Development and analysis of image sorting methods for single-particle cryo-electron microscopy

Dissertation

presented for the degree of
Dr. rer. nat.

to the Faculty of Mathematics, Informatics and Natural Sciences
University of Hamburg

submitted to the Department of Informatics by

Anna Theresa Cavasin
born in Essen

Hamburg, 24.03.2025

1. Reviewer: Prof. Dr. Matthias Rarey
2. Reviewer: Prof. Dr. Holger Stark
3. Reviewer: Dr. Ashwin Chari
Date of thesis defense: 27.10.2025

Abstract

Single-particle cryo-electron microscopy (cryo-EM) has evolved into a standard method for determining the 3D structure of proteins. As a single-particle technique, it allows for the elucidation of large macromolecular complexes, provides information on protein dynamics, and gives access to proteins that are difficult to crystallize. For this purpose, molecules in aqueous solution are rapidly frozen and then analyzed by transmission electron microscopy. The resulting 2D images are processed in computationally intensive pipelines to finally reconstruct a 3D density map which can be used for building an atomic model. As a result of the low signal-to-noise ratio in the images, thousands to millions of 2D projection views are necessary to reconstruct a single density map. These images can be of varying quality due to several reasons that include beam-induced motion, optical aberrations and sample heterogeneity. Thus, sorting procedures are required to create high-quality datasets.

In this thesis, new approaches for image sorting in cryo-EM were developed and analyzed. First, the potential of quality parameters derived from the reconstruction workflow was assessed. To this end, the final dataset was divided into subsets in an automated fashion and the gold-standard FSC resolutions achieved for reconstructions from these subsets were compared to the expected resolutions from a Rosenthal-Henderson plot. Consecutively, image filtering based on the most promising quality parameters was carried out using single-parameter filtering as well as combined methods. As a second approach, a new filtering parameter based on the consistency of determining the 3D orientation of the 2D images was proposed. For each image, orientations were computed in three different ways including reference-based alignment as well as reference-free angular reconstitution. The average distance between the resulting three orientations was computed with a unit quaternion-based distance measure under consideration of structural symmetry. Finally, a new filtering approach is presented, which bins all images based on their assigned projection before filtering

within each bin, leading to more uniform orientation distributions. The methods were validated on simulated data and a high-resolution dataset of *S. cerevisiae* fatty acid synthase.

Kurzfassung

Die Einzelpartikel-Kryoelektronenmikroskopie (Kryo-EM) hat sich zu einer Standardmethode für die Bestimmung der 3D-Struktur von Proteinen entwickelt. Sie ermöglicht die Analyse großer makromolekularer Komplexe, liefert Informationen zu Proteindynamik und gewährt Zugang zu Proteinen, die aufgrund schlechter Kristallisierbarkeit für die Röntgenstrukturanalyse ungeeignet sind. Zu diesem Zweck werden einzelne Proteine in wässriger Lösung in kürzester Zeit eingefroren und dann mittels Transmissionselektronenmikroskopie analysiert. Die resultierenden 2D-Bilder werden durch rechenintensive Verfahren prozessiert, um eine 3D-Dichtekarte zu rekonstruieren, die zur Generierung eines atomaren Strukturmodells verwendet werden kann. Aufgrund des schlechten Signal/Rausch-Verhältnisses der Bilder müssen mehrere Tausend bis mehrere Millionen von ihnen kombiniert werden, um eine einzelne 3D-Dichtekarte zu erhalten. Die 2D-Bilder können dabei – bedingt durch multiple Faktoren wie strahlinduzierte Bewegungen, Aberrationen im Mikroskop und strukturelle Heterogenität der Probe – unterschiedliche Qualität aufweisen. Daher sind Sortierungsmethoden notwendig, um hochqualitative Bilddatensätze zu erzeugen.

In dieser Dissertation wurden neue Methoden zur Sortierung von Partikelbildern entwickelt und analysiert. Zunächst wurde das Potential von Qualitätsparametern untersucht, welche aus der Prozessierung der Kryo-EM-Bilddaten abgeleitet sind. Hierfür wurde der finale Bilddatensatz anhand der Parameter in Teildatensätze aufgeteilt und die Auflösung der hieraus rekonstruierten Dichtekarten mit der erwarteten Auflösung gemäß eines Rosenthal-Henderson-Plots verglichen. Anschließend wurde der Datensatz anhand der vielversprechendsten Qualitätsparameter gefiltert, sowohl einzeln per Parameter als auch mittels kombinierter Methoden. Als zweiter Ansatz wurde ein neuer Filterparameter entwickelt, der auf der konsistenten Bestimmung von 3D-Orientierungen basiert. Für jedes Bild wird hierbei die 3D-Orientierung auf drei verschiedene Arten bestimmt, wobei sowohl referenzbasierte als auch referenzfreie Algorithmen verwendet werden. Die durchschnittliche Distanz der drei Orientierungen

wird danach durch eine quaternionenbasierte Distanzmetrik unter Berücksichtigung der strukturellen Symmetrie des Proteins berechnet. Abschließend wurde in dieser Arbeit eine neue Filtermethode konstruiert, bei der alle Bilder zunächst nach ihrer Projektionsrichtung gruppiert werden. Daraufhin wird das Filtern innerhalb der Gruppen ausgeführt, was zu einer gleichmäßigeren Orientierungsverteilung des gefilterten Datensatzes führt. Die Methoden wurden mithilfe von simulierten Daten sowie mit einem hochauflösenden Datensatz der Fettsäure-Synthase aus der Hefe *S. cerevisiae* evaluiert.

Contents

1. Introduction

1.1. Resolution limiting factors in cryo-EM

In the last decade, single-particle cryo-electron microscopy (cryo-EM) has evolved into a high resolution structure elucidation technique due to numerous advances in sample preparation methods, microscope hardware and data processing software, as summarized in [1, 2]. Figure 1.1 shows the resolution distribution of the density maps in the Electron Microscopy Data Bank (EMDB) [3]. Of the currently around 40,000 entries, more than half have a resolution below 4 Å [4], making secondary structure elements visible and *de novo* model building possible [5].

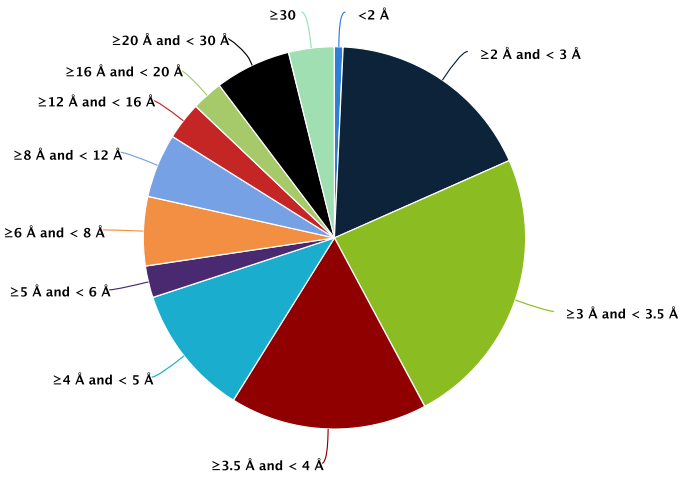


Figure 1.1.: Resolution distribution of released entries in the EMDB. Image created with the EMDB Chart Builder [6] based on [4].

In 2020, true atomic resolution of around 1.2 Å was reported by two groups [7, 8] in independent proof-of-concept studies on apoferritin. At this resolution, individual atom positions, single-atomic chemical modifications, and even density for hydrogen atoms could be observed. With the increase in resolution capacity, cryo-EM has significantly gained relevance for application in structure-based drug design [9, 10, 11], granting access to large macromolecular complexes and membrane proteins like ion channels and G protein-coupled receptors (GPCRs), which are difficult to analyze by X-ray crystallography. In recent studies, cryo-EM has been used for solving protein-ligand complexes of the human CDK-activating kinase with a series of inhibitors [12] and even for fragment-based drug discovery [13].

While the resolution achievable by cryo-EM has substantially increased, only 0.7 percent of the maps reach the very high resolution regime of below 2 Å [4]. Beyond this threshold, individual amino acids and chemical transformations can directly be visualized [2], opening up new possibilities to understand biochemical mechanisms and observing interactions relevant for computational drug design. These prospects make high resolution cryo-EM a desirable goal. In practice, there are however a multitude of factors that make true atomic resolution difficult to achieve.

In a typical cryo-EM project (see figure 1.2), molecules in an aqueous solution are rapidly frozen in a thin layer and then analyzed by transmission electron microscopy. This produces thousands to millions of very noisy 2D projection views of the protein of interest, which are then processed in computationally intensive pipelines to finally reconstruct a single consensus density map or a discrete ensemble of a few maps representing different structural states. Here, one must keep in mind that the thousands of 2D projections originate from individual instances of the macromolecule. These projection images therefore not only represent different projection directions but also show structural heterogeneity. There might be contamination by macromolecules other than the one to be imaged, macromolecular complexes might be missing subunits (compositional heterogeneity), and atom positions may vary due to the dynamic movement of macromolecules at ambient temperature in solution (conformational heterogeneity). Measuring conformational dynamics is in fact considered as one of the great opportunities of cryo-EM. Classification approaches are well-established in separating discrete conformers [15], and other approaches are currently developed to analyze the dynamic behavior in a continuous way [16, 17, 18]. However, discrete classification has limitations and residual structural heterogene-

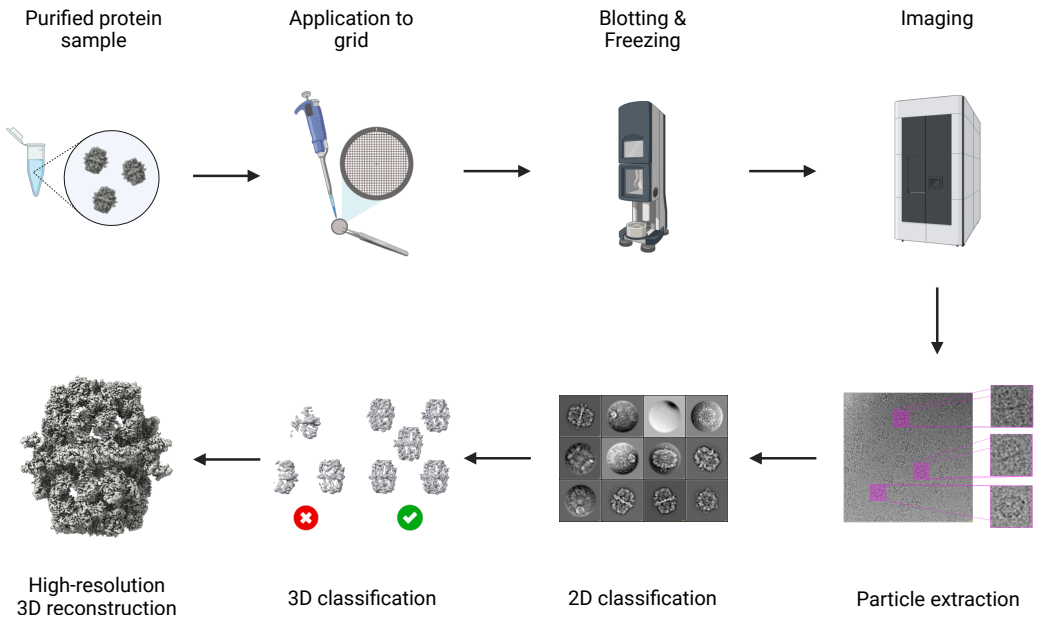


Figure 1.2.: Overview of a typical cryo-EM workflow. A drop of purified protein solution containing many copies of the structure of interest is applied onto a metal grid with a holey film. After blotting of excess liquid and rapid freezing, the specimen is inserted into the electron microscope for imaging. Individual 2D particle images are extracted from the resulting 2D micrograph images. The particle images are filtered by 2D and 3D classification before a 3D density map is reconstructed from them. [14]

ity might remain within one class, causing uncertainty in the atom positions of the consensus reconstruction.

Additional limitations occur in the process of sample preparation and data collection. During sample preparation, the macromolecule in solution is first applied onto a metal grid with a perforated foil. Then, excess sample is blotted to create a thin layer and the grid is rapidly plunge-frozen in liquid ethane to embed the sample in vitreous ice [19]. In the process, some proteins might come into contact with hydrophobic surfaces, i.e., the air-water interface or the support, and consequently denature [19, 20]. Another problem is that particles can have a preferred orientation in the grid plane, which leads to an uneven distribution of views of the 3D structure and can hinder the reconstruction procedure [21, 22]. Contamination of the grids can occur in the form of

dust or evaporated carbon and is counteracted by performing cleaning protocols [19]. An important quality factor of vitrified specimen is the thickness of the ice layer containing the proteins. Studies have shown that thicker ice can have a limiting effect on resolution, presumable due to an increase in elastic scattering events [23]. Finally, insufficient rapid freezing and devitrification can lead to hexagonal or cubic ice crystal formation in the specimen and frost ice crystals arising from atmospheric water vapor can build up on the vitrified grids [24].

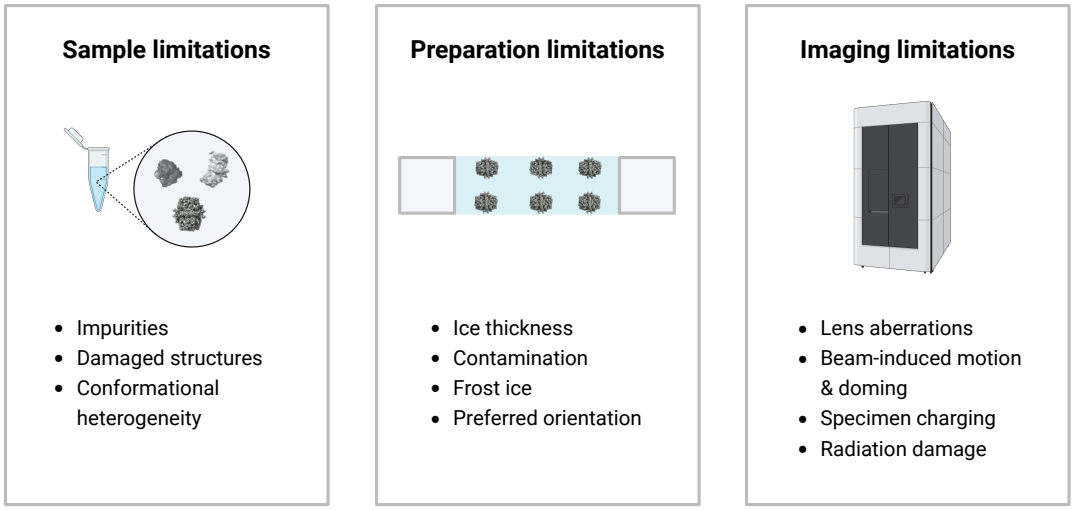


Figure 1.3.: Summary of typical cryo-EM limitations related to sample, preparation and imaging. [25]

The imaging process in the transmission electron microscope encompasses several limitations, which are mainly related to lens aberrations and have varying importance over different resolution ranges [2]. While defocus and two-fold astigmatism are generally limiting, coma and higher-order aberrations become increasingly important when aiming for atomic resolution [2, 26]. Furthermore, chromatic aberration needs to be limited by improving the temporal coherence of the electron beam [2, 7, 8]. Another factor that comes into play at high resolution is Ewald sphere curvature [27]. In the history of cryo-EM, custom solutions have been developed to account for these limitations both on the hardware [7, 8, 2] and software [28, 29, 26, 27] side.

Besides the microscope-related challenges, limitations of the imaging process in cryo-EM are caused by the interaction of the ionizing electron beam with the specimen.

Glaeser summarized several resulting effects in a comprehensive review [30]. A well-known aspect is that the electron beam causes radiation damage that accumulates during the exposure and leads to a loss of high-resolution features. Another effect is the blurring of the image by beam-induced motion that is presumably caused by radiation-induced relief of mechanical stress and generation of additional stress by radiolysis. In order to compensate for the lateral beam-induced motion in the image plane, computational tools have been developed that align the movie frames generated by a direct electron camera for each recorded micrograph (e.g. [31]). These tools also allow for the weighting of the individual frames according to the applied electron dose to compensate for radiation damage. Beam-induced motion also appears in the form of doming or bulging perpendicular to the image plane which causes a z height change in the sample [32], leading to a defocusing of the sample and thus image blurring. As an additional contribution, blurring can be caused by electric charging of the specimen, leading to the deflection of electrons during image formation. The different cryo-EM limitations discussed in this section are summarized in figure 1.3.

1.2. State-of-the-art methods for particle sorting

As a result of the limitations, the individual particle images can largely differ in their quality and it is necessary to eliminate images with structural defects, artifacts, or substantial blurring before performing 3D reconstruction. Several approaches have been developed to address this task. A conceptual summary of the most important method types and targets is given in figure 1.4. A first selection can already be carried out on the aligned and averaged micrograph images, which are the recorded outputs from the microscope and show a region of the specimen optimally containing multiple protein particles. In practice, this selection is often performed in the form of visual inspection or by excluding outliers in terms of common quality parameters like the maximum resolution of the contrast transfer function (CTF) fit. The first opportunity to perform a selection on the particle level is when clipping the individual particle images from the micrograph images (particle picking). The direct identification of high quality particles is, however, a difficult task due to the extremely low signal-to-noise-ratio of the images resulting from the necessary electron dose limitation imposed by the damaging nature of the electron beam. It is therefore common practice to use a rather unspecific picking approach that identifies particles by comparing image

positions to a reference that shows the rough shape of the expected particle of interest (e.g. Gautomatch [33]) and then leave the particle sorting – the separation of good and bad particles – to the consecutive step in the cryo-EM processing workflow. For an overview of the processing pipeline, the reader is referred to section 2.1.2.

The standard way to select good particle images after picking is through 2D classification and 3D classification, where particles are clustered into a fixed number of classes based on cross-correlation similarity [15]. Only the particles corresponding to class averages showing high resolution features will be kept while the particles belonging to blurred out class averages will be discarded. This procedure comes with a number of disadvantages. Firstly, it is inherently subjective as the selection of good and bad classes is performed by manual user interaction. Secondly, it is computationally costly as scores are calculated for many combinations of rotation, translation and class. Thirdly, it does not always provide a sufficient separation of good and bad particles, as some particles might be assigned to classes that do not represent their quality. In practice, biologists often experience a trade-off between keeping good particles and excluding bad ones.

In order to improve the standard reconstruction procedure, several approaches have been developed that aim at improving the initial dataset before classification, automating the selection process or even replacing the 2D/3D classification scheme. A fast and relatively easy approach is the Z-Score filtering implemented in RELION [34]. As a first step, basic image features like mean, standard deviation, skewness and kurtosis of pixel values are calculated for each image after subtraction of an aligned template. For each feature and difference image, a Z-Score is calculated as

$$Z = \frac{x - \mu}{\sigma} \tag{1.1}$$

where μ and σ are the empirical mean and standard deviation over all images for this particular feature, respectively. The particles images are sorted by their average Z-Scores over all features and presented to the user for selection. The hope is that good particles concentrate at the first list positions and the user can decide to discard the last particles at some point without having to visually inspect them. Vargas *et al.* [35] proposed a variant of Z-Score sorting that makes use of more sophisticated features which are combined by principal component analysis. The Z-Score above is replaced by a score based on the Mahalanobis distance and a fixed threshold is pro-

posed to avoid the need for user intervention. The machine-learning-based MAPPOS approach [36] automatically selects particles with a bagging approach that combines the results of elementary classifiers, such as decision trees and k-nearest neighbors (KNN) classifiers. The images are represented by seven calculated features. The approach has to be trained by the user on approximately 1000 manually selected examples before it can be applied to the entire dataset. While the approaches above are computationally cheaper than 2D classification, they are often not adopted in practice, presumably because they do not yield substantially better results than 2D classification and – in the case of MAPPOS and RELION Z-Scoring – require user intervention.

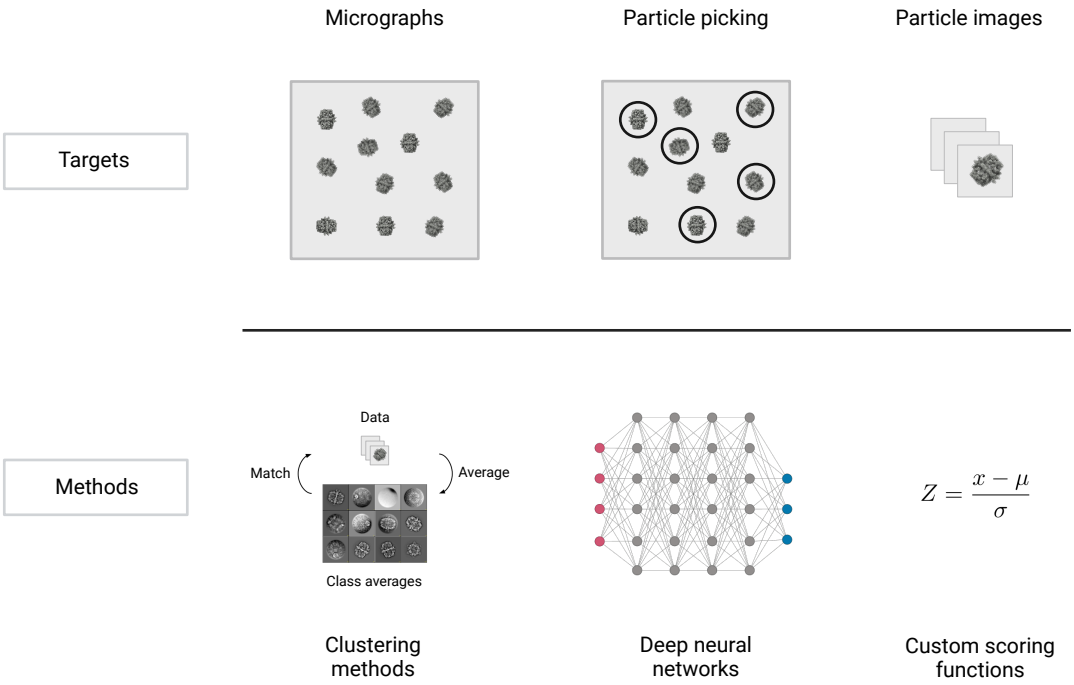


Figure 1.4.: Summary of targets and methods for the cryo-EM filtering approaches discussed in this section. The approaches are applied on full micrographs, during the particle extraction from the micrographs (picking) or on the extracted particle images. The filtering methods are based on clustering, deep neural networks or custom scoring functions that were specifically tailored for cryo-EM applications. Note that the micrograph and particle images are schematic for better visualization and do not accurately represent the usually much higher noise level in the data. [37]

In recent years, deep learning approaches have become increasingly popular which aim at automating the data processing pipeline [38]. Multiple particle pickers based on neural networks [39, 40] have been published to improve the quality of the particles selected from the micrographs and avoid picking artifacts that show contamination rather than actual macromolecules. Other methods use convolutional neural networks (CNNs) for automated micrograph and 2D class average selection [41, 42, 43], or to segment micrographs into regions suitable and unsuitable for picking e.g. due to carbon contamination [44]. The segmentation results can be used for the filtering of picking results by discarding particles that were picked in unfavorable areas. The approaches were all trained on manually curated training datasets. In the *DeepConsensus* approach [45], a CNN is used to combine the results from multiple particle picking algorithms applied on the same input data. The network is trained with the particles picked by all pickers as positive examples and picks from random coordinates that were not chosen by any pickers, optionally enriched by contamination examples showing e.g. ice or carbon, as negative examples. After training, the network is used for classification of the union of picking results from all pickers.

An example for a modification of 3D classification is a procedure proposed by Gong *et al.* [46]. Using a modified version of RELION, they perform iterative 3D classification with two classes, of which the first one is the unmodified reconstruction and the other one a 3D map where the phases were randomized above a specific resolution. After each iteration, particles assigned to the modified class are discarded. The resolution limit for randomization is gradually increased over the iterations. An approach by Zhou *et al.* [47] seeks to completely omit 2D and 3D classification. Instead, an iterative procedure is proposed that combines 3D refinement and particle sorting. The input is a raw candidate particle set. In a projection matching step, the particles are compared to 2D projections of a reference structure to determine the alignment parameters. At the same time, each particle is assigned a cross-correlation score describing the quality-of-fit. For the first round, 2D projections are generated from a given reference map, originating, for instance, from a stochastic gradient descent ab initio method [48, 49]. Zhou *et al.* [47] show that their quality-of-fit score has a distinct distribution over the particles and define a threshold value by fitting a bimodal mixture of Gaussians model to the data. The equal probability point between the Gaussians is defined as the threshold and for the following 3D reconstruction step, only particles with scores above that threshold are used. This procedure is iterated until convergence. In contrast to the supervised 2D/3D classification strategy, the approach was able to

successfully reconstruct maps without any user intervention. It was, however, only applied for reconstructing a single map and conformational heterogeneity was not considered.

Méndez *et al.* [50] proposed a method to estimate the reliability of particle orientation assignments using a graph signal processing approach. The method can be applied after any 3D refinement to separate the dataset into two groups, which are the reliably (C) and unreliably (W) assigned particles. The authors show that refinements from subsets of the C group have better Fourier shell correlation (FSC) curves than refinements from equally sized subsets of the W group. Refinements from the entire C group showed approximately the same FSC curve as refinements of the entire unfiltered dataset. However, the exclusion of the W datasets did not lead to an improvement of the FSC curve with respect to the unfiltered dataset.

A recent approach by Zhu *et al.* [51] investigates the question of how many particles from a final stack of images are necessary for 3D reconstruction. They rigorously filter out around 70 % to 80 % of the particles for eight datasets from the EMPIAR [52] database. To this end, their method CryoSieve computes the similarity between the particle images and the assigned reference projection above a given frequency threshold that gradually increases over several iterations of filtering. The authors demonstrate that refinements from the filtered datasets yield maps of equal or better FSC resolutions than refinements from the unfiltered datasets in most cases. The reported FSC resolution improvements were, however, rather small. CryoSieve is compared to particle exclusion by the above described procedure by Zhou *et al.* [47], filtering by the cisTEM [53] per-particle score and random exclusion, using half-map and map-to-model FSC, Q-scores [54] and Rosenthal-Henderson B-factors [55] as validation measures. In another experiment, they extracted particles from groups of raw cryo-EM movie frames representing different amounts of electron exposure to simulate images with different levels of radiation damage. In comparison to the approaches by Zhou *et al.* [47], Méndez *et al.* [50], the cisTEM [53] score and RELION [56] classification without alignment, CryoSieve retained the most particles derived from early movie frames representing lower levels of radiation damage while discarding the largest amount of particles from later movie frames.

In conclusion, a number of approaches have been proposed who aim at altering the state-of-the-art image processing scheme (figure 2.4), where particle sorting is mainly

carried out by subjective and computationally expensive 2D and 3D classification. Z-Score and machine learning methods based on simple image features for filtering after particle picking are rarely used in practice, presumably due to a bad trade-off between performance and user intervention. Deep neural networks (DNNs) are becoming increasingly popular for improved particle picking and automated masking of unsuitable picking areas, leading to better initial datasets and a reduced need for particle sorting later on. At the same time, there is a trend towards a fully automatized reconstruction procedure by using DNNs for classification tasks [41, 42, 43]. In the recent (non-DNN) approach by Zhou *et al.* [47], 2D/3D classification was even completely omitted. Recent image sorting methods [50, 51] were able to reduce the number of particles in processed cryo-EM datasets, but the reported FSC resolution improvements were mostly rather small. Deep learning methods will likely play an important role for making cryo-EM more accessible through automated procedures in the future and help to address open challenges like structural heterogeneity [17, 18, 57]. In practice, however, cryo-EM projects often have individual challenges and the custom workflows with manual intervention are currently the method of choice when pushing resolution boundaries [7, 8].

1.3. New particle sorting methods based on metadata and orientation consistency

The aim of the research project presented in this thesis was to develop and analyze new methods for particle image sorting. The first idea explored here was to assess the potential of metadata from the cryo-EM reconstruction workflow for particle filtering. To this end, a dataset of fatty acid synthase (FAS) was processed in state-of-the-art manner while storing and calculating quality parameters during different steps of the processing pipeline. To assess the usefulness of the parameters, the final image set was divided into subsets based on metadata value ranges and the gold-standard FSC resolutions achieved for reconstructions from these subsets were compared to the expected resolutions from a Rosenthal-Henderson plot [55]. Consecutively, it was investigated whether the exclusion of certain particle images from the whole dataset leads to an improvement of the overall FSC resolution. To this end, both single-parameter filtering and methods which combine multiple parameters were applied.

The idea of using metadata for particle filtering is widespread in the cryo-EM community. As mentioned in the previous section, it is for example common practice to exclude micrographs for which the maximum resolution of the CTF fit is low. The use of metadata was previously investigated by Stagg *et al.* [58]. By excluding particles based on certain metadata they were not able to improve the resolution, but could reduce the number of required particles to reach a certain resolution. The approach in this thesis is similar to the one by Stagg *et al.* in the sense that it compares resolutions from metadata subsets to expected resolutions and uses metadata for filtering. However, instead of their newly introduced ResLog plots, the approach here uses the more common Rosenthal-Henderson plot, contains a way to automatically determine metadata subsets and outliers from the data distribution, and – most importantly – investigates different quality parameters that only become available during data processing.

As a second approach, a new particle scoring method based on consistency of particle orientations when determined in different ways was developed. In this approach, particle orientations are calculated by reference-based alignment as well as reference-free angular reconstitution [59]. The underlying concept is that the particles with consistent orientations are less likely to introduce errors into the reconstruction. The newly developed method uses unit quaternions to calculate orientation distances and takes structural symmetry into account. It was assessed on simulated data as well as on the FAS dataset from the metadata approach.

As a third approach, a method for filtering particles with respect to their assigned projection direction is presented. Here, particle images are divided into discrete direction bins at a user-defined HEALPix [60] level. Filtering is then carried out within each bin, either by discarding a certain percentage of the worst scoring particles or by determining outliers based on boxplot statistics. The method was assessed on the FAS dataset in comparison to standard value-based filtering and random particle exclusion as a baseline.

Finally, the methods developed in this thesis were compared to the state-of-the-art filtering method CryoSieve [51] in terms of FSC resolution and Q-scores [54]. It was further analyzed whether the application of CryoSieve filtering would be feasible on data subsets to enable on-the-fly processing while recording the data. Additionally,

the relationship between per-subset filtering rates and subset quality in terms of FSC resolution was examined.

2. Theoretical and practical background

2.1. Cryo-electron microscopy (cryo-EM)

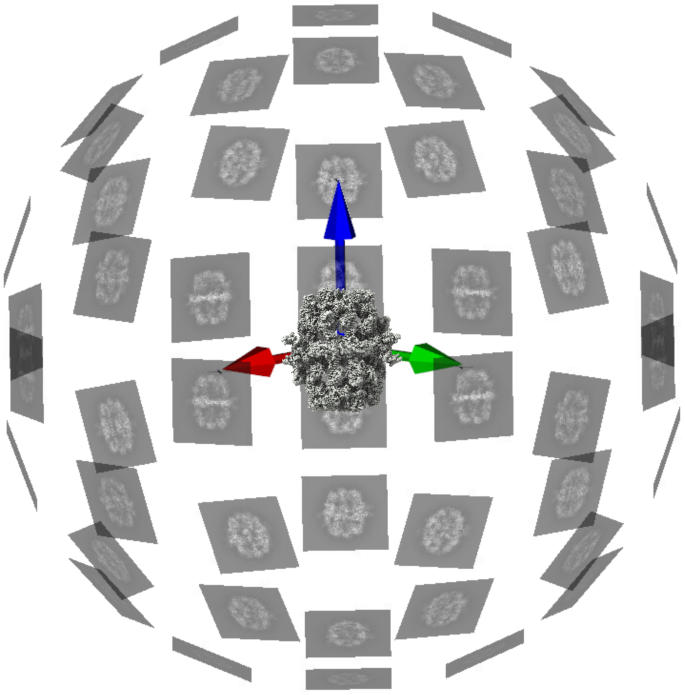


Figure 2.1.: A 3D cryo-EM map of the fatty acid synthase surrounded by the corresponding 2D projection images. Image created with the COW halo viewer [61].

Cryo-EM solves the inverse problem of reconstructing the 3D structure of an object of interest from 2D projection images of several instances of that object. The 2D projec-

tion images are generated in a transmission electron microscope (TEM) from carefully prepared samples. Section 2.1.1 describes how the raw image data is produced and section 2.1.2 explains how this data is processed to yield a 3D reconstruction.

2.1.1. Generation of cryo-EM data

Sample preparation

The first challenge of every cryo-EM project is the preparation of a sample in a suitable way for imaging in the microscope. In the context of high-resolution macromolecular cryo-EM, the goal is to produce a specimen that consists of a thin layer of frozen aqueous solution on a supporting material. [19] Figure 2.2 summarizes the key steps of specimen preparation. The molecule of interest, often a protein, first needs to be purified by a suitable biochemical technique so that an aqueous solution is produced that ideally only contains several instances of this specific molecule and is free from contamination. A small amount of this solution is applied onto a metal grid with a holey film of e.g. carbon or gold. A typical way to do this is pipetting a small drop of solution and then blotting away excess liquid so only a thin film remains on the support. Modern sample preparation techniques also allow for direct application of the right amount of solution onto the support by either spraying small droplets onto the EM grid or scribing with a hovering element over the grid [62]. Creating a thin film of the sample is crucial because the fraction of inelastically scattered electrons increases with increasing sample thickness, which leads to a deterioration of image quality [63]. The prepared grid is then vitrified by rapid plunging into liquid ethane and stored in liquid nitrogen until being inserted into a TEM for imaging. The resulting specimen is not only stable in the vacuum of the microscope, the vitreous ice embedding and cooling of the specimen to liquid nitrogen temperature also counteracts beam damage during imaging [64].

The transmission electron microscope (TEM)

After sample preparation, the frozen specimen containing the molecule of interest is inserted into a TEM where 2D projection images of areas inside the holes of the

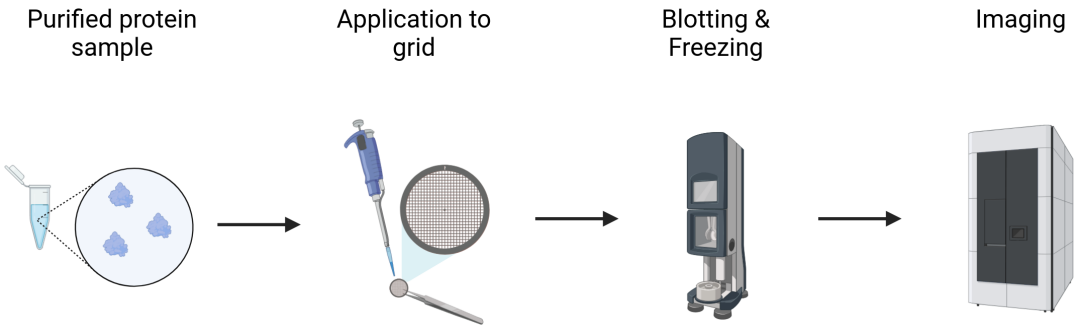


Figure 2.2.: Schematic overview of specimen preparation for cryo-EM. A drop of purified protein solution is applied onto a metal grid with a holey film. The grid is inserted into an automated device, which blots away excess liquid and rapidly freezes the grid by plunging it into liquid ethane. The frozen specimen is stored in liquid nitrogen until imaging in the electron microscope. [65]

support are produced as an electron beam passes through them in perpendicular direction. Figure 2.3 shows a schematic representation of a TEM based on the review by Orlova & Saibel [66], which also serves as the main source for the rest of this section on data generation.

The electron beam originates from an electron source, which is in the case of high-resolution cryo-EM typically a field emission gun (FEG). Electrons are extracted from a very sharp tungsten crystal tip coated with zirconium dioxide by a strong electric field and accelerated to voltages of 100-300 kV. While tips used to be heated to enable electron emission, cold FEGs operated at room temperature have been developed in recent years. These electron guns promise higher temporal coherence of the electron beam and have been successfully employed in high-resolution studies [68, 69, 8].

In order to illuminate the specimen, the diverging electron beam from the source is converted into a parallel beam by a set of condenser lenses. The cryo-EM equivalent of optical lenses are electromagnetic lenses containing copper coils, which deflect the electrons by a magnetic field [70]. After passing through the specimen, the beam is deflected by an objective lens leading to the first magnification of the image. At the back focal plane of this lens, electrons scattered at high angles are eliminated by the

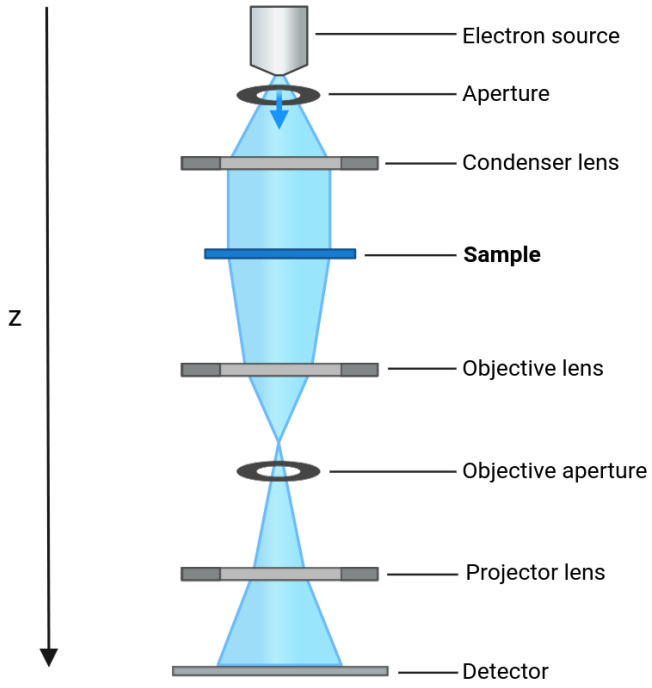


Figure 2.3.: Simplified illustration of the TEM based on the depiction by Orlova & Saibel [66]. The electron beam is shown in blue and the projection direction for image generation (z) is indicated. [67]

objective aperture before the beam passes through a set of projector lenses for full magnification.

The image is then recorded by a detector. For a long time, cryo-EM images were recorded on photographic film that was later digitized for computational analysis. This procedure was eventually replaced by digital detectors based on charge-coupled device (CCD) sensors combined with a scintillator. In this detector type, the scintillator converts incident electrons to photons, which are transmitted via fiber optics to a CCD chip that produces an electronic signal. The conversion step in the scintillator adds noise to the image, but is necessary to prevent CCD sensor damage by the high-energy electrons. Nowadays, high-end microscopes contain direct electron detectors, where electrons interact directly with a sensor without electron-photon-conversion [71]. The invention of these directors has been a landmark in the history of cryo-EM, as they not only yield better signal-to-noise ratios but also allow for

recording movies with several frames of an exposure. This enables the correction of motion occurring during the exposure by frame alignment (see also section 2.1.2), which further improves image quality.

Besides these essential building blocks, cryo-electron microscopes can contain additional components that optimize image quality. In a 2020 high-resolution study [7], a monochromator was used to improve the temporal coherence of the electron beam and a spherical aberration corrector was used to reduce axial and off-axial coma and linear magnification distortions. Another common way to improve image quality are energy filters, which remove electrons that lost energy by inelastic scattering in the sample. These electrons are focused differently by the microscope lenses and therefore lead to blurring and background noise in the image. A newly developed energy filter was employed in another 2020 high resolution study [8].

Image formation and optical aberrations

Since biological samples consist mainly of light organic atoms, the incident electrons are in general not absorbed at the atom positions, but only scattered at different angles. Therefore, the exit wave leaving the sample does not contain strong amplitude contrast, i.e., areas of strongly varying intensity which could directly be interpreted as image features. Instead, a location-dependent phase shift is introduced (phase contrast). For weak phase objects like biological specimen, which scatter the electrons by small angles and thus only introduce a small phase shift, the exit wave can be approximated by

$$\Psi_{\text{exit}}(\vec{r}) \approx \Psi_0(\vec{r})(1 + i\sigma\phi_{\text{pr}}(\vec{r})) \tag{2.1}$$

where Ψ_0 is the incident wave, $\sigma = m_{\text{e}}\lambda/(2\pi\hbar)$, m_{e} is the electron mass, λ is the electron wavelength, $\hbar = h/2\pi$, h is the Planck constant, and $\phi_{\text{pr}}(\vec{r}) = \int_{-t/2}^{+t/2} \phi_{\text{pr}}(\vec{r}, z)\mathrm{d}z$ is the projection of the specimen's electron potential along the projection direction z (see fig. 2.3).

Assuming that $|\Psi_0| = 1$ everywhere, the intensity distribution in the image corresponding to pixel values and image features can be expressed in terms of the exit wave as

$$I(\vec{r}) = \Psi_{\text{exit}}(\vec{r})\Psi^*_{\text{exit}}(\vec{r}) \approx 1 + (\sigma\phi_{\text{pr}}(\vec{r}))^2. \tag{2.2}$$

The second term in this equation is very small, which leads to very limited contrast in the image. An increase in contrast can be achieved by a 90° phase shift of the scattered part of the beam, changing equation 2.1 to

$$\Psi_{\text{exit}}(\vec{r}) \approx \Psi_0(\vec{r})(1 - \sigma\phi_{\text{pr}}(\vec{r})) \tag{2.3}$$

where, again assuming that $|\Psi_0| = 1$, the corresponding intensity distribution can be Taylor-approximated to first order as

$$I(\vec{r}) = \Psi_{\text{exit}}(\vec{r})\Psi^*_{\text{exit}}(\vec{r}) \approx 1 - 2\sigma\phi_{\text{pr}}(\vec{r}). \tag{2.4}$$

Compared to equation 2.2, the influence of the second term of the equation is increased, which signifies an increase in contrast.

Imaging in cryo-EM is subject to several imperfections such as lens aberrations and limited temporal coherence of the electron beam. The combined effect of these imperfections will become visible as blurring and signal loss in the recorded image and can be described by the impulse-response or point spread function (PSF) of the electron microscope. The microscope imperfections influence different spatial frequencies in the image to a different extent. This can be described in Fourier space by

$$F\{\Psi_{\text{obs}}(\vec{r})\} = F\{\Psi_{\text{exit}}(\vec{r})\} \cdot \text{CTF}(\vec{R}) \cdot \text{E}(\vec{R}) \tag{2.5}$$

where $\vec{R}$ is the spatial frequency. The $\text{CTF}(\vec{R})$ is the contrast transfer function of the microscope and $\text{E}(\vec{R})$ is an envelope functions that accounts for an increasing decay of contrast with higher spatial frequency. Together, they are the Fourier transform of the point spread function $F\{\text{PSF}(\vec{r})\} = \text{CTF}(\vec{R}) \cdot \text{E}(\vec{R})$. The CTF is given by $\text{CTF}(\vec{R}) = \exp(i\gamma(\vec{R}))$ with

$$\gamma(\vec{R}) = -2\pi \left(\frac{1}{2}\Delta\lambda\vec{R}^2 - \frac{1}{4}C_{\text{s}}\lambda^3\vec{R}^4 \right) \tag{2.6}$$

where C_s is the coefficient of spherical aberration and Δ the defocus of the image. γ has the effect of a phase shift based on defocus and spherical aberration on the exit wave. In practice, a variation of the image defocus can therefore induce a phase shift of the scattered part of the exit wave that leads to an increase in contrast.

An alternative method to increase image contrast is negative staining, where the macromolecules of interest are embedded in a heavy metal solution. The signal is then predominantly caused by the heavy atoms, which cover reachable areas of the macromolecules' surfaces and thereby give information on their rough size and shape. Since high-resolution details are not accessible via this method, it is primarily used for the optimization of the sample preparation process [19] rather than for high-resolution reconstructions. Besides this, phase plates have been used for single-particle cryo-EM to visualize phase contrast for some time. However, it was shown recently that the Volta phase plate [72], which was specifically designed for single-particle cryo-EM, can decrease data quality and thereby limit the resolution of reconstructions from the data [73]. Alternative phase plate designs such as laser-based phase plates are currently under development [74].

2.1.2. Data processing & Reconstruction

The aim of cryo-EM data processing is to reconstruct a single Coulomb potential density map [75] that accurately represents the 3D structure of the object of interest or to reconstruct an ensemble of such maps showing different functional states. To achieve this, three main challenges have to be met: The correction for imaging limitations, the extraction and filtering of individual high-quality particle images from the microscope micrographs, and finally, 3D reconstruction while estimating the relevant missing information on particle orientation. Figure 2.4 shows an exemplary workflow for a typical structure determination project. The individual steps are explained in the following in order of their position in the workflow.

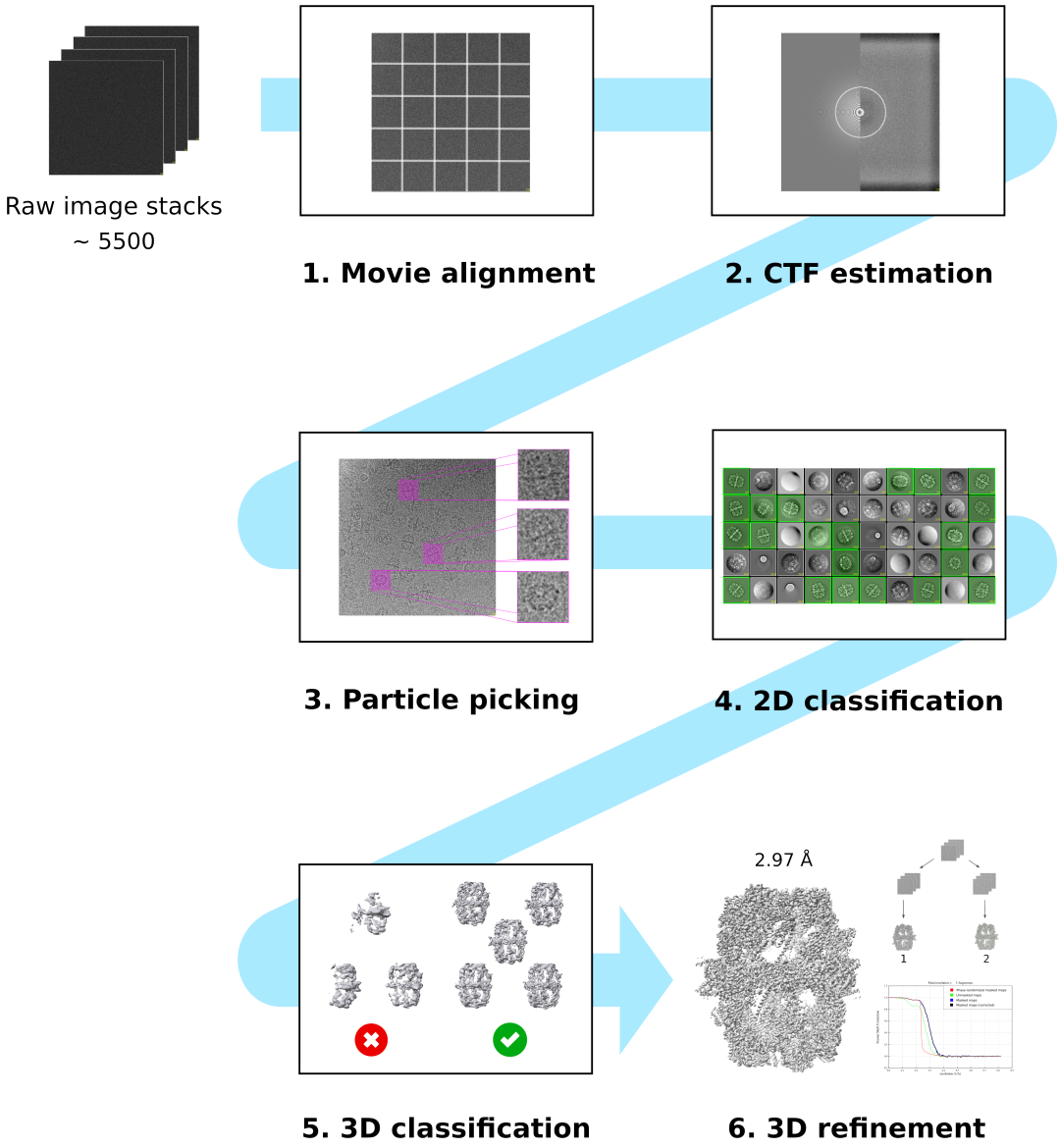


Figure 2.4.: Basic example of a cryo-EM processing workflow. The raw movies from the microscope are aligned and averaged (1). On the micrographs, the CTF parameters are estimated (2). Individual particle images are cropped from the micrographs (3). The particle image set is filtered by 2D and 3D classification (4/5). Finally, a 3D density map is reconstructed from the curated particle set (6).

