmikroskopischen Abbildung*, 1966. (Cited on page 9.)

[Tombola *et al.* 2006] Francesco Tombola, Medha M. Pathak and Ehud Y. Isacoff. *How Does Voltage Open an Ion Channel?* Annual Review of Cell and Developmental Biology, vol. 22, no. 1, pages 23–52, 2006. (Cited on page 47.)

[Tonggu & Wang 2018] Lige Tonggu and Liguo Wang. *Broken symmetry in the human BK channel.* bioRxiv, page 494385, jan 2018. (Cited on page 89.)

[van Heel & Frank 1981] Marin van Heel and Joachim Frank. *Use of Multivariate Statistics in Analysing the Images of Biological Macromolecules.* Ultramicroscopy, vol. 6, pages 187–194, 1981. (Cited on page 12.)

[van Heel & Hollenberg 1980] M van Heel and J Hollenberg. *On the Stretching of Distorted Images of Two-Dimensional Crystals.* In Wolfgang Baumeister and Wolrad Vogell, editeurs, Electron Microscopy at Molecular Dimensions: State of the Art and Strategies for the Future, pages 256–260. Springer Berlin Heidelberg, Berlin, Heidelberg, 1980. (Cited on pages 37 and 57.)

[van Heel & Schatz 2005] Marin van Heel and Michael Schatz. *Fourier shell correlation threshold criteria.* Journal of structural biology, vol. 151, no. 3, pages 250–62, sep 2005. (Cited on pages 13, 41, 50 and 97.)

[van Heel *et al.* 1996] Marin van Heel, George Harauz, Elena V. Orlova, Ralf Schmidt and Michael Schatz. *A new generation of the IMAGIC image processing system.* Journal of structural biology, vol. 116, no. 1, pages 17–24, 1996. (Cited on page 57.)

[van Heel *et al.* 2000] Marin van Heel, B Gowen, R Matadeen, E V Orlova, R Finn, T Pape, D Cohen, H Stark, R Schmidt, M Schatz and A Patwardhan. *Single-particle electron cryo-microscopy: towards atomic resolution.* Quarterly reviews of biophysics, vol. 33, no. 4, pages 307–69, nov 2000. (Cited on page 12.)

[van Heel *et al.* 2012] Marin van Heel, Rodrigo Portugal, A Rohou, C Linnemayr, C Bebeacua, R Schmidt, T Grant, M Schatz, M Van Heel, Rodrigo Portugal, A Rohou, C Linnemayr, C Bebeacua, R Schmidt, T Grant and M Schatz. *Four-dimensional cryo-electron microscopy at quasi-atomic resolution: IMAGIC 4D.* International Tables for Crystallography, vol. F., pages 624–628, 2012. (Cited on page 13.)

[van Heel *et al.* 2016] Marin van Heel, Rodrigo V. Portugal, Micheal Michael Schatz, Marin Van Heel, Micheal Michael Schatz, Marin van Heel, Rodrigo V. Portugal and Micheal Michael Schatz. *Multivariate Statistical Analysis of Large Datasets: Single Particle Electron Microscopy.* Open Journal of Statistics, vol. 6, no. 4, pages 701–739, 2016. (Cited on page 12.)

[Wade 1992] R. H. Wade. *A brief look at imaging and contrast transfer.* Ultramicroscopy, vol. 46, no. 1-4, pages 145–156, 1992. (Cited on page 9.)

[Watson & Crick 1953] James D Watson and Francis H C Crick. *Molecular structure of nucleic acids.* Nature, vol. 171, no. 4356, pages 737–738, 1953. (Cited on page 7.)

[Welch 1984] Welch. *A Technique for High-Performance Data Compression.* Computer, vol. 17, no. 6, pages 8–19, 1984. (Cited on page 24.)

[White 2009] Stephen H White. *Biophysical dissection of membrane proteins.* Nature, vol. 459, page 344, may 2009. (Cited on pages 5 and 6.)

[Williams & Carter 2009] David B Williams and C Barry Carter. Transmission Electron Microscopy: a Textbook for Materials Science. Springer, 2009. (Cited on pages 8 and 11.)

[Wlodawer & Dauter 2017] Alexander Wlodawer and Zbigniew Dauter. *'Atomic resolution': a badly abused term in structural biology.* Acta Crystallographica Section D Structural Biology, vol. 73, no. 4, pages 381–383, apr 2017. (Cited on page 3.)

[Word *et al.* 1999] J.Michael Word, Simon C Lovell, Jane S Richardson and David C Richardson. *Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation11Edited by J. Thornton.* Journal of Molecular Biology, vol. 285, no. 4, pages 1735–1747, 1999. (Cited on page 51.)

[Zhang & Hong Zhou 2011] Xing Zhang and Z Hong Zhou. *Limiting factors in atomic resolution cryo electron microscopy: No simple tricks.* Journal of Structural Biology, vol. 175, no. 3, pages 253–263, sep 2011. (Cited on page 14.)

[Zhang *et al.* 2010] Xing Zhang, Lei Jin, Qin Fang, Wong H Hui and Z Hong Zhou. *3.3 Å Cryo-EM Structure of a Nonenveloped Virus Reveals a Priming Mechanism for Cell Entry.* Cell, vol. 141, no. 3, pages 472–482, 2010. (Cited on page 11.)

[Zhang 2016] Kai Zhang. *Gctf: Real-time CTF determination and correction.* Journal of Structural Biology, vol. 193, no. 1, pages 1–12, jan 2016. (Cited on page 60.)

[Zheng *et al.* 2017] Shawn Q Zheng, Eugene Palovcak, Jean-Paul Armache, Kliment A Verba, Yifan Cheng and David A Agard. *MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy.* Nat Meth, vol. advance on, feb 2017. (Cited on pages 11, 23, 38, 40, 49 and 83.)

[Zhou *et al.* 2015] Jing Zhou, Agnieszka Bronowska, Johanne Le Coq, Daniel Lietha and Frauke Gräter. *Allosteric Regulation of Focal Adhesion Kinase by PIP2 and ATP.* Biophysical Journal, vol. 108, no. 3, pages 698–705, feb 2015. (Cited on pages 75 and 79.)

[Zivanov *et al.* 2018] Jasenko Zivanov, Takanori Nakane, Björn O Forsberg, Dari Kimanius, Wim JH Hagen, Erik Lindahl and Sjors HW Scheres. *New tools for automated high-resolution cryo-EM structure determination in RELION-3.* eLife, vol. 7, page e42166, nov 2018. (Cited on pages 58 and 87.)

[Zivanov *et al.* 2019] Jasenko Zivanov, Takanori Nakane and Sjors H W Scheres. *A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis.* IUCrJ, vol. 6, no. 1, jan 2019. (Cited on page 11.)