1. Movie alignment

The raw output from a TEM equipped with a direct electron detector is usually a movie of e.g. 40 frames displaying the same area inside a hole of a cryo-EM grid. Within this area, several copies of the object of interest are contained. During illumination with the electron beam, the specimen undergoes a significant amount of motion. This motion consists of a horizontal drift of the sample holder, referred to as stage drift, and a vertical drum-like doming of the ice layer that has been associated with mechanical stress relief [30, 32].

To correct for this motion, the movie frames have to be aligned before being averaged into a single 2D image, called a micrograph. The alignment is done globally on the whole frames to correct for stage drift. Newer approaches like MotionCor2 [31] include an additional local alignment step based on a user-defined number of image patches to account for non-uniform local deformations. After alignment, the frames are summed up accordingly. A strategy to produce a weighted frame sum based on the exposure of the frames was introduced by Grant and Grigorieff [76] and is widely available in later developed software, e.g. [31].

2. CTF estimation

The averaged 2D micrograph from the motion correction step has sufficient contrast to estimate and correct the spatial-frequency dependent contrast variations that derive from imperfections in the microscope and are described by the CTF and an envelope function (equations 2.5 and 2.6). The CTF depends on the spherical aberration as well as on the defocus during imaging. Since the spherical aberration is usually treated as a constant for the microscope, the CTF is estimated by fitting the defocus parameters to the experimental images. Common programs for CTF estimation are for example CTFFIND4 [28] and GCTF [29]. In these programs, the defocus is described as an ellipse characterized by a defocus maximum, a perpendicular defocus minimum and a 2D rotation angle to account for potential astigmatism in the image. The three parameters are fitted by maximizing the cross-correlation to the intensity oscillations in the Fourier transform of the 2D micrograph, also know as Thon rings. Instead of correcting the 2D micrographs with the fitted CTF right away, the estimated defocus

parameters are stored and CTF correction is incorporated in the classification and refinement algorithms further below.

3. Particle picking

A 2D micrograph depicts an area on the cryo-EM grid with several 2D projections of individual copies of the macromolecule of interest oriented in different ways. Particle picking is the task of cropping these individual views from the micrograph image, thus creating quadratic single-particle images of a user-defined box size. It is important to choose a box size that is larger than the particle itself because the particle's signal is delocalized in the micrographs, which are not yet CTF-corrected at this point. Delocalization depends on the defocus and the resolution of the image features of interest [77]. Therefore, a too small box size is a limiting factor for high-resolution reconstruction.

The identification of particles in the micrographs is a difficult task since the signal-to-noise ratio in the image is very low. For a long time, picking was a manual task where a user would select the particle positions on the micrograph. As raw datasets easily contain tens of thousands to a million particles today to achieve high resolution reconstructions, manual picking is hardly feasible anymore.

There are several different approaches for automated particle picking. If the macromolecule of interest is already known, a few projections of a strongly low-pass filtered existing 3D map can be used as picking templates that represent the rough shape of the macromolecule. These templates can be compared to the micrographs by one of many available template-matching algorithms [33, 78]. There are also tools for automated particle picking in cases where no reference is available. One example is Laplacian-of-Gaussian (LoG) auto-picking in RELION [49]. In this approach, a Laplacian-of-Gaussian filter is applied to the Fourier transform of the micrograph in order to detect blobs of a user-defined particle diameter. In recent years, several approaches based on deep neural networks [79, 39, 40, 80] have been proposed for automated particle picking. This is still an active area of research and new methods are developed that seek to improve the state of the art, for example in terms of generalization to unseen data [81].

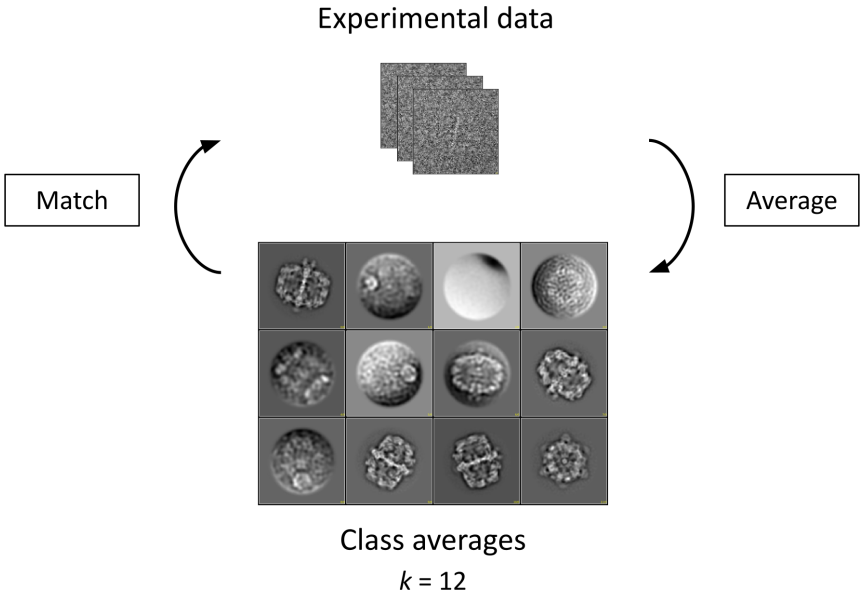
4./5. Classification in 2D and 3D

As outlined in chapter 1, particle images can be of low quality for several reasons and particle picking algorithms do not differentiate sufficiently between good and bad particles. Therefore, a sorting procedure is required after picking. The state-of-the-art strategy to do so are classification algorithms in 2D and 3D.

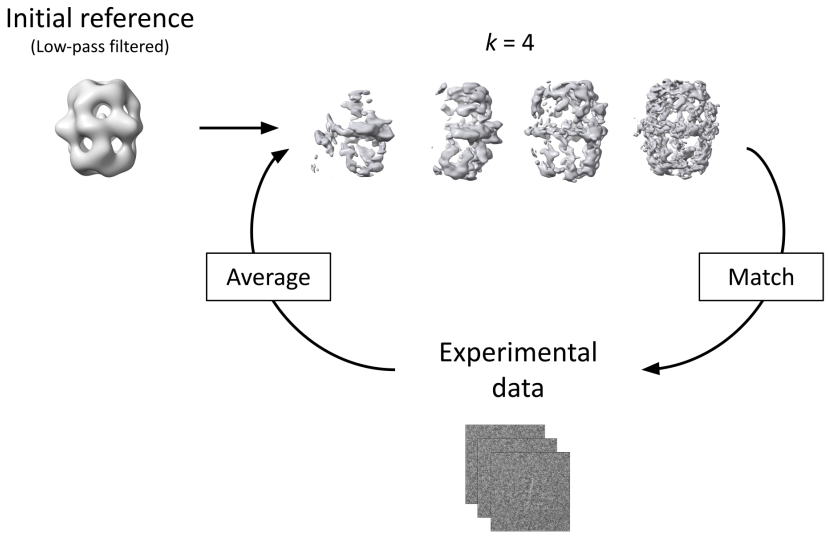
Cryo-EM reconstruction is an incomplete inverse problem. The particle images miss information that is necessary to average them in 2D or 3D. The missing parameters are the orientation of the 2D image in 3D space with respect to the original 3D structure, the translation that describes the offset from the center and the class membership if the image data corresponds to a set of different 3D structures. This can happen, for example, if there are multiple conformations in the sample, which is usually the case. The missing information must thus be estimated. This is realized by the algorithms explained in the following. Figure 2.5 gives a visual overview of the iterative procedures.

The common basis for most classification algorithms in cryo-EM is an iterative procedure that resembles the k-means clustering algorithm [82]. Here, the user has to define the desired number of classes k. In the case of 2D classification (see figure 2.5a), the particle images are first randomly assigned to k bins. An average is computed of each bin. Afterwards, the iterative procedure begins where in each iteration, the experimental particles are aligned by in-plane rotation and x/y translation to each of the previous class averages, and then assigned to the class where the cross-correlation to the average image is maximized. New averages are formed for each class by averaging the assigned images at the optimal rotation and x/y shift. Note that in the context of 2D classification, class averages can originate from different 3D structures, but also represent different projection directions of the same 3D structure. The algorithm proceeds for a user-defined number of iterations. The final set of class averages ideally contains good ones resembling the macromolecule of interest as well as bad ones showing only noise or contamination. The user will then make a selection of classes to keep and only the particle images that were assigned to these classes are used for further processing.

Classification algorithms are prone to errors in cryo-EM because of the very low signal-to-noise ratio in the data, which can cause false peaks in the cross-correlation



(a) 2D classification



(b) 3D classification

Figure 2.5.: Overview of the most common iterative procedures for 2D classification (a) and 3D classification (b). The algorithms alternate between comparing the experimental images to the k current class averages to estimate the class membership, orientation and translation and averaging the images based on the latest parameter estimates. The parameters can be assigned by selecting an optimum or by calculating a probability distribution over the possible values (maximum likelihood approach).

function. In 1998, Sigworth proposed to solve cryo-EM alignment problems by a statistical maximum likelihood approach instead [83]. This idea has since been adopted by the community and maximum likelihood expectation maximization algorithms are the basis for classification and refinement algorithms in the most common cryo-EM software packages [56, 48]. The most striking difference to the traditional cross-correlation based method is that rather than estimating a single optimum for the rotations, translations and class assignments, the maximum likelihood approach calculates a probability-weighted average over these parameters [84]. Thus, it explicitly considers uncertainty in the hidden variable assignment and makes the approach better suited for noisy cryo-EM data.

In 2007, Scheres proposed a maximum likelihood-based 3D classification approach with the aim of separating conformational states in the data [85]. It is visualized in figure 2.5b. The same iterative procedure as for the 2D classification approach is employed, but the class averages are now 3D maps and the hidden variables are class membership, 3D rotation, and translation. A low-pass filtered initial reference that represents the rough 3D shape of the macromolecule of interest is required. It is used in the first iteration of the algorithm, where a likelihood optimization is performed separately on k randomly sampled subsets of the data [85, 86]. This gives k different initial references, which are then used for performing the classification algorithm on the entire dataset.

An alternative way to classify particles is by multivariate statistical analysis (MSA) [66, 87]. The particle images are first aligned e.g. by matching to 2D projection images of a reference structure. Afterwards, each image is represented as a point in pixel space and a principal component analysis is carried out to reduce the dimensionality of that space. This results in an image representation in the form of compressed feature vectors, on which hierarchical clustering can be performed. The clustering procedure can be agglomerative or divisive and different metrics can be employed to compute the distances between the images. The reader is referred to [66] for a more detailed description of the approach.

6. Refinement

After cleaning the dataset and separating conformational states, 3D refinement is performed to reconstruct a final 3D density map. The algorithm employed here is in principle equivalent to performing a 3D classification with a single class. The iterative procedure starts with an initial low-pass filtered reference and then alternates matching of the experimental data to reference projections with averaging of the particles images into a new reference. An illustration of the process is provided in figure 2.6. While classification and refinement algorithms are closely related, the aim is a different one. Instead of clustering and filtering the experimental data, the goal is to reconstruct a 3D map that resembles the true 3D structure of the macromolecule in the most detailed and accurate manner. Therefore, a number of practical differences and optimizations have evolved for the refinement procedure. While 3D classification is carried out at a fixed granularity of orientation sampling, the sampling becomes finer and finer in the course of the refinement to determine the rotation angles as accurately as possible. For computational efficiency reasons, at later stages of the refinement a local search around the best previous orientations is usually performed instead of a global one [56].

A central problem of cryo-EM refinement is overfitting due to the noise in the data. Shatsky *et al.* [88] showed that pure noise images could reconstruct an image of Einstein, when aligned to it. In practice, this could lead to reconstructed features in the 3D map that are not present in the experimental data. In 2012, Scheres proposed a Bayesian regularization approach [89] which automatically applies a Fourier filter on the 3D map based on estimates for the power of the signal and the noise in the data and thereby suppresses high-resolution details that are not sufficiently supported. This approach replaced previously widespread manual filtering procedures. Instead of running for a defined number of iterations, the refinement algorithm automatically terminates according to a convergence criterion, i.e., when the estimated resolution (see section 2.1.3) and the optimal orientation, position and class assignments for the particles stop changing [56].

In recent years, refinement algorithms for cryo-EM have been and still are developed further. The algorithm in the software cryoSPARC [48] (2017) offers a substantial speed-up by using a branch-and-bound procedure and overcomes the need for an external initial reference by generating the reference directly from the data using

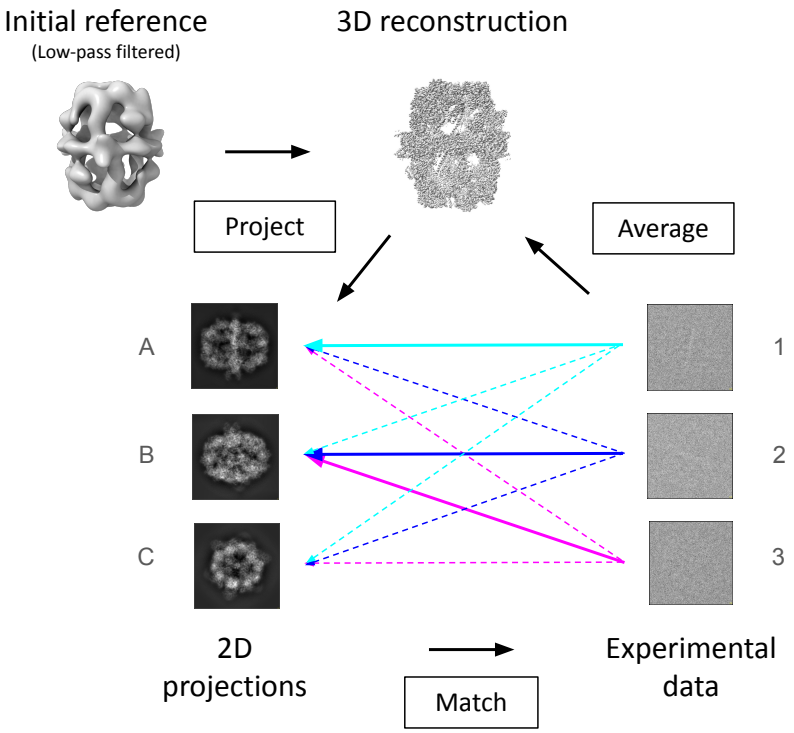


Figure 2.6.: 3D refinement of a cryo-EM map by iterative projection matching. In each iteration, 2D projection images representing different projection directions are generated based on the current 3D reconstruction. In the first iteration, a strongly low-pass filtered external reference is used. The experimental images are compared to the 2D projections to estimate 3D orientation and x-y shift. The parameter assignments can be optimum- or probability-based. Finally, a new 3D map is reconstructed based on the latest parameter estimates. The sampling of orientation and shift parameters usually starts relatively coarse and becomes finer in the course of the refinement.

stochastic gradient descent. Hu *et al.* [90] address the refinement problem with a particle-filter algorithm that includes per-particle defocus refinement and promises robustness to bad particles.

Further optimization

After refinement, additional corrections can be performed on the experimental images to account for different effects that limit resolution. These corrections include per-particle motion correction (Bayesian Polishing) [91], per-particle CTF refinement [49] and higher-order aberration corrections [26]. They can only be carried out after refinement, because they require a high-quality reconstruction as input. It is common practice to perform multiple iterations of particle corrections and refinements, since the quality of the reconstruction affects the accuracy of the corrections and vice versa. It should be noted at this point that the corrections modify the pixel intensities in the images, which might introduce errors or correlations that can cause an over-estimation of the final reconstruction's resolution. Another optimization that can be beneficial at high resolution is Ewald sphere curvature correction [27].

Angular reconstitution

Angular reconstitution [59] is an alternative method to determine the orientations of particle images with respect to the 3D structure, so that a reconstruction can be performed afterwards. The method does not require an initial reference and was therefore a popular method for generating these before stochastic gradient descend methods [48] were abundant. Nowadays, it is not part of the standard workflow in the most popular software packages for cryo-EM structure determination [92, 56, 48]. It is, however, used in this work as a reference-free method to validate orientation assignments (see section 3.4).

The method is based on the principle that, in Fourier space, two 2D central sections through the Fourier transform of a 3D density will always share a common line. Correspondingly in real space, two 2D projections of a 3D density share a common 1D projection [59]. The common 1D projection of two 2D projections fixes their

relative orientation in 3D space, except for the rotation around the common line. This rotation can be determined by a third 2D projection, assuming that it does not share the common 1D projection with the first two.

A computational approach to determine the common 1D projections in real space and derive the particle orientations is described in [59]. For each particle image, a sinogram is created where each row corresponds to a 1D projection along a fixed axis while the image is rotated in-plane from 0 to 180°. For all sinogram pairs, a 2D sinogram correlation function (SCF) containing all cross-correlation coefficients between the lines of the first and the second sinogram is computer over the full 0 to 360° interval. This can be achieved based on the 0 to 180° sinograms, since the lines from 180° onward can be obtained by mirroring the previous ones. Some implementations also store the full 0 to 360° sinograms. The common 1D projections and the corresponding relative in-plane rotations are deduced from the SCFs. In [59], this is done by least-squares fitting of a paraboloid function. Based on the common 1D projections, the 3D rotations of all images are calculated as Euler angles (see section 2.2) after placing the common line projection between the first two images on the z axis and the projection direction of the first image along the x axis of the 3D coordinate system. Note that this initial placement is arbitrary and may vary in different implementations of the algorithm. To make the rotation angles fit a 3D structure with a fixed orientation in 3D space, some implementation offer the option to provide a few particle images with know orientations as an *anchor set* and the first images will be placed accordingly. Before determining orientations by angular reconstitution, the input images need to be centered correctly and, in general, the algorithm is more reliable when performed on class averages, e.g. from 2D classification, because these have an enhanced signal-to-noise ratio [66].

2.1.3. Validation measures for data and map quality

Resolution: Gold-standard Fourier shell correlation

The resolution of a reconstructed cryo-EM is usually determined by gold-standard Fourier shell correlation. The term gold-standard refers to a procedure, where the dataset is split randomly into two halves before refinement [93]. Afterwards, both half-sets are refined independently, yielding two reconstructions referred to as half-maps. Between the 3D Fourier transforms of these two half-maps, the Fourier shell correlation (FSC) [94, 95] is calculated, which is the normalized cross-correlation over spherical shells r_i with increasing radius from the center:

$$\mathrm{FSC}_{12}(r_i) = \frac{\sum_{r \in r_i} F_1(r) \cdot F_2(r)^*}{\sqrt{\sum_{r \in r_i} F_1^2(r) \cdot F_2^2(r)}} \tag{2.7}$$

For each shell, an average correlation over all contained voxels is computed. As the outer shells, which contain the higher spatial frequencies, correspond to finer details in the maps, the FSC typically declines with increasing shell indices. Although the correct resolution threshold is controversial in the cryo-EM community [95], the resolution is typically determined as the inverse of the spatial frequency at which the FSC falls below 0.143 [93].

The FSC resolution gives a single value for the entire density map. However, maps often contain regions of lower and higher flexibility, which leads to better or less defined regions and an overall heterogeneous resolution distribution. It is therefore common in structure determination projects to additionally compute a local resolution for different map regions, for example with the ResMap tool [96].

B-factor plots

In cryo-EM applications, the question might arise how many particle images are necessary to reconstruct a 3D density map at a desired resolution. In general, the resolution improves as more images are averaged, since the consistent signal is added while the random noise cancels out as it can be assumed to be zero on average. How

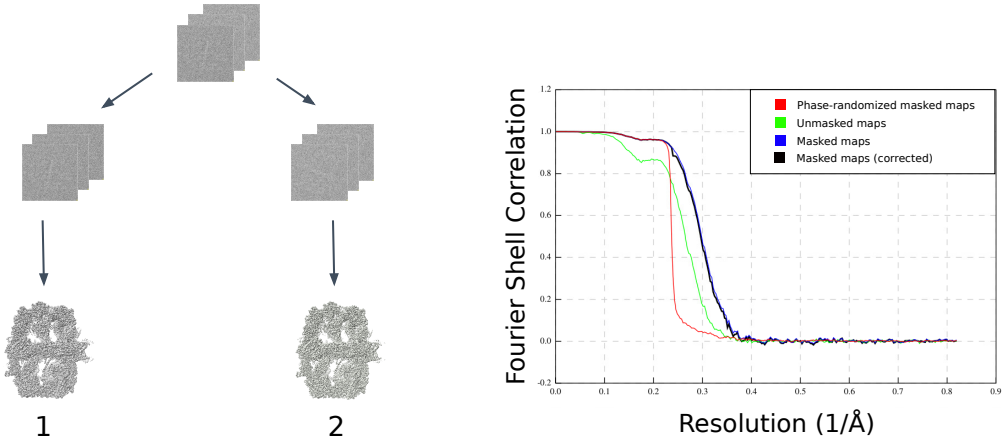


Figure 2.7.: Gold standard procedure [93] for determining the resolution of a cryo-EM map. The dataset is randomly split into two half-sets. From each set, a reconstruction (half-map) is refined independently. The resolution is the spatial frequency, at which the Fourier shell correlation (FSC) between the two half-maps falls below a specific threshold, usually 0.143. An example curve for the FSC with respect to spatial frequency is displayed on the right.

fast the resolution improves with the number of images can be used as a criterion for dataset quality. In 2003, Rosenthal and Henderson [55] introduced a way to correlate the resolution of a reconstruction with the number of particles used. This has inspired the use of so-called B-factor or Rosenthal-Henderson plots in the cryo-EM community. An example is given in figure 2.8. Refinements are carried out for increasingly smaller subsets of a dataset and the squared inverse resolution of the resulting maps is plotted against the natural logarithm of the respective number of particles. A linear fit through the points allows for inter- and extrapolating the necessary number of particles for a given resolution. The B-factor is usually reported, which is defined as two times the inverse slope of the linear fit. Lower B-factors thus correspond to faster resolution improvements and better data quality. B-factor plots are often used in benchmark studies, e.g. to assess the advantages of new or improved microscope components [7, 8].

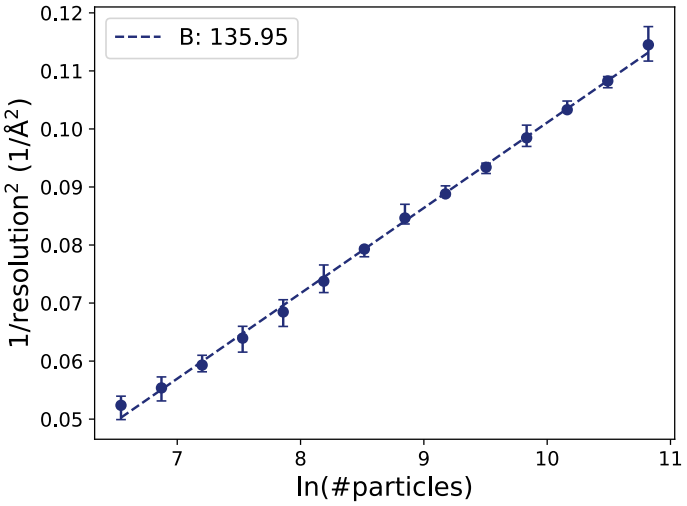


Figure 2.8.: Example of a B-factor or Rosenthal-Henderson plot. Refinements were calculated from random data subsets of different size and the squared inverse of the FSC resolution was plotted against to natural logarithm of the number of particles. Here, each point is the average resolution of five sampling and refinement repetitions for a given subset size. The B-factor is calculated as two over the slope of the linear fit of these points.

Map and model quality

The gold-standard FSC resolution is, strictly speaking, just a measure of consistency in the data and does not directly evaluate the biological accuracy and interpretability of the map. This can only be fully assessed after an atomic model has been built into the map. At resolutions of around 3.8 Å and below, *de novo* model building becomes possible [5]. The resulting atomic models can be assessed – similarly to models from X-ray crystallography – in terms of chemically plausible geometry (bond lengths, angles etc.) and structural clashes. A comprehensive review by Lawson *et al.* [97] based on the 2019 EMDataResource challenge summarizes a number of available measures. After all, good maps should lead to reasonable atomic models. In addition to coordinate-based measures, it is of course important to ensure that the atomic model fits the density well, which can be measured by different metrics [97]. At a very high resolution of 2 Å and below, the number of reliably modeled water molecules can be used as a quality measure for the map [7].

One option to assess map-to-model conformity is the Q-score [54], which measures the resolvability of individual atoms belonging to the protein of interest, ligands or solvent molecules. The Q-score Q of a specific atom is calculated as the normalized about-the-mean cross-correlation between a vector $\mathbf{u}$ of sampled density values in the vicinity of the atom in the map and a vector $\mathbf{v}$ of values sampled from a reference 3D Gaussian centered at the atom's position. This Gaussian describes the expected density decline around a well-resolved atom. Q hence quantifies how well the density distribution around an atom corresponds to this expected density decline, and is calculated as

$$Q = \frac{\langle \mathbf{u} - \mathbf{u}_{\text{mean}} \rangle \langle \mathbf{v} - \mathbf{v}_{\text{mean}} \rangle}{|\mathbf{u} - \mathbf{u}_{\text{mean}}||\mathbf{v} - \mathbf{v}_{\text{mean}}|} \,. \tag{2.8}$$

This way, placement errors, missing densities and blurred out features can be detected. Q-scores generally decrease with lower map resolution.

2.2. Description of 3D orientations

2.2.1. Orientation representations

As described in the previous sections, the correct orientations of the 2D projection images in 3D space have to be estimated for back-projection into the cryo-EM reconstruction. There are different options to define rotations of objects in 3D space, which are explained in the following.

Euler angles

Euler angles describe a 3D rotation as a series of three consecutive rotations around the axes of a Cartesian coordinate system. The Euler convention denotes around which axes the rotations take place and in which order, e.g. XYZ. The rotations can either be intrinsic, i.e. refer to a coordinate system that rotates with the object, or extrinsic, i.e refer to a static coordinate system. In this thesis, the Euler convention and the coordinate system of the COW software suite are used [98]. As shown in

figure 2.9, this coordinate system is right-handed with counter-clockwise rotations. The Euler convention is ZYZ and encompasses three static rotations, applied in the following order:

1. Rotation around the z axis by angle $\alpha \in [-180, 180[$
2. Rotation around the y axis by angle $\beta \in [0, 180]$
3. Rotation around the z axis by angle $\gamma \in [-180, 180[$

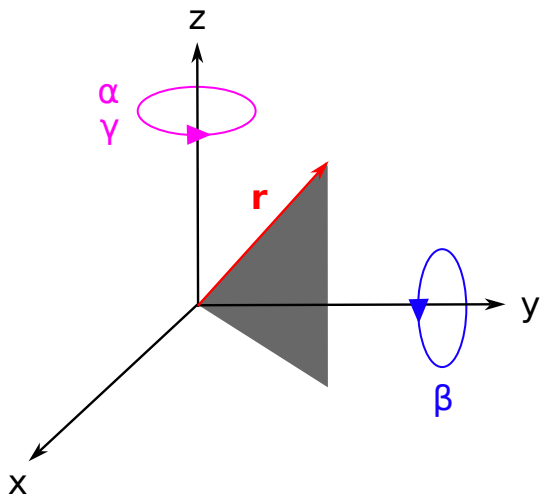


Figure 2.9.: The COW coordinate system including the rotation directions for the extrinsic ZYZ Euler convention. The red arrow is an example of a view vector (or projection direction) **r**.

Euler angles offer a compact description of rotations and are easily understandable from a human perspective, but also have a substantial downside. They are not unique, i.e. two different sets of angles can lead to the same rotation [99]. This makes Euler angles unsuitable for distance calculations. This task is better solved by using either rotation matrices or quaternions.

Rotation matrices

In many cryo-EM software packages, the Euler angles displayed to the user are internally converted into rotation matrices. 3D rotation matrices are 3x3 orthogonal matrices with real coefficients and determinant 1 [99]. The conversion from Euler an-

gles is achieved by constructing the equivalent rotation matrices for the three Euler rotations and then combining them via matrix multiplication [100]. For the COW coordinate system and the extrinsic ZYZ convention, this is given by:

$$R(\alpha, \beta, \gamma) = R(\alpha)R(\beta)R(\gamma) \tag{2.9}$$

$$= \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix} \begin{pmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.10}$$

$$= \begin{pmatrix} \cos\alpha\cos\beta\cos\gamma - \sin\alpha\sin\gamma & \cos\alpha\cos\beta\sin\gamma + \sin\alpha\cos\gamma & -\cos\alpha\sin\beta \\ -\sin\alpha\cos\beta\cos\gamma - \cos\alpha\sin\gamma & -\sin\alpha\cos\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\sin\beta \\ \sin\beta\cos\gamma & \sin\beta\sin\gamma & \cos\beta \end{pmatrix} \tag{2.11}$$

Note that the order of rotations in the extrinsic case described here is inverted in comparison to the intrinsic case described in [99, 100]. The third row of the combined rotation matrix R indicates the projection direction of a rotated image [99].

Quaternions

A third way to describe rotations are unit quaternions. A recent review by Hu *et al.* [101] offers a comprehensive introduction of unit quaternions for cryo-EM by explaining the basic concept, defining distance and geodesic measures, and describing methods for statistical analysis and rotational sampling. The following information is based on this review.

A quaternion $\mathbf{q}$ is a hypercomplex number i.e. an extended complex number that consists of one real part and three imaginary parts with

$$\mathbf{q} = q_0 + q_1\mathbf{i} + q_2\mathbf{j} + q_3\mathbf{k}, \tag{2.12}$$

which is often also written as a 4D vector $\mathbf{q} = (q_0, q_1, q_2, q_3)$. The complex conjugate of a quaternion is given by

$$\mathbf{q}^* = q_0 - q_1\mathbf{i} - q_2\mathbf{j} - q_3\mathbf{k} \tag{2.13}$$

and its norm is defined as

$$|\mathbf{q}| = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2}. \tag{2.14}$$

Quaternions with norm $|\mathbf{q}| = 1$ are called unit quaternions. All unit quaternions together form a 3-dimensional hypersphere embedded in $\mathbb{R}^4$ and can be used to describe 3D spatial rotations. The rotation of a vector $\mathbf{v}$ by a unit quaternion $\mathbf{q}$ is given by

$$R_{\mathbf{q}}(\mathbf{v}) = \mathbf{q} \otimes \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \mathbf{q}^* \tag{2.15}$$

where $\otimes$ is the quaternion product that follows specific rules as described in [101]. The unit quaternion that rotates an object by the angle ϕ about the 3D rotation axis $\mathbf{n_q}$ can directly be obtained by

$$\mathbf{q} = \begin{pmatrix} \cos\frac{\phi}{2} \\ \mathbf{n_q} \sin\frac{\phi}{2} \end{pmatrix}. \tag{2.16}$$

Unit quaternions offer a more compact rotation representation than rotation matrices and provide elegant mathematical tools for distance calculation and statistics. The relevant distance and statistical measures and their application in this thesis are described in chapter 3.

View vectors

The orientation of an image in 3D space can also be described as a view vector with a corresponding rotation angle as proposed by Heymann *et al.* [100]. The view vector is defined as a 3-dimensional vector $\mathbf{r} = (x, y, z)$ with norm $|\mathbf{r}| = 1$ and the corresponding view angle ϕ as a rotation about $\mathbf{r}$. The view vector and angle can

be derived from Euler angles by the following equations for the Euler convention and coordinate system in this thesis, adapted from [100]:

$$x = \sin\beta\cos\gamma \tag{2.17}$$

$$y = \sin\beta\sin\gamma \tag{2.18}$$

$$z = \cos\beta \tag{2.19}$$

$$\phi = \alpha + \gamma \tag{2.20}$$

Note that this conversion corresponds to the third row of the rotation matrix in equation 2.11. The view vector also describes the projection direction of the image and can be interpreted as a normal vector on the image. The standard orientation is usually defined as $\{x, y, z, \phi\} = \{0, 0, 1, 0°\}$ [100], so that the 2D image is embedded in the x-y plane and the normal vector is positioned on the z axis in the positive direction.

2.2.2. Rotation sampling: HEALPix

As described in section 2.1.2, cryo-EM refinement is an iterative algorithm that compares projections of the current density map estimate with the experimental images. These projections are taken from sampled directions, which are ideally uniformly distributed on a sphere surrounding the 3D density map. The sampling should be coarse at the beginning and become finer in the course of the refinement.

There are multiple ways of sampling directions from the sphere, for instance based on different geometric shapes (see [60] for a discussion) or based on quaternions [101]. A popular choice in cryo-EM [56] is the HEALPix [60] framework. HEALPix stands for Hierarchical Equal Area isoLatitude Pixelization. It creates a tessellation of the sphere, i.e., it discretizes it into N_{Pix} curvilinear quadrilateral pixels of equal area. The centers of these pixels are located on isolatitude rings with uniform distances between centers on the same ring. They form an approximately uniform distribution of points on the sphere. The HEALPix framework can create grids of varying resolution as shown in figure 2.10. The base resolution consists of 12 pixels, of which four each are located on three isolatitude rings. The higher resolutions are reached by dividing

each pixel evenly into four sub-pixels. The number of pixels for a resolution level k is thus given by $N_{\mathrm{Pix}} = 12 \cdot 4^k$. The fact that each higher resolution pixel is derived from a subdivision of the parent pixels creates a hierarchical ordering, which allows for storing data in tree-like data structures and fast nearest neighbor searches.

The approximately uniform sampling and the hierarchical structure of the distributed points make HEALPix a suitable choice for cryo-EM refinement [56]. It is used to sample the two Euler angles that describe the projection direction, called β and γ in this thesis. The third Euler angle α, which describes an in-plane rotation, is sampled linearly with an accordingly adjusted resolution.

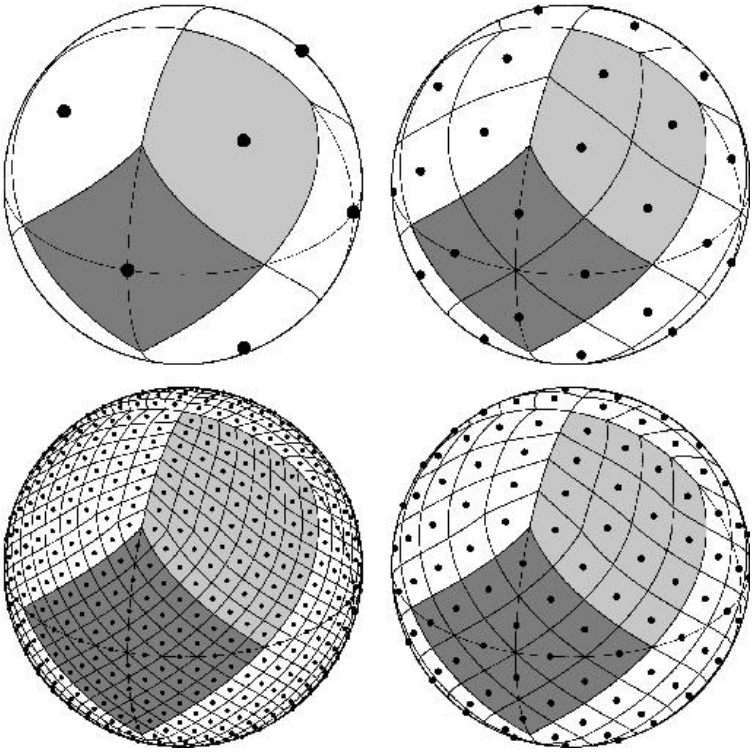


Figure 2.10.: HEALPix tessellations of the sphere corresponding to resolution level $k = 0$ (upper left), 1 (upper right), 2 (lower right) and 3 (lower left). Each pixel is divided into 4 sub-pixels as the resolution level increases by one. Image courtesy NASA/JPL-Caltech, reproduced in accordance with the JPL Image Use Policy. [102]

3. Methods

In this chapter, the newly developed methods for particle sorting are described. Section 3.3 focuses on the acquisition of metadata from the cryo-EM workflow, in section 3.4 a new method for measuring consistency of orientation assignment is described and in section 3.5, different strategies for filtering are explained.

3.1. Primer on particle image sorting

All methods presented in this thesis serve a common goal, namely the identification and removal of low-quality particle images from cryo-EM datasets. This improves the overall quality of the filtered dataset and ideally leads to a better reconstructed density map. An alternative objective is the reduction of the number of images in the dataset while maintaining the reconstruction quality. This increases processing speed and feasibility for very large datasets. The basic principle is visualized in figure 3.1.

The challenges in particle sorting are related to finding and evaluating predictors for image quality, and to developing suitable filtering strategies that also take competing objectives like the orientation distribution into account. The respective new approaches are presented in the following sections. In this thesis, quality parameters are collected while or directly after processing a cryo-EM dataset in a state-of-the-art workflow. The final dataset and final 3D density map from this workflow serve as the baseline for the consecutive filtering experiments.

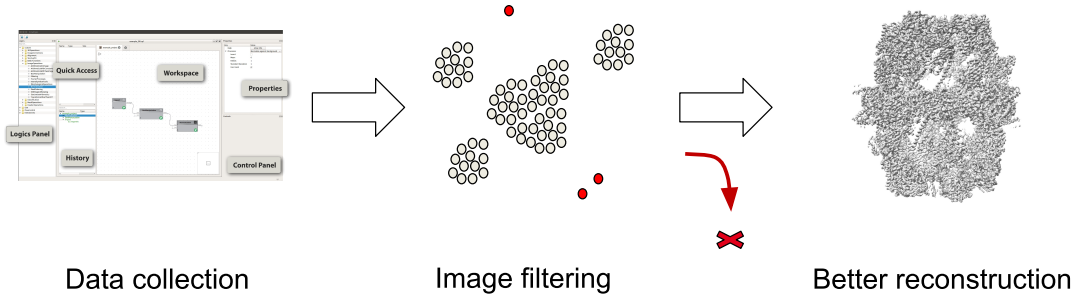


Figure 3.1.: Basic principle of particle sorting in this thesis. Data is collected during or after image processing. Based on this data, low-quality images are removed from the dataset, which ideally leads to better reconstruction quality.

3.2. Prerequisite: The COW software suite

The new methods for parameter calculation and filtering are implemented within the COW suite [61], a software package for single-particle cryo-EM image processing developed by the department of Structural Dynamics at the Max-Planck-Institute for Multidisciplinary Sciences.

The main tool is a graphical user interface for visual programming, that allows the user to put together custom image processing workflows. Figure 3.2 shows the graphical user interface. Workflows can be built by dragging calculation units – called *logics* – from the *logics panel* on the left into the main *workspace* window. Icons on the logics show the calculation status, e.g., completed, running, pending or failure. Necessary calculation parameters can be specified in the *properties* window on the right. The *history* window lists all logics currently used in the project. The output of each logic can be examined by clicking on the logic's output arrow, or via the history list. This opens a viewer window which shows the images and their associated metadata. A single project can consist of multiple workflows which are organized as different tabs. Logics can be copied into other tabs by dragging them from the history panel into the workspace tab. This also allows for transferring the output of a logic from one tab to another for further processing. Workflows can be stored as templates (without the calculation results) for transfer between projects.

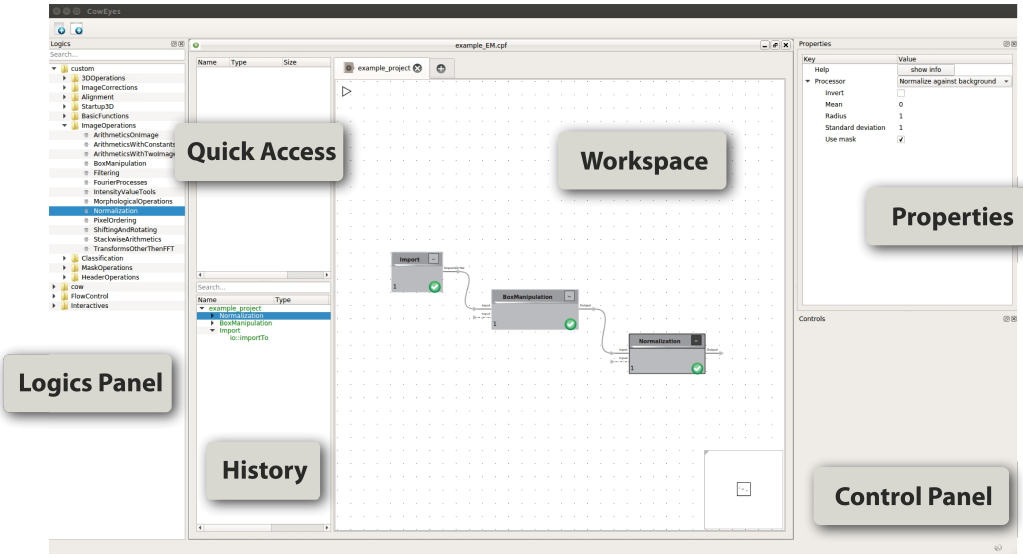


Figure 3.2.: The visual programming graphical user interface in COW. Reproduced with permission from [103].

The COW backend contains a large number of image processing functionalities organized in libraries, which are used by the logics for the calculation routines. A logic base class defines the general logic architecture. New logics can easily be added by inheriting from the base class and specifying a logic name, calculation routine and expected user parameters. External programs are included via custom wrappers. Most calculation routines in COW are performed on *Data* objects which broadly speaking consist of an *Image* object storing the pixel values and a *Parameter* object storing the image metadata in a map, i.e., an associative array with keys and values. This metadata is also referred to as *header* data, as it is usually contained in image file headers. Data is passed on to downstream logics by *DataIO* containers, which point to files storing the output of the previous logic. A typical calculation routine inside a logic iterates over the stored image or respective header data. For each image, it reads the corresponding file content into a *Data* or *Parameter* object, performs calculations, and then writes the modified content, i.e., image and header data or only header data, to the specified output file.

3.3. Metadata from the processing workflow

The first filtering approach in this thesis is particle sorting based on metadata from the processing workflow. This metadata is acquired in a COW workflow, which includes the image processing logics from the standard reconstruction workflow as depicted in figure 2.4, as well as custom logics for parameter calculation. For each processing step described below, a logic was implemented in COW, which calculates the respective quality parameters. A more detailed description of the logics and their usage is given in the appendix (section B.3). The calculated results are stored in the header of each corresponding image. This way, the workflow yields both a 3D reconstruction and a set of quality parameters. These parameters can be directly exported as CSV (comma-separated values) files and used for further analysis. Table 3.1 summarizes all collected and calculated parameters, grouped by the corresponding workflow steps. In the following, these parameters are explained in detail. The parameter names are highlighted in bold font.

3.3.1. Motion correction

During motion correction, which is carried out using MotionCor2 [31], shifts between the movie frames are calculated in a cross-correlation based procedure [104] in order to align them for averaging. This is done globally on the entire micrograph as well as locally for image patches, which are derived by splitting the image into n equally sized sub-images in x and y direction (yielding n^2 sub-images in total). The local shift is calculated in an extra step on the globally corrected micrographs [31], so that both global and local shifts are reported separately.

Even though the motion in the images is corrected for, it is unclear whether the correction is powerful enough to fully compensate for it. Unusually large shifts could also be a sign for alignment errors due to, e.g., inconclusiveness in the image features. Large shifts could therefore be an indicator for poor image quality. For the parameter calculation, three MotionCor2 jobs are carried out for n=3, 5, 7 patches respectively. The average global shift over all movie frames (**avgGlobalShift**) is stored as well as the average patch shift over all movie frames (**avgPatchShift**). The average shifts

Table 3.1.: Overview of quality parameters collected and calculated during cryo-EM data processing sorted by the corresponding workflow steps

#	**Processing step**	**Parameter**	**Description**
1	Motion correction	avgGlobalShift	Average global shift over all movie frames
2		avgPatchShift	Average local shift of the particle's patch over all movie frames
3	CTF estimation	ΔGlobalU	Deviation of the patch maximum defocus z_{U} from the global value
4		ΔGlobalV	Deviation of the patch minimum defocus z_{V} from the global value
5		ΔGlobalAst	Deviation of the patch defocus angle from the global value
6		ΔNeighborU	Average deviation of the patch maximum defocus z_{U} from the neighbor patch values
7		ΔNeighborV	Average deviation of the patch minimum defocus z_{V} from the neighbor patch values
8		ΔNeighborAst	Average deviation of the patch defocus angle from the neighbor patch values
9		ctfFitResolution	Resolution up to which the CTF fit is valid
10	Particle picking	pickingCounter	Number of times the particle was found by different pickers
11		pickerQualityScore	Cross-correlation-based similarity to picking template
12	3D classification	classREI	Relative entropy index of distribution over the classes
13		σShiftClass3D	Standard deviation of shift (distance from center)
14		σRotClass3D	Standard deviation of rotation

per patch are later associated with every particle that lies within each respective patch.

Figure 3.3.: Definition of neighboring patches. A micrograph was divided into 4x4=16 patches. The neighbors of a patch are defined as all other patches sharing an edge with this patch. Inner patches therefore have more neighbors than outer ones.

3.3.2. CTF estimation

CTF parameters are estimated with GCTF [29]. As explained in section 2.1.2, the fitted CTF parameters describe a defocus ellipse by its maximum z_{u}, the perpendicular minimum z_{v} and an in-plane rotation angle θ_{ast}.

The accurate estimation of the CTF parameters is a prerequisite for good reconstruction quality, since wrong CTF parameters would falsely modify the signal from the corresponding images. The parameters below were designed to describe the reliability of the CTF parameters. The CTF parameters can vary over the micrographs in a dataset as they are usually recorded within a defocus range instead of a fixed value. Within a single micrograph, however, the defocus values should show slight variations at worst. Strong variations could be a sign for errors in the parameter assignment or an indication for local deformations in the specimen [30], which in turn could limit the image quality.

To measure the stability of the CTF parameter estimation, the CTF parameters are once again determined globally as well as on patches for n=5. For each patch, the deviations of z_{u}, z_{v} and θ_{ast} from the global values are calculated and stored (**ΔGlobalU**, **ΔGlobalV**, **ΔGlobalAst**). Before calculating the parameters, it is ensured that z_{u} always contains the maximum value, which is not necessarily the case in the GCFT output. If z_{v} is larger than z_{u}, the two values are switched and θ_{ast} is adjusted accordingly by adding 90 degrees. The deviation of θ_{ast} is computed as the cosine distance between the local and global angle, i.e.

$$\Delta\mathrm{GlobalAst} = \cos|\theta_{\mathrm{ast,global}} - \theta_{\mathrm{ast,local}}|. \tag{3.1}$$

In addition to the deviations from the global values, the average deviations from all neighbor patches (**ΔNeighborU**, **ΔNeighborV**, **ΔNeighborAst**) are calculated and stored to detect local variations of the CTF parameters. See figure 3.3 for the definition of neighbor patches. The values per patch are later passed on to the corresponding particles. Finally, the coefficient describing the resolution until which the CTF could be estimated (**ctfFitResolution**) is stored.

3.3.3. Particle picking

As outlined in section 1.2, the picking, i.e. cropping, of particle images from micrographs is a difficult task and often leads to false results. If a particle is identified by multiple picking algorithms, this might reduce the chances of false picking. Therefore, particle picking is carried out in three different variants and for each particle, it is counted by how many of the three approaches it was found (**pickingCounter**). The picking results from a Gautomatch [33] run with template images derived from a former FAS reconstruction are the *reference* picking set. These are the particles which will be passed on to the next processing steps. For each particle in the reference set, the results from a reference-free Gautomatch run and a RELION Laplacian-of-Gaussian [49] run are queried. If picking coordinates with a distance below a threshold to the reference particle coordinates are found, the picking counter is increased by one. As a threshold, 10 percent of the particle size are used, as done in the *DeepConsensus* [45] approach. As an additional parameter, the particle quality score calculated by Gautomatch, which is based on the cross-correlation between template and particle,

is stored (**pickerQualityScore**). In the picking analysis step, the particles are also extracted from the micrographs at a user-defined box size. The header parameters from the previous workflow steps are passed on from the micrographs to the corresponding particles. The patch-based parameters from motion correction and CTF estimation are associated with the particles by determining the corresponding patch value for each pixel in the particle box and averaging over all values. As a result, the particle inherits the local shift or CTF deviation values from the micrograph patches weighted by the number of particle pixels in the respective patches.

3.3.4. 3D classification

During 3D classification, a discrete set of 3D reconstructions is generated from the particle images. In this work, the 3D classification algorithm from RELION [85, 56] is used. It runs for a user-defined number of iterations. In each iteration, the algorithm compares every experimental image to a sampled set of 2D projections of the 3D volumes. For each image, weights are calculated for each combination of rotation, x-y shift and 3D class (expectation step). In the maximization step, a new 3D volume is calculated from the experimental images by probability-weighted back-projection under a regularization prior. In each iteration and for each image, the shift, rotation and class for which this image has the highest weight are tracked.

From the tracked data, parameters that quantify the consistency of the assignments are derived. The idea behind these parameters is that low-quality images, e.g. due to structural damage of the particles, might show an instable behavior terms of shift, orientation and class assignment. The quality parameters below therefore describe the stability of these assignments over the classification iterations. The stability of orientation assignments has previously been used as a filter criterion to remove particle images from a GroEL cryo-EM dataset [105].

Since class memberships are nominal – class numbers have no specific order – the variability is measured in terms of entropy [106]. Class membership tracking over the algorithm yields a vector of n observations, each belonging to one of the I classes.

The relative frequency per class is calculated as $p_i = n_i/n$. The relative entropy index (REI) can then be calculated as

$$\mathrm{REI} = \frac{-\sum_{i}^{I} p_i \log p_i}{\log I} \tag{3.2}$$

and describes the distribution of particle assignment over the classes [106]. The REI takes values between 0 and 1, where 0 results from assignment to one class only and 1 stands for a uniform distribution over the classes. The entropy parameter over the class assignments is stored (**classREI**).

In the case of the x-y shift, the standard deviation of the magnitudes of the shifts over the iterations is calculated. This is done by first transforming the separately stored x and y offsets into n distances from the center by

$$d_{0,i}(x_i, y_i) = \sqrt{x_i^2 + y_i^2} \tag{3.3}$$

and then calculating the mean of those distances as

$$\mu_{\mathrm{shift}} = \frac{\sum_{i=1}^{n} d_{0,i}(x_i, y_i)}{n}. \tag{3.4}$$

From there, the standard deviation is calculated as

$$\sigma_{\mathrm{shift}} = \sqrt{\frac{\sum_{i=1}^{n} (d_{0,i}(x_i, y_i) - \mu_{\mathrm{shift}})^2}{n}}. \tag{3.5}$$

The standard deviation of the distances is stored (**σShiftClass3D**).

Mean and standard deviation are also calculated for the rotation parameter, but the calculation is less trivial in this case. In cryo-EM, rotations are usually described by Euler angles, which are unsuitable for statistical calculations as multiple Euler angles can describe the same rotation. In a recent review, Hu *et al.* [101] describe the advantages of rotation representation by unit quaternions including the feasibility of

statistical analysis. They show that the distance between two rotations represented by unit quaternions can be calculated as

$$d_{SO(3)}(\mathbf{q}_1, \mathbf{q}_2) = 2 \arccos(|\mathbf{q}_1 \cdot \mathbf{q}_2|), \tag{3.6}$$

where $SO(3)$ denotes the 3D rotation group and $\cdot$ the scalar product of the quaternion vectors. The weighted geometric mean for a set of n rotations is given by

$$\underset{R_\mathbf{q} \in SO(3)}{\arg\min} \sum_{i=1}^{n} w_i d^2_{SO(3)}(\mathbf{q}, \mathbf{q}_i). \tag{3.7}$$

According to a solution by Horn [107] and Markley *et al.* [108], this mean can be approximated by the normalized principal eigenvector $\bar{\mathbf{q}}$ of the 4x4 matrix $\mathbf{T}$,

$$\mathbf{T} = \sum_{i=1}^{n} w_i \mathbf{q}_i \mathbf{q}_i^{\mathrm{T}} \tag{3.8}$$

with $w_i = 1$ in this case of an unweighted mean. While Hu *et al.* approximate $\bar{\mathbf{q}}$ by Von Mises iteration [109], here it is determined with the LAPACK solver for real symmetric matrices (ssyev) [110].

To determine the standard deviation of the rotation assignments over the classification algorithm for each particle, the tracked Euler angle rotations are converted to unit quaternions and $\bar{\mathbf{q}}$ is calculated as described above. Using equation 3.6, the standard deviation σ_{rot} can easily be computed as

$$\sigma_{\mathrm{rot}} = \sqrt{\frac{\sum_{i=1}^{n} d^2_{SO(3)}(\mathbf{q}_i, \bar{\mathbf{q}})}{n}}. \tag{3.9}$$

The standard deviation of the assigned particle rotations is then stored (**σRotClass3D**).

3.4. Orientation consistency

The second approach for quality filtering in this thesis aims at quantifying the consistency of orientation assignments by different methods. Here, a quality parameter is calculated by a special methodology after a first density map has been reconstructed,

while the parameters in the previous section are deducted from the standard processing workflow. The correct determination of particle orientations is a prerequisite for high quality reconstructions. The back-projection of image data from a wrong direction will not beneficially contribute to the average reconstruction, but add noise and false signal. Since in asymmetric proteins, there is only one correct orientation with respect to a fixed 3D structure in space, the particle image should be assigned this orientation no matter what method was used to determine it. Therefore, in this approach, 3D orientations for all particle images are determined in three different ways and then the distance between those orientations is determined. When particles have a low distance, this indicates that the orientations could be reliably determined independent of the method used. These particles should have a higher probability to yield high quality reconstructions. This new method is independent from the standard cryo-EM processing workflow, but requires the existence of an already decent 3D reconstruction that should ideally be improved further by the method.

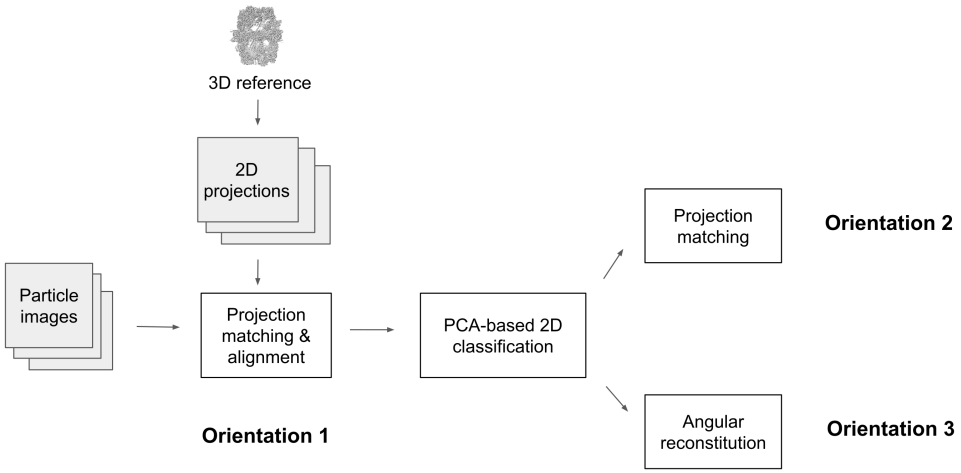


Figure 3.4.: Procedure for determining particle orientations in three different ways. The resulting Euler angles are converted into quaternions and distances are computed to quantify the consistency between the orientations for each particle.

The workflow to determine the orientations is shown in figure 3.4. First, the particles are aligned and matched to 2D projections of the existing 3D reconstruction. For each particle, the rotation Euler angles of the most similar projection in terms of

cross-correlation are stored. This is the first orientation. Afterwards, the particles are clustered by a COW-internal variant of an MSA classification [66] (see also section 2.1.2). The class averages are again matched to the reference projections and each particle is assigned the rotation angles of the most similar projection to the class average it belongs to. This is the second orientation. The third orientation is determined with the COW-internal implementation of angular reconstitution [59] (see also section 2.1.2) based on the class averages. This method only takes a few projections as an anchor set and, apart from that, determines the rotation angles in a reference-free manner.

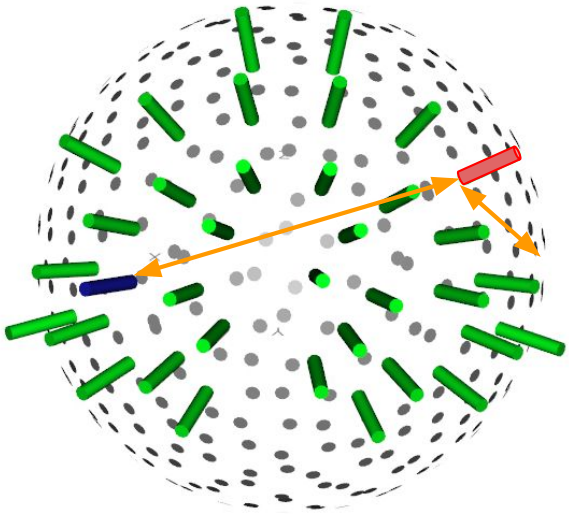


Figure 3.5.: Illustration of the challenges when computing orientation distances in the context of structural symmetry. The cylinders indicate different projection directions on the sphere, in this case all within an asymmetric region corresponding to the D3 point group. In this group, each projection direction has six equivalents on the sphere. If the distance between the red and the blue direction is computed, the closest distance is not the the one inside the asymmetric region (left orange arrow), but the one from red to a symmetry equivalent of blue (right orange arrow). For illustration purposes, the in-plane rotation around the projection direction is omitted in this example.

Now the distance between the three orientations needs to be determined. The first challenge to meet is the definition of a suitable distance measure. For projection

directions, the distance could simply be determined as the distance of two points on the unit sphere. However, the cryo-EM orientation also encompasses an in-plane rotation around the projection direction, i.e., two images can have the same projection direction but show differently rotated 2D views of the protein and should therefore get a distance greater than zero. Hu *el al.* describe quaternions for cryo-EM in terms of their swing-twist decomposition [101]. The orientation workflow (figure 3.4) yields the orientations as Euler angles. These are converted into quaternions. Then, pairwise distances according to $d_{SO(3)}$ (equation 3.6) are calculated. Finally, the average over the three pairwise distances is computed, which is referred to as d_{total} in the following. The second challenge to handle is symmetry in the structure of interest. Many proteins – especially the large ones traditionally targeted by cryo-EM – consist of multiple sub-units that are arranged in a symmetric way [111]. This creates an internal symmetry that is described by one of five point-groups: cyclic, dihedral, tetrahedral, octahedral and icosahedral [112]. The practical consequence of structural symmetry is that different orientations – namely the symmetry equivalents – will show equivalent views of the 3D structure. This is an advantage for cryo-EM refinement, since a single image can be back-projected into the reconstruction multiple times, but a hurdle for distance measurement. Not considering symmetry would lead to a significant overestimation of the distances for symmetry equivalents. Projecting symmetry equivalents to a fixed asymmetric region would not fully solve the problem, since near the boundaries, distances to symmetry equivalents just outside of the boundaries could be smaller than to the projections inside of the asymmetric region, see figure 3.5. Therefore, the minimal pairwise distance in the presence of structural symmetry for two orientations represented by quaternions $\mathbf{q}_a$ and $\mathbf{q}_b$ is determined as

$$d_{\text{min}} = \min_{T \in S} d_{SO(3)}(\mathbf{q}_a, T(\mathbf{q}_b)) \tag{3.10}$$

where S is the symmetry and T is each associated transform of the symmetry, including the identity (no transform). The final value of d_{min} is reported as the distance of the orientation pair $\mathbf{q}_a$ and $\mathbf{q}_b$. The minimum pairwise distances of all orientations are then averaged like in the non-symmetric case (d_{total}). Since the number of symmetry transforms is usually small and quaternion distance calculation is fast, the algorithm is not expected to lead to a significant practical overhead compared to non-symmetric distance calculation.

Two new custom COW logics were implemented for the orientation consistency approach, one for mapping the class average results to the particle images and one for distance calculation. The other logics for orientation determination – in particular the ones for MSA classification and angular reconstitution – had already been available in COW. A workflow template for the orientation calculation procedure shown in figure 3.4 including mapping and distance calculation was generated and deposited in COW. A detailed description of the workflow template is given in the appendix, section B.4.

3.5. Filtering methods and validation strategy

The central problem for the evaluation of particle sorting methods is that the large amount of noise makes it impossible to judge the quality of individual images. For real data, there is hence no ground truth telling if an image is good or bad. Only if multiple images are averaged – in 2D or 3D –, the quality of the whole set can be evaluated based on features or resolution of the average image or density map. In this thesis, quality parameters are evaluated by comparing the resolution of data subsets to the expected resolution from a B-Factor plot (section 3.5.1) and by comparing FSC resolutions before and after filtering out a number of presumably bad particles from a baseline dataset. The filtering can be done by different strategies. Next to single-parameter filtering based on a percentage of particles or box plot criteria (section 3.5.2), filtering strategies that combine multiple parameters (section 3.5.3) and a new method for direction-based filtering (section 3.5.4) that aims at preserving the orientation distribution are used.

3.5.1. Subset evaluation

The following subset validation approach is used to validate the parameters in table 3.1 as well as the average orientation distance d_{total} and is generally applicable to any potential quality parameter. The particle images are divided into subsets based on the parameter to be validated. The resulting subsets should yield reconstructions with varying resolutions, which reflects quality differences among the subsets that can optimally be explained by the subset's parameter values.

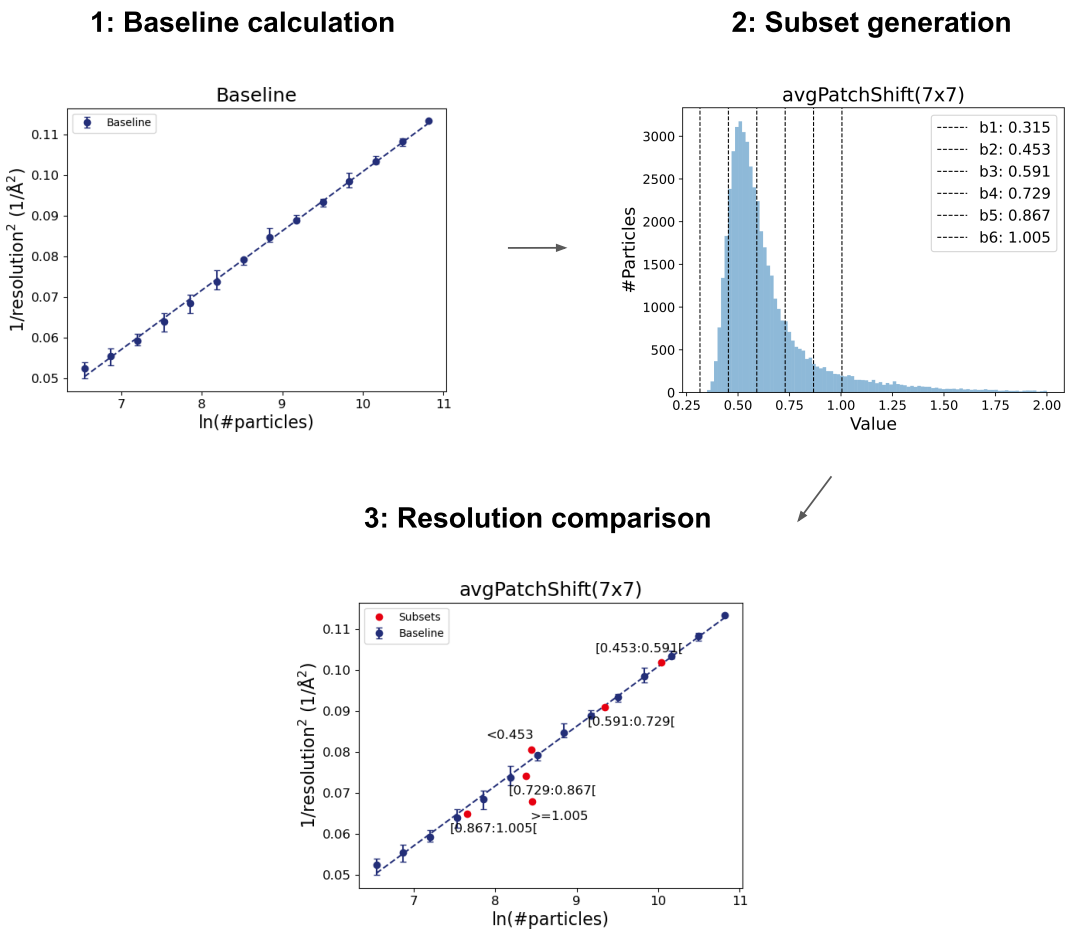


Figure 3.6.: Overview of the subset evaluation approach. First, a B-factor plot is generated based on random data subset as described in 2.1.3. Then, the dataset is split into subsets based on the value distribution of the parameter to be validated. Here, a histogram of the avgPatchShift(7x7) parameter is shown with the black vertical lines indicating the subset bin borders. Finally, refinements are carried out for all subsets and the resulting FSC resolutions are plotted together with the baseline B-factor curve. The validation approach is similar to the parameter evaluation with ResLog plots by Stagg *et al.* [58].

Figure 3.6 summarizes the subset evaluation procedure in this thesis. Since the achievable resolution depends on the number of images, one cannot simply compare resolutions of differently sized data subsets. Instead, a B-factor plot [55] (see section 2.1.3) is constructed for the whole dataset. A linear fit of the points allows for interpolating expected resolutions for subsets of variable size. To evaluate the meaningfulness of a quality parameter, the image dataset is sorted by this parameter. Then, standard box plot statistics are applied to identify the outlier limits of the distribution. The lower limit is 1.5 times the inter-quartile ratio below the first quartile and the upper limit is 1.5 times the inter-quartile ratio above the third quartile. Particles with a parameter value outside of the limits are considered outliers. Five equidistant bins are defined in between the limits to further split up the non-outliers. This leads to seven data subsets in total (lower outliers, upper outliers and five sets within the limits). Data subsets that are smaller than a minimum particle number are merged with their smallest neighboring set. Finally, refinements are calculated for all subsets individually. The resulting FSC resolutions are compared to the expected resolutions from the B-Factor plot. In figure 3.6, the subset resolutions are plotted into the B-factor plots. Since the y axis shows the square of the inverted resolution, points above the B-factor line correspond to lower (and thus better) resolutions than expected and those below the line to worse resolutions. The procedure above resembles an evaluation with ResLog plots by Stagg *et al.* [58], but in this thesis B-Factor plots are used instead of ResLog plots and an automated procedure for subset generation is described.

3.5.2. Single-parameter filtering

The most basic filtering strategy in this thesis is filtering based on a single parameter. Particles images are kept or discarded based on a user-defined threshold value. This functionality is available in COW via the *Extract* logic, where the user can define the filtering parameter, the threshold value and whether images above or below the threshold should be kept. A suitable threshold value can be determined, for example, by the boxplot outlier criteria described in the previous section. Alternatively to threshold-based filtering, a given number of particles can be removed based on the parameter of interest. This can also be achieved in COW. To this end, the images are first sorted by the parameters of interest in the *SortImages* logic, where the user can decide between ascending and descending sorting. Afterwards, the user can filter

images via the *Range* option in the *Extract* logic by defining the first index and the number of images to be extracted from the sorted list.

3.5.3. Combined filtering

As an extension of the single-parameter filtering approach, multiple parameters can be combined for filtering. In this thesis, two approaches for combined parameter filtering are tested. The initial image stack is either subsequently filtered by different parameters (filtering cascade) or a ranking is computed by considering multiple parameters at once (combined ranking).

Filtering cascade

In a filtering cascade, the initial particle image stack in subjected to multiple filtering steps. In each step, the images are filtered by a different parameter, either based on a threshold value or by excluding a certain percentage or number of particles. The remaining particles are passed on to the next filtering step. After each round, the filtering success can be evaluated by computing a refinement based on the current filtered image stack and comparing the FSC resolution or other quality parameters to the unfiltered baseline. The principle of a filtering cascade is visualized in figure 3.7. A corresponding workflow can easily be constructed in COW by stacking the logics used for single-parameter filtering as described in section 3.5.2.

Combined ranking

The combined ranking approach is based on the assumption that low-quality particles might score badly for multiple quality parameters. Therefore, the particles are filtered by multiple parameters simultaneously. Particle images that show an unfavorable behavior with respect to a single quality parameter are not as easily filtered out if they show a favorable behavior with respect to the other parameters. This might make this filtering approach more robust than the filtering cascade described above.

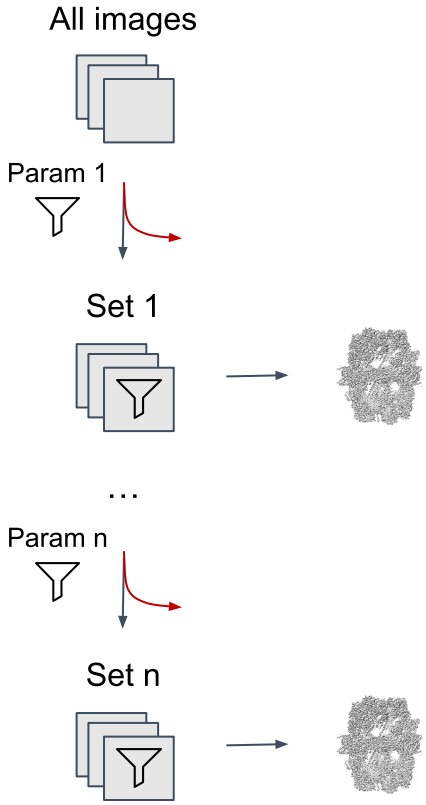


Figure 3.7.: Principle of a parameter filtering cascade. The input image stack is consecutively filtered by multiple parameters based on a given threshold or filtering percentage for the current parameter. After each filtering step, a refinement is carried out with the remaining particle images to evaluate the filtering performance.

A simple example of the combined ranking approach is given in figure 3.8 for five particle images and five parameters. There is no restriction in principle on the number of parameters that can be used. The input particle image stack is sorted by each of the p user-defined sorting parameters individually. Each image is then assigned a rank with respect to each parameter. This is the position in each respective sorted list. The ranking procedure is dense, i.e., if multiple images have the same parameter value, they will all be assigned the same rank and the next images with a different value will be assigned the following ranks without skipping any numbers. The p ranks are then averaged into a single combined rank by computing the arithmetic mean of the ranks. As stated in [113], this corresponds to ranking the images by Borda's

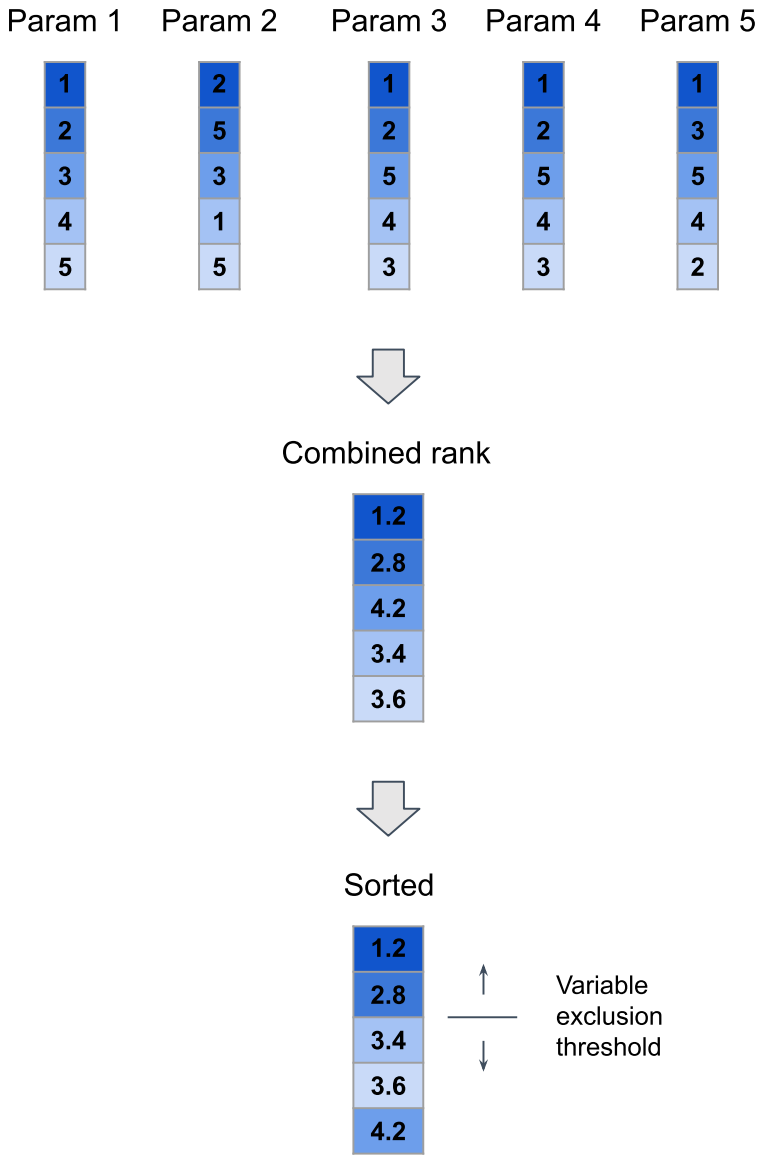


Figure 3.8.: Example of the combined ranking approach for five particle images which are sorted by five parameters simultaneously. The images are first sorted by each parameter individually. They are then assigned a rank with respect to each parameter, which is the position in the respective sorted list. For each image, the five ranks are averaged into a combined rank. The combined rank can then be used for filtering, e.g. by sorting the images by the combined rank and only keeping the n first ones.

method [114] with the L_1-norm of the vector of all p ranks as the sorting criterion. Alternative aggregation techniques include sorting by the median or geometric mean of the rank vector, but these are not evaluated in this thesis. The combined rank constitutes a new quality parameter, by which the particle images can be filtered as described in section 3.5.2. That is, either an exclusion threshold can be defined or the particles are sorted by the combined rank and only the first n ones are kept.

A new logic for computing the combined rank was implemented in COW for this thesis. In the *CombinedRanking* logic, the user can define the filtering parameters as a comma-separated list of strings. When running the logic, a new *combinedRank* parameter is written into the header of each input image based on the calculated procedure described above. The output images are not sorted by the combined rank, but kept in the same order as in the input. Sorting and filtering of the images can be achieved afterwards with the logics *SortImages* and *Extract* as described in section 3.5.2.

3.5.4. Directional filtering

Good cryo-EM reconstructions require image data that shows the 3D object from as many different directions as possible. It is a well-known practical problem that proteins can have a preferred orientation in the prepared specimen. This leads to an uneven distribution of views during imaging and in the worst case makes reconstruction impossible.

In figure 3.9, the preferred orientation problem is schematically illustrated for the FAS structure. Experimental cryo-EM datasets for the FAS consist primarily of side views, presumably due to a preferred orientation where the FAS protein is oriented side-wise towards the air-water interface and thus perpendicular to the electron beam. Since a full distribution of views around the z axis is sufficient for 3D-reconstruction and the FAS structure shows D3 symmetry, high-resolution reconstruction are nevertheless possible in this case.

However, the removal of orientations from this already limited distribution might cause severe problems for the 3D reconstruction. If a parameter used for filtering is correlated with orientation, the filtering might introduce a more uneven orientation

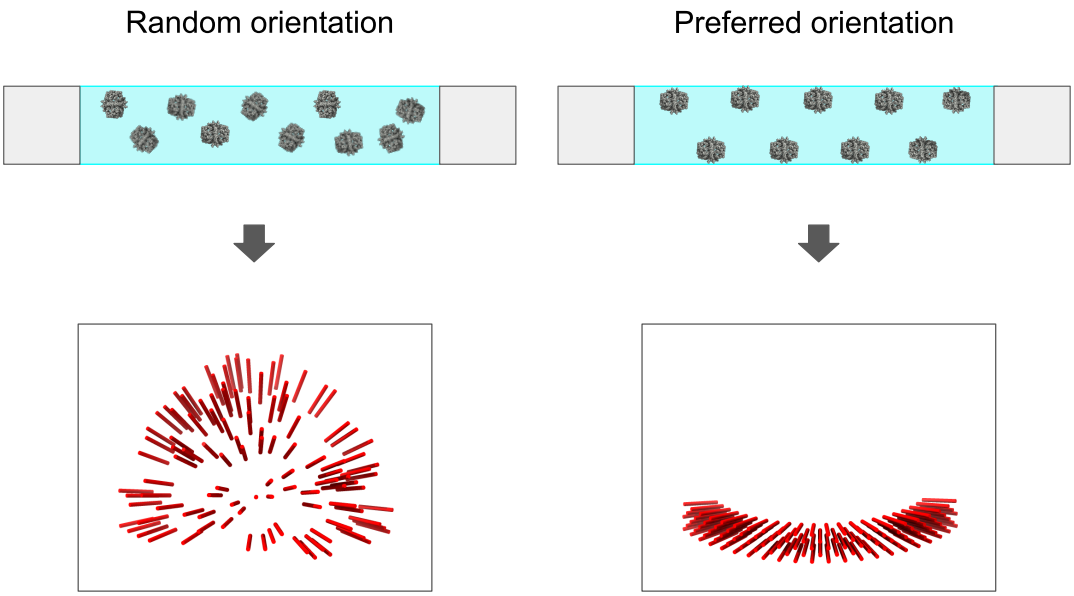


Figure 3.9.: Illustration of the preferred orientation problem in cryo-EM. Particles line up at the air-water interface of the specimen while preferably adopting certain orientations. The resulting 2D projection views are no longer randomly distributed, but show the protein of interest only from specific angles making 3D reconstruction more difficult and in some cases impossible. The lower images show example distributions of views on a sphere around the 3D reconstruction. The views are in this case limited to a triangular region in the upper half of the sphere that corresponds to the asymmetric unit of the D3 symmetry group, which is sufficient for reconstruction of the FAS structures in this thesis.

distribution that counteracts the positive effects of removing low-quality particles. In the worst case, information from certain projection directions might be deleted completely. Preferred orientation and the resulting rarity of certain views can be observed in many datasets, even for successful projects. The protection of rare views and the maintenance of a uniform orientation distribution therefore has potential practical relevance in many cases.

For this reason, an orientation-based filtering approach was developed. In this approach, the particle images are binned by their projection direction, and filtering only takes place within these bins. The binning is based on proximity to projection directions from HEALPix [60] (see section 2.2.2). All directions for a user-defined

HEALPix resolution and symmetry are generated. Then for each particle image, a view vector describing the projection direction of the image is computed based on equations 2.17-2.19. This vector is compared to all HEALPix directions, also in view vector form, by computing the distance as

$$d(\mathbf{v_1}, \mathbf{v_2}) = \arccos(\mathbf{v_1} \cdot \mathbf{v_2}). \tag{3.11}$$

and assigned to the bin with the closest distance. Once all particle images are assigned, filtering is carried out by sorting the images within each bin and then – based on the user's choice – either discarding a user-defined percentage of particles per bin or discarding each bin's box plot outliers (as defined in section 3.5.1). To additionally protect rare views, the user can define a minimum number of particles to be kept per bin.

If applicable, the structural symmetry can be defined. In this case, the HEALPix reference orientations are only created within the asymmetric unit of the defined symmetry, which speeds up the binning. This option should, however, only be applied if the orientations of the input particles were restricted to the respective asymmetric unit beforehand; otherwise, particles with orientations outside of the asymmetric unit would erroneously be assigned to the nearest bin on the border of the asymmetric unit.

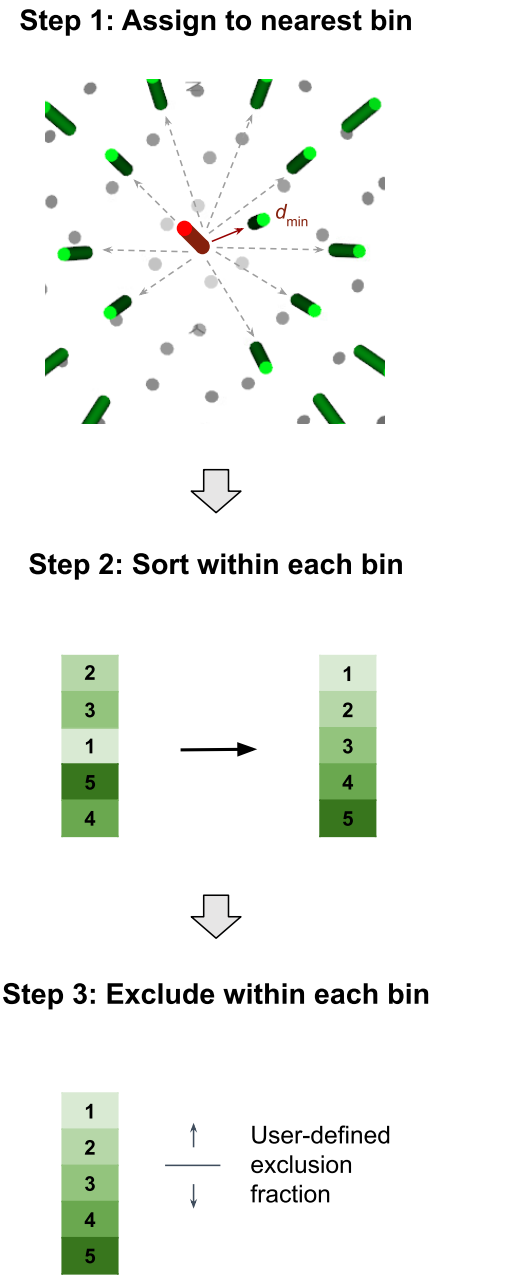


Figure 3.10.: Overview of the directional filtering method. The particle images are binned by comparing their orientation to all HEALPix reference orientations and the assigning them to the closest one (Step 1). Within each orientation bin, the images are sorted by the user-defined filtering parameter (Step 2). Finally, a user-defined fraction of images is removed from each bin (Step 3). Alternatively, image exclusion can be carried out based on boxplot criteria within each bin.

4. Data

In this chapter, the data for evaluating the filtering methods is described. In section 4.1, the generation of simulated image data is presented. In section 4.2, the image datasets derived from real FAS data and their respective baseline processing before filtering are outlined.

4.1. Simulated data

A number of simulated datasets were created based on the cryo-EM map with EMDB ID 4578 and the atomic model with Protein Data Bank (PDB) [115] ID 6QL6 of the $\Delta\gamma$-FAS complex reconstructed by Singh *et al.* [116]. The datasets are summarized in table 4.1.

Table 4.1.: Overview of simulated datasets

#	**Positive**	**Negative**	**Size (Ratio)**	**Box size**	**Pixel size (Coarse factor)**
1	FAS (4578)	Disk	6,144 (1:1)	80 px	4.24 Å/px (4)
2	FAS (4578)	Ribosome (2847)	6,144 (1:1)	80 px	4.24 Å/px (4)
3	FAS (4578)	FAS (4577)	6,144 (1:1)	80 px	4.24 Å/px (4)
4	FAS (6QL6)	FAS (6QL6-A,G)	6,144 (1:1)	80 px	4.24 Å/px (4)
5	FAS (6QL6)	FAS (6QL6-A,G d1)	6,144 (1:1)	80 px	4.24 Å/px (4)
6	FAS (6QL6)	FAS (6QL6-A,G d2)	6,144 (1:1)	80 px	4.24 Å/px (4)

The goal was to create datasets that consist of $\Delta\gamma$-FAS images (positive) as well as images from a different source that are unsuited for reconstruction (negative). In contrast to real data, the simulated data then contains a label for each image based

on its origin. This label can be used for benchmarking. The second and third column in table 4.1 state the structures the positive and negative images were created from.

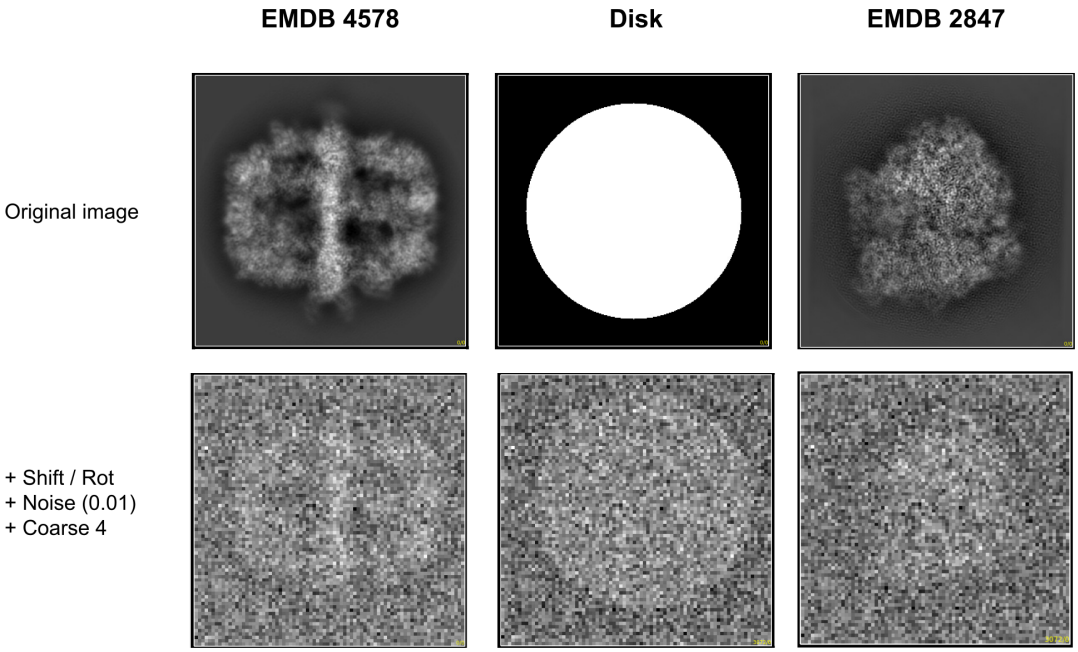


Figure 4.1.: Example images from dataset 1 and 2 (table 4.1). The top row shows the original 2D images (after 2D projection in case of the EMDB maps). The lower row shows the same images after application of random x/y shift and in-plane rotation, normalization and addition of noise to a signal-to-noise ratio of 0.01 as well as coarsening by factor 4.

In dataset 1, the negative images are a simple white disk on black background of approximately the same diameter as the FAS. In dataset 2, negative images were created from a ribosome cryo-EM map (EMDB 2847) [117]. The white disk image and an example ribosome projection are displayed in figure 4.1. In dataset 3, the negative images were created from a cryo-EM map of the FAS/γ-subunit complex (EMDB 4577). This map shows an alternative FAS conformation which is characterized by a 15° rotation at the top of the dome area [116].

Datasets 4 to 6 were created to simulate structural damage of the FAS molecule. The procedure shown in figure 4.2 was carried out in UCSF Chimera [118]. Starting with the atomic model (PDB 6QL6) of the Δγ-FAS complex, the two β-subunits

A and G were removed from the lower dome area to simulate a damaged structure (6QL6-A,G). This structure was deformed by moving the subunits H and I from their original position (6QL6-A,G d1). Finally, an even stronger deformation was produced by moving the subunits B, C, D, H, I from their original position (6QL6-A,G d2). From all of these atomic models, artificial density maps were created with the UCSF Chimera molmap command, which generates a 3D Gaussian blob for each atom. The resolution (2.9 Å), box size (320 px) and pixel size (1.06 Å/px) were chosen to match the EMDB 4578 map. Instead of the EMDB 4578 map, the artificial map from the undamaged PDB 6QL6 model was used for creating the positive images in dataset 4 to 6 in order to rule out the artificial Gaussian map creation as a difference factor between the positive and negative maps.

The artificial images were created in COW. For all positive and negative structures except the already 2-dimensional disk image in dataset 1, projections were taken at HEALPix level k=4 with C1 symmetry producing 3072 images. The disk image was duplicated a corresponding number of times. The projections or duplicates were further processed in the COW logic *TestImage*, which applied a random in-plane rotation and x-y shift to each image. The images were normalized and Gaussian noise was added. The simulation did not include any CTF modulations. To produce the datasets in table 4.1, the respective positive and negative projected and modified images were concatenated, resulting in a total number of 2x3,072 = 6,144 images with a positive to negative ratio of 1:1. The images originally had a pixel size of 1.06 Å/px, but were binned by factor 4 which resulted in a new pixel size of 4.24 Å/px and a new box size of 80 px. Figure 4.1 shows example images for a projection of the EMDB 4578 (FAS) map, the artificial disk and a projection the EMDB 2847 (Ribosome) map, each before and after the application of random shift and rotation, noise, and coarsening.

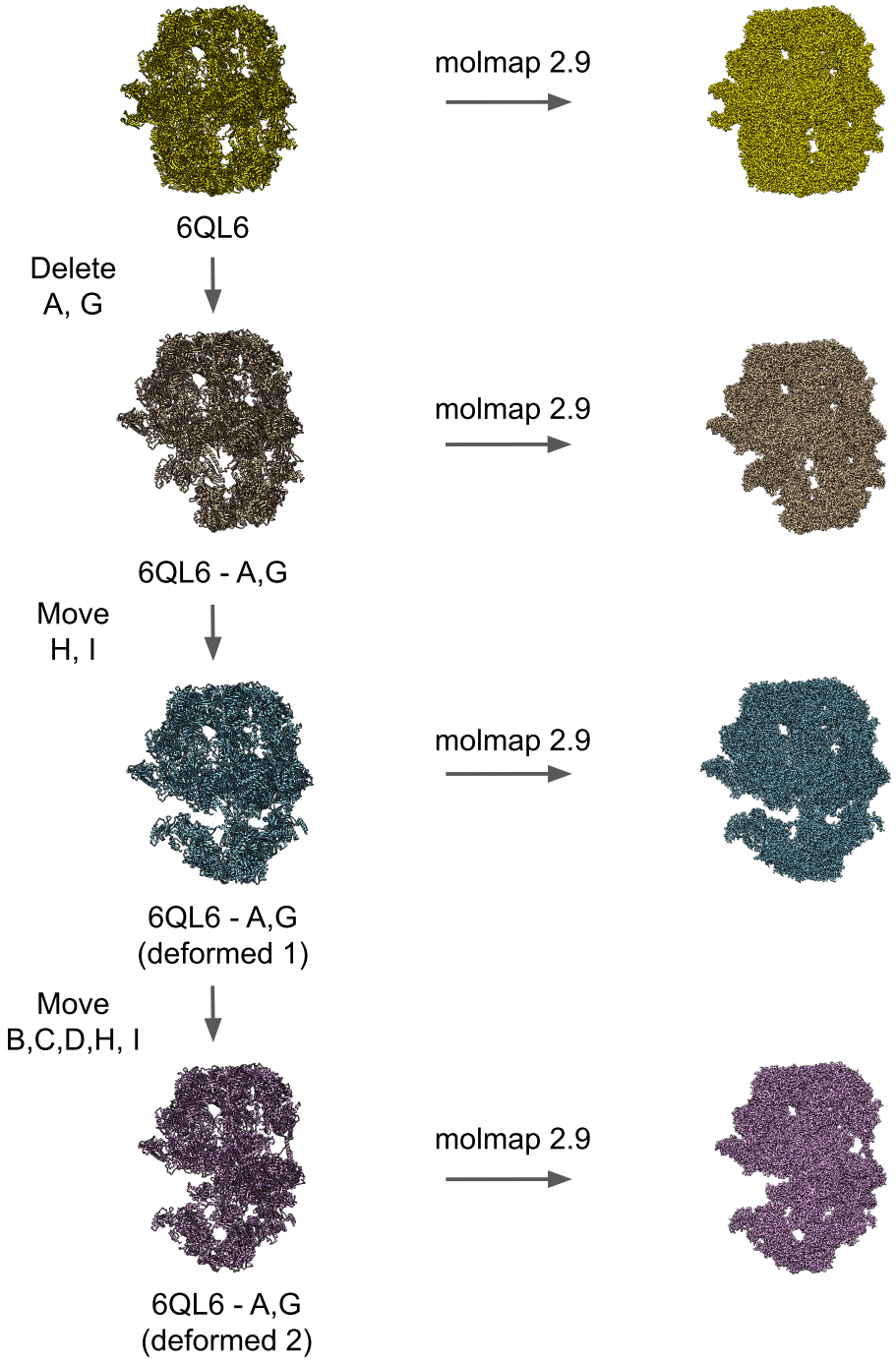


Figure 4.2.: Workflow to create the FAS maps for datasets 3 to 5 (table 4.1). Starting from the atomic model PDB 6QL6, subunits A and G were deleted and the remaining atomic model was deformed by moving subunits H and I (deformed 1) and then subunits B,C,D,H and I (deformed 2). Gaussian blob maps were created for all models using the molmap command. The procedure was carried out with UCSF Chimera [118].

4.2. Real data

For evaluation on real data, a high-resolution dataset of *S. cerevisiae* fatty acid synthase (FAS) [119] was chosen. Fatty acids serve as building blocks for cell membranes, energy storage and precursors to second messenger molecules in living organisms [120]. Consequently, FAS complexes are a potential target for antibiotic drug discovery and offer opportunities for industrial fatty acid synthesis [120]. The *S. cerevisiae* FAS is an example of a type I FAS system, where the component enzymes are integrated in an $\alpha_6\beta_6$ oligomer [121]. Figure 4.3 displays the *S. cerevisiae* FAS complex. The subunits form a D3-symmetric barrel-like structure with an equatorial central wheel of six α subunits and two domes above and below the central wheel, which consist of three β-subunits each [116, 119].

The cryo-EM data was collected with a customized high-end 300 kV Titan Krios electron microscope containing a monochromator and an aplanatic image corrector and previously allowed for a reconstruction at 1.9 Å resolution [119]. Two datasets were used in this thesis. The first one was the whole FAS dataset of 30,926 movies ("large dataset"). This dataset is structured in 15 subsets based on the recording days of the movies. The second dataset used in this thesis corresponds to subset number 10, which consists of 5,437 movies ("small dataset").

Both datasets were processed in a similar way to the original processing in [119] in workflows built in COW [61]. These workflows combined the state-of-the-art processing outlined below with the collection of the quality parameters described in section 3.3. An example of such a processing workflow is given in the appendix (section B.3). The main steps of the processing schemes are summarized in figure 4.5. For the small dataset, the movies were subjected to motion correction with MotionCor2 [31] using local correction with 5x5 patches and dose weighting. CTF estimation was performed with GCTF [29] on the non-dose-weighted frame sums, but the dose-weighted micrographs were used for further processing. Template-based particle picking was carried out with Gautomatch [33], yielding approximately 190,000 particle images. The particle images were extracted at a box size of 720 px at the original pixel size of 0.61 Å/px and then binned by factor 8 before 2D classification (90 px, 4.88 Å/px). Three rounds of RELION 2D classification [56] were carried out.

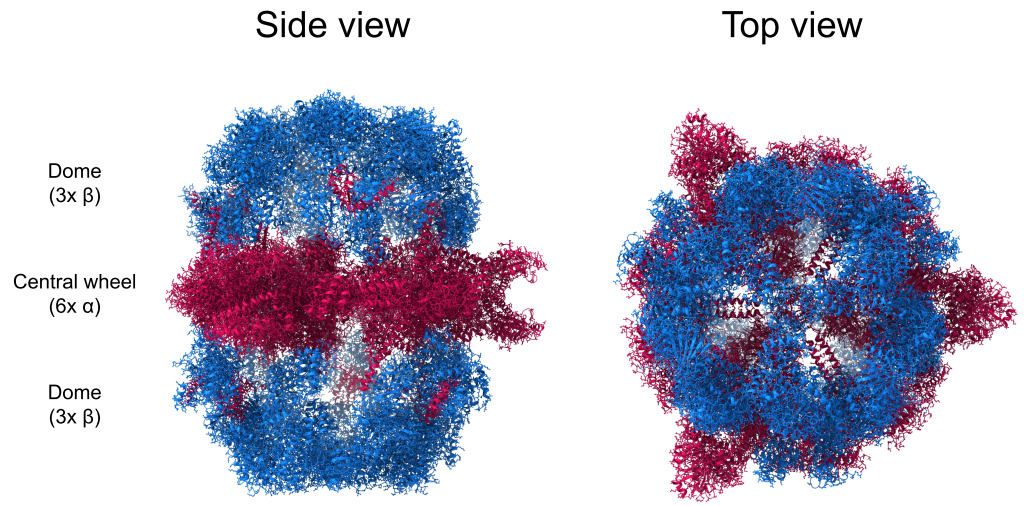


Figure 4.3.: Structure of the *S. cerevisiae* FAS. The D3-symmetric barrel-shaped oligomer consists of six α subunits forming an equatorial central wheel (red) and six β subunits, of which three each form the two domes (blue) above and below the central wheel. Image generated with USCF ChimeraX [122] with PDB 6QL6 based on the depiction in [116].

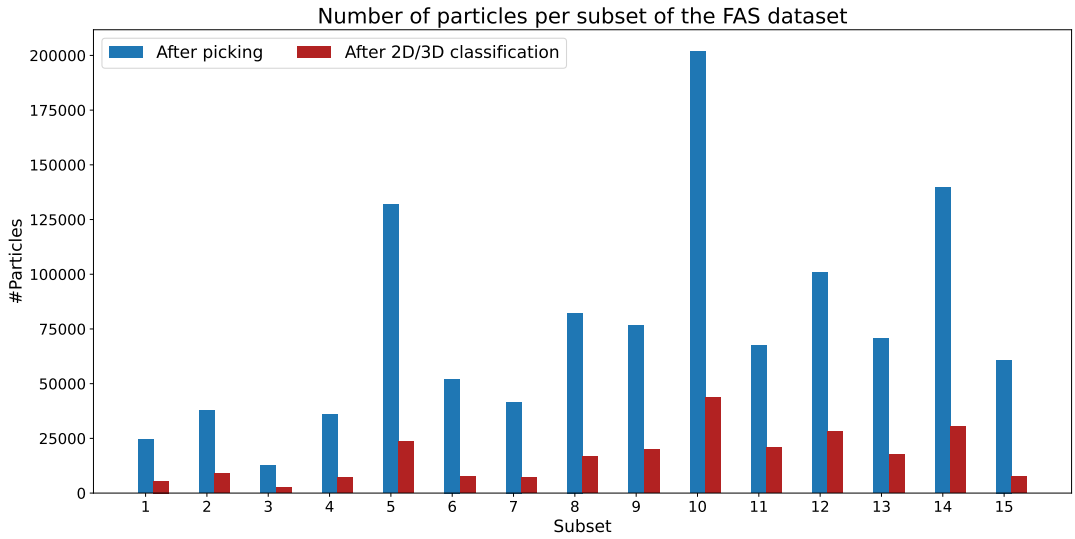


Figure 4.4.: The *S. cerevisiae* FAS dataset [119] consists of 15 subsets corresponding to different recording days. The bars show the particle numbers for each subset directly after picking and after particle selection by 2D and 3D classification.

Afterwards, the remaining particles were filtered by one round of RELION 3D classification [85, 56] with C1 symmetry at factor 4 binning (180 px, 2.44 Å/px). After 2D and 3D classification, approximately 50,000 particles remained for the small dataset. These particles were re-extracted at a box size of 560 px at original pixel size and then refined to 2.97 Å[1] resolution using RELION auto-refine [56]. No further processing was done for the small dataset.

For the large dataset, the processing of the movies was carried out analogously to the small dataset processing using MotionCor2 and GCTF. The template-based particle picking with Gautomatch yielded approximately 1,140,000 particle images. 2D classification and 3D classification were performed as described above, but on each of the 15 subsets individually. An overview of the particle numbers over the subsets before and after classification is given in figure 4.4. In total, approximately 250,000 particle images remained after 2D and 3D classification. These particles were re-extracted at a box size of 672 px at original pixel size and refined together to 2.58 Å resolution using a solvent mask. CTF refinement [49] was performed, followed by two additional rounds of Bayesian Polishing [91] and CTF refinement. 3D refinement of the final dataset yielded a resolution of 1.93 Å.

[1]This is the resolution of the first refinement. The refinement was repeated afterwards yielding resolutions of 2.96 ± 0.04 Å(see section 5.2).

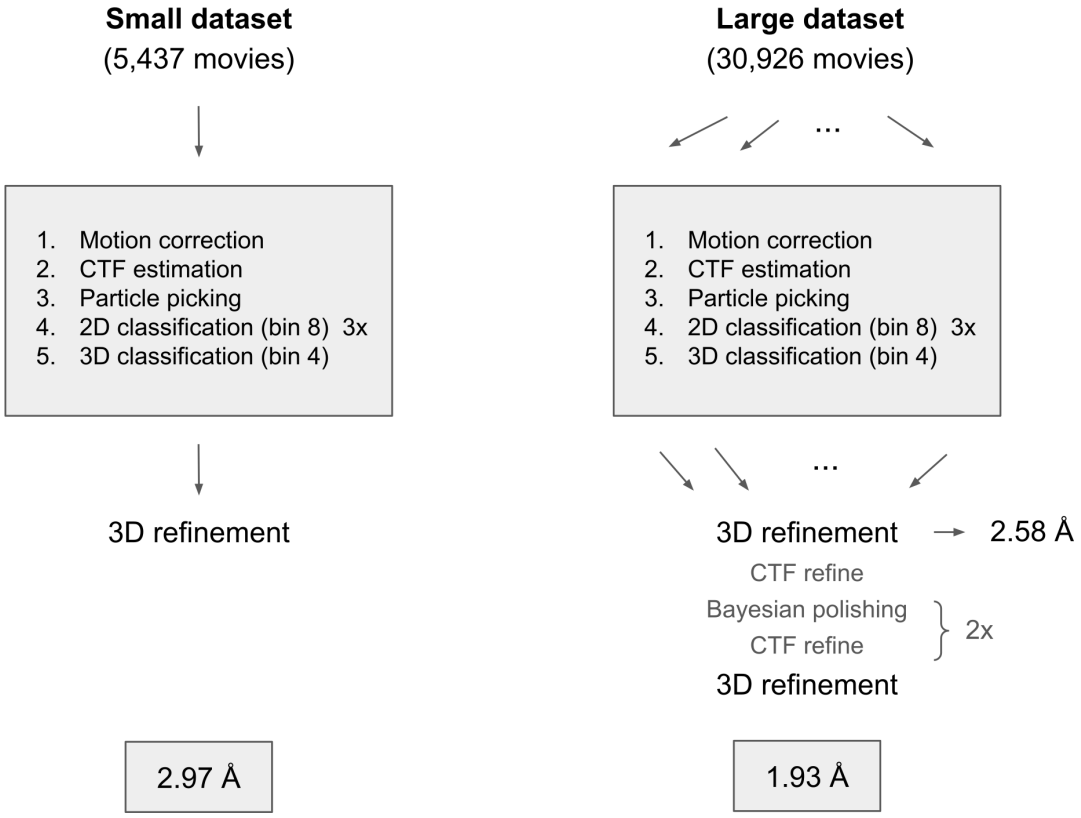


Figure 4.5.: Processing scheme of the small and large high-resolution FAS dataset. The small dataset is a subset of the large one and was processed without per-particle CTF refinement and Bayesian polishing, see section 4.2 for details. The procedures are based on the dataset's original processing in [119].

5. Results and discussion

In this chapter, the methods developed in this thesis (chapter 3) are evaluated on the simulated and experimental data described in chapter 4. First, the relationship between the workflow metadata parameters and reconstruction quality is analyzed (section 5.1). Then, the effect of excluding presumably low-quality particle images from the FAS datasets based on single-parameter (section 5.2) or combined filtering (section 5.3) is evaluated. The orientation consistency approach is evaluated for filtering on simulated images and experimental FAS data (section 5.4). Afterwards, the effect of combining the orientation consistency approach with directional filtering is analyzed (section 5.5). Finally, all filtering methods are compared to the state-of-the-art filtering approach CryoSieve (section 5.6) and the effect of applying CryoSieve on-the-fly on recording day subsets is investigated (section 5.7). Additionally, the relationship between per-subset filtering rates and particle quality assessed (section 5.8). The experiments in this chapter were guided by the research questions below.

1. **Analysis of metadata influence on particle image quality**
    a) Are certain metadata values connected to lower or higher image quality in terms of subset FSC resolutions?
    b) Do the relationships remain after the application of per-particle corrections such as CTF refinement and Bayesian Polishing?
2. **Particle filtering based on image metadata**
    a) How does the FSC resolution change when excluding low-quality images according to metadata parameters?
3. **Combined filtering based on image metadata**
    a) Are the best-performing metadata parameters correlated to each other?

b) How many particles can be excluded by a metadata outlier filtering cascade and what are the effects on the resolution?

c) Is the combined rank based on multiple quality parameters related to image quality as measured by subset FSC resolution?

d) How do FSC resolution and Q-scores change when excluding images based on the combined rank?

4. **Orientation consistency**

a) Can the orientation consistency method separate positive and negative images from a simulated dataset?

b) What is the effect of taking structural symmetry into account?

c) Is the orientation consistency parameter related to image quality in terms of subset FSC resolution on real FAS data?

5. **Directional filtering**

a) Does the directional filtering approach mitigate filtering-induced orientation bias?

b) How do FSC resolution and Q-scores evolve when excluding images based on the orientation consistency parameter with or without directional filtering?

6. **Comparison to a state-of-the-art filtering method**

a) How do the methods developed in this thesis compare to the state-of-the-art method CryoSieve in terms of FSC resolution and Q-scores?

7. **On-the-fly application of CryoSieve filtering**

a) Can CryoSieve effectively reduce the number of particles when applied on recording day subsets instead of on the whole FAS dataset?

8. **Relationship of subset filtering rates and particle quality**

a) Are the per-subset filtering rates of the methods from this thesis and CryoSieve related to subset quality in terms of FSC resolution?

5.1. Analysis of metadata influence on particle image quality

The metadata collected during cryo-EM processing described in section 3.3 were evaluated for their potential to separate high and low quality particles. To this end, the small FAS dataset was processed in a COW workflow as depicted in figure 4.5 and described in section 4.2. At the same time, the quality parameters listed in table 3.1 were collected.

To generate baseline resolution estimates for particle subsets of variable size, a B-factor plot (see section 2.1.3) was created by randomly sampling 13 subsets of size 695 to 36,004 from the final dataset and plotting the squared inverse of the gold-standard FSC resolutions of maps against the natural logarithm of the subset size. The subset sizes were chosen so that the points in the plot were evenly spaced on the logarithmic x axis. Sampling and refinement were repeated five times for each subset and for each point, the mean resolution of the five points was plotted together with error bars stretching from minimum to maximum resolution of the five refinements. The B-factor plot also contains a point for the refinement of the whole dataset, which was repeated five times.

In order to validate the quality parameters, subsets of the final dataset corresponding to different parameter ranges were created as described in 3.5.1 and the FSC resolutions from refinements of these subsets were plotted into the B-factor plot. The resolutions from the metadata-based subsets can now be compared to the expected resolutions from the B-factor plot. Due to the inverted resolution on the y axis, a point below the B-factor line signifies a lower (i.e. worse) resolution than expected, and a point above the line a higher (i.e. better) one. In the following, the plots are presented and discussed for all parameters.

5.1.1. Motion correction

The parameter validation plots for the average global micrograph shift (avgGlobalShift) and the local shift of the patch the particle was located in (avgPatchShift) are displayed in figure 5.1. The local shift was determined for splitting the micro-

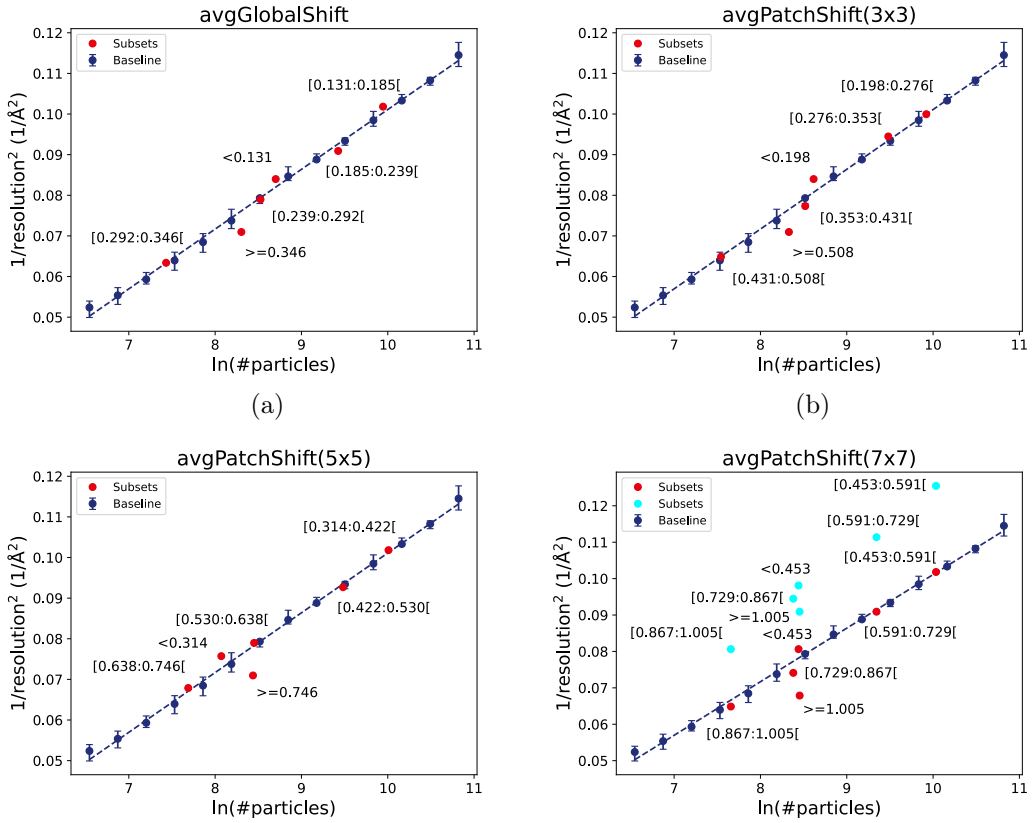


Figure 5.1.: FSC resolutions from motion parameter subsets (red) of the small FAS dataset plotted next to the B-factor plot indicating the expected resolutions for randomly sampled particle sets (blue). Each subset point is labeled with the corresponding parameter value range. The plot for the avgPatchShift(7x7) parameter contains additional points (light blue) for the same subsets after the application of Bayesian polishing [91].

graph into three, five and seven patches in x and y direction. Abnormally large local shifts can be a sign for image blurring and thus are associated with bad image quality. Therefore, it is expected that small shift values lead to high quality images while large shift values lead to low quality images. In all four plots in figure 5.1, the particle subsets corresponding to the largest shifts yield a resolution that is worse than expected from the linear fit of the B-factor points and even below the error bar of the next lower B-factor point. The difference becomes more pronounced with an increasing number of patches for the average patch shift parameter. The subsets corresponding to the smallest shifts all yield resolutions slightly above the B-factor line, but the difference is not as significant.

It was shown that FSC resolutions can be improved by Bayesian per-particle motion correction (Bayesian polishing) after a first high-resolution map is refined [91]. Therefore, Bayesian polishing was applied to the small FAS dataset and subsets based on the unchanged avgPatchShift(7x7) parameter were formed containing the same particles as before. New refinements were performed and the FSC resolutions were plotted into figure 5.1d. It can be observed that FSC resolution was improved for all datasets and is now better than the B-factor expectation from the unpolished data. Interestingly, the relative order of the subset results seems to have been preserved. This is best visible for the three almost vertically stacked points in the middle of the avgPatchShift(7x7) plot ($<$0.453, [0.729:0.867[, $\geq$1.005). Even though the points lie closer together on the y axis, the set with the smallest shift still yields the best result. These results suggest that wrong shifts in motion correction can be corrected for during polishing, but that particle sorting might still be beneficial as quality differences between the subsets remain.

5.1.2. CTF estimation

Figure 5.2 contains the validation plots for the parameters that describe the deviation of the defocus parameters (maximum z_{u}, minimum z_{v} and an in-plane rotation angle θ_{ast}) from the patch containing the respective particle to the entire micrograph or the neighboring patches. The patches were generated by splitting the micrographs into 5 pieces in the x and y directions analogously to patch splitting in MotionCor2 [31]. In general, the defocus should only vary slightly among neighboring patches and also not

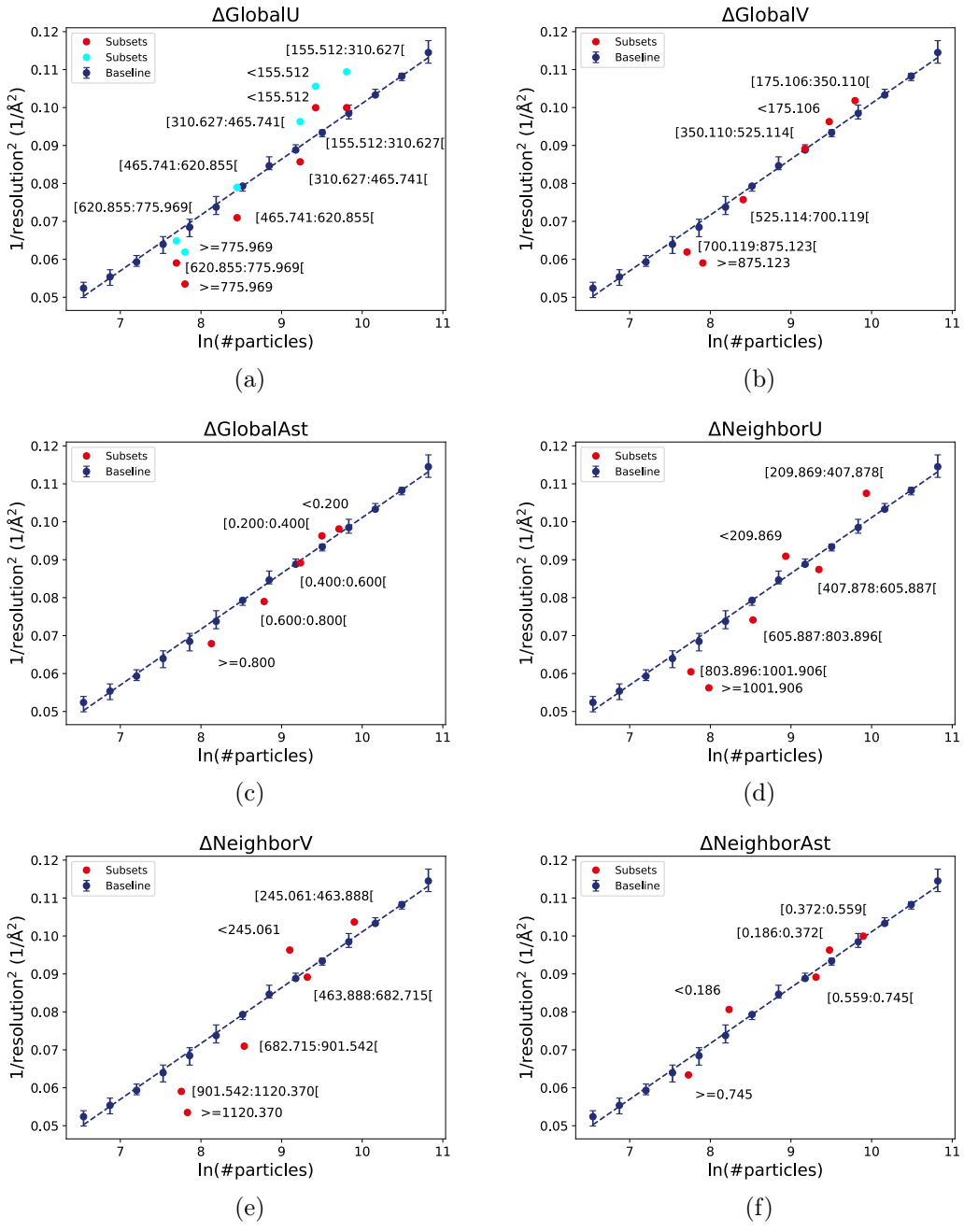


Figure 5.2.: FSC resolutions from subsets of the small FAS dataset based on CTF differences of patches to the global micrograph or neighboring patches (red) plotted next to the B-factor plot indicating the expected resolutions for randomly sampled particle sets (blue). Each subset point is labeled with the corresponding parameter value range. The plot for the ΔGlobalU parameter contains additional points (light blue) for the same subsets after the application of per-particle CTF refinement [49].

differ too much from local to global. Stronger deviations could be a sign for wrong parameters or deformation of the specimen.

All six plots show similar results. The subsets corresponding to larger deviations of the CTF parameters perform worse than expected while the subsets with the smallest deviations yield resolutions better than expected from the B-factor plot. This is less pronounced for the angle difference than for the maximum/minimum difference for both the local-global and the local-neighbor comparison. Like for the particle motion, there are methods available to optimize defocus parameters for each particle individually once a first high-resolution map is available. The small FAS dataset was subjected to RELION CTF refinement [49] to fit defocus and astigmatism for each particle. The refined particles were split into the same subsets as before by the unchanged ΔGlobalU parameter. New refinements were performed for each subset and the resolution plotted into the ΔGlobalU validation plot. As with the Bayesian polishing, the CTF refinement leads to a resolution improvement for all subsets, but once again the relative order of the subset resolutions remains similar. The subsets with the largest differences do not reach the expected resolution values from the unrefined B-factor plot in this case, suggesting that filtering of these particles might be beneficial. The performance of the best subsets can be further improved. The validation plot for the CTF fit resolution is found in figure 5.3a. The two subsets corresponding to the worst CTF fit resolution values yield FSC resolutions slightly below expected in the refinements, but the deviation is not significant.

5.1.3. Particle picking

The validation plot for the picking counter indicating how often a particle was picked (section 3.5) is shown in figure 5.3b. The expectation was that particles picked by all three picking methods are more likely to be actual particles rather than contamination. In contrast, all three subsets yield FSC resolutions very close to the expected resolution of an equally sized random particle set indicating no predictive value for particle quality. Figure 5.3c contains the plot for the picker quality score reported by Gautomatch [33]. The references for picking were a number of projections from a strongly low-pass filtered FAS map. Therefore, the meaningfulness of this parameter in terms of describing similarity to the reference is limited for high-resolution applications. This may be a reason for the surprising result that particles with a higher

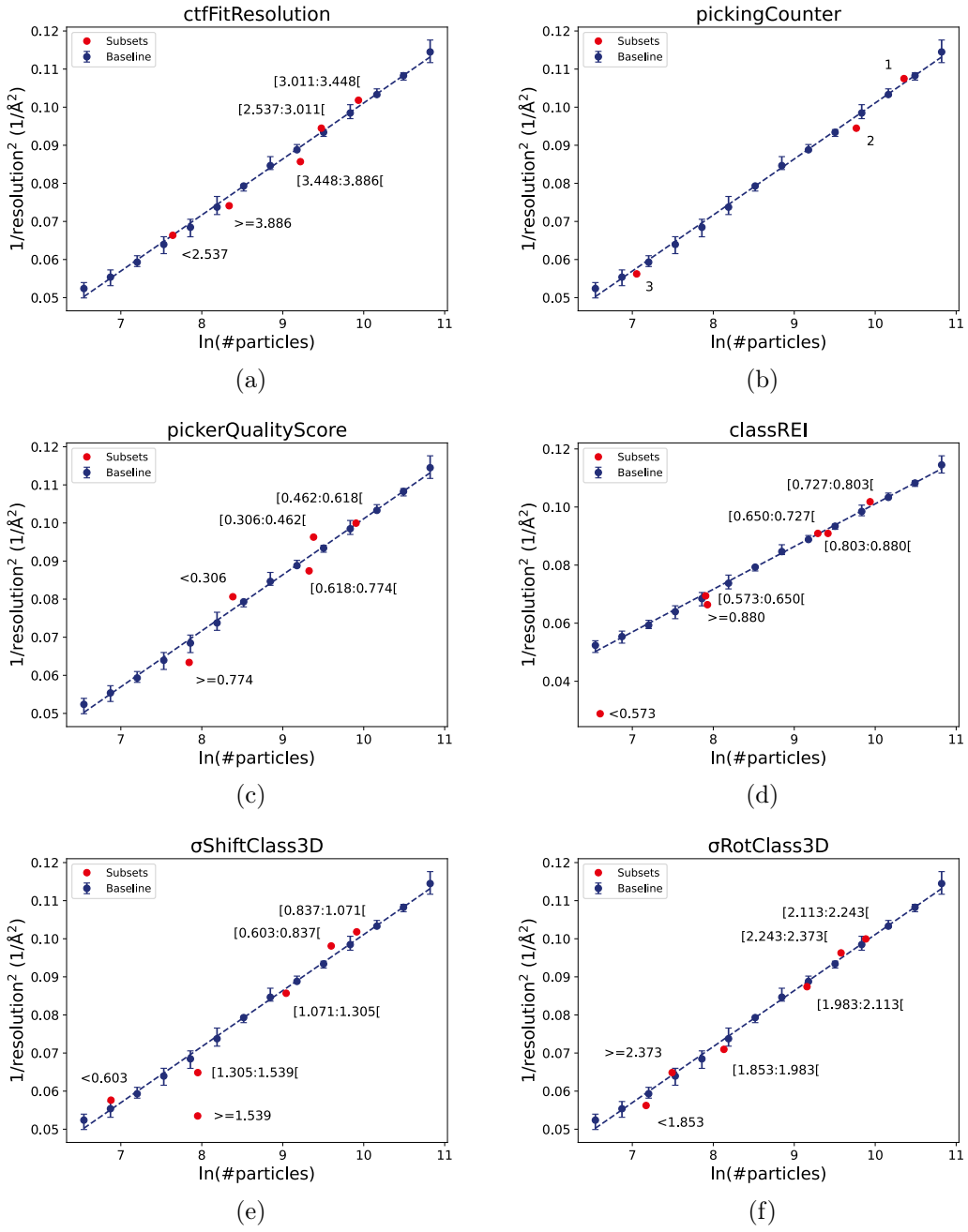


Figure 5.3.: FSC resolutions from subsets of the small FAS dataset based on CTF fit resolution, picking parameters or metadata from 3D classification (red) plotted next to the B-factor plot indicating the expected resolutions for randomly sampled particle sets (blue). Each subset point is labeled with the corresponding parameter value range. See table 3.1 for parameter explanations.

score actually yield worse FSC resolutions and particles with a lower score yield better resolutions. The picker quality score does not seem to be a helpful parameter for sorting high-resolution particle images.

5.1.4. 3D classification

The validation plots for the parameters from 3D classification are shown in figure 5.3d - 5.3f. The idea for these parameters was that high-quality particles should be reliably assignable and thus show less variation in optimal class, x-y shift and 3D orientation. For the class entropy (classREI) and the standard deviation of the particle orientation (σRotClass3D), nearly all subsets yield FSC resolutions that are very close the B-factor line i.e. very close to the expected values from random sampling. The subset with the smallest class variation even performs significantly worse than expected. However, the subset size is very small and the refinement might have experienced issues due to an unfortunate distribution of CTF zeros or 3D orientations. For the standard deviation of the applied x-y shift during 3D classification (σShiftClass3D), the subset corresponding to the largest standard deviations performs significantly worse than the expectation from the B-factor plot. The subset with the second largest deviations yields a resolution closer to the expected value but still worse than the lower error bound of the next smaller B-Factor point. The parameters classREI and σRotClass3D therefore seem unsuitable for particle filtering, while σShiftClass3D may have predictive value.

5.2. Particle filtering based on image metadata

In the validation above (section 5.1), the parameters avgPatchShift(7x7), ΔGlobalU, ΔNeighborV and σShiftClass3D showed the most promising correlation with particle quality. It was investigated whether the exclusion of the particle subset corresponding to the boxplot outliers, which yielded bad FSC resolution in the previous experiments, could lead to an FSC resolution improvement for the remaining small FAS dataset. The results are shown in table 5.1 together with the average result and error margins from the five refinement runs of the whole small dataset and an expected resolution for the given number of particles from the B-factor plot linear fit.

Table 5.1.: FSC resolutions of the small FAS dataset before and after filtering based on different quality parameters. Additional results after a single round of Bayesian Polishing and after a single round of CTF refinement are shown in table (b) and (d).

Data	**#Images**	**Resolution**
All	50,030	2.96 ± 0.04 Å
Expected	45,335	2.99 Å
<1.005 px	45,335	2.95 Å

(a) avgPatchShift(7x7)

Data	**#Images**	**Resolution**
All	50,030	2.69 Å
Expected	45,335	n.a.
<1.005 px	45,335	2.71 Å

(b) avgPatchShift(7x7)*

Data	**#Images**	**Resolution**
All	50,030	2.96 ± 0.04 Å
Expected	47,585	2.98 Å
<775.969 Å	47,585	2.94 Å

(c) ΔGlobalU

Data	**#Images**	**Resolution**
All	50,030	2.82 Å
Expected	47,585	n.a.
<775.969 Å	47,585	2.82 Å

(d) ΔGlobalU*

Data	**#Images**	**Resolution**
All	50,030	2.96 ± 0.04 Å
Expected	47,509	2.98 Å
<1120.370 Å	47,509	2.92 Å

(e) ΔNeighborV

Data	**#Images**	**Resolution**
All	50,030	2.96 ± 0.04 Å
Expected	47,197	2.98 Å
<1.539 Å	47,197	2.94 Å

(f) σShiftClass3D

For all four parameters, the resolution achieved upon removing the outlier particles is better than the expected resolution from the B-Factor linear fit and the average resolution of the whole dataset. The best resolution is achieved for the ΔNeighborV parameter with a resolution of 2.92 Å. However, the improvements are not significant as the resolutions after filtering all lie within the error bounds representing the minimum and maximum refinement results of the whole dataset. A single round of Bayesian polishing (table 5.1b) leads to a lower resolution for the whole small dataset. Here, the resolution cannot be further improved, but becomes worse when excluding the particles with the strongest local shift from the previous patch-based motion correction. After a single round of CTF refinement, the resolution for the whole small dataset improves as well, although not as much as through Bayesian polishing. Here, the resolution is unchanged when excluding the particles with the strongest deviation from the global values during initial CTF estimation. After all, the results suggest that despite the subset-based quality differences shown in the previous sec-

tion, metadata-based particle filtering does not lead to significant improvements in map resolution and that Bayesian polishing and CTF refinement might be able to mitigate the effects of strong local shifts and locally differing CTF parameters.

5.3. Combined filtering based on image metadata

After evaluating the filtering capacity of the quality parameters individually, the effect of combining them was evaluated. The same four parameters as in the previous section (avgPatchShift(7x7), ΔGlobalU, ΔNeighborV, σShiftClass3D) were chosen, as they showed the most promising results during subset evaluation. Additionally, the avgGlobalShift parameter was used for the combined evaluation to make the selection more balanced in terms of processing steps. The avgGlobalShift parameter also showed some effect during subset evaluation, although weaker than the other parameters. First of all, correlations of the parameters were analyzed (section 5.3.1). Then, a filtering cascade (section 5.3.2) and filtering based on combined ranking (section 5.3.3) were carried out.

5.3.1. Correlation of parameters

In figure 5.4, all pairwise correlation plots for the five quality parameters for combined filtering are presented. The overview also contains histograms for each parameter. For a more detailed view of the distributions, the reader is referred to the appendix (section B.3), where the histograms are plotted individually.

In most cases, there appears to be little correlation between the parameters. Especially among parameters derived from different processing steps, no correlation is visible. This is favorable for combined filtered. If the parameters were strongly correlated, combined filtering would lead to the exclusion of the same particles as in the single-parameter case. Since they are, however, not strongly correlated, combined filtering might unveil additional low-quality particles.

There is some correlation between the avgGlobalShift and the avgPatchShift(7x7) parameter. Particles with large global shifts seem to have higher local shifts as well.

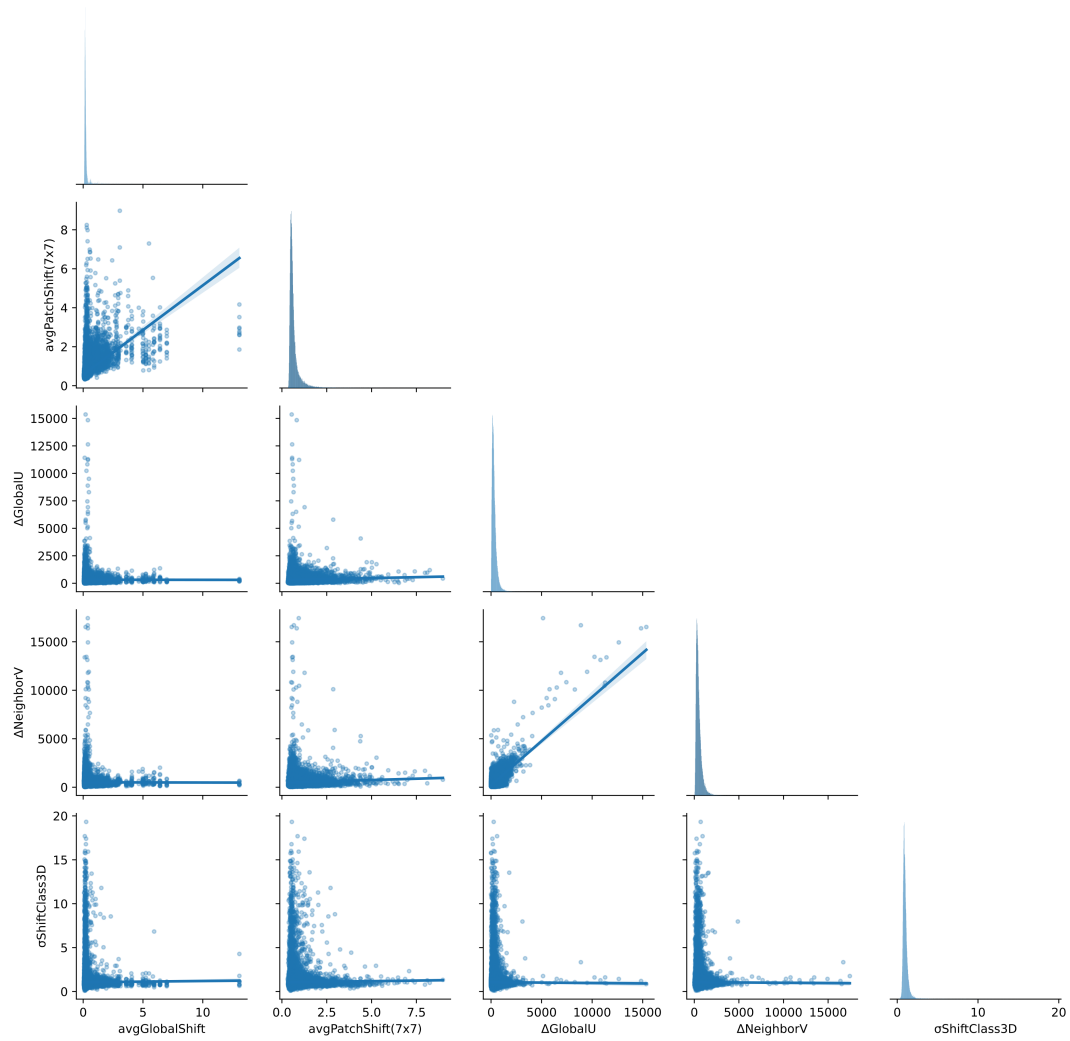


Figure 5.4.: Pairwise correlations of the five parameters used for combined filtering (avgGlobalShift, avgPatchShift(7x7), ΔGlobalU, ΔNeighborV, σShiftClass3D). On the diagonal, which would show the correlation of the parameter with itself, histograms are plotted instead. The upper triangle is left empty as the plots would be redundant to the lower triangle. In each pair plot, a linear regression line is shown. Created with seaborn [123].

There is, however, still a strong variation of the local shift among the particles with relatively low global shifts. The vertical bands in the plot arise from the fact that the avgGlobalShift is the same for all particles of a micrograph. The CTF-based parameters ΔGlobalU and ΔNeighborV also seem correlated, in particular with regard to the relatively few large outlier particles. In the area below 5,000 Å, where most particle images lie, the correlation is less sharp. This justifies the use of both parameters for combined filtering.

5.3.2. Filtering cascade

For the filtering cascade, the input images from the final baseline small FAS dataset were subjected to filtering by the following parameters and respective thresholds:

1. avgGlobalShift <0.346 px
2. avgPatchShift(7x7) <1.005 px
3. ΔGlobalU <777.969 Å
4. ΔNeighborV <1120.370 Å
5. σShiftClass3D <1.539 Å

The order of parameters in the cascade is arbitrary. It was chosen according to their order of appearance in the standard cryo-EM processing workflow. In each case, the thresholds were determined on the whole dataset to remove the boxplot outliers. Due to the computation on the whole dataset, a change in the filtering order will not have an influence on the number and choice of filtered particles after the entire cascade.

After each filtering step, a refinement was carried out with the remaining particles. In table 5.2, the FSC resolutions and remaining particle numbers are listed. It should be noted that numbers of filtered out particles in the individual filtering steps might be lower than if the filtering had been carried out on the whole dataset (see table 5.1). These numbers should therefore not be used to compare the parameters. Table 5.2 also contains resolution estimates based on the B-factor-plot for the small FAS dataset and the deviations between the predicted and experimentally determined values.

For all filtering steps, the actual FSC resolution is better than expected. However, the deviation is relatively small. The resolution only declines with ongoing filtering. It never improves significantly upon the baseline resolution of 2.96 ± 0.04 Å from the five refinement runs of the whole small dataset. At the same time, it also does not decline significantly with respect to the margins of the baseline resolution. After the last filtering step, the particle number has been reduced to 76 % of the initial number. This indicates that filtering by multiple parameters could potentially be used to reduce the number of particles for applications where computation power is a bottleneck. However, the amount of particle reduction achieved in this cascade with the boxplot outlier reduction was relative small. Therefore, also the predicted FSC degradations were relatively small, making the effect of the filtering difficult to interpret. A more rigorous particle filtering is performed with the combined ranking method in the following section.

Table 5.2.: Particle image numbers as well as predicted and experimentally determined FSC resolutions after each step of the filtering cascade on the small baseline FAS dataset. The resolution difference is calculated by subtracting the experimental from the predicted values.

Step	**#Images**	$\mathbf{FSC_{pred}}$ (Å)	$\mathbf{FSC_{exp}}$ (Å)	$\mathbf{\Delta FSC}$ (Å)
1	45,991	2.99	2.94	0.05
2	43,822	3.00	2.94	0.06
3	41,706	3.01	2.97	0.04
4	40,439	3.01	3.00	0.01
5	38,153	3.03	3.00	0.03

5.3.3. Combined ranking

The combined ranking approach was evaluated with the same five parameters as in the previous sections on the final baseline small FAS dataset. First, the combined rank was computed for all the images with the corresponding COW logic based on the five parameters. Afterwards, the images were sorted by their combined rank. The sorted dataset was subdivided into 10 equal-sized subsets and a refinement was carried out for each of them. The resulting FSC resolutions are listed in table 5.3 along with the expected FSC resolution for the subset size and the deviations from the expected value. The FSC deviations show a clear picture: The first subsets with the high-rank particles show significantly better results than the later subsets with the

low-rank particles. In fact, the FSC deviations consistently decline over the subsets. This suggests that the combined rank is a suitable predictor for image quality.

Table 5.3.: Predicted and experimentally determined FSC resolutions for subsets of the small baseline FAS dataset corresponding to different ranges of the combined rank. The subsets were derived from the dataset sorted in ascending order by combined rank, so that the first subsets contain small-rank images. The resolution difference is calculated by subtracting the experimental from the predicted values. Negative deviations are highlighted in red.

Step	**#Images**	**FSC$_{pred}$** (Å)	**FSC$_{exp}$** (Å)	**ΔFSC** (Å)
1	5,003	3.55	3.28	0.27
2	5,003	3.55	3.42	0.13
3	5,003	3.55	3.45	0.10
4	5,003	3.55	3.45	0.10
5	5,003	3.55	3.49	0.06
6	5,003	3.55	3.60	-0.05
7	5,003	3.55	3.63	-0.08
8	5,003	3.55	3.67	-0.12
9	5,003	3.55	3.80	-0.25
10	5,003	3.55	4.02	-0.47

In table 5.4, FSC resolutions are presented for increasingly larger subsets of size n, each corresponding to the first n particle images of the dataset sorted by combined rank. The subset resolutions are plotted next to the B-factor plot for the small FAS dataset in figure 5.5. The deviations from the expected resolutions are always positive, but decline as more and more large-rank particles are included. Subset 4 is an outlier with respect to that trend. Here, the FSC deviation is more positive than in the previous step. Remarkably, the resolution achieved with subset 4, which consists of only 40 % of the particle images, is within the margins of the resolution from five runs with the whole baseline set (2.96 ± 0.04 Å). In other words, the dataset size could be reduced significantly by filtering by the combined rank causing only a small loss of resolution. However, also the filtering by the combined rank did not lead to any resolution improvements.

While the results on the small FAS dataset look promising in terms of particle reduction, it should be investigated whether the effect remains after more rigorous processing with per-particle corrections like CTF refinement and particle polishing.

Table 5.4.: Particle image numbers as well as predicted and experimentally determined FSC resolutions for increasingly larger subsets of the small baseline FAS dataset, which was sorted by the combined rank. The subsets contain the first n particles of the sorted dataset. The resolution difference is calculated by subtracting the experimental from the predicted values.

Step	**#Images**	$\mathbf{FSC_{pred}}$ (Å)	$\mathbf{FSC_{exp}}$ (Å)	**ΔFSC** (Å)
1	5,003	3.55	3.28	0.27
2	10,006	3.34	3.19	0.15
3	15,009	3.24	3.13	0.11
4	20,012	3.17	3.00	0.17
5	25,015	3.12	3.02	0.10
6	30,018	3.08	3.00	0.08
7	35,021	3.04	3.00	0.04
8	40,024	3.02	3.00	0.02
9	45,027	2.99	2.97	0.02

The combined ranking approach was therefore additionally evaluated on the large FAS dataset, which was processed in full state-of-the-art manner including multiple rounds of per-particle corrections and yielded a high-resolution reconstruction.

In table 5.5, the results of a filtering series on the large FAS dataset are listed. The series was created by repeatedly removing 10 % of the original number of images from the dataset sorted by the combined rank. Note that the filtering series is in reverse order with respect to the series on the small dataset (table 5.4) to comply better with later results on the large FAS dataset. Table 5.5 contains the FSC resolutions after each filtering step as well as predicted resolutions and the difference between the two. Up to filtering step 4, the experimental FSC resolutions are very close to the expected values. In the following filtering steps, the resolution difference consistently increases up to a difference of 0.8 Å.

A visualization of the filtering series is given in figure 5.6. Based on the refined maps, the average Q-scores for the protein and the water molecules as well as the number of waters with a Q-score of at least 0.8 were computed. For the computation of the Q-scores, a single consensus model PDB-file was used. This model had previously been manually fitted to a refined map from the experimental FAS data and contains a total number of 7,680 water molecules. Consequently, the Q-score results in this thesis are only a measure of the resolvability of the atoms with respect to this specific model.

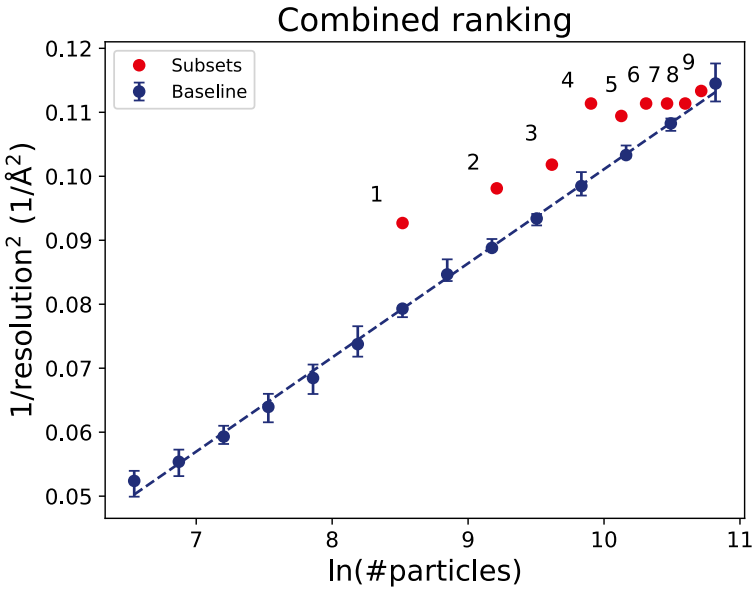


Figure 5.5.: FSC resolutions for increasingly larger subsets of the small FAS dataset sorted by the combined rank (red) plotted next to the B-factor plot indicating the expected resolutions for randomly sampled particle sets (blue). Each subset point is labeled with the subset number. The corresponding particle numbers and resolution values are listed in table 5.4.

Fitting an individual atomic model for each map might describe the individual map quality more accurately, however, this would not have been practically feasible for the large number of maps in this thesis. In addition, individual manual modeling might introduce bias, especially for waters, which appear as more or less spherical density blobs that can be confused with noise and become harder and harder to detect as resolution declines in the filtering series.

Next to the combined ranking series, a filtering series based on random particle exclusion is shown. The random series does not represent B-factor plot predictions, but was created by randomly removing particles from the dataset and calculating refinements afterwards. This was necessary to allow for the computation of Q-scores with the refined maps.

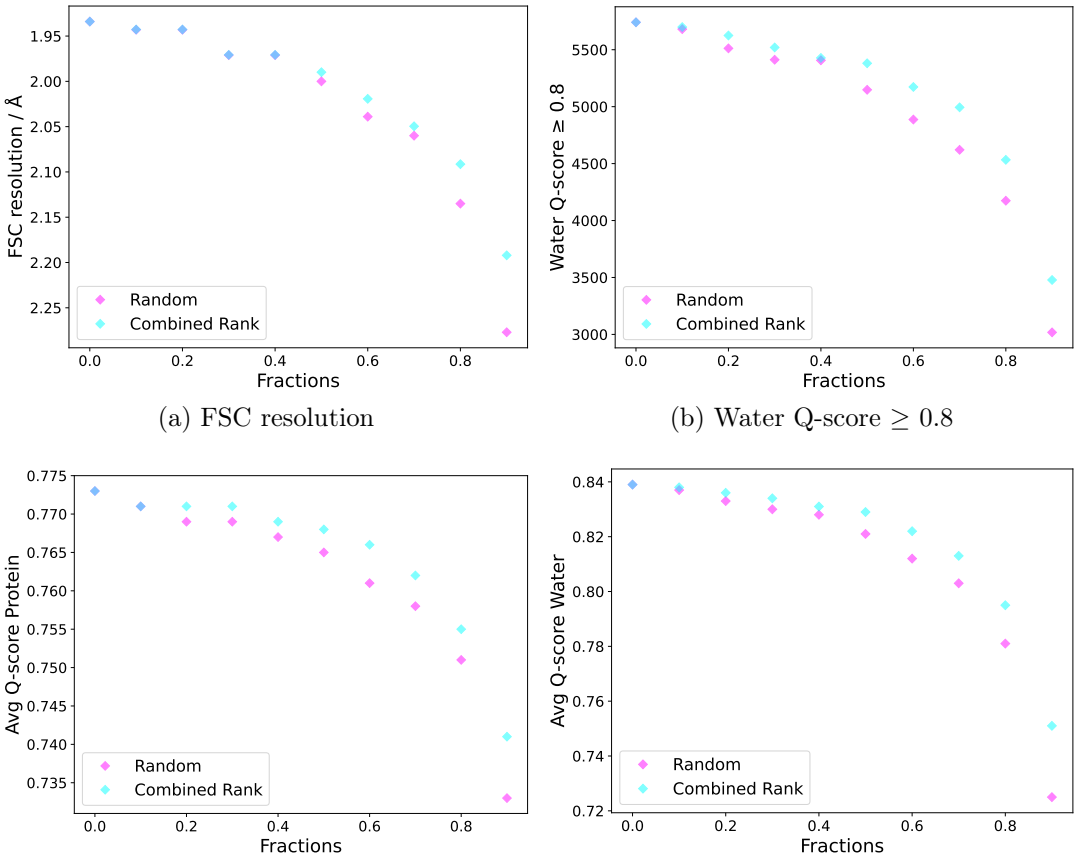
Like described above for the predicted values in table 5.5, the FSC resolution stays close to random up to filtering step 4. Afterwards, the FSC resolution from the combined rank filtering outperforms random exclusion. The difference is especially

Table 5.5.: Particle image numbers as well as predicted and experimentally determined FSC resolutions when filtering an increasing amount of images from the large baseline FAS dataset, which was sorted by the combined rank. The subsets contain the first n particles of the sorted dataset. The resolution difference is calculated by subtracting the experimental from the predicted values.

Step	**Filtering Fraction**	**#Images**	**FSC_{pred}** (Å)	**FSC_{exp}** (Å)	**ΔFSC** (Å)
0	-	246,726	-	1.93	-
1	0.1	221,995	1.93	1.94	-0.01
2	0.2	197,322	1.94	1.94	0.00
3	0.3	172,648	1.96	1.97	-0.01
4	0.4	147,981	1.98	1.97	0.01
5	0.5	123,332	2.01	1.99	0.02
6	0.6	98,613	2.04	2.02	0.02
7	0.7	73,960	2.08	2.05	0.03
8	0.8	49,272	2.15	2.09	0.05
9	0.9	24,613	2.27	2.19	0.08

distinct for the last two filtering steps. In terms of the Q-scores, the combined rank filtering series yields values that are consistently better than or equal to the random filtering values. The number of water molecules with a Q-score of at least 0.8 behaves similar to the FSC resolutions. The average Q-scores for the protein and the water molecules of the combined rank filtering series already start improving over the random filtering values at step 2.

The combined ranking approach was also tested in combination with the directional filtering method, where the particle images are binned by orientation and then filtering takes place within each bin (see section 3.5.4). The results were almost identical to regular filtering. The FSC resolutions for the filtering series were exactly the same and the Q-scores differed only by a small amount. As the differences were so small, the directional filtering results were not explicitly included in figure 5.6. This allows for a clearer comparison between the regular combined ranking and the random filtering series. A plot with all three filtering series can be found in the appendix (figure A.17).



(a) FSC resolution

(b) Water Q-score ≥ 0.8

(c) Avg Q-score Protein

(d) Avg Q-score Water

Figure 5.6.: Summary of different measures to estimate the quality of the refined maps when repeatedly filtering 10 % of the original images of the large FAS data set by the combined rank (cyan). As a reference, the plot shows the results for random exclusion (pink).

In summary, the results on the large FAS show a similar picture to small dataset results. The more particles with a large combined rank are removed, the more the FSC resolution improves over the expected value. At first glance, the deviations are more pronounced for the small dataset than for the large one. This may indicate that the per-particle corrections make up for the quality differences due to particle shifts and CTF errors to some extend. It should be considered, however, that the particle numbers and FSC resolutions are in general higher for the large FAS dataset. To achieve a resolution improvement at this regime, many more particles are required as can be deduced from the exponential nature of the Rosenthal-Henderson plot. Resolution differences at different resolution levels are thus difficult to compare and further experiments beyond the scope of this thesis would be needed to make a clear statement on the effect of the per-particle corrections. The fact that the FSC improvements at high filtering rates remain for the large dataset indicates that there are quality differences among the particles even after applying per-particle corrections. The combined rank-based filtering approach might be helpful in future scenarios if one has to extract the best particle images from huge amounts of data that are not processable anymore. At this point, however, the assumed quality differences do not outweigh the signal loss from excluding the lower ranking particles and filtering does not lead to an improvement over the whole baseline dataset in terms of FSC resolution or Q-scores.

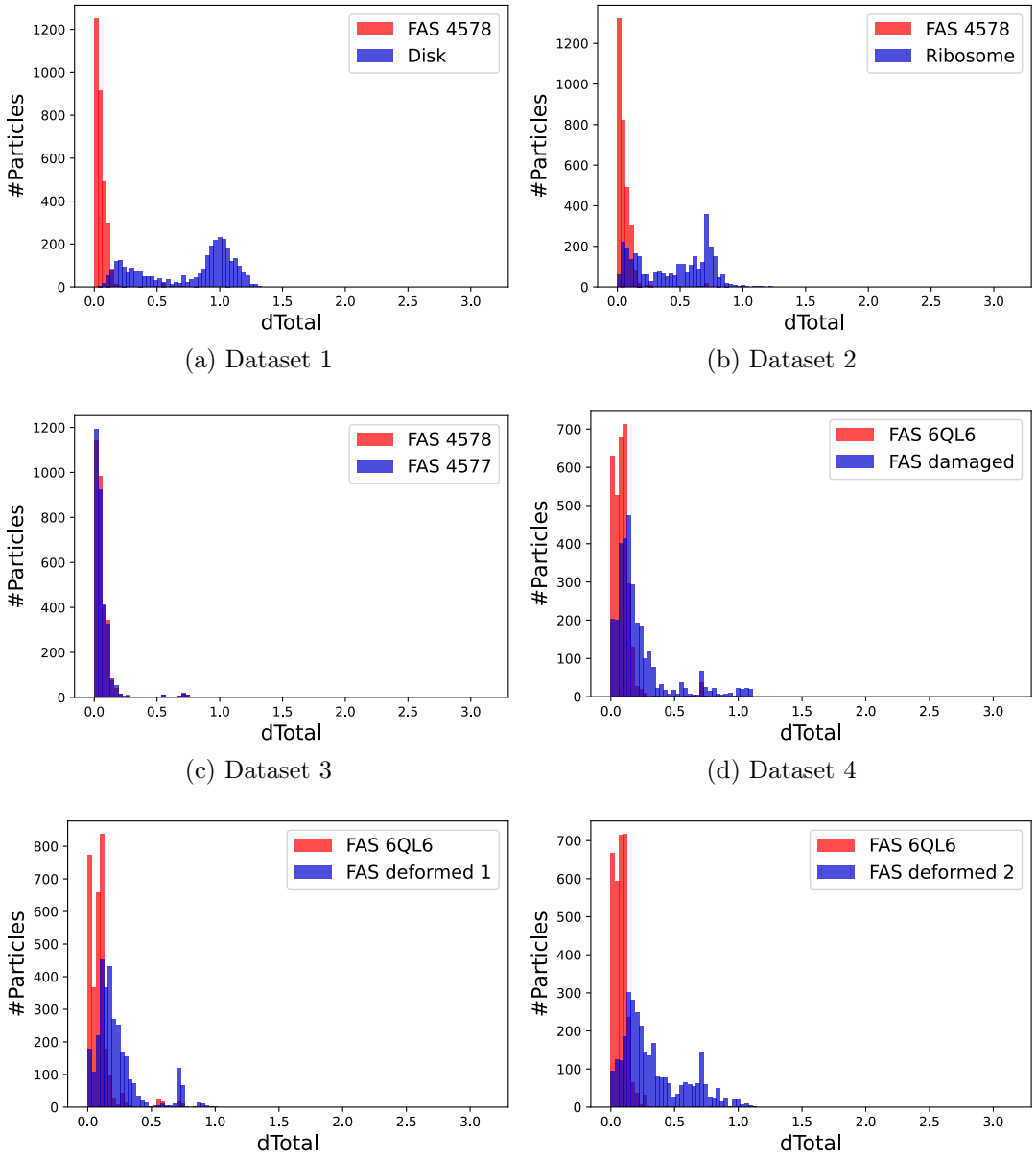
5.4. Orientation consistency

5.4.1. Simulated data: Separation of positive and negative images

The orientation consistency approach was first tested on the simulated datasets described in section 4.1. Each dataset was processed according to the workflow in figure 3.4 to determine three orientations for each image and the average pairwise distance between these orientations d_{total} was calculated. As reference for the projections, the respective positive map was used. The map was binned by factor 4 so that the pixel size was identical to the pixel size of the images. The symmetry group D3, which is known to be the symmetry group of the yeast FAS, was applied in all logics and during distance calculation.

It was tested, whether the d_{total} parameter could be used to separate positive and negative images. Figure 5.7 contains histograms of the d_{total} values for all datasets in table 4.1. The d_{total} values of the FAS images derived from the positive map are displayed in red and the d_{total} values of the negative images are displayed in blue. The histograms 5.7a and 5.7b, which compare the FAS images to the disk and the ribosome images, show that the majority of the d_{total} values for the positive images is below 0.2 while the values for the disk and ribosome images are spread between 0 and approximately 1.3. Splitting the dataset by the d_{total} values with a threshold of around 0.2 would eliminate nearly all of the disk images or the majority of the ribosome images respectively. In histogram 5.7c, there is barely any difference between the distributions of the d_{total} values for the positive images derived from the unrotated FAS map (EMDB 4578) and negative images derived from the rotated FAS map (EMDB 4577). The histograms 5.7d to 5.7f show the d_{total} distributions for the simulated FAS maps that were first damaged (5.7d), then deformed by moving two subunits (5.7e) and finally further deformed by moving additional subunits (5.7f) in contrast to the original structure (FAS 6QL6). In all three plots, the positive images mainly show d_{total} values of below 0.2, while the values for the negative image are more spread out. There is some overlap of positive and negative images in the area below 0.2, which makes their separation more difficult than e.g. in the disk case. However, the distribution of the d_{total} values for the negative images progressively shifts to the right from histogram 5.7d to histogram 5.7f. The separation of positive and negative images thus becomes easier the more the original FAS map is altered.

In addition to the histograms, receiver operating characteristic (ROC) plots were generated for all simulated datasets. For these plots, the datasets were sorted by d_{total} in ascending order. Then, the true positive rate was plotted against the false positive rate for different points originating from moving the positive/negative threshold along the sorted list. All images above the threshold are predicted positive, all images below are predicted negative. The lower left area of the plot thus describes the enrichment of true positive images in the first entries of the sorted list. An optimal curve should first rise vertically as more and more true positive images are included in the predicted positive images by the moving threshold. At the same time, the false positive rate remains zero as there are no negative images in the upper part of the optimally sorted list. Once all positive images are detected (true positive rate = 1), the optimal curve should go horizontally to the right as more and more negative images are predicted positive, which increases the false positive rate. The plots also



(a) Dataset 1 (b) Dataset 2

(c) Dataset 3 (d) Dataset 4

(e) Dataset 5 (f) Dataset 6

Figure 5.7.: Histograms of the orientation consistency d_{total} values for the simulated datasets 1 to 6 in table 4.1. The positive FAS images are displayed in red, the negative images in blue.

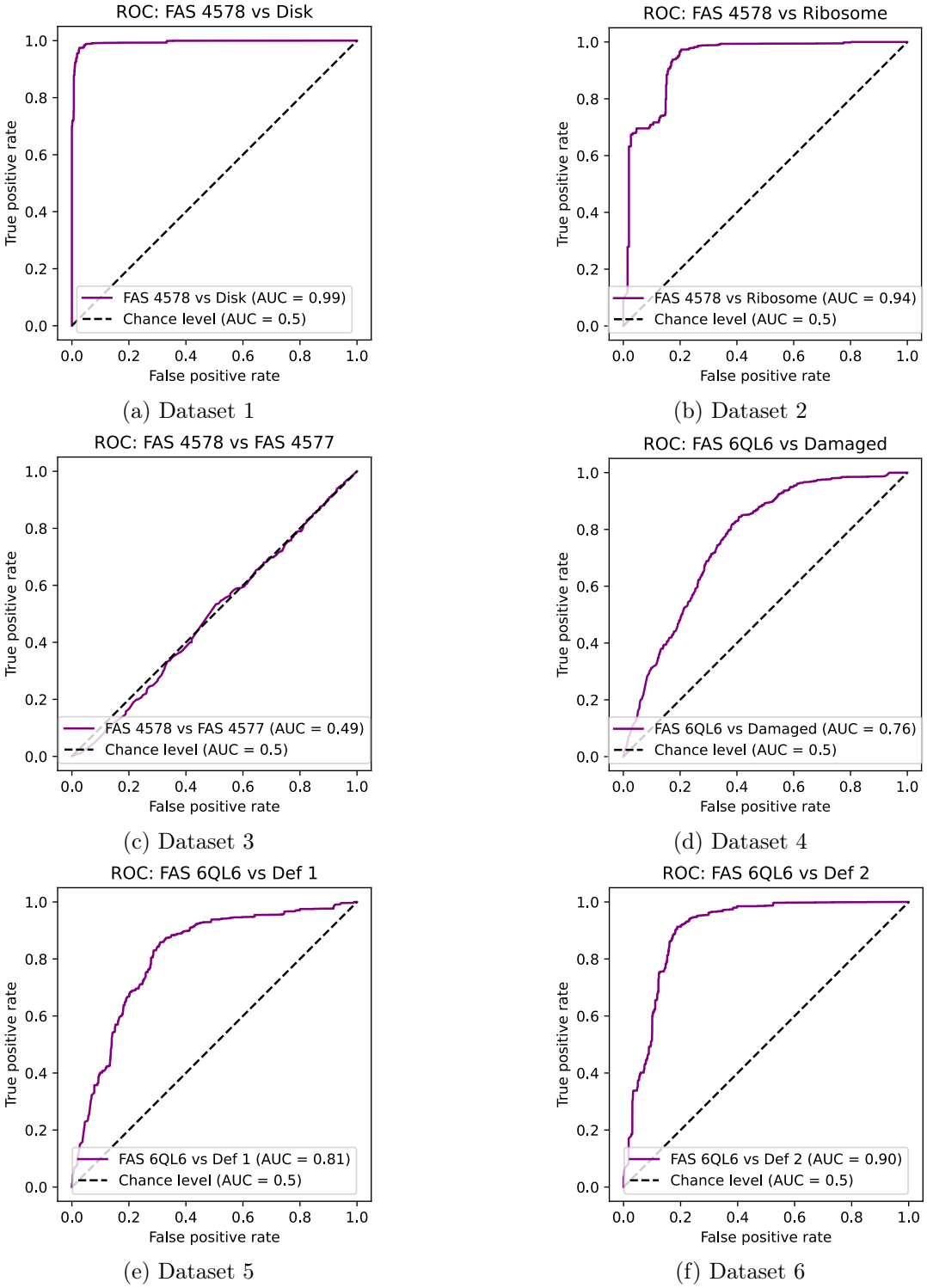
contain a curve describing random distribution of positive and negative images in the sorted list. This curve is a straight line from the lower left to the upper right corner as true positive and false positive rate increase simultaneously. The area under the ROC curve (AUC) can be used as a numerical measure for the separability of positive and negative images. It is 1 for the optimal ROC curve and 0.5 for the random curve.

The ROC curves are displayed in figure 5.8 and confirm the impressions from the histograms. The separability of the positive FAS images from the disk images is nearly perfect with an AUC of 0.99. The separability of the ribosome images is still good, but there is some mixing up of positive and negative images at intermediate d_{total} values leading to a dent in the ROC curve. The AUC is still quite high at 0.94. The images from the unrotated (4578) and rotated (4577) FAS maps are indistinguishable. The ROC curve runs along the random curve and the AUC is basically random at 0.49. The separability of images from the simulated datasets based on the damaged and deformed FAS (figure 5.8d to 5.8f) is significantly above random, but not as good as in the disk or ribosome case. The curves get closer to optimal and the AUC rises from 0.76 to 0.81 and finally 0.90 as the amount of deformation in the maps increases.

Overall, the experiments on the simulated datasets suggest that the d_{total} parameter can be used to separate positive images, which belong to the cryo-EM map of interest, from negative images, which do not show a protein structure, show a different protein structure or stem from a damaged and deformed structure. The quality of the separation improves with increasing deviation from the original map. Relatively small structural changes like the 15° dome rotation from FAS map 4578 to 4577 were not detectable. This limitation might be caused by the relatively strong coarsening applied to the images in these experiments.

5.4.2. Simulated data: Influence of symmetry consideration

When calculating the orientation consistency, the symmetry group can be specified in the workflow as well as in the distance calculation between the orientations. In the workflow, the main difference is whether 2D projections are generated for projection directions all over a sphere surrounding the 3D map (see figure 2.10) or just for the asymmetric region corresponding to the symmetry group. During pairwise distance calculation, the minimum distance between the first orientation and all transformed



(a) Dataset 1

(b) Dataset 2

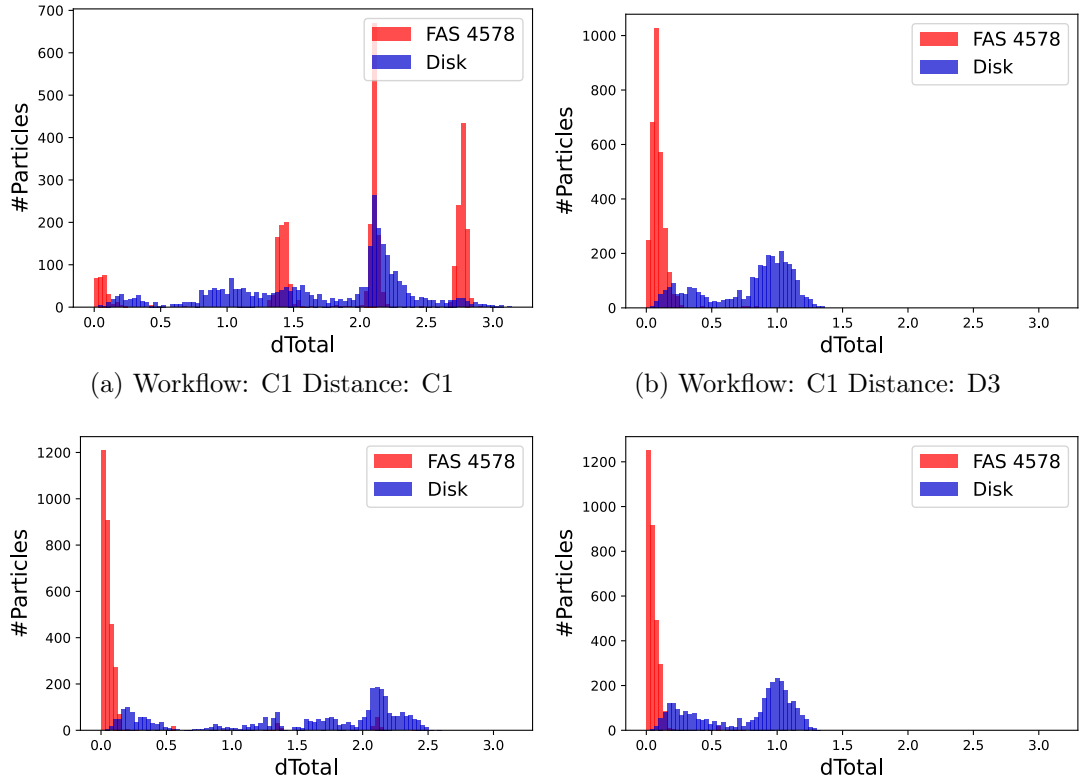
(c) Dataset 3

(d) Dataset 4

(e) Dataset 5

(f) Dataset 6

Figure 5.8.: Receiver operating characteristic (ROC) curves for separating the positive and negative images of the simulated datasets in table 4.1 by the orientation consistency parameter d_{total}. In each plot, the area under the curve (AUC) is displayed. The dashed line indicates the random distribution ROC curve.



(a) Workflow: C1 Distance: C1

(b) Workflow: C1 Distance: D3

(c) Workflow: D3 Distance: C1

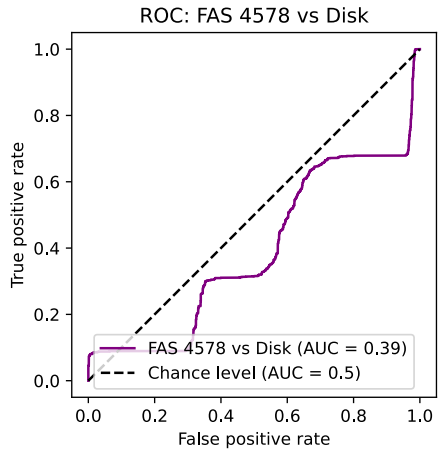
(d) Workflow: D3 Distance: D3

Figure 5.9.: Histograms of the orientation consistency d_{total} values for simulated dataset 1 (see table 4.1). The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. The positive images are displayed in red, the negative images in blue.

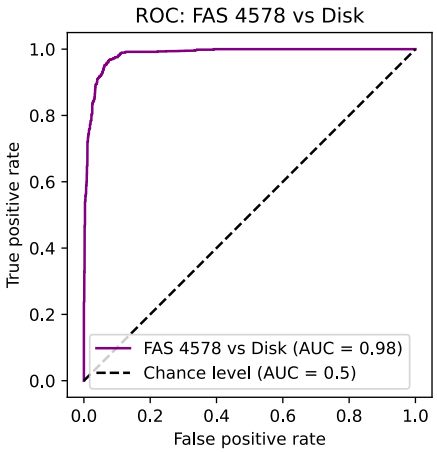
variants of the second orientation based on the symmetry group is reported. In this section, it is investigated how different choices of symmetry affect the distance values.

In figure 5.9 and 5.10, histograms of the d_{total} values for simulated dataset 1 and the corresponding ROC curves are displayed for different combinations of symmetry group C1 and D3 in the workflow and during distance calculation. When using C1 (no symmetry) in both workflow and distance calculation, the distribution has several peaks, especially in the case of the positive FAS images. The explanation for this is that the three methods for orientation determination assigned different orientations to the images which are, however, symmetry transformed variants of each other. Since the FAS map has in fact D3 symmetry, the symmetry-related 2D projections are indistinguishable for the algorithms. The assignments are thus not wrong and would give a correct result during reconstruction, but the d_{total} values are overestimated. This makes a threshold-based separation of positive and negative images impossible, as the ROC curve (figure 5.10a) with an AUC of 0.39 clearly shows. When using D3 symmetry in the distance calculation on the same orientations from the C1 workflow, there is no longer the observation of multiple peaks for the positive images. Due to the minimization over all symmetry transforms, the d_{total} values for the positive images are folded into a single peak close to zero. The values for the negative images are also reduced, but not to the same degree leading to a peak around a distance of 1.0. The positive and negative images can now be separated by a d_{total} threshold and the AUC of the ROC curve rises to 0.98.

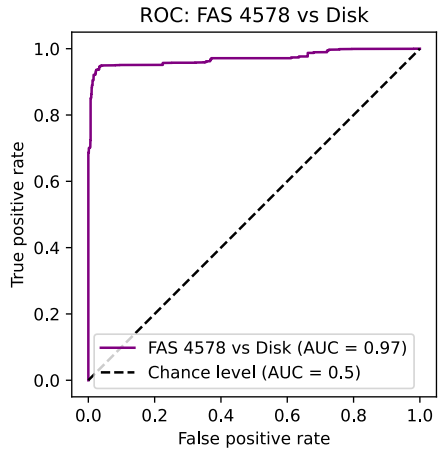
When D3 symmetry is applied directly in the workflow, the majority of distance values for the FAS images is close to zero even when not considering symmetry during distance calculation (figure 5.9c). Since only a single variant of the symmetry related projection images was generated during 2D projection, the alignments of the images and class averages cannot assign a symmetry transformed orientation. The angular reconstitution logic in COW is not restricted in that sense, which probably leads to the few larger d_{total} values for the positive images visible when closely examining the distribution e.g. at the d_{total} values of around 2.1. This probably leads to the slightly reduced ROC-AUC of 0.97. Interestingly, the d_{total} values for the negative images are almost as spread out as in case of using C1 symmetry in the workflow, which suggests that many orientations outside of the asymmetric triangle corresponding to the D3 symmetry were assigned during angular reconstitution. When using D3 symmetry



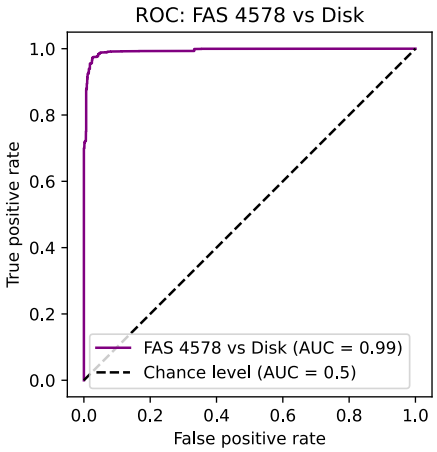
(a) Workflow: C1 Distance: C1



(b) Workflow: C1 Distance: D3



(c) Workflow: D3 Distance: C1



(d) Workflow: D3 Distance: D3

Figure 5.10.: Receiver operating characteristic (ROC) curves for separating the positive and negative images of simulated dataset 1 (see table 4.1) by the orientation consistency parameter d_{total}. The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. In each plot, the area under the curve (AUC) is displayed. The dashed line indicates the random distribution ROC curve.

during distance calculation with the orientations from the D3 workflow, the outlier d_{total} values for the positive images disappear and the ROC-AUC rises to 0.99 which is the maximum of the four symmetry cases discussed here.

The different symmetry combinations were tested for all six datasets in table 4.1 and the ROC-AUC values are summarized in table 5.6. The corresponding histograms and ROC curves can be found in the appendix (section A.2). In all cases except for dataset 3, the best results are achieved when applying the correct symmetry (here D3) during distance calculation. This shows that the distance calculation methodology developed in this thesis as presented in section 3.4 is reasonable. In the case of dataset 3, no meaningful deductions are possible since positive and negative images behaved equally and the image separation was close to random (0.5) in all cases. For two datasets in table 5.6, a slightly better AUC is achieved when using the D3 distance calculation on orientations from the C1 workflow. This may be caused by a different orientation distribution of the negative images. The additional 2D projections in the C1 case decrease the probability that the images are assigned three times the same orientation by chance. For three datasets, the best separability was achieved when using the D3 distance calculation on D3 workflow orientations. Therefore, no clear recommendation which symmetry to use in the workflow can be deduced. After all, the corresponding AUC values are close so that both options seam reasonable. In real world examples, the datasets are much larger than in the simulated case. Here, it saves substantial computing costs to apply the correct symmetry in the workflow, as fewer 2D projections have to be tried during alignment. In case the symmetry of the map is unknown and C1 was used in the orientation consistency workflow, multiple peaks can be a sign that a different symmetry group is valid for the data. Once this symmetry group is correctly determined, the distance calculation can quickly be repeated with the correct symmetry without the need to repeat the entire orientation determination workflow.

5.4.3. Real data results

After the analysis on the simulated datasets, the orientation consistency method was applied to the large FAS dataset. The dataset was processed as shown in figure 4.5 yielding a FAS map at a resolution of 1.93 Å based on a final stack of 246,726 particles. The final map and the final particle set were used as inputs for the orientation

Table 5.6.: Area under the curve (AUC) of the d_{total} ROC plots for all simulated datasets in table 4.1 using symmetry groups C1 and/or D3 in the workflow and during distance calculation. The best value per row is highlighted.

Dataset	Workflow C1		Workflow D3	
	Distance C1	Distance D3	Distance C1	Distance D3
1	0.39	0.98	0.97	**0.99**
2	0.62	**0.98**	0.91	0.94
3	**0.52**	0.51	0.51	0.49
4	0.52	**0.77**	0.72	0.76
5	0.57	0.79	0.77	**0.81**
6	0.56	0.88	0.86	**0.90**

determination workflow in figure 3.4 in order to calculate orientation consistency parameter d_{total} for all particles. For computational reasons, the particle images were coarsened by factor 8 and D3 symmetry was applied in the workflow. D3 symmetry was also specified for distance calculation. Based on d_{total}, the particles were divided into subsets as described in section 3.5.1 and refinements were computed for each subset.

Table 5.7.: Predicted and experimentally determined FSC resolutions for subsets of the large FAS dataset corresponding to different ranges of the orientation consistency parameter d_{total}. The resolution difference is calculated by subtracting the experimental from the predicted values. Negative deviations are highlighted in red.

Data	#Particles	FSC_{pred} (Å)	FSC_{exp} (Å)	ΔFSC (Å)
$<$0.011	2,532	2.94	3.76	-0.82
[0.011, 0.032[	82,543	2.06	2.06	0.00
[0.032, 0.053[	86,536	2.06	2.05	0.01
[0.053, 0.074[	35,928	2.20	2.19	0.01
[0.074, 0.096[	21,167	2.30	2.30	0.00
[0.096, 0.117[	8,571	2.52	2.53	0.01
$\geq$0.117	9,449	2.49	2.55	-0.06

In figure 5.11, the subset resolutions are plotted next to a B-factor plot of the large FAS dataset. The resolution values and subset sizes are listed in table 5.7. There is no evident relationship between the value ranges and the particle set quality. Most subsets yield resolutions close to the expected value for a randomly sampled particle set of the same size. The only outlier is in fact the set of particles with the smallest

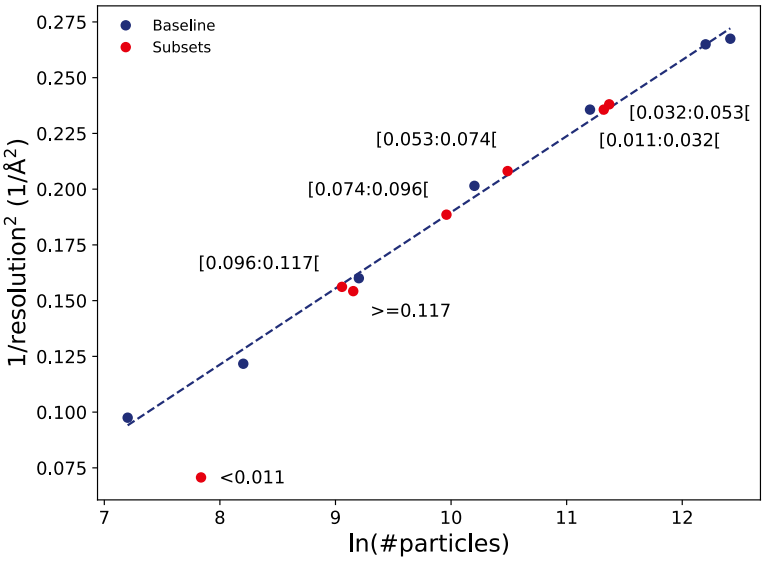


Figure 5.11.: FSC resolutions from subsets of the large FAS dataset based on the orientation consistency parameter d_{total} (red) plotted next to a B-factor plot indicating the expected resolutions for randomly sampled particle sets (blue). Each subset point is labeled with the corresponding parameter value range.

d_{total} values which was expected to contain the most reliably assigned images and therefore perform well. Instead, the resolution is much lower than expected. The particle images with a d_{total} value of 0.117 and above were removed from the dataset and a refinement was computed with the remaining 237,277 particles. The FSC resolution of this refinement was 1.94 Å, which is slightly worse than the 1.93 Å achieved in the baseline refinement.

5.5. Directional filtering

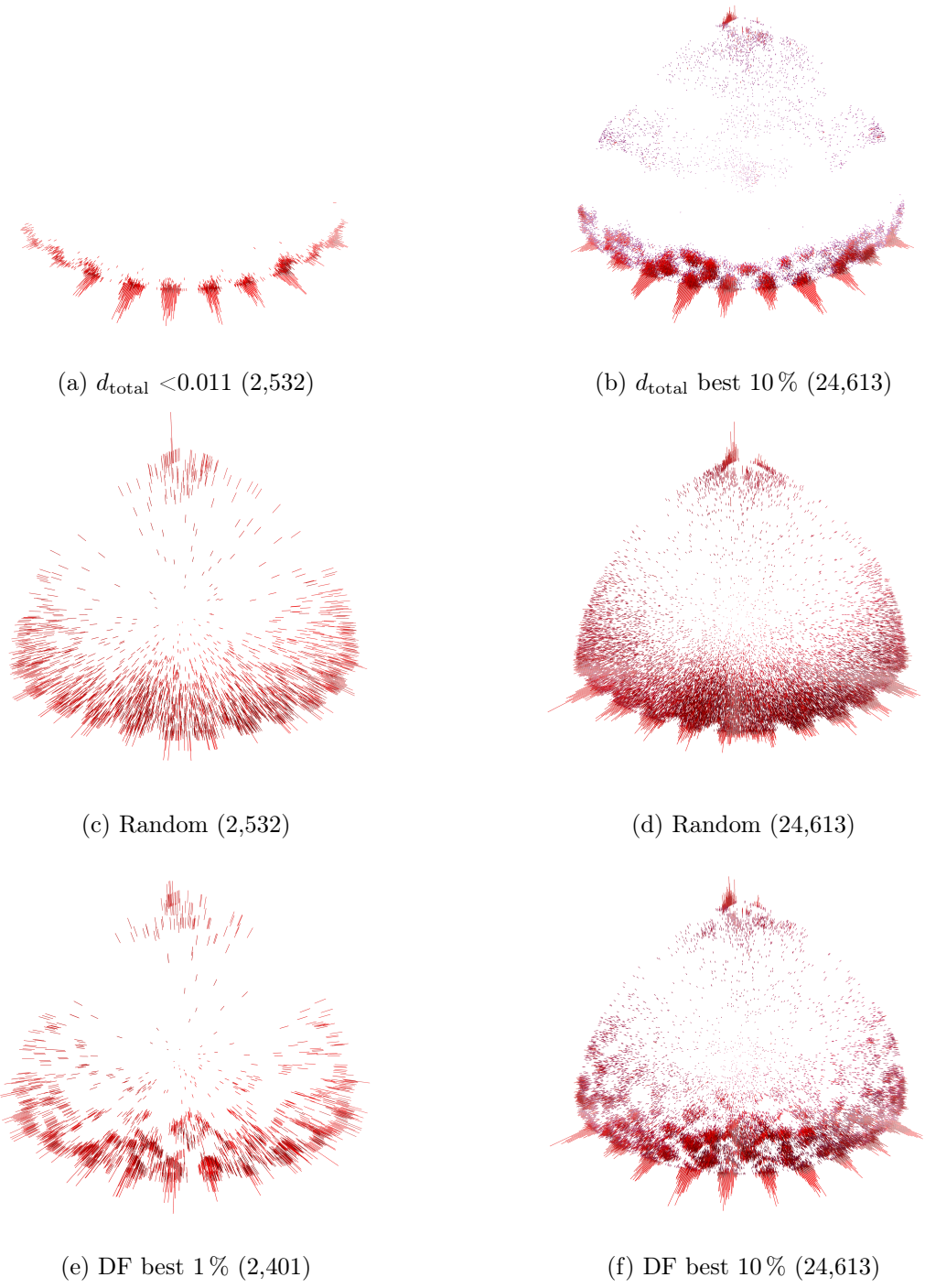
5.5.1. Orientation plots show filtering bias

In figure 5.12, the distribution of the computed projection directions is plotted for all particles of the d_{total} <0.011 subset (figure 5.12a) as well as for a random subset of the same size (figure 5.12c). Evidently, the distribution of directions is nonuniform for the d_{total} <0.011 subset. The particles orientations mainly correspond to side views of the FAS. This can limit the 3D reconstruction and potentially explains the unexpectedly bad FSC resolution of the subset. In figure 5.12e, the orientation distribution is shown for a similarly sized subset generated by the directional filtering (DF) approach with HEALPix level 3. The filtering parameter was d_{total} and the best 1 % of particles were kept in each bin. The resulting orientation distribution resembles the random sampling. The directional filtering approach thus seems to circumvent the orientation bias created when simply filtering by the d_{total} parameter. A 3D refinement of the particles after directional filtering gave an FSC resolution of 3.04 Å, which is significantly better than the 3.76 Å of the d_{total} <0.011 subset.

In the right column of figure 5.12, projection direction distributions are shown for selecting the best 10 % of the data by single-parameter d_{total} filtering, directional filtering and random sampling. Also in this case, the distribution is non-uniform for the single-parameter filtering, showing an area of missing orientations above the strongly populated equatorial ring. The directional filtering distribution is more uniform and resembles the random sampling distribution. Overall, the results suggest that the directional filtering strategy can prevent artificially imposed orientation bias in particle filtering scenarios and thus lead to better FSC resolutions in the filtered datasets.

5.5.2. Application to the large FAS dataset

The directional filtering approach was applied to the large FAS dataset. A series of gradually filtering an increasing amount of particles (10 % of the original number in each step) was created and the FSC resolution was monitored. In figure 5.13, the results are displayed for directional filtering with the d_{total} parameter and HEALPix level 3 as well as for single-parameter d_{total} filtering and random exclusion of particles.



(a) d_{total} <0.011 (2,532)

(b) d_{total} best 10 % (24,613)

(c) Random (2,532)

(d) Random (24,613)

(e) DF best 1 % (2,401)

(f) DF best 10 % (24,613)

Figure 5.12.: Distributions of the projection directions of particles selected by d_{total}, random sampling and directional filtering (DF). Orientations originate from a refinement of the whole large FAS dataset with D3 symmetry. The plots were created with UCSF pyem [124] and visualized with UCSF ChimeraX [122].

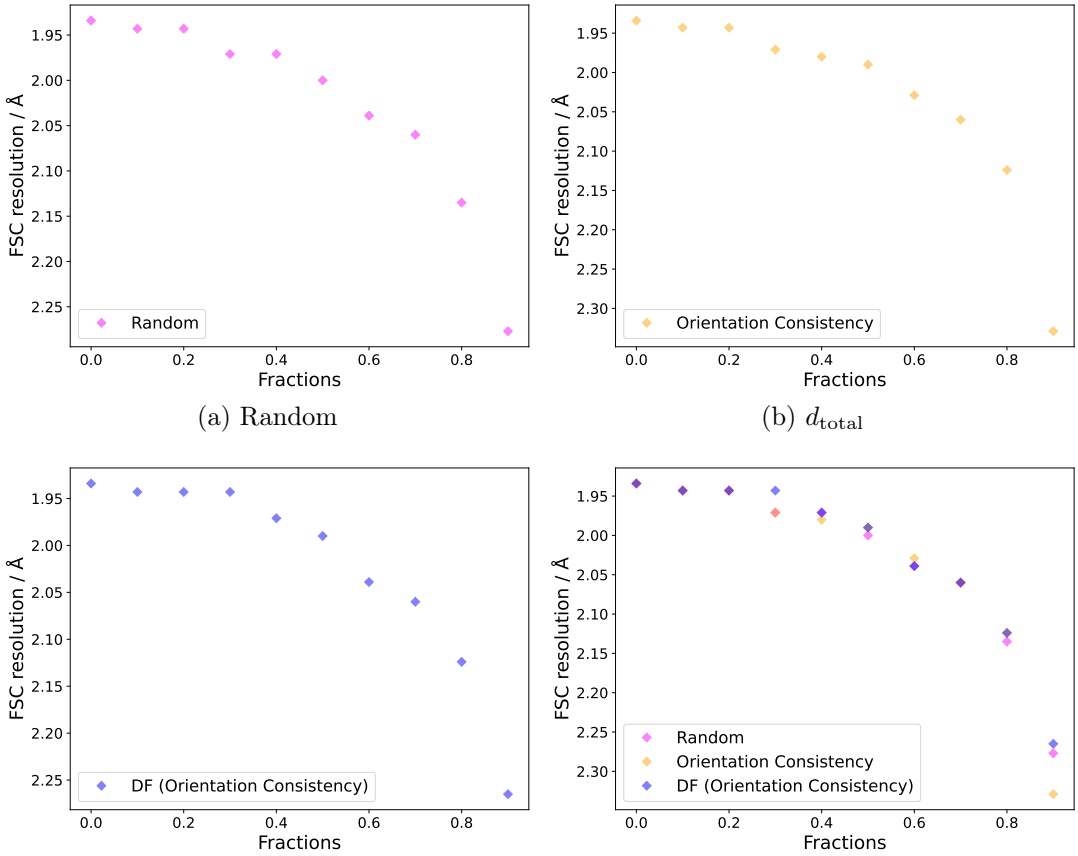
The exact particle numbers and FSC resolutions for each step are summarized in the appendix (section A.3).

In all three cases, the resolution worsens as the number of particle images is reduced. The best resolution is always achieved using the entire particle set. For the directional filtering approach, the resolution remains close to the original one when filtering up to 30 % of the particles, which is one filtering step longer than in the case of random or single-parameter d_{total} filtering. In the comparison plot (figure 5.13d), the directional filtering approach performs best in all but one filtering step. However, the deviations are small and after all, there is no significant improvement over random exclusion of particles neither for the single-parameter nor for the directional d_{total} filtering. In conclusion, the application of directional filtering with the d_{total} parameter to the experimental large FAS dataset does not lead to an improved density map in terms of FSC resolution.

5.6. Comparison to a state-of-the-art filtering method

In this section, the filtering approaches developed in this thesis are compared to the external filtering software CryoSieve [51] (see also section 1.2). It was shown that CryoSieve can significantly reduce the number of particles in a cryo-EM dataset while maintaining the resolution of the refined 3D map [51]. For comparison, a filtering series of repeatedly filtering 10 % of the original number of particles was carried out on the large FAS dataset using the CryoSieve algorithm. CryoSieve usually takes a fixed filtering percentage which is applied on the remaining particles in each round. Instead of filtering a fixed number of particles per round, the number of filtered particles therefore gradually decreases. To ensure the filtering steps contain approximately the same number of particles as in the filtering series based on the d_{total} parameter, the CryoSieve script was adjusted accordingly. This adjustment was a prerequisite for fair comparison since resolution and map quality depend on the number of particles used for refinement.

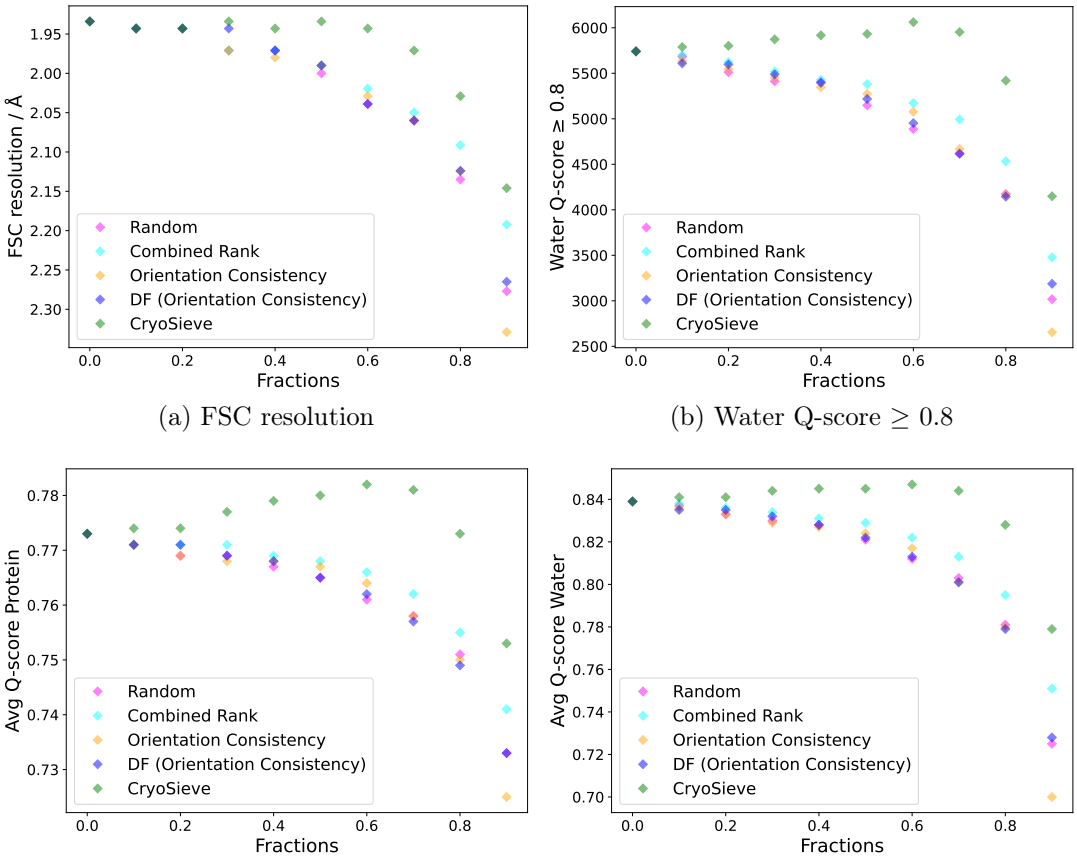
The results of the 3D refinements from the particle sets of the CryoSieve filtering series were compared to the previous results in section 5.5.2 and 5.3.3. Comparison plots are presented in figure 5.14. In addition to the FSC resolution, Q-scores [54]



(a) Random (b) d_{total}

(c) DF (d_{total}) (d) Comparison

Figure 5.13.: Development of the FSC resolution when repeatedly filtering 10 % of the original data of the large FAS data set. The images (a) to (c) show the filtering series for random exclusion (pink), single-parameter filtering by the d_{total} parameter (orange) and directional filtering by the d_{total} parameter (blue). In image (d), all three series are plotted together for better comparison.



(a) FSC resolution

(b) Water Q-score ≥ 0.8

(c) Avg Q-score Protein

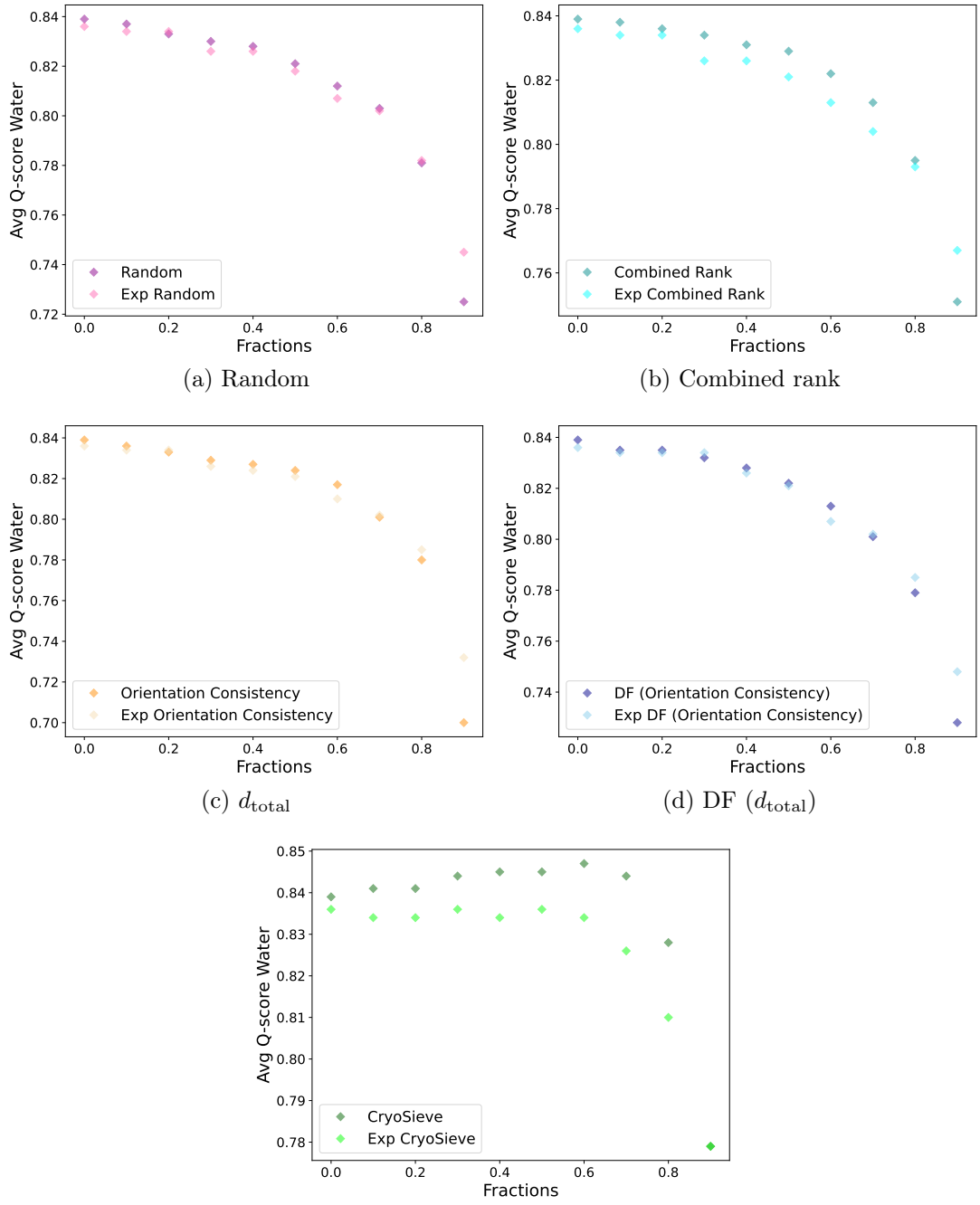
(d) Avg Q-score Water

Figure 5.14.: Summary of different measures to estimate the quality of the refined maps when repeatedly filtering 10 % of the original images of the large FAS data set by different filtering methods. Each plot shows the results for random exclusion (pink), single-parameter filtering by the combined rank (cyan) and the d_{total} parameter (orange), directional filtering by the d_{total} parameter (blue) and the external software CryoSieve [51] (green).

were used as a quality criterion. The Q-scores were computed using the same atomic model as described in section 5.3.3, which contained a total number of 7,680 modeled water molecules. The average Q-scores for the protein atoms and for the water atoms are reported. Furthermore, the number of waters with a Q-score of 0.8 or higher was determined. In the original Q-score publication [54], Q-scores at or above 0.8 are considered high, the corresponding waters can thus be considered reliable. All values corresponding to the figures in this section are summarized in tables in the appendix (section A.3).

CryoSieve clearly outperforms the other filtering methods with respect to all metrics. While single-parameter and directional filtering based on the d_{total} parameter yield results close to random, the FSC resolution in the CryoSieve series stays close to the whole particle set resolution when filtering up to 60 % of the original images. The combined ranking approach yields slightly better results than orientation consistency variants, especially in the later filtering rounds, but is still outperformed by CryoSieve. Also in the last step of the filtering series, where only the supposedly best 10 % of particle images remain, the CryoSieve set clearly outperforms the sets of the other methods. The resolution in the filtering series, however, never improves over the resolution of the entire dataset. In contrast to that, the Q-score values improve in the course of the CryoSieve filtering series. In the baseline map, 5,741 of the 7,680 waters in the map are supported with a Q-score of 0.8 or above. This number consistently increases up to a filtering fraction of 0.6, where 6,062 waters are supported with a Q-score of 0.8 or higher. From there, the number decreases again. The CryoSieve map at a filtering fraction of 0.6 also shows the best average Q-score for the protein and the waters. Based on Q-scores, this map consequently has the highest quality even though it was derived from only 40 % of the original data.

The Q-scores for the filtering series by random exclusion and single-parameter or directional filtering based on the d_{total} parameter consistently decrease or remain constant and thereby show a progression similar to the FSC resolution in the previous section 5.5.2. As described before, the filtering series based on the single-parameter or directional filtering with the d_{total} parameter behave very similarly to random exclusion regardless of the quality measure used. Once again, the performance of the combined ranking approach was slightly better than for random exclusion and the orientation consistency variants, especially in the later filtering rounds. Still, no improvement of Q-scores was observed in the combined ranking filtering series.



(a) Random

(b) Combined rank

(c) d_{total}

(d) DF (d_{total})

(e) CryoSieve

Figure 5.15.: Development of the average water Q-score of the refined maps when repeatedly filtering 10 % of the original images of the large FAS data set. The plots show the filtering series for random exclusion (pink), single-parameter filtering by the combined rank (cyan) and the d_{total} parameter (orange), directional filtering by the d_{total} parameter (blue) and the external CryoSieve software [51] (green). In each case, the expected water Q-score at the given map resolution is indicated in a lighter color.

The Q-score software reports expected Q-score values for each atom based on the estimated map resolution provided by the user. In figure 5.15, the average water Q-scores are plotted next to the expected water Q-scores for all filtering series individually. In the case of random exclusion and single-parameter or directional d_{total} filtering, the average water Q-scores are very close to the expected values throughout the filtering series with the exception of the last step, where the average Q-score is lower than expected. For the combined ranking approach, the water Q-scores are slightly above the expected values between the filtering steps 0.3 and 0.7. In the case of CryoSieve filtering, the average water Q-scores significantly exceed the expected values. Only in the last filtering step, the values are on par with the expectation.

In conclusion, the CryoSieve filtering approach can reduce the number of particles in the large FAS dataset to 40 % of the original data without a significant resolution loss and thereby even increases the Q-score values for protein and water atoms. It outperforms the filtering methods developed in this thesis, which give a slight improvement over random exclusion for the FSC resolution and Q-scores at best.

5.7. On-the-fly application of CryoSieve filtering

In the previous sections, particle filtering was applied as final step in the processing pipeline after a high resolution map of the FAS protein had already been reconstructed. In practice, it is beneficial to identify and remove low-quality particle images as early as possible in the processing pipeline, since this reduces time and resource requirements for all downstream processing tasks, especially 3D refinements. It was therefore investigated, whether CryoSieve filtering – which gave the only clear improvement on the FAS dataset – could also be applied on subsets of the data. A graphic representation of the subset filtering procedure is given in figure 5.16. The final large FAS dataset was divided into 15 subsets corresponding to different recording days (some subsets include multiple days). The subsets were the same as those used for 2D and 3D classification (see section 4.2). To make the results comparable to the filtering results on the whole final dataset, the particle images from Bayesian Polishing were used as well as the locally refined CTF parameters from CTF refine.

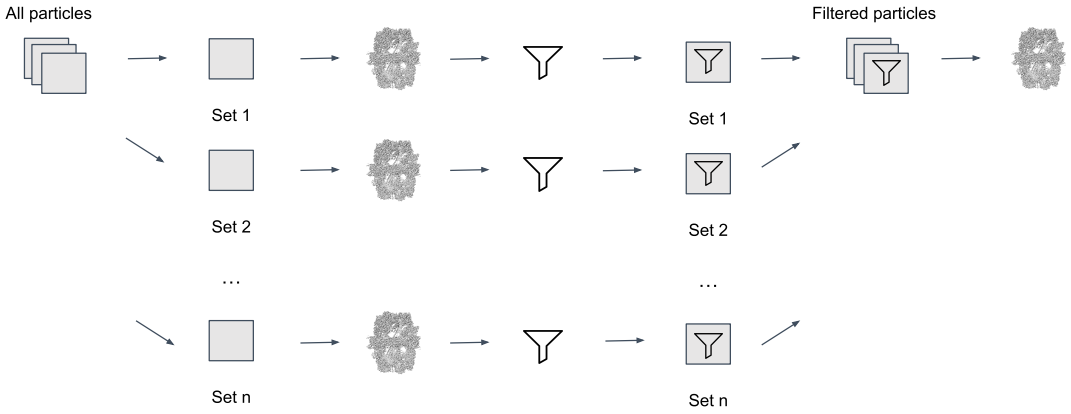


Figure 5.16.: Overview of the subset filtering procedure. The particle images are divided into subsets based on the recording days. A 3D map is refined and a filtering algorithm is applied to each subset individually. The filtered subsets are combined into a single final particle set. A consensus 3D refinement is carried out with the final particle stack.

The CryoSieve script, which had been adjusted to filter a constant number of particles in each iteration, was then applied to each particle subset individually and a filtered subset from one of the CryoSieve filtering iterations was selected. The 15 selected filtered subsets were combined into a single filtered particle set, which was subjected to 3D refinement. The subset filtering procedure was carried out twice. In the first version, the smallest of the filtered sets with the best resolution (first choice) was selected for each subset. In the second version, the second choice was selected for each subset to exclude more particles. The second choice is here defined as the smallest set with the best resolution among the sets smaller than the first choice set. The rationale behind this definition is that the second choice set should always be smaller than the first choice set.

As an example, in table 5.8, the FSC resolutions from the reconstructions during CryoSieve filtering are depicted for subset 3. The filtered sets which were selected for the consensus refinements are highlighted. In this filtering run, the resolution declines monotonically and at earlier iterations than for the whole set filtering described in section 5.6. This is the case for most of the subsets. All 15 subset tables can be found in the appendix (section A.4). Table A.16 shows an example of oscillating

resolution behavior, where the definition of the second choice set becomes useful. As a consequence of the faster resolution decline during subset filtering, fewer particles were filtered out overall. While the number of particles was reduced to 40 % when CryoSieve was applied on the whole dataset (section 5.6), the combined first choice filtered set contained approximately 82 % of the original particles (see table 5.9). The combined set of the second choice subsets still contained approximately 65 % of the images.

Table 5.8.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 3. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	**% Particles**	**FSC (Å)**
0	100	2.99
1	90	2.99
2	80	3.01
3	70	3.01
4	60	3.04
5	50	3.06
6	40	3.20
7	30	3.39
8	20	3.50
9	10	4.18

Table 5.9.: Results statistics of the combined filtered particle sets derived from the smallest subsets with the best resolution (first choice) and the smallest subsets with the second best resolution after the optimum (second choice) in comparison with the baseline final large FAS dataset.

Particle Set	**#Particles**	**FSC (Å)**	**#Waters(Q-score≥0.8)**
Baseline	246,726	1.93	5,741
First choice	202,316	1.92	5,803
Second choice	160,191	1.93	5,864

The 3D maps from the combined filtered particle sets were compared to the unfiltered final large FAS dataset in terms of FSC resolution and the number of waters with a Q-score of 0.8 or higher. The results are summarized in table 5.9. The FSC resolution of the combined first choice dataset improved over the baseline resolution while the resolution of the combined second dataset was equal to the baseline one. The number of reliable waters with a high Q-score of 0.8 or above increased for the first choice set, and then again for the second choice dataset. For the 40 % dataset derived by

CryoSieve filtering on the whole dataset, this number was even higher at 6,062 waters (see table A.6).

In conclusion, the experiments in this section suggest that CryoSieve filtering is more powerful in reducing the particle number when applied on the whole dataset. However, applying CryoSieve filtering on the subsets lead to a significant reduction of the number of particles without loss of FSC resolution and with an improvement of the water Q-scores. CryoSieve might therefore be a useful tool for reducing the number of particles early on in the processing workflow for future cryo-EM projects.

5.8. Relationship of subset filtering rates and particle quality

In the previous sections, different methods were applied to filter the large FAS dataset. For the external software CryoSieve, this filtering was additionally carried out individually on the 15 subsets of the large FAS dataset before combining the filtered subsets again (section 5.7). The 15 subsets correspond to different recording sessions on the microscope and could therefore potentially show different data quality. In this section it is analyzed, how many images were removed from each subset by the different global and per-subset filtering approaches and whether the filtering rates are connected to particle quality in terms of subset FSC resolutions.

In figure 5.17, the ratios of remaining particles with respect to the raw data directly after particle picking are plotted for each FAS subset before and after applying the different filtering strategies. The corresponding particle numbers are given in the appendix (table A.22). It is important to note that the subsets are quite variable in their absolute particle size, as visualized in figure 4.4. The ratios of remaining particles are displayed for the final baseline dataset after 2D and 3D classification and after filtering this final set by combined ranking, directional filtering with the d_{total} parameter or CryoSieve. Additionally, the ratios after applying CryoSieve individually on the subsets and selecting the first or second choice filtered subsets are shown.

The final FAS dataset after 2D and 3D classification contains particle images from all recording subsets. This is related to the fact that 2D and 3D classification were

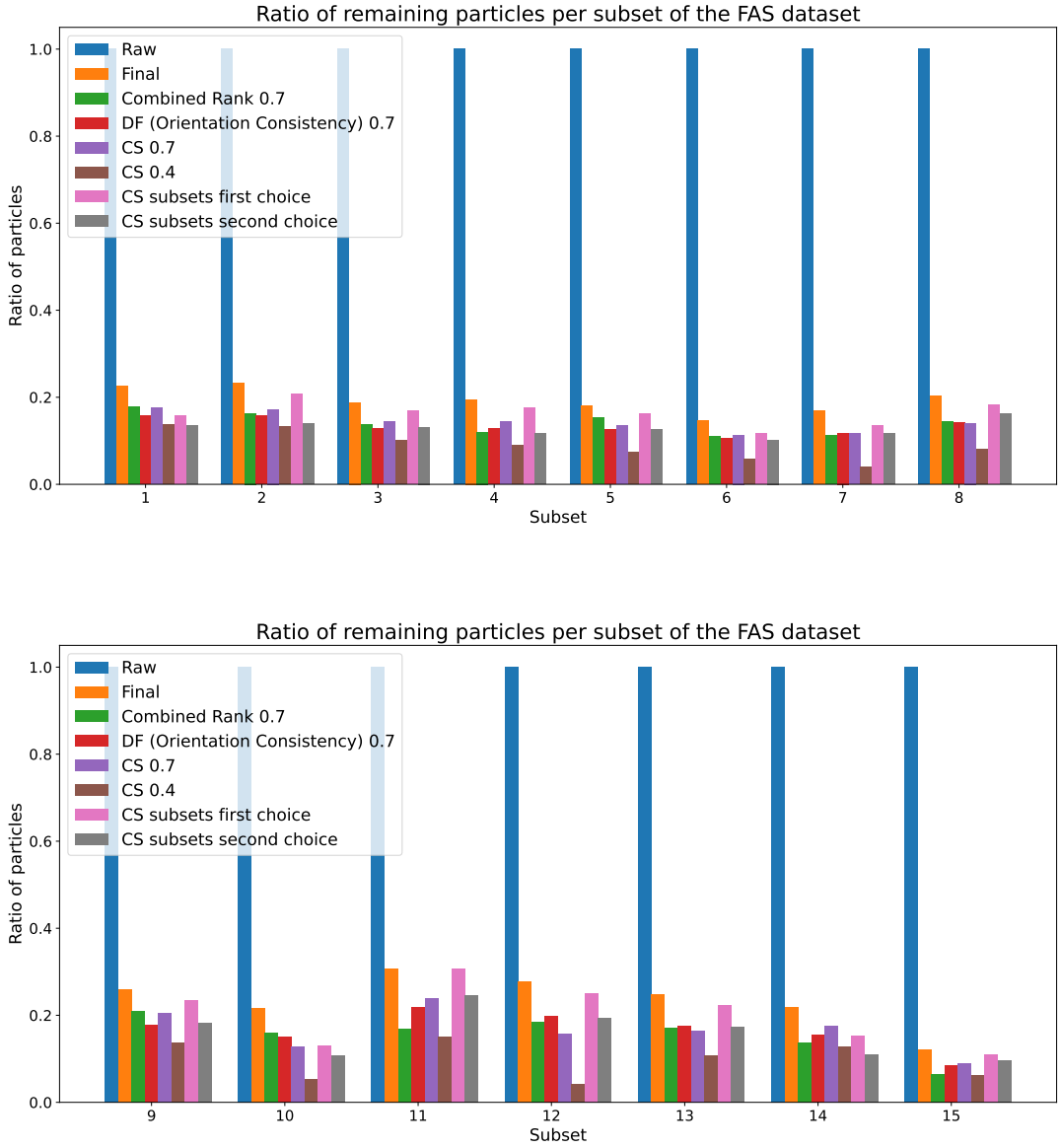


Figure 5.17.: Ratio of remaining particles of the large FAS dataset directly after picking (blue), after particle selection by 2D and 3D classification (orange), after reducing the final baseline dataset to 70 % by combined rank filtering (green) or directional filtering with the d_{total} parameter (red), filtering to 70 % (purple) and 40 % (brown) with CryoSieve and filtering each subset individually with CryoSieve and selecting the first (pink) and second (gray) choice subset in each case. The ratios were derived from the particle numbers by dividing by the raw particle number of the respective subset.

carried out per subset and all subsets showed meaningful class averages. Also after filtering by combined ranking, directional filtering based on the d_{total} parameter or CryoSieve on the whole final baseline dataset no subset is eliminated completely. This suggests that meaningful data is present in all subsets.

To provide a better overview of the filtering effects of the different methods, the kept ratios were also calculated relative to the final baseline dataset after 2D and 3D classification and listed in table 5.10. In the following, the term *kept ratio* refers to the relative amount of remaining particles after filtering, which is equal to $1-f$ where f is the *filtering fraction* used in previous sections to describe the relative amount of filtered particles. When the final dataset is filtered to 70 % of the original size using directional filtering, the particle number in each individual subset is also reduced to approximately 70 % (range: 66-72 %). This behavior might suggest that particles are effectively excluded in a random manner with this method.

Table 5.10.: Kept ratios with respect to the final baseline large FAS dataset for the 15 subsets when filtering by combined ranking to 70 % (CR 0.7), by directional filtering with the d_{total} parameter to 70 % (DF 0.7), with CryoSieve to 70 % (CS 0.7) and 40 % (CS 0.4), or with CryoSieve per subset while selecting the first (CS 1.) or second (CS 2.) choice filtered subsets. Additionally, the FSC deviations from the predicted values from table 5.11 are listed. The subsets with the best FSC deviations are highlighted in blue, and the ones with the worst FSC deviations are highlighted in red.

Set	**CR 0.7**	**DF 0.7**	**CS 0.7**	**CS 0.4**	**CS 1.**	**CS 2.**	**ΔFSC** (Å)
1	0.79	0.70	0.78	0.61	0.70	0.60	-0.03
2	0.71	0.68	0.74	0.57	0.90	0.60	-0.07
3	0.73	0.68	0.77	0.54	0.90	0.70	-0.03
4	0.62	0.66	0.74	0.46	0.90	0.60	0.07
5	0.85	0.70	0.75	0.41	0.90	0.70	0.12
6	0.75	0.72	0.78	0.40	0.80	0.70	0.07
7	0.67	0.70	0.69	0.24	0.80	0.70	0.02
8	0.71	0.70	0.69	0.39	0.90	0.80	-0.09
9	0.80	0.69	0.78	0.52	0.90	0.70	0.03
10	0.74	0.69	0.59	0.24	0.60	0.50	-0.02
11	0.55	0.71	0.77	0.49	1.00	0.80	0.06
12	0.66	0.72	0.57	0.15	0.90	0.70	0.02
13	0.69	0.70	0.66	0.43	0.90	0.70	-0.09
14	0.63	0.71	0.80	0.59	0.70	0.50	0.01
15	0.53	0.69	0.73	0.52	0.90	0.80	-0.12

In contrast, filtering to 70 % of the original size using CryoSieve leads to stronger variations of the filtered particle number among the subsets. Here, between 57 % and 80 % of the particles remain in each subset with respect to the particle number after 2D/3D classification. When filtering down to 40 %, the variation of the kept ratio among the subsets becomes even more pronounced (range: 15-61 %). There seems to be some consistency in the CryoSieve filtering: the subsets for which relatively more particles were filtered out when globally filtering to 70 % were most of the time also filtered more strongly when globally filtering to 40 %. See figure A.18 in the appendix for a visualization of the correlation. When applying CryoSieve to the subsets individually, however, that effect is no longer visible.

An interesting question is whether the percentage of filtered particles per subset is related to the particle quality in these subsets. To this end, the FSC resolutions from refinements of the individual subsets were compared to their expected resolution from a B-factor plot. The results are summarized in table 5.11 and visualized in figure 5.18. Additionally, the deviations from the expected FSC resolutions are listed in table 5.17 and the subsets with the three largest positive and negative deviations are highlighted. In the appendix, correlations between the filtering rates and the FSC deviations are plotted (figure A.19). The per-subset kept ratios for the directional filtering with the orientation consistency parameter showed too little variation to deduce meaningful trends.

The largest negative deviations, i.e. FSC performances worse than expected, are exhibited by subsets 8, 13 and 15. In subset 15, the number of kept particles after 2D and 3D classification is generally low (see figure 5.17). Consequently, the amount of particles from this subset is also low in the CryoSieve-filtered datasets. The relative percentage of kept particles from the final dataset is, however, not particularly low. At a global kept ratio of 0.7, the relative amount of kept particles from subset 15 is 73 % and at 0.4, it is even 52 %. Also for subsets 8 and 13, the kept ratio for these subsets is not significantly below the overall kept ratio. In the per-subset CryoSieve runs, there is no specific behavior towards the low-performing subsets. The subsets from which CryoSieve filtered relatively the most particles in the global run were 7, 10, and 12. All of these subsets showed FSC resolutions close to the expected values. The largest positive FSC deviations, i.e. performances better than expected, are shown by subsets 4, 5, and 6, but no particular behavior is exhibited by CryoSieve for these subsets either. In the correlation plots (figure A.19), no CryoSieve filtering version

shows a meaningful correlation to the FSC deviations. Consequently, no relationship between the kept ratios for the subsets and the underlying data quality could be demonstrated in this experiment.

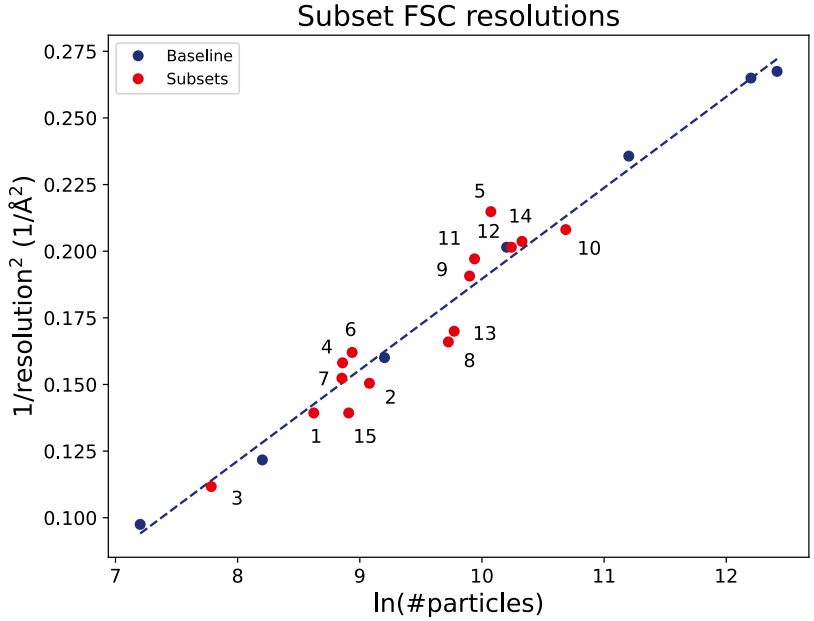


Figure 5.18.: FSC resolutions from the 15 subsets of the large FAS dataset based on different recording days (red) plotted next to a B-factor plot indicating the expected resolutions for randomly sampled particle sets (blue). Each subset point is labeled with the corresponding subset number.

The combined ranking approach also shows stronger variations in the per-subset kept ratios, which range from 53 % to 85 % of the final baseline set. The five parameters used for ranking were calculated within the subset-based processing workflow part, so the values would not differ if the subsets were filtered individually. However, the ranking was performed on the whole large final baseline dataset. The strong variations in the filtering rates could therefore potentially indicate quality differences among the subsets. However, the comparison of kept ratios and per-subset FSC resolutions does not give a conclusive picture here. For subset 4 for example, the percentage of remaining particles was rather low while the per-subset FSC was better than expected. Even though the strongest positive correlation between the FSC deviations and kept ratios was reached with the combined rank approach in comparison to the other methods (see figure A.19), the overall correlation with an R^2 of 0.07 is too low to be considered meaningful. At the same time, subset 5 with the best FSC difference and subset 15 with the worst FSC difference showed the highest and lowest percentage

Table 5.11.: Predicted and experimentally determined FSC resolutions for the 15 subsets of the large FAS dataset corresponding to different recording days. The resolution difference is calculated by subtracting the experimental from the predicted values. Negative deviations are highlighted in red.

Subset	**#Particles**	**FSC_{pred}** (Å)	**FSC_{exp}** (Å)	**ΔFSC** (Å)
1	5,558	2.65	2.68	-0.03
2	8,768	2.51	2.58	-0.07
3	2,401	2.96	2.99	-0.03
4	7,036	2.58	2.51	0.07
5	23,691	2.28	2.16	0.12
6	7,609	2.55	2.48	0.07
7	6,997	2.58	2.56	0.02
8	16,732	2.36	2.45	-0.09
9	19,907	2.32	2.29	0.03
10	43,754	2.17	2.19	-0.02
11	20,730	2.31	2.25	0.06
12	28,029	2.25	2.23	0.02
13	17,550	2.34	2.43	-0.09
14	30,567	2.23	2.22	0.01
15	7,397	2.56	2.68	-0.12

of remaining particles respectively, indicating some connection to real experimental variations in particle quality.

In conclusion, the analysis in this section showed that the final large FAS dataset comprises data from all 15 recording day subsets, but the per-subset kept ratios can be variable dependent on the filtering method used. In contrast to the other methods, the orientation consistency approach with directional filtering showed little variation among the subsets, so that this behavior could not be related to subset quality. For CryoSieve, the varying kept ratios could not be related to better or worse FSC resolutions of the subsets. It hence remains unclear, whether CryoSieve filtering is based on data quality or removes particles for other reasons like structural heterogeneity. Dependent on the desired downstream processing of the data, CryoSieve should therefore be used with caution. The combined ranking approach showed the smallest kept ratio for the subset with the least favorable FSC deviation from the expected value and the largest kept ratio for the subset with the most favorable FSC deviation. Still, no meaningful correlation was visible when plotting all kept ratios against the FSC deviations.

6. Conclusion

6.1. Summary

In this thesis, different approaches for particle image sorting were examined. The goal of all methods was to eliminate low-quality particles from the unfiltered datasets in order to improve the quality of the reconstructed 3D maps. The experiments were carried out on simulated data and on high-resolution fatty acid synthase data. From the experiments, the following conclusions could be drawn:

Meaningfulness of workflow quality parameters

With the subset validation procedure, it could be shown that parameters related to particle motion, CTF consistency and standard deviation of the estimated shift in 3D classification are potential indicators for particle quality. Particle subsets with unfavorable values for these parameters yield FSC resolutions worse than expected from a Rosenthal-Henderson plot. Per-particle corrections improve the absolute resolution values of the parameter subsets, but relative relationships are maintained.

Limited effect of particle sorting by quality parameters

When the particle images with unfavorable values for the most promising quality parameters were eliminated from the small FAS dataset, the FSC resolution only improved by a small margin within the resolution range of multiple refinement runs on the whole small dataset. The quality parameter filtering could therefore not show a practical benefit in this application scenario.

Consistent decline of FSC resolution over combined ranking subsets

A correlation analysis showed that quality parameters from different workflow steps were uncorrelated. On the small FAS dataset, the combined ranking approach with the five most promising quality parameters could substantially reduce the particle number at a relatively small FSC loss. A combined rank-based split of the small dataset into equal-sized subsets revealed a trend of consistently worse FSC results going from the high-rank to the low-rank subsets, suggesting a correlation between combined rank and particle quality.

No FSC resolution improvement on the large FAS dataset by combined rank-based filtering

On the large dataset, differences between random exclusion and combined rank-based exclusion became visible when excluding larger amounts of data. The use of directional filtering instead of standard filtering based on the combined rank lead to almost identical results. After all, no resolution improvement over the whole baseline dataset could be achieved and the potential positive effect of excluding lower ranking particles could not compensate for the associated signal loss. A potential future use is rigorous filtering for very large amounts of data.

Orientation consistency parameter separates simulated data

On simulated data, the orientation consistency parameter gave different values for positive and negative images. A (partial) separation of FAS images from a simple disk, a ribosome and damaged FAS structures would be possible by setting a threshold. In contrast, simulated images of similar FAS structures differing by a 15° dome rotation could not be separated.

Directional filtering mitigates orientation bias

The orientation consistency parameter showed an orientation bias on the large FAS dataset. The particles with the 1 % best values were only side views and a strap-like region of orientations was missing from the orientation plot of the 10 % best value particles. The use of directional filtering led to more uniform orientation plots.

Limited effect of orientation consistency and directional filtering on real data
When filtering the large FAS dataset based on the orientation consistency parameter using either naive or directional filtering, no improvement with respect to the baseline reconstruction could be achieved. The FSC resolution remained almost unchanged when eliminating up to 20 % of the particles by naive and up to 30 % of the particles by directional filtering. The Q-scores consistently decreased upon particle removal. Overall, the behavior was similar to random exclusion of particle data. Hence, no significant practical benefit of these methods could be proven.

CryoSieve maintains FSC resolution and improves Q-scores
The external particle sorting software CryoSieve [51] could approximately maintain the FSC resolution upon filtering up to 60 % of the particles in the large FAS dataset. Up to this filtering step, the Q-Scores for the protein as well as for the modeled waters improved. An FSC improvement could, however, not be observed.

Applicability of CryoSieve in subset filtering scenarios
When applied separately on 15 subsets of the large FAS dataset corresponding to different recording days, the CryoSieve approach could reduce the particle number in the combined dataset by 35 % while maintaining the FSC resolution and increasing the number of waters with a Q-Score of at least 0.8. CryoSieve might therefore be beneficial for early filtering and refinement speed-up when large amounts of data are collected. However, no direct relationship between the per-subset filtering rate of CryoSieve and subset quality measured by the corresponding FSC resolution could be established.

6.2. Limitations & Potential improvements

There are a number of potential improvements for the methods investigated and developed in this thesis. Regarding the quality parameters collected in the workflow, the parameters presented in this thesis are an exemplary selection of many possible options. Other such parameters could be evaluated with the standardized procedure presented here. The combined ranking method is readily available in COW and can

easily be applied to a variable number of other numerical parameters. The presented parameters from this thesis might also be improvable. The pickingCounter parameter for instance might yield a better filtering performance if more selective and sophisticated picking methods are employed. The parameters describing the consistency of shift and 3D orientation in the course of the 3D classification could also be evaluated during 3D refinement, which may unveil more detailed differences due to the finer sampling. In this case, it might be beneficial to restrict the analysis to later refinement iterations, which is already implemented as an option in the respective COW logic.

An additional possibility to boost the filtering performance could be the development of machine learning-based filtering methods that consider multiple parameters simultaneously. There are, however, obstacles for machine learning on cryo-EM data. First, the noise in the data is much stronger than the signal. This makes a manual annotation of positive and negative images impossible. Second, alternative methods for annotating particles images e.g. based on the resolution of refined maps from random subsets of the data would require substantial computational resources since 3D refinements are computationally costly.

In case of the new orientation consistency score, there are several parameters in the calculation workflow that could be optimized. These include for example the amount of coarsening applied to the images, the HEALPix level of the projection image sampling and several fine-tuning parameters of the alignment logic and the classification logic. Currently, a bottleneck of the workflow is the speed of the alignment logic, which could be improved by implementing a parallel computation routine over multiple GPU nodes. A speed-up of the alignment would facilitate parameter optimization and make a finer projection image sampling and the resulting increase in reference projection images practically feasible. A shorter calculation time would also be beneficial for using the orientation consistency filtering as an early filtering step in the processing pipeline.

The directional filtering approach could be implemented in a more efficient way. For binning, the images are currently compared to all orientations of the user-defined HEALPix level. This is currently not a practical bottleneck: for example, the binning of the large FAS dataset is feasible within minutes. However, an improved solution would be to make use of the hierarchical structure of HEALPix. This way, the binning

algorithm would start by comparing the current image orientation to all base level HEALPix orientations and then repeatedly compare the image orientation to the 4 sub-pixels of the closest HEALPix orientation until the user-defined HEALPix level is reached. Another potential improvement is the development of further filtering strategies for the directional filtering. The current implementation with fractions per bin maintains the relative amounts of images over the bins. Another strategy could for example aim at filtering the less occupied bins less strongly in order to achieve a more uniform orientation distribution. The concept of orientation-aware filtering is transferable to other methods and it would be interesting to see if CryoSieve [51] could benefit from this idea. The implementation of directional filtering in COW already allows for using it on any numerical filtering parameter.

Finally, the results in this thesis were – with the exception of the simulated data – only generated on data from a high-resolution FAS dataset. Since the computational expenses related to cryo-EM made further evaluation on other datasets infeasible within the time frame of this thesis, it remains to be seen whether the presented approaches could prove beneficial in other application scenarios. The directional filtering approach might show a stronger effect on datasets where preferred orientation is a stronger bottleneck than for the FAS data. The orientation consistency or combined rank procedures could be evaluated on more challenging datasets, where the current state-of-the-art processing scheme fails to produce any meaningful reconstructions, to see whether an improvement can be achieved by these complimentary techniques.

6.3. Outlook

The ability to filter particle images and create datasets with a low B-factor is gaining importance as the acquisition speed, and therefore the amount of acquired data, of single-particle cryo-EM increases. With ever larger datasets, computationally intensive procedures such as 3D refinements can become a bottleneck in the way of scientific discovery. In this sense, it is an important question how much the individual images of a dataset actually contribute to the 3D reconstruction. The CryoSieve [51] software was able to reduce the number of particles in the large FAS dataset by 60 % without a significant loss of resolution and Q-scores. A reason for this might be the maximum-likelihood down-weighting of particles with low similarity to the current

reconstruction estimate, which is employed in state-of-the-art 3D refinement algorithms. In the future, particle sorting strategies like CryoSieve should be considered to reduce the number of particles for downstream tasks such as refinement repetitions after CTF refinement and Bayesian Polishing. Furthermore, it should be questioned whether carrying along all particle images up to the very costly last refinement iterations is worthwhile or whether weakly contributing particles could be identified and filtered early on within the refinement algorithm.

An important caveat of particle filtering is the potential to remove structural heterogeneity from the dataset. While structural heterogeneity related to damaged particles or contamination should be removed, there is also a desired type that encompasses, for example, different functional states of the protein. Methods based on a single reference map, like the orientation consistency approach developed in this thesis and CryoSieve, are prone to filter out particles corresponding to other functional states than the reference. The key problem of the CryoSieve approach regarding this issue is that the filtering is based on a simple consistency score between reference map and particle images. Consistency of map and images is rewarded regardless if it arises from image quality, structural homogeneity or even consistent noise in the images. Features already present in the reference map are reinforced even if they arose from systematic errors in the previous data processing. Approaches like CryoSieve are not suited to remove particle images for specific biochemical or physical reasons. In order to precisely filter bad particles without compromising heterogeneity, a better understanding of the actual reasons of limited image quality is needed. Filtering approaches tailored exactly to these limitations could be an impactful addition to the state-of-the-art workflows in high-resolution single-particle cryo-EM.

Finally, the development and analysis of new filtering approaches is impeded by the limited availability of suitable evaluation criteria. A number of evaluation metrics exist as summarizes by Lawson *et al.*][97], but many of them rely on the availability of an atomic model fitted to the map. Model building is a labor-intensive process that requires manual human intervention and is therefore highly impractical for large-scale evaluations. Certain criteria like counting of modeled water molecules have the additional limitation that they can only be performed at high resolutions of around 2 Å or below. Consequently, the choice of the evaluation method – as in this thesis – often comes down to the FSC resolution, even though it is known that there are considerable issues [2, 7]. First, the FSC resolution is only a measure of consistency

between the two half-maps reconstructed during a refinement and can therefore not detect systematic errors that are present in both. This can lead to an overestimation of the resolution e.g. in the presence of optical aberrations [125]. Second, the meaning of FSC resolution differences is resolution-dependent. Due to the exponential relationship between particle number and resolution [55], further improvements at already high resolution are more difficult to achieve and require much more additional particles than at low resolution. At the same time, small improvements of a fraction of an angstrom at high resolution can make the difference of achieving true atomic resolution or not. This makes it hard to define what meaningful FSC resolution differences are and to compare filtering effects at different resolution levels. For these reasons, there is a need for better quality criteria to quantify map quality and interpretability for biochemical and drug design-related questions. New advances in the automation of atomic model building [126] and automated water placement methods might help towards the development of practically applicable evaluation criteria for particle filtering studies and their widespread use in the cryo-EM community.

Bibliography

[1] Frank, J. Advances in the field of single-particle cryo-electron microscopy over the last decade. *Nature Protocols* **12**, 209–212 (2017).

[2] Chari, A. & Stark, H. Prospects and limitations of high-resolution single-particle cryo-electron microscopy. *Annual Review of Biophysics* **52**, 391–411 (2023).

[3] The wwPDB Consortium. EMDB—the Electron Microscopy Data Bank. *Nucleic Acids Research* **52**, D456–D465 (2023).

[4] Electron Microscopy Data Bank (EMDB). EMDB entry resolution in shells distribution. `https://www.ebi.ac.uk/emdb/statistics/emdb_resolution_distribution` (Accessed: 20.11.2024).

[5] Casañal, A., Shakeel, S. & Passmore, L. A. Interpretation of medium resolution cryoEM maps of multi-protein complexes. *Current Opinion in Structural Biology* **58**, 166–174 (2019).

[6] Electron Microscopy Data Bank (EMDB). Chart Builder documentation. `https://www.ebi.ac.uk/emdb/documentation/builder` (Accessed: 20.11.2024).

[7] Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020).

[8] Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020).

[9] Robertson, M. J., Meyerowitz, J. G. & Skiniotis, G. Drug discovery in the era of cryo-electron microscopy. *Trends in Biochemical Sciences* **47**, 124–135 (2022).

[10] Van Drie, J. H. & Tong, L. Cryo-EM as a powerful tool for drug discovery. *Bioorganic & Medicinal Chemistry Letters* **30**, 127524 (2020).

[11] Wigge, C., Stefanovic, A. & Radjainia, M. The rapidly evolving role of cryo-EM in drug design. *Drug Discovery Today: Technologies* **38**, 91–102 (2020).

[12] Cushing, V. I. *et al.* High-resolution cryo-EM of the human CDK-activating kinase for structure-based drug design. *Nature Communications* **15**, 2265 (2024).

[13] Saur, M. *et al.* Fragment-based drug discovery using cryo-EM. *Drug Discovery Today* **25**, 485–490 (2020).

[14] Cavasin, T. Created in BioRender. `https://BioRender.com/l801467` (2025).

[15] Scheres, S. H. W. Chapter six - Processing of structurally heterogeneous cryo-EM data in RELION. In Crowther, R. (ed.) *The Resolution Revolution: Recent Advances In cryoEM*, vol. 579 of *Methods in Enzymology*, 125–157 (Academic Press, 2016).

[16] Sorzano, C. O. *et al.* Survey of the analysis of continuous conformational variability of biological macromolecules by electron microscopy. *Acta Crystallographica Section F: Structural Biology Communications* **75**, 19–32 (2019).

[17] Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods* **18**, 176–185 (2021).

[18] Punjani, A. & Fleet, D. J. 3DFlex: determining structure and motion of flexible proteins from cryo-EM. *Nature Methods* **20**, 860–870 (2023).

[19] Passmore, L. A. & Russo, C. J. Chapter three - Specimen preparation for high-resolution cryo-EM. In Crowther, R. (ed.) *The Resolution Revolution: Recent Advances In cryoEM*, vol. 579 of *Methods in Enzymology*, 51–86 (Academic Press, 2016).

[20] Seigel, R. R. *et al.* On-line detection of nonspecific protein adsorption at artificial surfaces. *Analytical Chemistry* **69**, 3321–3328 (1997).

[21] Boisset, N. *et al.* Overabundant single-particle electron microscope views induce a three-dimensional reconstruction artifact. *Ultramicroscopy* **74**, 201–207 (1998).

[22] Sorzano, C. *et al.* Algorithmic robustness to preferred orientations in single particle analysis by cryoEM. *Journal of Structural Biology* **213**, 107695 (2021).

[23] Neselu, K. *et al.* Measuring the effects of ice thickness on resolution in single particle cryo-EM. *Journal of Structural Biology: X* **7**, 100085 (2023).

[24] Bhella, D. Cryo-electron microscopy: an introduction to the technique, and considerations when working to establish a national facility. *Biophysical Reviews* **11**, 515–519 (2019).

[25] Cavasin, T. Created in BioRender. `https://BioRender.com/c90u133` (2025).

[26] Zivanov, J., Nakane, T. & Scheres, S. H. W. Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in RELION-3.1. *IUCrJ* **7**, 253–267 (2020).

[27] Russo, C. J. & Henderson, R. Ewald sphere correction using a single side-band image processing algorithm. *Ultramicroscopy* **187**, 26–33 (2018).

[28] Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology* **192**, 216–221 (2015).

[29] Zhang, K. Gctf: Real-time CTF determination and correction. *Journal of Structural Biology* **193**, 1–12 (2016).

[30] Glaeser, R. M. Chapter two - Specimen behavior in the electron beam. In Crowther, R. (ed.) *The Resolution Revolution: Recent Advances In cryoEM*, vol. 579 of *Methods in Enzymology*, 19–50 (Academic Press, 2016).

[31] Zheng, S. Q. *et al.* MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature Methods* **14**, 331–332 (2017).

[32] Naydenova, K., Jia, P. & Russo, C. J. Cryo-EM with sub–1 Å specimen movement. *Science* **370**, 223–226 (2020).

[33] Zhang, K. Gautomatch. `http://www.mrc-lmb.cam.ac.uk/kzhang/` (Accessed: 10.05.2021).

[34] Scheres, S. H. Semi-automated selection of cryo-EM particles in RELION-1.3. *Journal of Structural Biology* **189**, 114–122 (2015).

[35] Vargas, J. *et al.* Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. *Journal of Structural Biology* **183**, 342–353 (2013).

[36] Norousi, R. *et al.* Automatic post-picking using MAPPOS improves particle image detection from cryo-EM micrographs. *Journal of Structural Biology* **182**, 59–66 (2013).

[37] Cavasin, T. Created in BioRender. `https://BioRender.com/j04b673` (2025).

[38] Stabrin, M. *et al.* TranSPHIRE: automated and feedback-optimized on-the-fly processing for cryo-EM. *Nature Communications* **11**, 5716 (2020).

[39] Wagner, T. *et al.* SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Communications Biology* **2**, 218 (2019).

[40] Bepler, T. *et al.* TOPAZ: A positive-unlabeled convolutional neural network cryoEM particle picker that can pick any size and shape particle. *Microscopy and Microanalysis* **25**, 986–987 (2019).

[41] Li, Y., Cash, J. N., Tesmer, J. J. G. & Cianfrocco, M. A. High-throughput cryo-EM enabled by user-free preprocessing routines. *Structure* **28**, 858–869.e3 (2020).

[42] SPHIRE team. Cinderella: Deep learning based binary classification tool. `http://sphire.mpg.de/wiki/doku.php?id=auto_2d_class_selection` (Accessed: 07.05.2021).

[43] Kimanius, D., Dong, L., Sharov, G., Nakane, T. & Scheres, S. H. W. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochemical Journal* **478**, 4169–4185 (2021).

[44] Sanchez-Garcia, R., Segura, J., Maluenda, D., Sorzano, C. O. S. & Carazo, J. M. MicrographCleaner: A python package for cryo-EM micrograph cleaning using deep learning. *Journal of Structural Biology* **210**, 107498 (2020).

[45] Sanchez-Garcia, R., Segura, J., Maluenda, D., Carazo, J. M. & Sorzano, C. O. S. Deep Consensus, a deep learning-based approach for particle pruning in cryo-electron microscopy. *IUCrJ* **5**, 854–865 (2018).

[46] Gong, X. *et al.* Structural insights into the Niemann-Pick C1 (NPC1)-mediated cholesterol transfer and Ebola infection. *Cell* **165**, 1467–1478 (2016).

[47] Zhou, Y., Moscovich, A., Bendory, T. & Bartesaghi, A. Unsupervised particle sorting for high-resolution single-particle cryo-EM. *Inverse Problems* **36**, 044002 (2020).

[48] Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **14**, 290–296 (2017).

[49] Zivanov, J. *et al.* New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).

[50] Méndez, J., Garduño, E., Carazo, J. M. & Sorzano, C. O. S. Identification of incorrectly oriented particles in cryo-EM single particle analysis. *Journal of Structural Biology* **213**, 107771 (2021).

[51] Zhu, J. *et al.* A minority of final stacks yields superior amplitude in single-particle cryo-EM. *Nature Communications* **14**, 7822 (2023).

[52] Iudin, A. *et al.* EMPIAR: the Electron Microscopy Public Image Archive. *Nucleic Acids Research* **51**, D1503–D1511 (2023).

[53] Grant, T., Rohou, A. & Grigorieff, N. cisTEM, user-friendly software for single-particle image processing. *eLife* **7**, e35383 (2018).

[54] Pintilie, G. *et al.* Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nature Methods* **17**, 328–334 (2020).

[55] Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology* **333**, 721–745 (2003).

[56] Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology* **180**, 519–530 (2012).

[57] Schwab, J., Kimanius, D., Burt, A., Dendooven, T. & Scheres, S. H. W. DynaMight: estimating molecular motions with improved reconstruction from cryo-EM images. *Nature Methods* **21**, 1855–1862 (2024).

[58] Stagg, S. M., Noble, A. J., Spilman, M. & Chapman, M. S. ResLog plots as an empirical metric of the quality of cryo-EM reconstructions. *Journal of Structural Biology* **185**, 418–426 (2014).

[59] van Heel, M. Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **21**, 111–123 (1987).

[60] Gorski, K. M. *et al.* HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal* **622**, 759–771 (2005).

[61] COW development team. COW - the official website. `https://www.cow-em.de/` (Accessed: 10.05.2021).

[62] Weissenberger, G., Henderikx, R. J. M. & Peters, P. J. Understanding the invisible hands of sample preparation for cryo-EM. *Nature Methods* **18**, 463–471 (2021).

[63] Glaeser, R. M. 2.4 Requirement to make very thin specimens for cryo-EM. In Glaeser, R. M., Nogales, E. & Chiu, W. (eds.) *Single-particle Cryo-EM of Biological Macromolecules*, 2.19–2.24 (IOP Publishing, 2021).

[64] Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Quarterly Reviews of Biophysics* **21**, 129–228 (1988).

[65] Cavasin, T. Created in BioRender. `https://BioRender.com/r01n470` (2024).

[66] Orlova, E. V. & Saibil, H. R. Structural analysis of macromolecular assemblies by electron microscopy. *Chemical Reviews* **111**, 7710–7748 (2011).

[67] Cavasin, T. Created in BioRender. `https://BioRender.com/c96b773` (2024).

[68] Kato, T. *et al.* CryoTEM with a cold field emission gun that moves structural biology into a new stage. *Microscopy and Microanalysis* **25**, 998–999 (2019).

[69] Hamaguchi, T. *et al.* A new cryo-EM system for single particle analysis. *Journal of Structural Biology* **207**, 40–48 (2019).

[70] Raimondi, V. & Grinzato, A. A basic introduction to single particles cryo-electron microscopy. *AIMS Biophysics* **9**, 5–20 (2021).

[71] McMullan, G., Faruqi, A. R. & Henderson, R. Chapter one - Direct electron detectors. In Crowther, R. (ed.) *The Resolution Revolution: Recent Advances In cryoEM*, vol. 579 of *Methods in Enzymology*, 1–17 (Academic Press, 2016).

[72] Danev, R. & Baumeister, W. Cryo-EM single particle analysis with the Volta phase plate. *eLife* **5**, e13046 (2016).

[73] Danev, R. *et al.* Routine sub-2.5 Å cryo-EM structure determination of GPCRs. *Nature Communications* **12**, 4333 (2021).

[74] Schwartz, O. *et al.* Laser phase plate for transmission electron microscopy. *Nature Methods* **16**, 1016–1020 (2019).

[75] Marques, M. A., Purdy, M. D. & Yeager, M. CryoEM maps are full of potential. *Current Opinion in Structural Biology* **58**, 214–223 (2019).

[76] Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4**, e06980 (2015).

[77] Sigworth, F. J. Principles of cryo-EM single-particle image processing. *Microscopy* **65**, 57–67 (2016).

[78] Zhu, Y. Automatic particle selection: results of a comparative study. *Journal of Structural Biology* **145**, 3–14 (2004).

[79] Wang, F. *et al.* DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM. *Journal of Structural Biology* **195**, 325–336 (2016).

[80] Tegunov, D. & Cramer, P. Real-time cryo-electron microscopy data preprocessing with Warp. *Nature Methods* **16**, 1146–1152 (2019).

[81] Zhang, X., Zhao, T., Chen, J., Shen, Y. & Li, X. EPicker is an exemplar-based continual learning approach for knowledge accumulation in cryoEM particle picking. *Nature Communications* **13**, 2468 (2022).

[82] Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**, 129–137 (1982).

[83] Sigworth, F. A maximum-likelihood approach to single-particle image refinement. *Journal of Structural Biology* **122**, 328–339 (1998).

[84] Sigworth, F. J., Doerschuk, P. C., Carazo, J. M. & Scheres, S. H. W. Chapter ten - An introduction to maximum-likelihood methods in cryo-EM. In Jensen, G. J. (ed.) *Cryo-EM, Part B: 3-D Reconstruction*, vol. 482 of *Methods in Enzymology*, 263–294 (Academic Press, 2010).

[85] Scheres, S. H. W. *et al.* Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods* **4**, 27–29 (2007).

[86] Scheres, S. H. W. Chapter eleven - Classification of structural heterogeneity by maximum-likelihood methods. In Jensen, G. J. (ed.) *Cryo-EM, Part B: 3-D Reconstruction*, vol. 482 of *Methods in Enzymology*, 295–320 (Academic Press, 2010).

[87] van Heel, M. Multivariate statistical classification of noisy images (randomly oriented biological macromolecules). *Ultramicroscopy* **13**, 165–183 (1984).

[88] Shatsky, M., Hall, R. J., Brenner, S. E. & Glaeser, R. M. A method for the alignment of heterogeneous macromolecules from electron microscopy. *Journal of Structural Biology* **166**, 67–78 (2009).

[89] Scheres, S. H. W. A Bayesian view on cryo-EM structure determination. *Journal of Molecular Biology* **415**, 406–418 (2012).

[90] Hu, M. *et al.* A particle-filter framework for robust cryo-EM 3D reconstruction. *Nature Methods* **15**, 1083–1089 (2018).

[91] Zivanov, J., Nakane, T. & Scheres, S. H. W. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *IUCrJ* **6**, 5–17 (2019).

[92] Electron Microscopy Data Bank (EMDB). EMDB current software usage distribution. `https://www.ebi.ac.uk/emdb/statistics/emdb_software_distribution` (Accessed: 08.12.2024).

[93] Scheres, S. H. W. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9**, 853–854 (2012).

[94] Harauz, G. & van Heel, M. Exact filters for general geometry three dimensional reconstruction. *Optik* **73**, 146–156 (1986).

[95] van Heel, M. & Schatz, M. Fourier shell correlation threshold criteria. *Journal of Structural Biology* **151**, 250–262 (2005).

[96] Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11**, 63–65 (2014).

[97] Lawson, C. L. *et al.* Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. *Nature Methods* **18**, 156–164 (2021).

[98] COW development team. The CowEyes Coordinate System. `https://www.cow-em.de/guide/doku.php?id=eyes:coordinatesandconventions` (Accessed: 18.01.2024).

[99] Sorzano, C. O. S. *et al.* Chapter 2 - Interchanging geometry conventions in 3DEM: Mathematical context for the development of standards. In Herman, G. T. & Frank, J. (eds.) *Computational Methods for Three-Dimensional Microscopy Reconstruction*, 7–42 (Birkhäuser New York, 2014).

[100] Heymann, J. B., Chagoyen, M. & Belnap, D. M. Common conventions for interchange and archiving of three-dimensional electron microscopy information in structural biology. *Journal of Structural Biology* **151**, 196–207 (2005).

[101] Hu, M., Zhang, Q., Yang, J. & Li, X. Unit quaternion description of spatial rotations in 3D electron cryo-microscopy. *Journal of Structural Biology* **212**, 107601 (2020).

[102] Courtesy NASA/JPL-Caltech. Original Image: `https://healpix.jpl.nasa.gov/healpixBackgroundPurpose.shtml`. Permission: `https://www.jpl.nasa.gov/jpl-image-use-policy`, Accessed: 12.04.2024.

[103] COW development team. COW Eyes Overview. `https://www.cow-em.de/guide/lib/exe/detail.php?id=gui&media=cow_eyes_overwiev.png` (Accessed: 06.03.2025).

[104] Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584–590 (2013).

[105] Stagg, S. M. *et al.* A test-bed for optimizing high-resolution single particle reconstructions. *Journal of Structural Biology* **163**, 29–39 (2008).

[106] Kvålseth, T. O. Measuring variation for nominal data. *Bulletin of the Psychonomic Society* **26**, 433–436 (1988).

[107] Horn, B. K. P. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A* **4**, 629–642 (1987).

[108] Markley, F. L., Cheng, Y., Crassidis, J. L. & Oshman, Y. Averaging quaternions. *Journal of Guidance, Control, and Dynamics* **30**, 1193–1197 (2007).

[109] Mises, R. V. & Pollaczek-Geiringer, H. Praktische Verfahren der Gleichungsauflösung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* **9**, 152–164 (1929).

[110] Anderson, E. *et al. LAPACK Users' Guide* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999), third edn.

[111] Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure* **29**, 105–153 (2000).

[112] Reboul, C. F., Kiesewetter, S., Elmlund, D. & Elmlund, H. Point-group symmetry detection in three-dimensional charge density of biomolecules. *Bioinformatics* **36**, 2237–2243 (2020).

[113] Dwork, C., Kumar, R., Naor, M. & Sivakumar, D. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, 613–622 (Association for Computing Machinery, New York, NY, USA, 2001).

[114] de Borda, J.-C. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences* (1781).

[115] Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology* **10**, 980 (2003).

[116] Singh, K. *et al.* Discovery of a regulatory subunit of the yeast fatty acid synthase. *Cell* **180**, 1130–1143.e20 (2020).

[117] Fischer, N. *et al.* Structure of the E. coli ribosome–EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. *Nature* **520**, 567–570 (2015).

[118] Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).

[119] Singh, K. *et al.* Reconstruction of a fatty acid synthesis cycle from acyl carrier protein and cofactor structural snapshots. *Cell* **186**, 5054–5067.e16 (2023).

[120] Günenc, A. N., Graf, B., Stark, H. & Chari, A. Fatty acid synthase: Structure, function, and regulation. In Harris, J. R. & Marles-Wright, J. (eds.) *Macromolecular Protein Complexes IV: Structure and Function*, 1–33 (Springer International Publishing, Cham, 2022).

[121] Schweizer, E. & Hofmann, J. Microbial type I fatty acid synthases (FAS): Major players in a network of cellular FAS systems. *Microbiology and Molecular Biology Reviews* **68**, 501–517 (2004).

[122] Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science* **27**, 14–25 (2018).

[123] Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021).

[124] Asarnow, D., Palovcak, E., Cheng, Y. UCSF pyem v0.5. Zenodo `https://doi.org/10.5281/zenodo.3576630` (2019).

[125] Bromberg, R., Guo, Y., Borek, D. & Otwinowski, Z. High-resolution cryo-EM reconstructions in the presence of substantial aberrations. *IUCrJ* **7**, 445–452 (2020).

[126] Jamali, K. *et al.* Automated model building and protein identification in cryo-EM maps. *Nature* **628**, 450–457 (2024).

[127] Pintilie, G. mapq. `https://cryoem.slac.stanford.edu/ncmi/resources/software/mapq` (Accessed: 25.11.2024).

[128] Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**, 90–95 (2007).

Acronyms

AUC Area under the curve.

CCD Charge-coupled device.

CNN Convolutional neural network.

Cryo-EM Cryo-electron microscopy.

CSV Comma-separated values.

CTF Contrast transfer function.

DF Directional filtering.

DNN Deep neural network.

EM Electron microscopy.

EMDB Electron Microscopy Data Bank.

EMPIAR Electron Microscopy Public Image Archive.

FAS Fatty acid synthase.

FEG Field emission gun.

FSC Fourier shell correlation.

GPCR G protein-coupled receptor.

GPU Graphics processing unit.

HEALPix Hierachical Equal Area isoLatitude Pixelization.

KNN k-nearest neighbors.

LoG Laplacian-of-Gaussian.

MSA Multivariate statistical analysis.

PDB Protein Data Bank.

PSF Point spread function.

REI Relative entropy index.

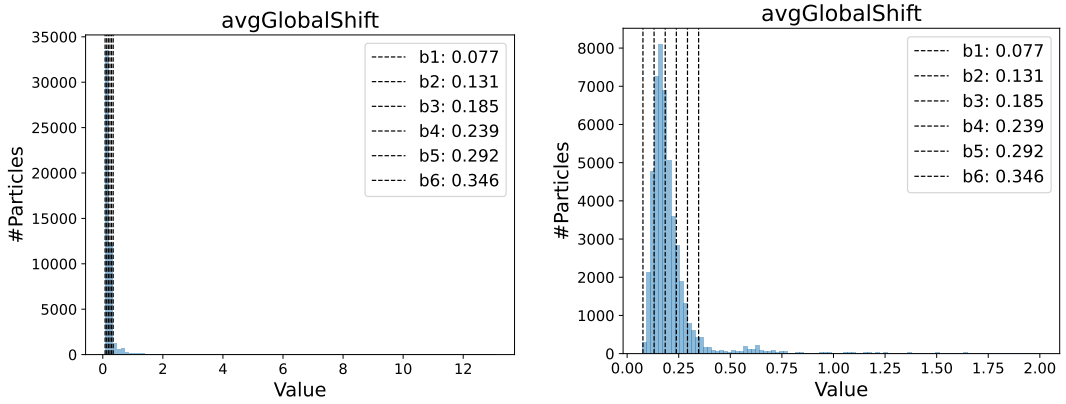
ROC Receiver operating characteristic.

SCF Sinogram correlation function.

TEM Transmission electron microscope.

A. Supplementary data

A.1. Validation plot histograms



(a) (b)

Figure A.1.: Distribution of the global motion quality parameter over the images in the small FAS dataset. The vertical lines indicate the subset bins used for validation. The right histogram zooms in on the area with the most values for better visualization.

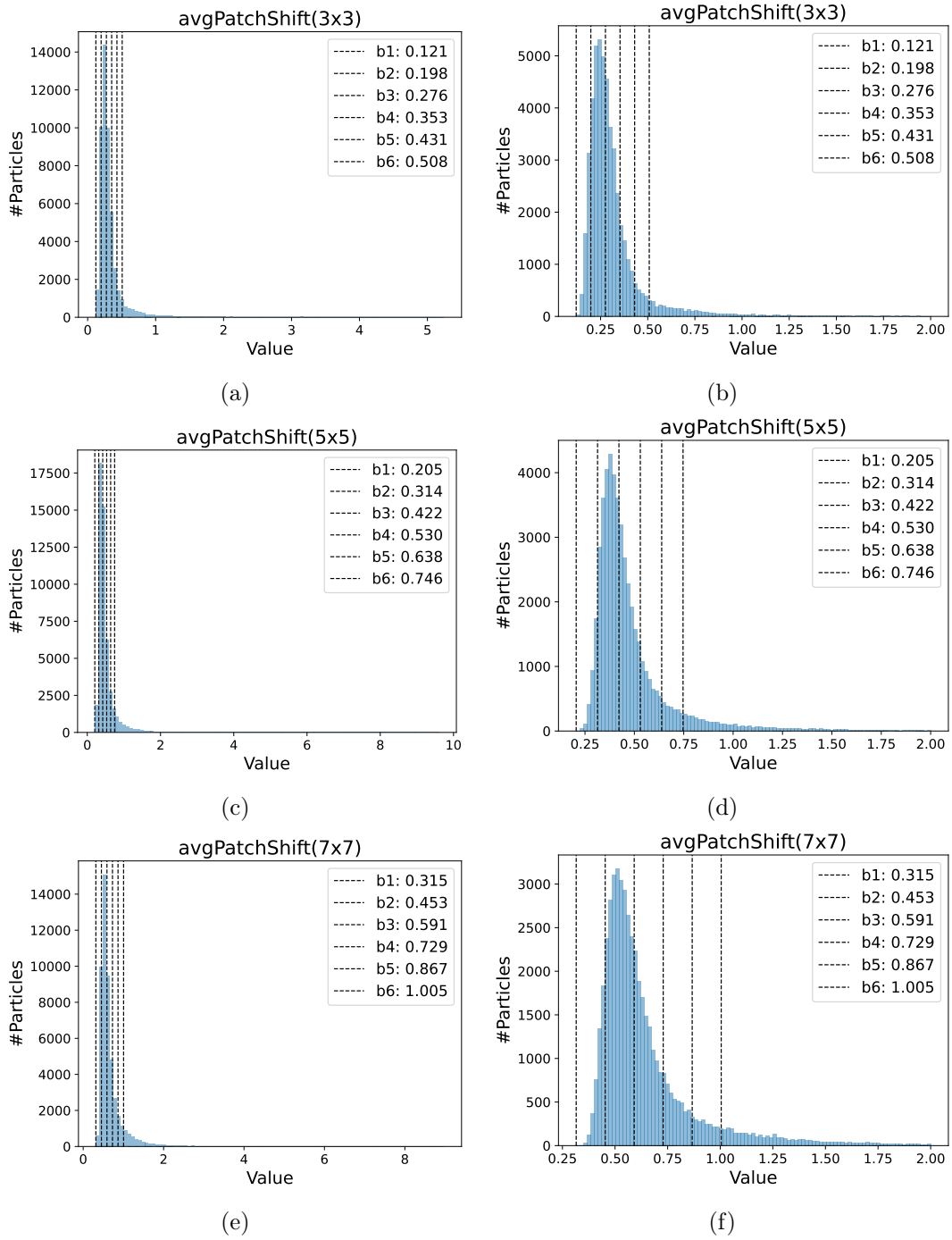


Figure A.2.: Distributions of the local motion quality parameter for different patch sizes over the images in the small FAS dataset. The vertical lines indicate the subset bins used for validation. The histograms in the right column zoom in on the area with the most values for better visualization.

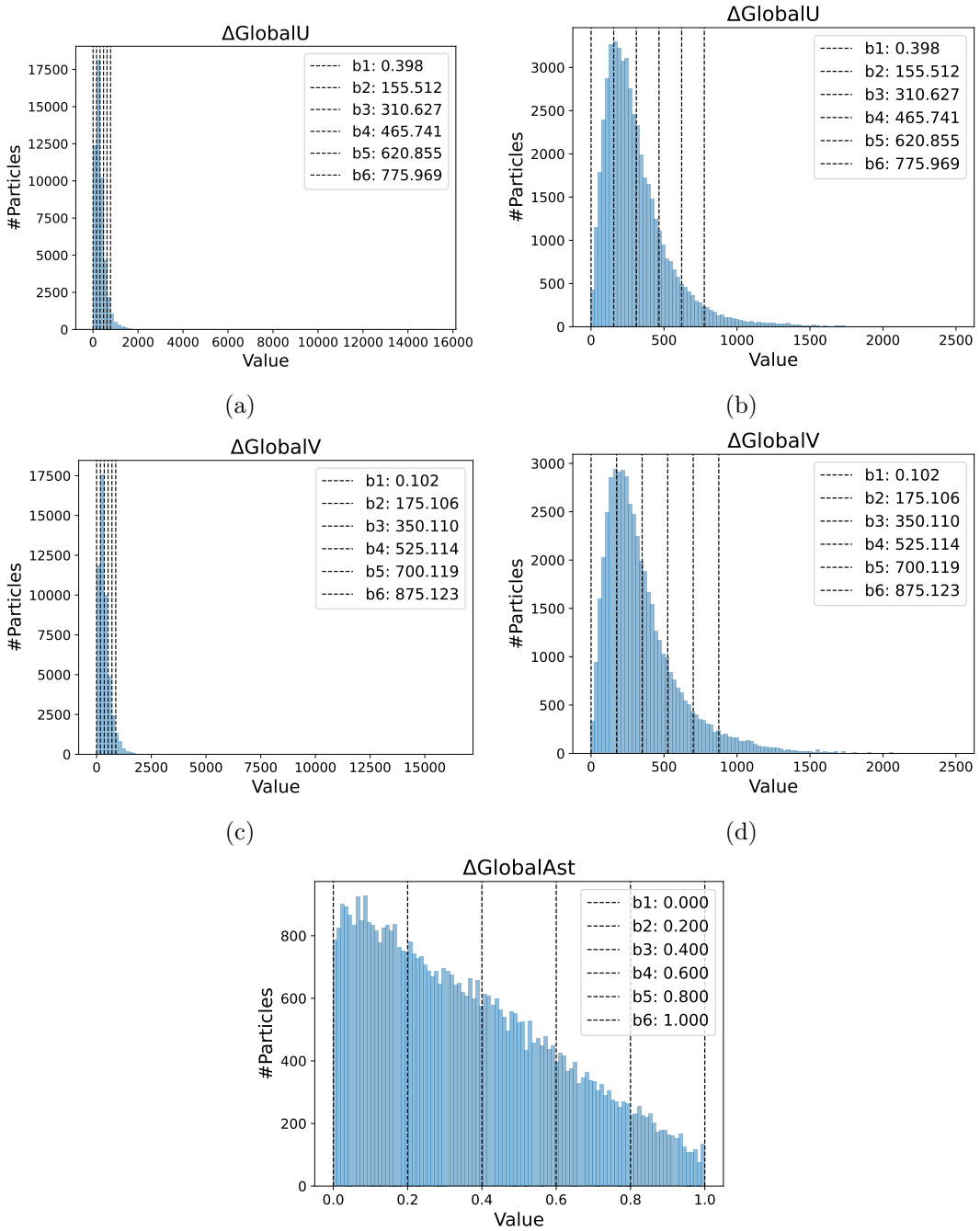


Figure A.3.: Distributions of the difference-to-global CTF quality parameters over the images in the small FAS dataset. The vertical lines indicate the subset bins used for validation. The right histograms in the first two rows zoom in on the area with the most values for better visualization.

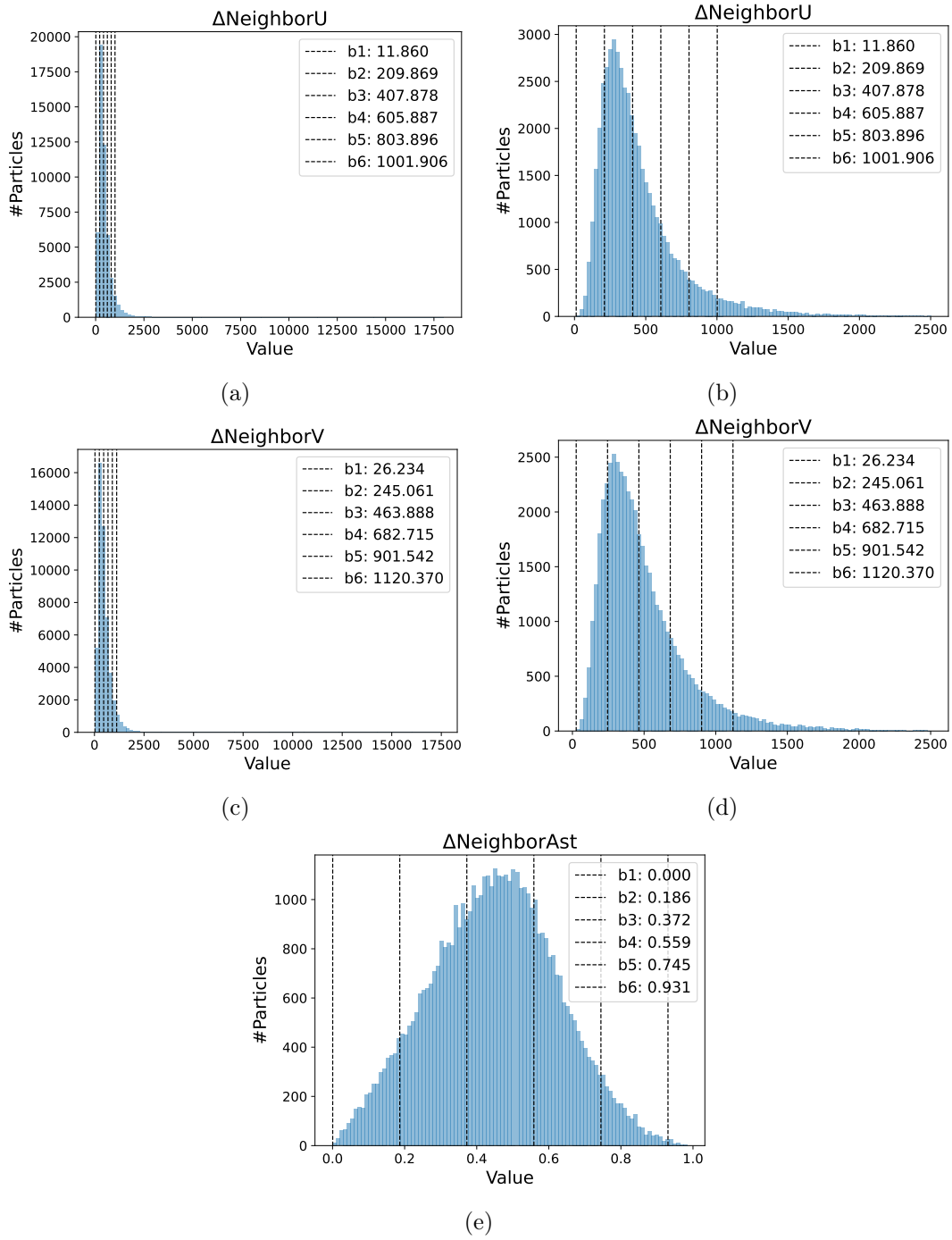


Figure A.4.: Distributions of the difference-to-neighbors CTF quality parameters over the images in the small FAS dataset. The vertical lines indicate the subset bins used for validation. The right histograms in the first two rows zoom in on the area with the most values for better visualization.

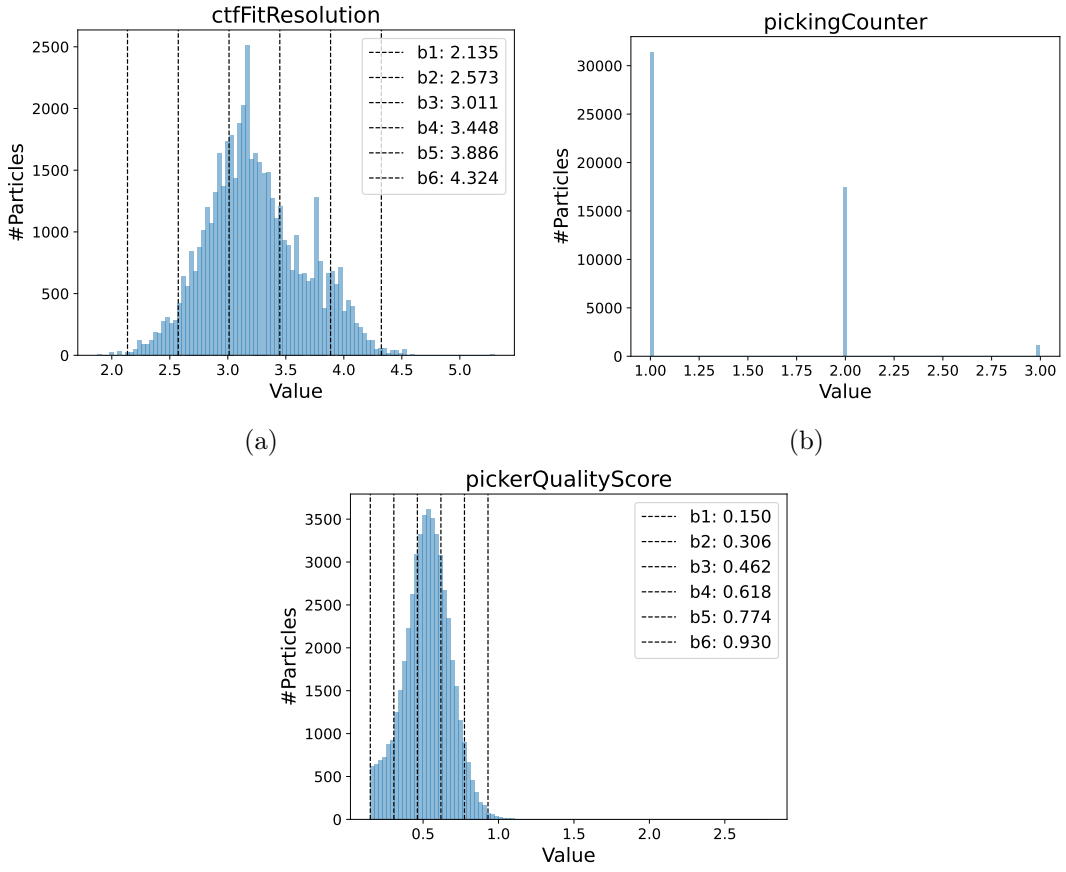


Figure A.5.: Distributions of the particle picking quality parameters and the CTF fit resolution over the images in the small FAS dataset. The vertical lines indicate the subset bins used for validation.

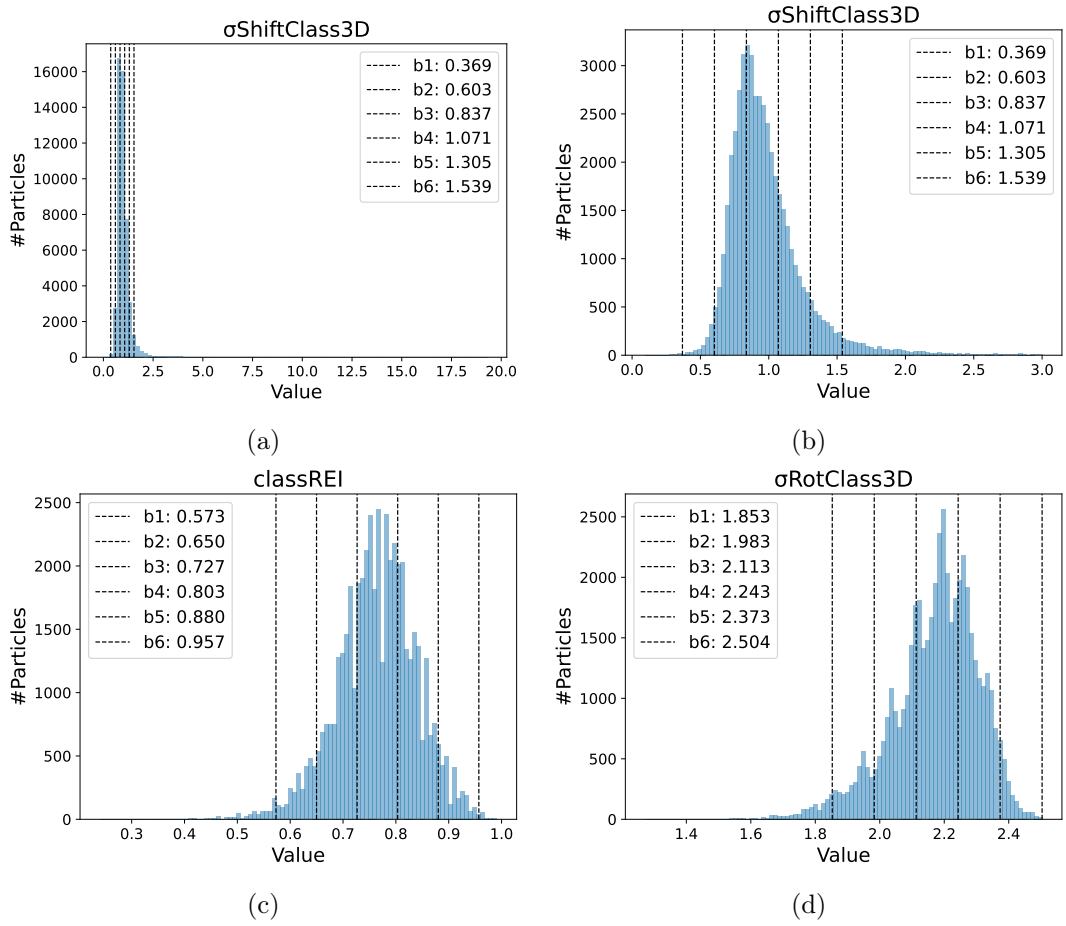
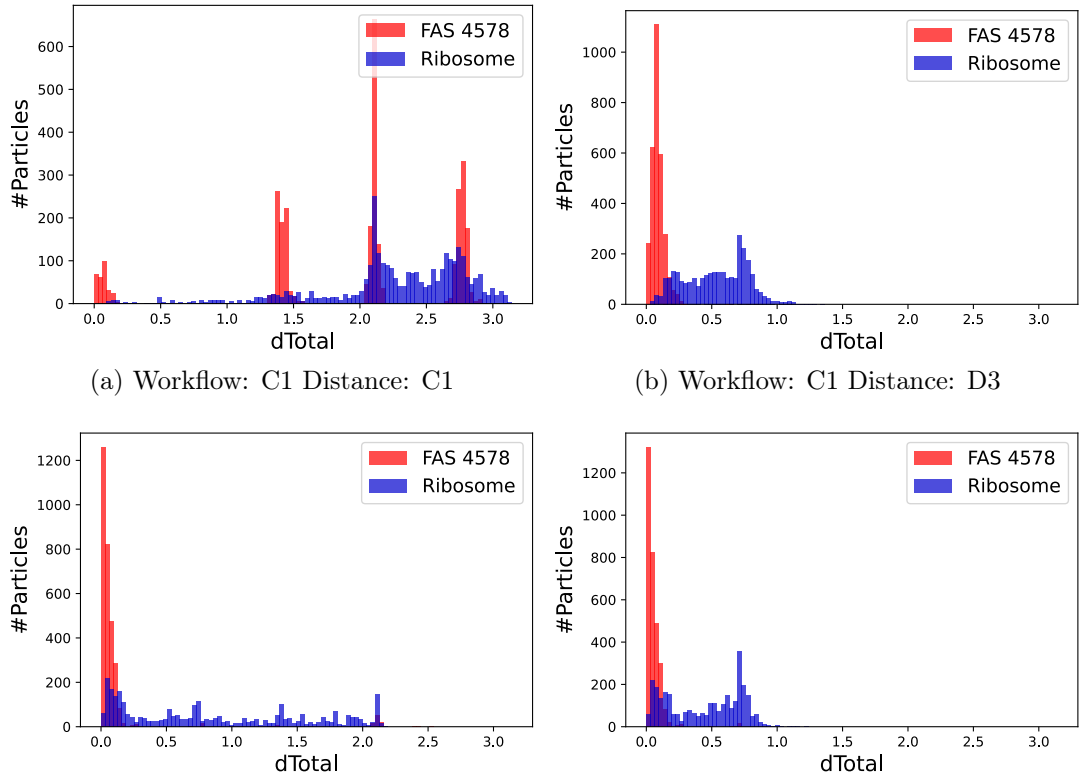


Figure A.6.: Distributions of the quality parameters collected during 3D classification over the images in the small FAS dataset. The vertical lines indicate the subset bins used for validation. The right histogram in the first row zooms in on the area with the most values for better visualization.

A.2. Symmetry plots for orientation consistency

Histograms

Simulated dataset 2



(a) Workflow: C1 Distance: C1

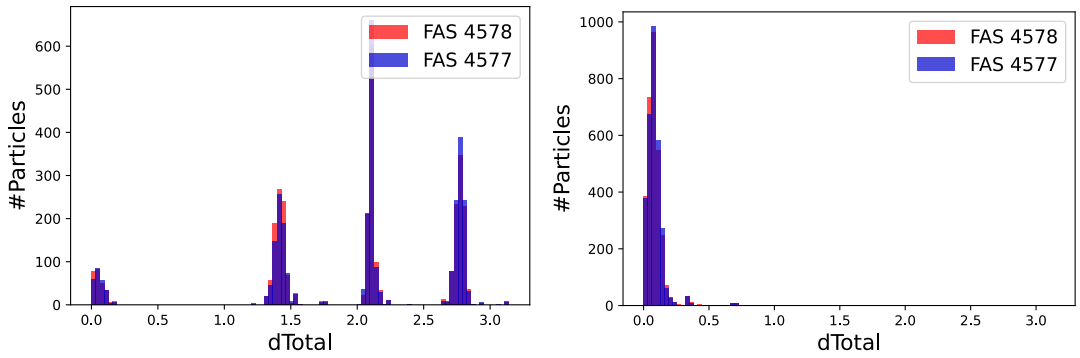
(b) Workflow: C1 Distance: D3

(c) Workflow: D3 Distance: C1

(d) Workflow: D3 Distance: D3

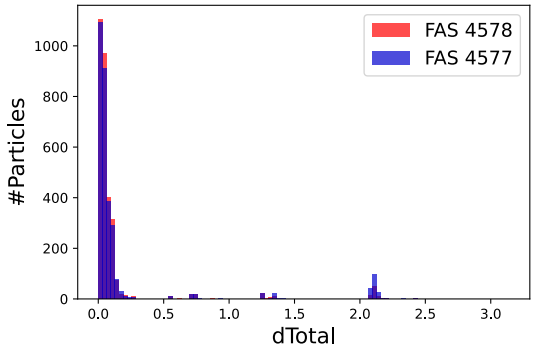
Figure A.7.: Histograms of the orientation consistency d_{total} values for simulated dataset 2 (see table 4.1). The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. The positive images are displayed in red, the negative images in blue.

Simulated dataset 3

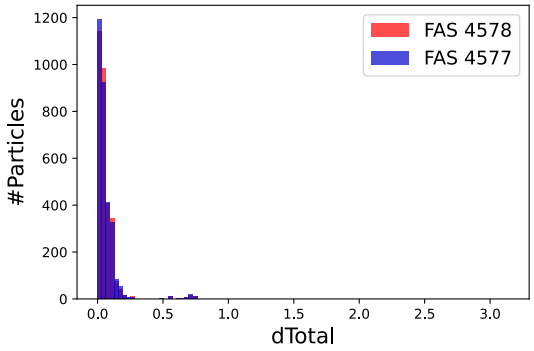


(a) Workflow: C1 Distance: C1

(b) Workflow: C1 Distance: D3



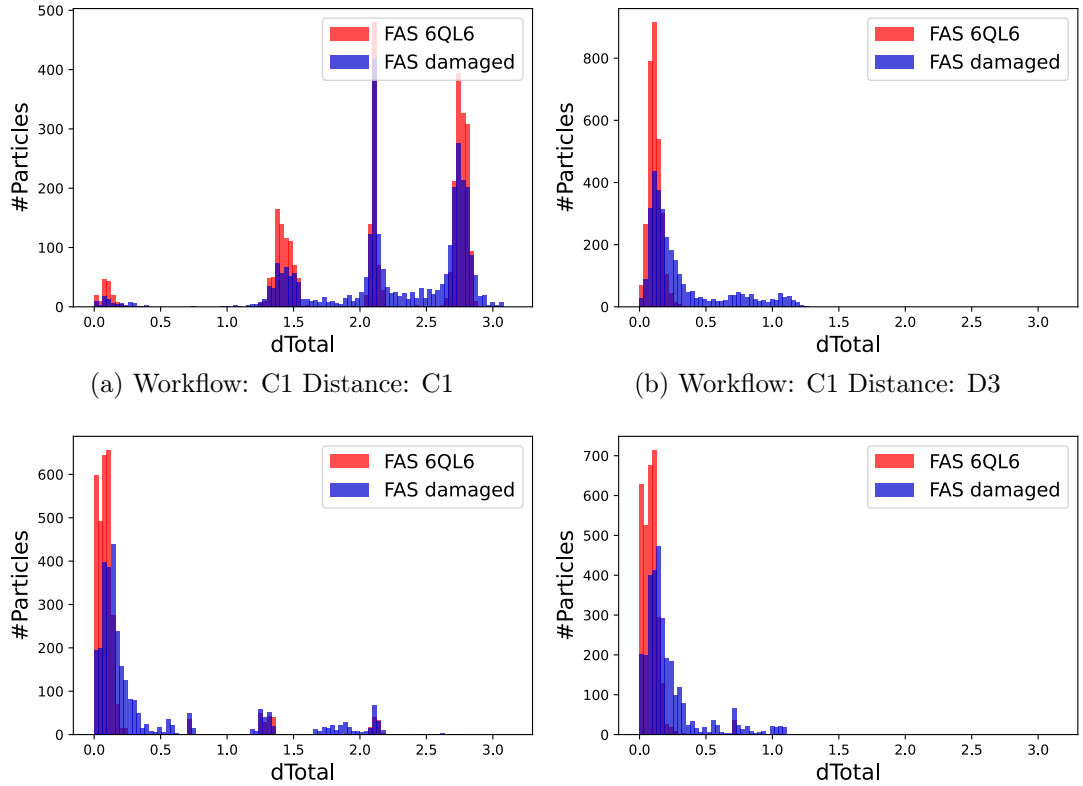
(c) Workflow: D3 Distance: C1



(d) Workflow: D3 Distance: D3

Figure A.8.: Histograms of the orientation consistency d_{total} values for simulated dataset 3 (see table 4.1). The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. The positive images are displayed in red, the negative images in blue.

Simulated dataset 4



(a) Workflow: C1 Distance: C1

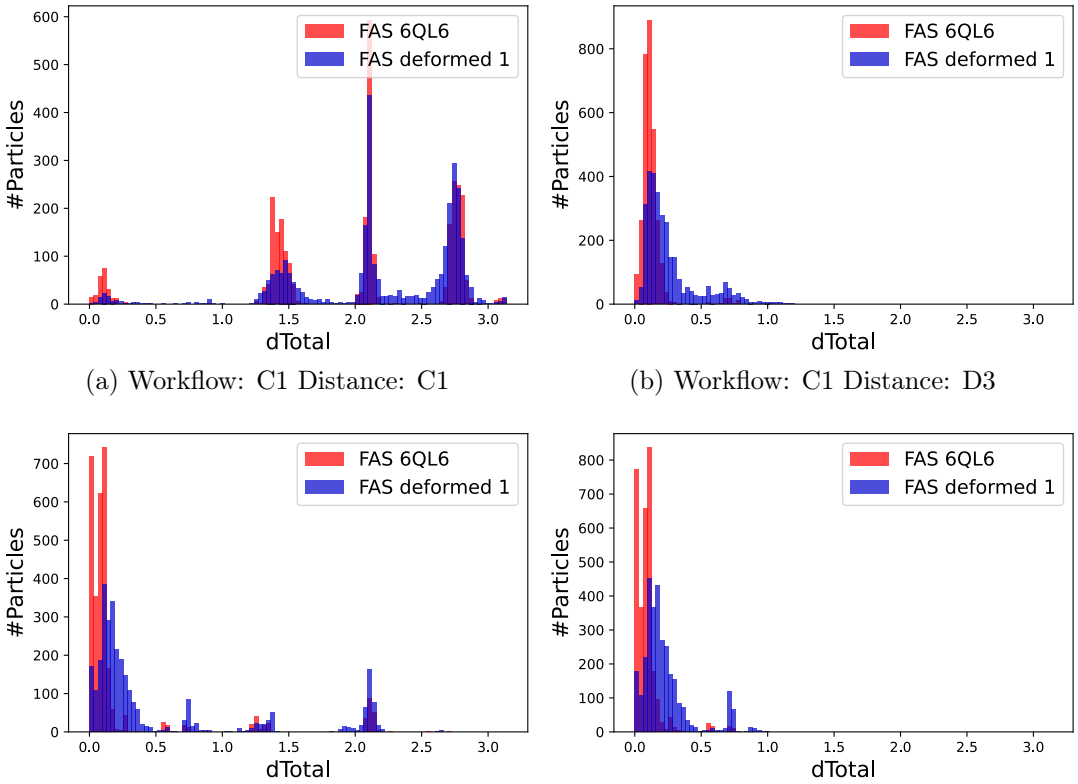
(b) Workflow: C1 Distance: D3

(c) Workflow: D3 Distance: C1

(d) Workflow: D3 Distance: D3

Figure A.9.: Histograms of the orientation consistency d_{total} values for simulated dataset 4 (see table 4.1). The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. The positive images are displayed in red, the negative images in blue.

Simulated dataset 5



(a) Workflow: C1 Distance: C1

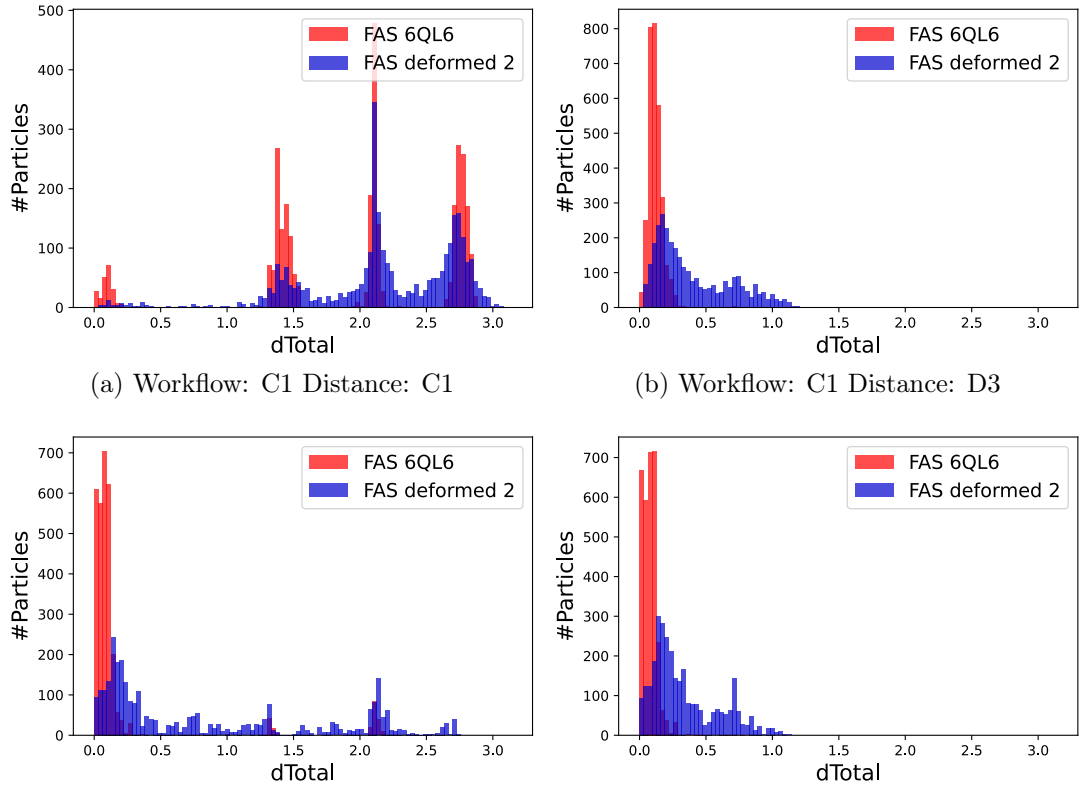
(b) Workflow: C1 Distance: D3

(c) Workflow: D3 Distance: C1

(d) Workflow: D3 Distance: D3

Figure A.10.: Histograms of the orientation consistency d_{total} values for simulated dataset 5 (see table 4.1). The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. The positive images are displayed in red, the negative images in blue.

Simulated dataset 6



(a) Workflow: C1 Distance: C1

(b) Workflow: C1 Distance: D3

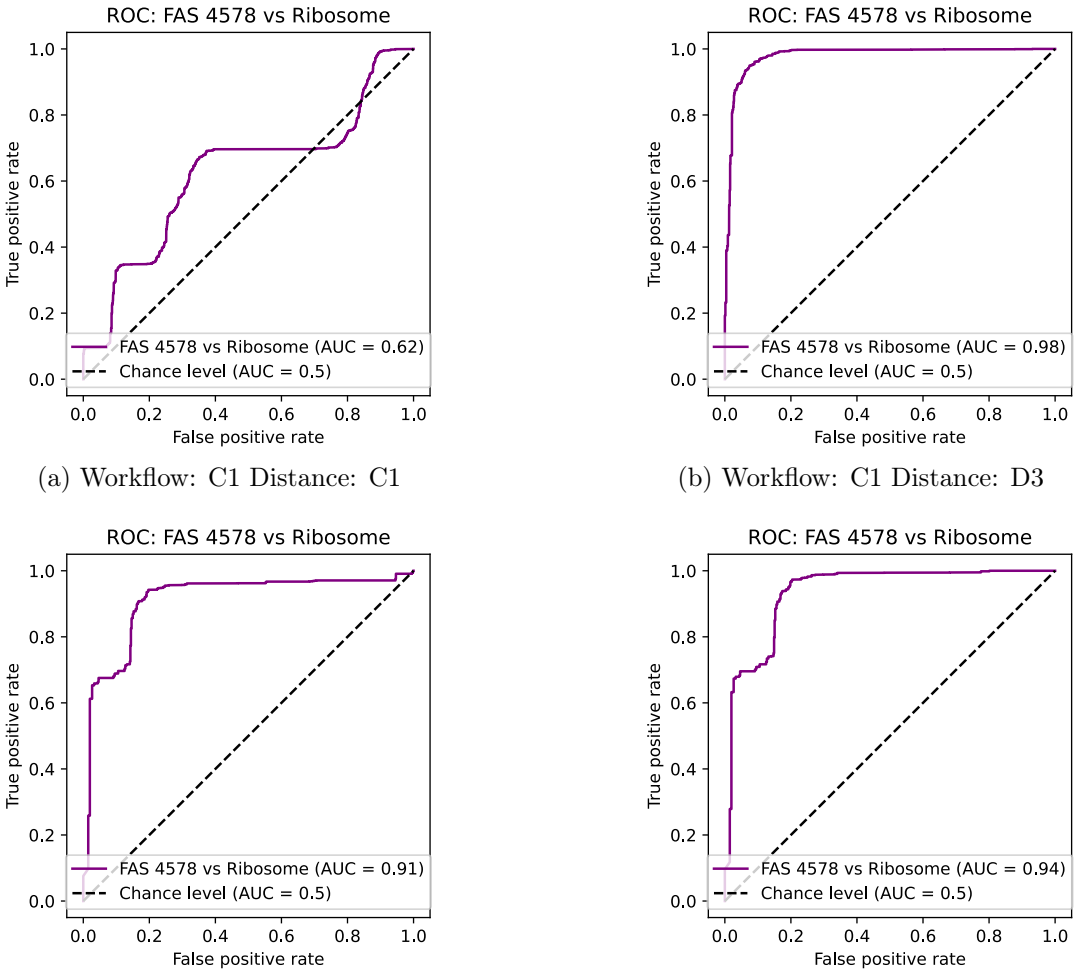
(c) Workflow: D3 Distance: C1

(d) Workflow: D3 Distance: D3

Figure A.11.: Histograms of the orientation consistency d_{total} values for simulated dataset 6 (see table 4.1). The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. The positive images are displayed in red, the negative images in blue.

ROC plots

Simulated dataset 2



(a) Workflow: C1 Distance: C1

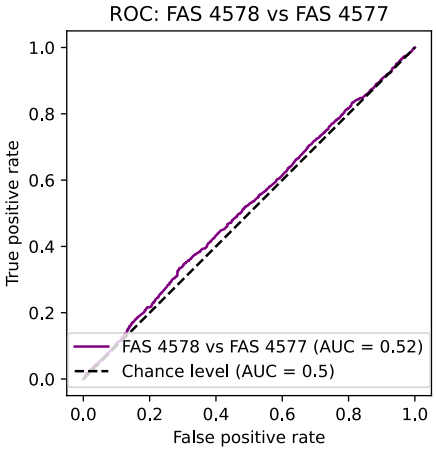
(b) Workflow: C1 Distance: D3

(c) Workflow: D3 Distance: C1

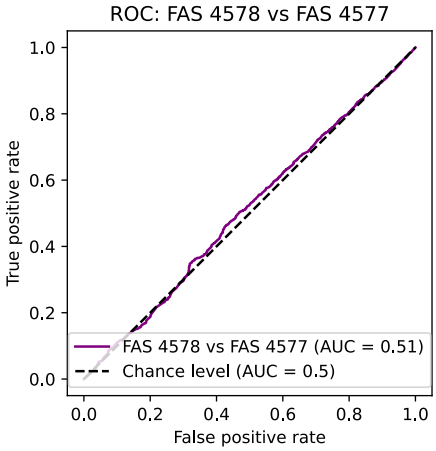
(d) Workflow: D3 Distance: D3

Figure A.12.: Receiver operating characteristic (ROC) curves for separating the positive and negative images of simulated dataset 2 (see table 4.1) by the orientation consistency parameter d_{total}. The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. In each plot, the area under the curve (AUC) is displayed. The dashed line indicates the random distribution ROC curve.

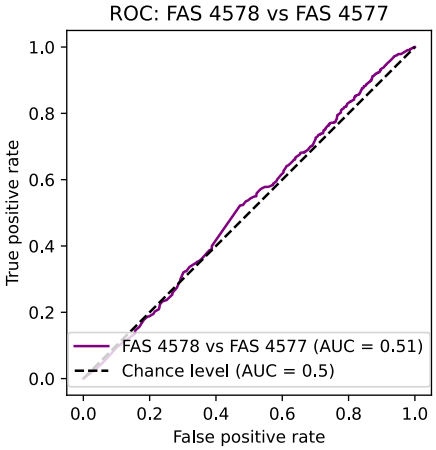
Simulated dataset 3



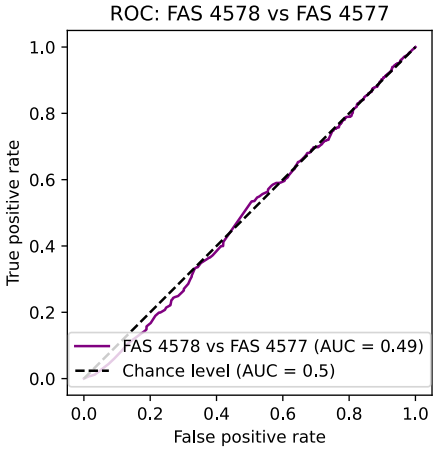
(a) Workflow: C1 Distance: C1



(b) Workflow: C1 Distance: D3



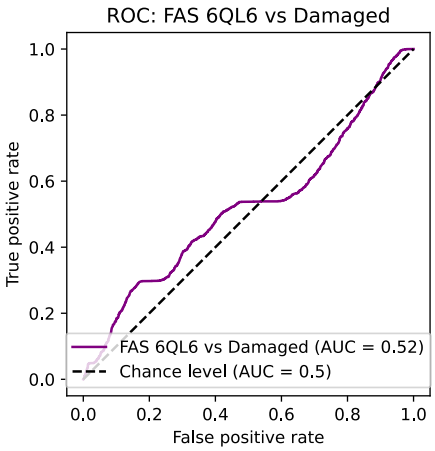
(c) Workflow: D3 Distance: C1



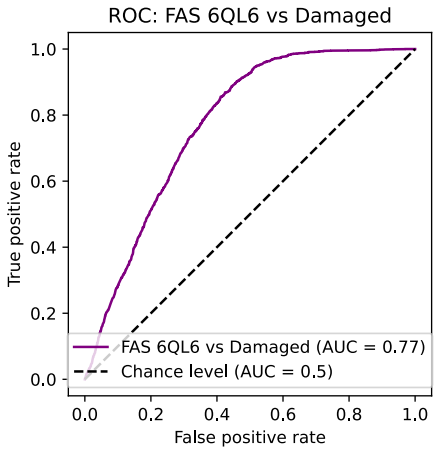
(d) Workflow: D3 Distance: D3

Figure A.13.: Receiver operating characteristic (ROC) curves for separating the positive and negative images of simulated dataset 3 (see table 4.1) by the orientation consistency parameter d_{total}. The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. In each plot, the area under the curve (AUC) is displayed. The dashed line indicates the random distribution ROC curve.

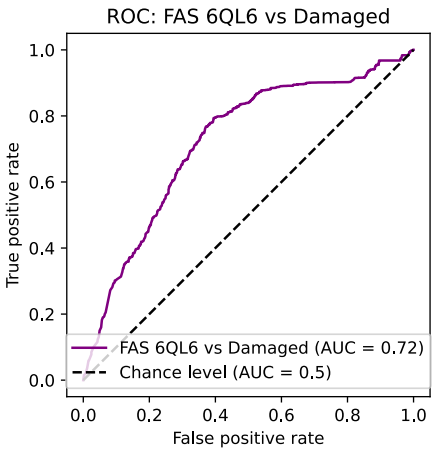
Simulated dataset 4



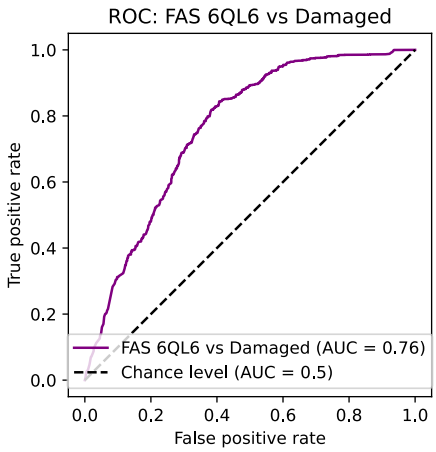
(a) Workflow: C1 Distance: C1



(b) Workflow: C1 Distance: D3



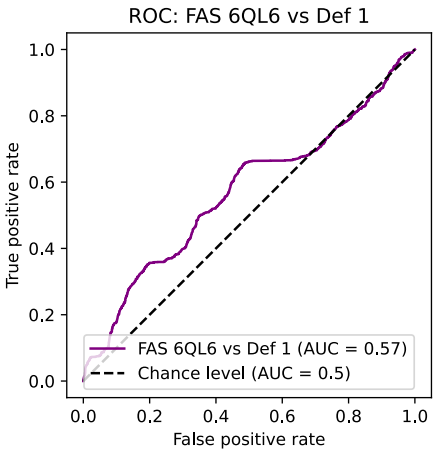
(c) Workflow: D3 Distance: C1



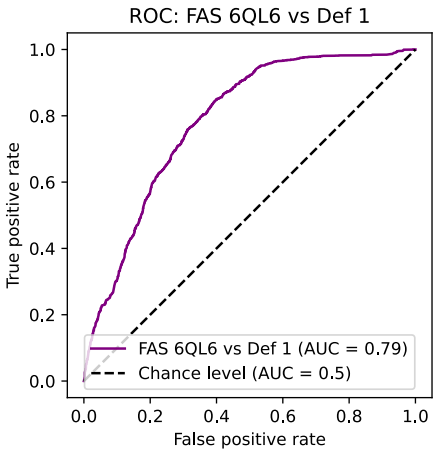
(d) Workflow: D3 Distance: D3

Figure A.14.: Receiver operating characteristic (ROC) curves for separating the positive and negative images of simulated dataset 4 (see table 4.1) by the orientation consistency parameter d_{total}. The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. In each plot, the area under the curve (AUC) is displayed. The dashed line indicates the random distribution ROC curve.

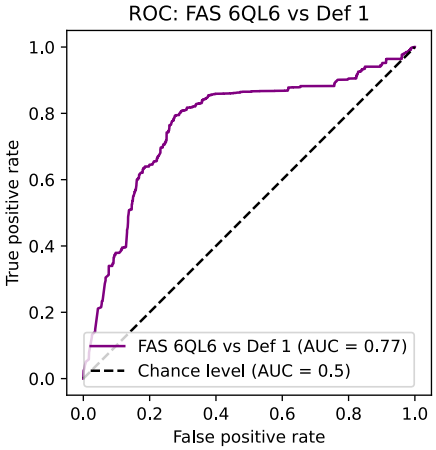
Simulated dataset 5



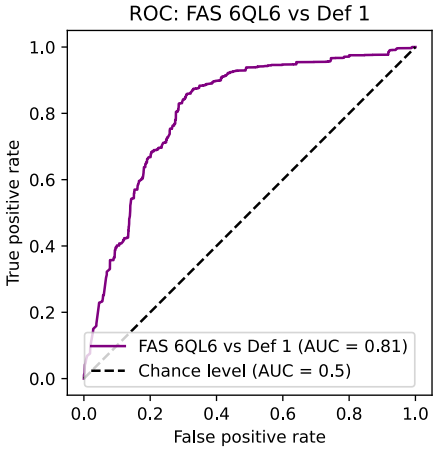
(a) Workflow: C1 Distance: C1



(b) Workflow: C1 Distance: D3



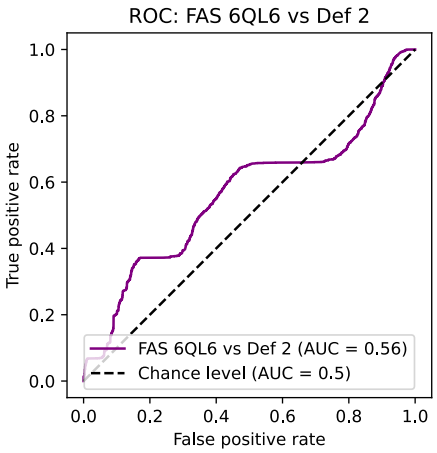
(c) Workflow: D3 Distance: C1



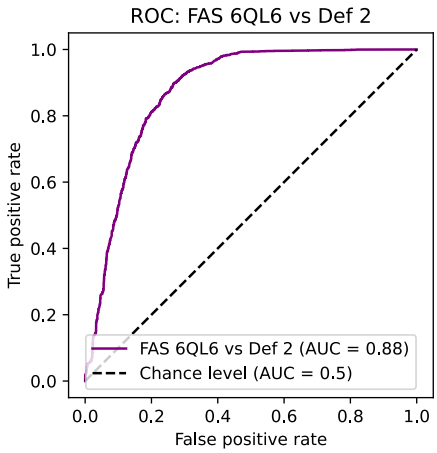
(d) Workflow: D3 Distance: D3

Figure A.15.: Receiver operating characteristic (ROC) curves for separating the positive and negative images of simulated dataset 5 (see table 4.1) by the orientation consistency parameter d_{total}. The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. In each plot, the area under the curve (AUC) is displayed. The dashed line indicates the random distribution ROC curve.

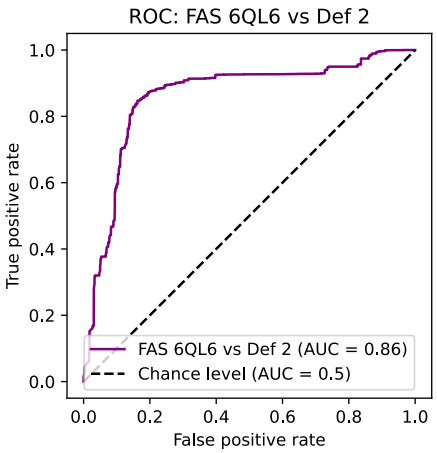
Simulated dataset 6



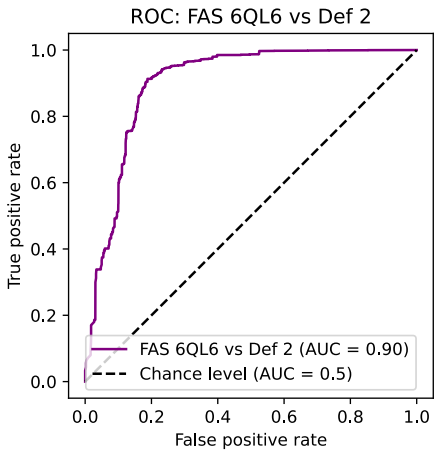
(a) Workflow: C1 Distance: C1



(b) Workflow: C1 Distance: D3



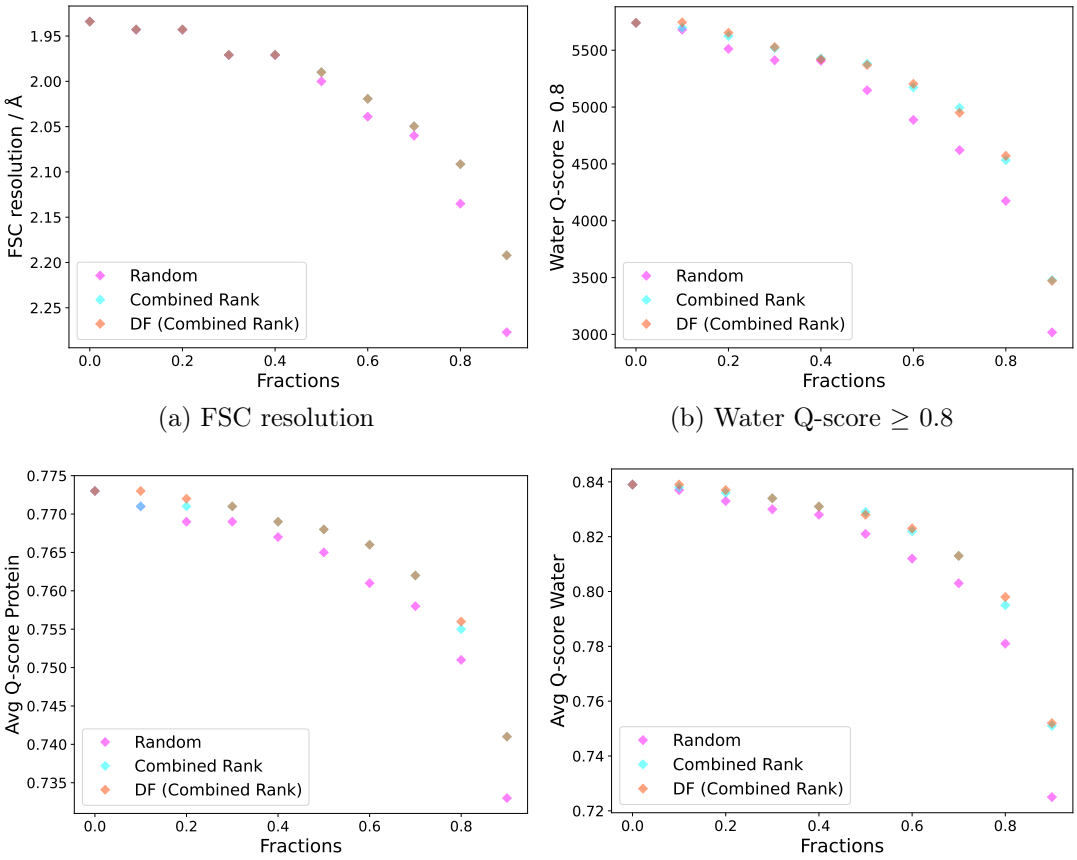
(c) Workflow: D3 Distance: C1



(d) Workflow: D3 Distance: D3

Figure A.16.: Receiver operating characteristic (ROC) curves for separating the positive and negative images of simulated dataset 6 (see table 4.1) by the orientation consistency parameter d_{total}. The symmetry groups C1 and D3 were applied in the workflow and during distance calculation as indicated in the image captions. In each plot, the area under the curve (AUC) is displayed. The dashed line indicates the random distribution ROC curve.

A.3. Additional data for filtering series



(a) FSC resolution

(b) Water Q-score ≥ 0.8

(c) Avg Q-score Protein

(d) Avg Q-score Water

Figure A.17.: Summary of different measures to estimate the quality of the refined maps when repeatedly filtering 10 % of the original images of the large FAS data set by the combined rank (cyan) or by directional filtering based on the combined rank (red). As a reference, the plot shows the results for random exclusion (pink).

Table A.1.: Particle numbers for filtering series on the large FAS dataset (CR=Combined ranking, DF=Directional filtering, OC=Orientation consistency, CS=CryoSieve).

Filtering Fraction	Random	CR	DF (CR)	OC	DF (OC)	CS
0.0	246,726	246,726	246,726	246,726	246,726	246,726
0.1	221,995	221,995	221,995	221,995	221,995	222,053
0.2	197,322	197,322	197,322	197,322	197,322	197,380
0.3	172,648	172,648	172,648	172,648	172,648	172,707
0.4	147,981	147,981	147,981	147,981	147,981	148,034
0.5	123,332	123,332	123,332	123,332	123,332	123,361
0.6	98,613	98,613	98,613	98,613	98,613	98,688
0.7	73,960	73,960	73,960	73,960	73,960	74,017
0.8	49,272	49,272	49,272	49,272	49,272	49,344
0.9	24,613	24,613	24,613	24,613	24,613	24,673

Table A.2.: FSC resolutions for filtering series on the large FAS dataset (CR=Combined ranking, DF=Directional filtering, OC=Orientation consistency, CS=CryoSieve). All resolution values in Å.

Filtering Fraction	Random	CR	DF (CR)	OC	DF (OC)	CS
0.0	1.93	1.93	1.93	1.93	1.93	1.93
0.1	1.94	1.94	1.94	1.94	1.94	1.94
0.2	1.94	1.94	1.94	1.94	1.94	1.94
0.3	1.97	1.97	1.97	1.97	1.94	1.93
0.4	1.97	1.97	1.97	1.98	1.97	1.94
0.5	2.00	1.99	1.99	1.99	1.99	1.93
0.6	2.04	2.02	2.02	2.03	2.04	1.94
0.7	2.06	2.05	2.05	2.06	2.06	1.97
0.8	2.14	2.09	2.09	2.12	2.12	2.03
0.9	2.28	2.19	2.19	2.33	2.27	2.15

Table A.3.: Average protein Q-scores for filtering series on the large FAS dataset (CR=Combined ranking, DF=Directional filtering, OC=Orientation consistency, CS=CryoSieve).

Filtering Fraction	**Random**	**CR**	**DF (CR)**	**OC**	**DF (OC)**	**CS**
0.0	0.773	0.773	0.773	0.773	0.773	0.773
0.1	0.771	0.771	0.773	0.771	0.771	0.774
0.2	0.769	0.771	0.772	0.769	0.771	0.774
0.3	0.769	0.771	0.771	0.768	0.769	0.777
0.4	0.767	0.769	0.769	0.768	0.768	0.779
0.5	0.765	0.768	0.768	0.767	0.765	0.780
0.6	0.761	0.766	0.766	0.764	0.762	0.782
0.7	0.758	0.762	0.762	0.758	0.757	0.781
0.8	0.751	0.755	0.756	0.750	0.749	0.773
0.9	0.733	0.741	0.741	0.725	0.733	0.753

Table A.4.: Expected water Q-scores for filtering series on the large FAS dataset (CR=Combined ranking, DF=Directional filtering, OC=Orientation consistency, CS=CryoSieve).

Filtering Fraction	**Random**	**CR**	**DF (CR)**	**OC**	**DF (OC)**	**CS**
0.0	0.836	0.836	0.836	0.836	0.836	0.836
0.1	0.834	0.834	0.834	0.834	0.834	0.834
0.2	0.834	0.834	0.834	0.834	0.834	0.834
0.3	0.826	0.826	0.826	0.826	0.834	0.836
0.4	0.826	0.826	0.826	0.824	0.826	0.834
0.5	0.818	0.821	0.821	0.821	0.821	0.836
0.6	0.807	0.813	0.813	0.810	0.807	0.834
0.7	0.802	0.804	0.804	0.802	0.802	0.826
0.8	0.782	0.793	0.793	0.785	0.785	0.810
0.9	0.745	0.767	0.767	0.732	0.748	0.779

Table A.5.: Average water Q-scores for filtering series on the large FAS dataset (CR=Combined ranking, DF=Directional filtering, OC=Orientation consistency, CS=CryoSieve).

Filtering Fraction	**Random**	**CR**	**DF (CR)**	**OC**	**DF (OC)**	**CS**
0.0	0.839	0.839	0.839	0.839	0.839	0.839
0.1	0.837	0.838	0.839	0.836	0.835	0.841
0.2	0.833	0.836	0.837	0.833	0.835	0.841
0.3	0.830	0.834	0.834	0.829	0.832	0.844
0.4	0.828	0.831	0.831	0.827	0.828	0.845
0.5	0.821	0.829	0.828	0.824	0.822	0.845
0.6	0.812	0.822	0.823	0.817	0.813	0.847
0.7	0.803	0.813	0.813	0.801	0.801	0.844
0.8	0.781	0.795	0.798	0.780	0.779	0.828
0.9	0.725	0.751	0.752	0.700	0.728	0.779

Table A.6.: Number of water molecules with a Q-score ≥ 0.8 for filtering series on the large FAS dataset (CR=Combined ranking, DF=Directional filtering, OC=Orientation consistency, CS=CryoSieve).

Filtering Fraction	**Random**	**CR**	**DF (CR)**	**OC**	**DF (OC)**	**CS**
0.0	5,741	5,741	5,741	5,741	5,741	5,741
0.1	5,681	5,699	5,746	5,633	5,609	5,789
0.2	5,512	5,625	5,654	5,547	5,597	5,802
0.3	5,412	5,520	5,528	5,454	5,489	5,873
0.4	5,407	5,428	5,421	5,346	5,395	5,918
0.5	5,148	5,381	5,370	5,274	5,218	5,932
0.6	4,887	5,173	5,204	5,077	4,953	6,062
0.7	4,621	4,994	4,950	4,668	4,615	5,953
0.8	4,174	4,533	4,571	4,165	4,144	5,419
0.9	3,018	3,478	3,471	2,655	3,187	4,149

A.4. Tables of CryoSieve subset filtering

Table A.7.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 1. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	% Particles	FSC (Å)
0	100	2.70
1	90	2.68
2	80	2.70
3	70	2.68
4	60	2.70
5	50	2.75
6	40	2.83
7	30	2.97
8	20	3.13
9	10	3.44

Table A.8.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 2. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	% Particles	FSC (Å)
0	100	2.63
1	90	2.63
2	80	2.66
3	70	2.66
4	60	2.66
5	50	2.71
6	40	2.75
7	30	2.83
8	20	3.01
9	10	3.31

Table A.9.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 3. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	% Particles	FSC (Å)
0	100	2.99
1	90	2.99
2	80	3.01
3	70	3.01
4	60	3.04
5	50	3.06
6	40	3.20
7	30	3.39
8	20	3.50
9	10	4.18

Table A.10.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 4. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	% Particles	FSC (Å)
0	100	2.53
1	90	2.53
2	80	2.56
3	70	2.56
4	60	2.56
5	50	2.58
6	40	2.70
7	30	2.83
8	20	2.93
9	10	3.23

Table A.11.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 5. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	**% Particles**	**FSC (Å)**
0	100	2.47
1	90	2.47
2	80	2.48
3	70	2.48
4	60	2.55
5	50	2.56
6	40	2.63
7	30	2.73
8	20	2.91
9	10	3.18

Table A.12.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 6. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	**% Particles**	**FSC (Å)**
0	100	2.81
1	90	2.81
2	80	2.75
3	70	2.87
4	60	2.91
5	50	2.93
6	40	3.08
7	30	3.20
8	20	3.47
9	10	3.94

Table A.13.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 7. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	% Particles	FSC (Å)
0	100	2.56
1	90	2.56
2	80	2.56
3	70	2.58
4	60	2.61
5	50	2.61
6	40	2.71
7	30	2.77
8	20	2.93
9	10	3.20

Table A.14.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 8. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	% Particles	FSC (Å)
0	100	2.45
1	90	2.45
2	80	2.48
3	70	2.50
4	60	2.56
5	50	2.56
6	40	2.64
7	30	2.68
8	20	2.81
9	10	3.01

Table A.15.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 9. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	**% Particles**	**FSC (Å)**
0	100	2.29
1	90	2.29
2	80	2.30
3	70	2.30
4	60	2.34
5	50	2.34
6	40	2.43
7	30	2.45
8	20	2.59
9	10	2.81

Table A.16.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 10. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	**% Particles**	**FSC (Å)**
0	100	2.20
1	90	2.20
2	80	2.20
3	70	2.19
4	60	2.19
5	50	2.22
6	40	2.24
7	30	2.30
8	20	2.38
9	10	2.59

Table A.17.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 11. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	% Particles	FSC (Å)
0	100	2.25
1	90	2.26
2	80	2.26
3	70	2.29
4	60	2.33
5	50	2.34
6	40	2.38
7	30	2.47
8	20	2.53
9	10	2.81

Table A.18.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 12. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	% Particles	FSC (Å)
0	100	2.23
1	90	2.23
2	80	2.24
3	70	2.24
4	60	2.28
5	50	2.29
6	40	2.32
7	30	2.34
8	20	2.44
9	10	2.70

Table A.19.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 13. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	**% Particles**	**FSC (Å)**
0	100	2.43
1	90	2.43
2	80	2.45
3	70	2.45
4	60	2.51
5	50	2.51
6	40	2.55
7	30	2.64
8	20	2.81
9	10	3.01

Table A.20.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 14. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	**% Particles**	**FSC (Å)**
0	100	2.24
1	90	2.23
2	80	2.24
3	70	2.22
4	60	2.24
5	50	2.24
6	40	2.32
7	30	2.34
8	20	2.44
9	10	2.71

Table A.21.: FSC resolutions over the iterations of the CryoSieve filtering run for subset 15. The first choice set is highlighted in blue, the second choice set is highlighted in purple.

Iteration	**% Particles**	**FSC (Å)**
0	100	2.70
1	90	2.70
2	80	2.71
3	70	2.75
4	60	2.77
5	50	2.77
6	40	2.87
7	30	2.93
8	20	3.11
9	10	3.36

A.5. Kept particle numbers and correlation with FSC deviations

Table A.22.: Particle numbers for the 15 subsets of the large FAS dataset directly after picking (Raw), after particle selection by 2D and 3D classification (Final), after reducing the final baseline dataset to 70 % by combined ranking (CR 0.7), to 70 % by directional filtering with the d_{total} parameter (DF 0.7), filtering to 70 % (CS 0.7) and 40 % (CS 0.4) with CryoSieve and filtering each subset individually with CryoSieve and selecting the first (CS 1.) and second (CS 2.) choice subset in each case.

Set	**Raw**	**Final**	**CR 0.7**	**DF 0.7**	**CS 0.7**	**CS 0.4**	**CS 1.**	**CS 2.**
1	24,572	5,558	4,398	3,875	4,321	3,395	3,890	3,334
2	37,857	8,768	6,186	6,005	6,532	4,990	7,891	5,261
3	12,761	2,401	1,758	1,628	1,848	1,289	2,161	1,681
4	36,048	7,036	4,337	4,632	5,216	3,236	6,333	4,223
5	131,792	23,691	20,165	16,629	17,750	9,691	21,322	16,583
6	52,134	7,609	5,722	5,459	5,908	3,053	6,087	5,326
7	41,619	6,997	4,712	4,913	4,848	1,647	5,597	4,897
8	82,286	16,732	11,797	11,707	11,566	6,595	15,058	13,384
9	76,506	19,907	16,018	13,676	15,621	10,409	17,916	13,934
10	201,646	43,754	32,363	30,238	25,788	10,718	26,254	21,878
11	67,280	20,730	11,316	14,688	16,063	10,206	20,730	16,584
12	100,843	28,029	18,639	20,095	15,854	4,162	25,226	19,620
13	70,584	17,550	12,146	12,370	11,565	7,548	15,795	12,285
14	139,702	30,567	19,189	21,636	24,410	17,930	21,399	15,284
15	60,778	7,397	3,902	5,097	5,417	3,819	6,657	5,917

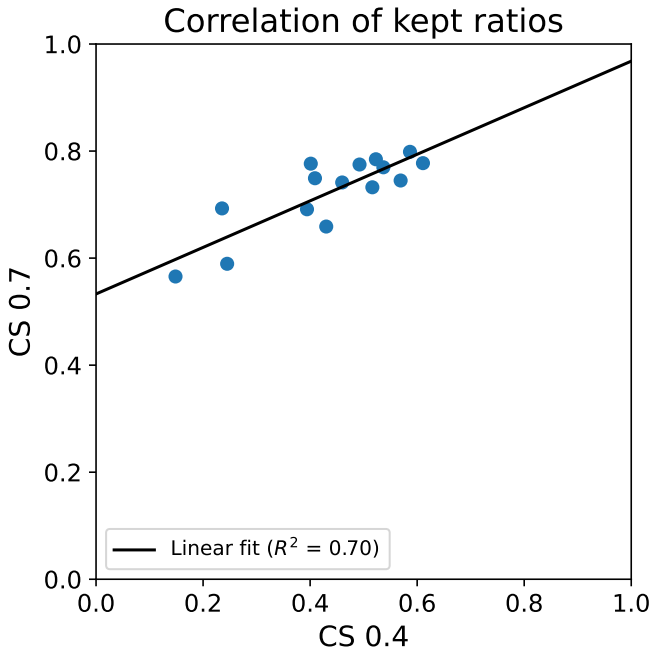


Figure A.18.: Correlation of the kept ratios with respect to the final baseline large FAS dataset over the 15 subsets when filtering the whole dataset with CryoSieve to 70 % (CS 0.7) and 40 % (CS 0.4) of the original data.

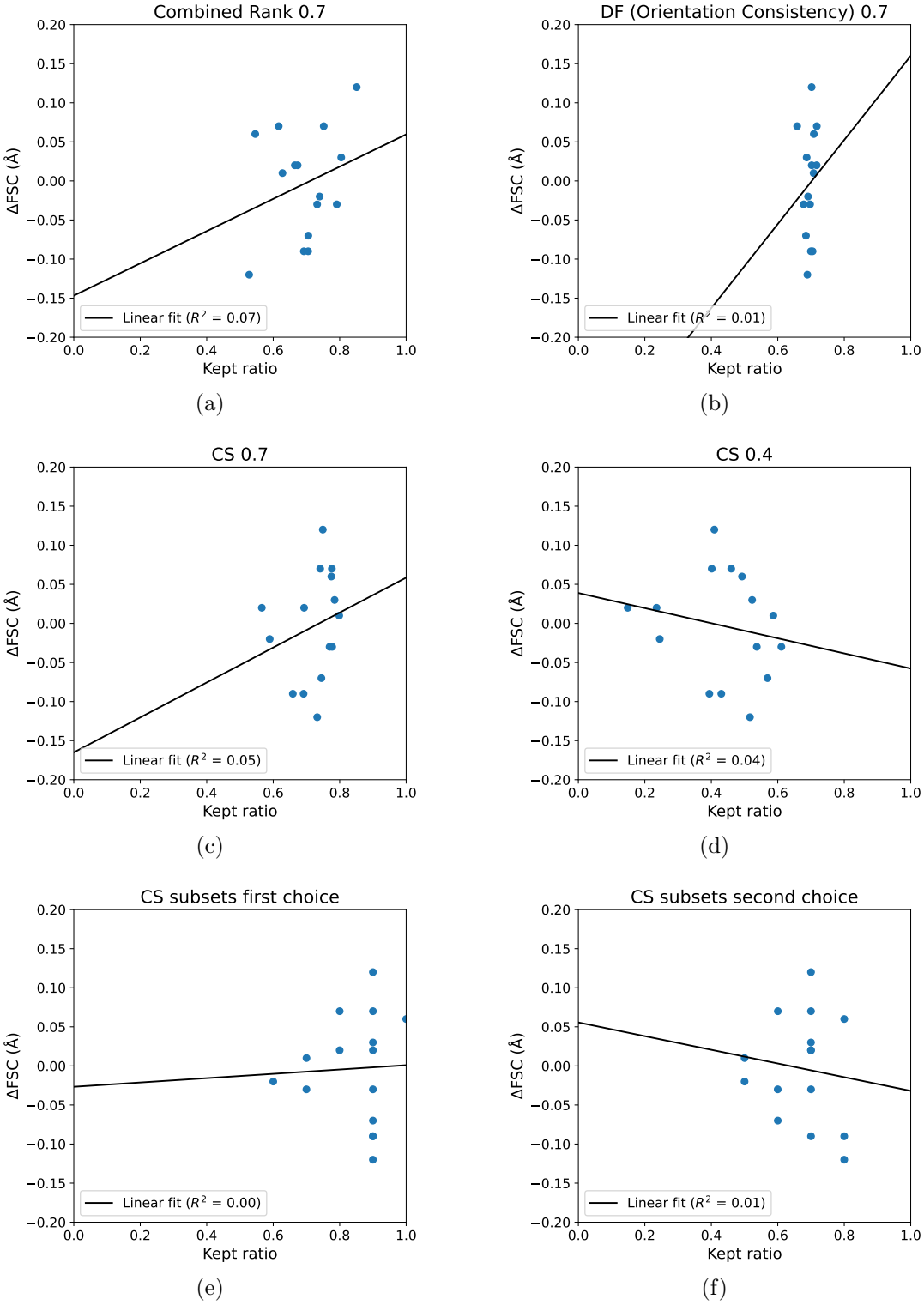


Figure A.19.: Correlations of the kept ratios with respect to the final baseline large FAS dataset over the 15 subsets with the per-subset FSC deviations of the refinement FSCs from the expected values for the different methods in section 5.7.

B. Processing details

B.1. Summary of scientific software

Table B.1.: Summary of the scientific software used in this thesis for data processing and analysis.

Application	**Software**
Workflow & Data management	COW [61]
Motion correction	MotionCor2 [31]
CTF estimation	GCTF [29]
Particle picking	Gautomatch [33], RELION 3.1 auto-picking [49]
2D/3D classification	RELION 3.1 [56]
3D refinement	RELION 3.1 [56]
Per-particle corrections	RELION 3.1 Bayesian Polishing [91] & CTF refine [49]
FSC calculation	RELION 3.1 [56]
Q-score calculation	mapq [127] plugin for UCSF Chimera [118]
Data simulation	COW [61], UCSF Chimera [118]
Visualization	UCSF Chimera [118], UCSF ChimeraX [122], pyem [124], matplotlib [128], seaborn [123]
Particle sorting (experiments)	CryoSieve [51]

B.2. COW: Basic use

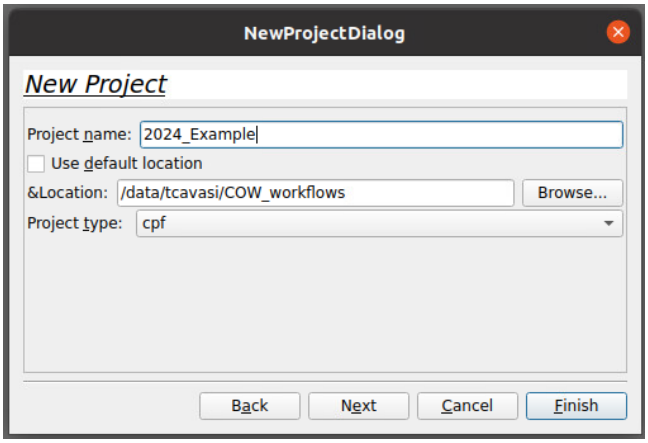


Figure B.1.: Creating a new project in COW [61].

The methods developed in this thesis are accessible via the COW [61] program described in 3.2. After starting the program, a new project can be created by clicking the quick access button in the top left corner, pressing Ctrl+N or via File→New Project. This induces the *NewProjectDialog* window, where name and path of the new project can be entered (see figure B.1). Pressing the *Finish* button completes the project creation. Once the COW project has been created, a new workflow tab can be added by clicking the plus sign on top of the workspace area. Now, workflows can be built by dragging logics from the *Logics* panel on the left into the workspace and connecting them. In figure B.2, this was done with the *Import* logic. The logics panel has a search function for quickly finding specific logics. All logics in the workspace are listed in the history window, sorted by their respective workflow tab. This allows for also dragging logics from other workflows into the current workflow tab. Logics can be connected with the mouse by pressing an inward/outward arrow of the first logic and then dragging a line to an outward/inward arrow of a second logic before releasing the mouse button. Connections can be removed by right-clicking on the connection line. The *Property* panel on the right shows all input parameters for the logic. In the case of the *Import* logic, this is for example the input file type and the path to the file. Another window shows the command line output after the logic is run. Running the workflow is achieved by pressing the play button in the top left corner of the workspace window. This triggers running of all logics, which did not run and successfully finish before, in their order in the workflow. If the option *Pause before executing* is turned on, the logic will pause once it is reached in the workflow

and the execution can be started by right-clicking on the paused logic and selecting *Start*. Most of the windows in the COW GUI can be adjusted to individual preferences. If they were closed by accident, they can be re-opened as one of the options under *Window* in the top bar. Below the *Logics* list, two extra tabs can be extended. The *Workflows* tab contains pre-built or user-created workflow templates, for example a cryo-EM image processing workflow using RELION logics. These templates can be dragged into the workspace just like logics. The *Tutorials* tab contains a tutorial video and an FAQ document.

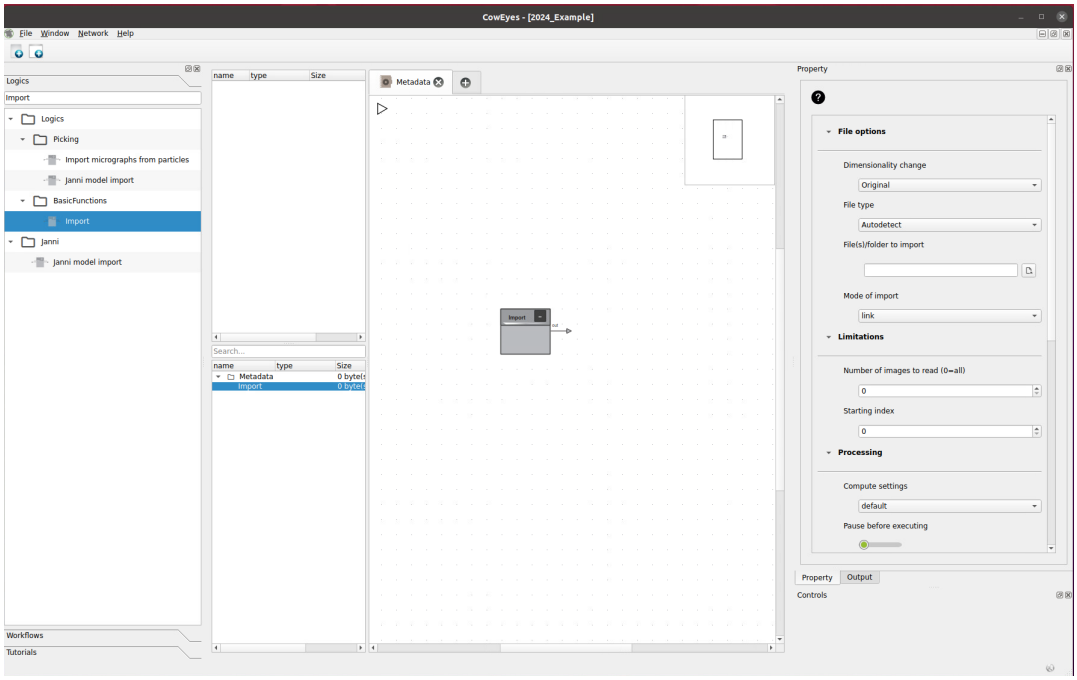


Figure B.2.: Overview of the COW [61] GUI, after a single *Import* logic was added to the workflow.

B.3. COW: Metadata collection

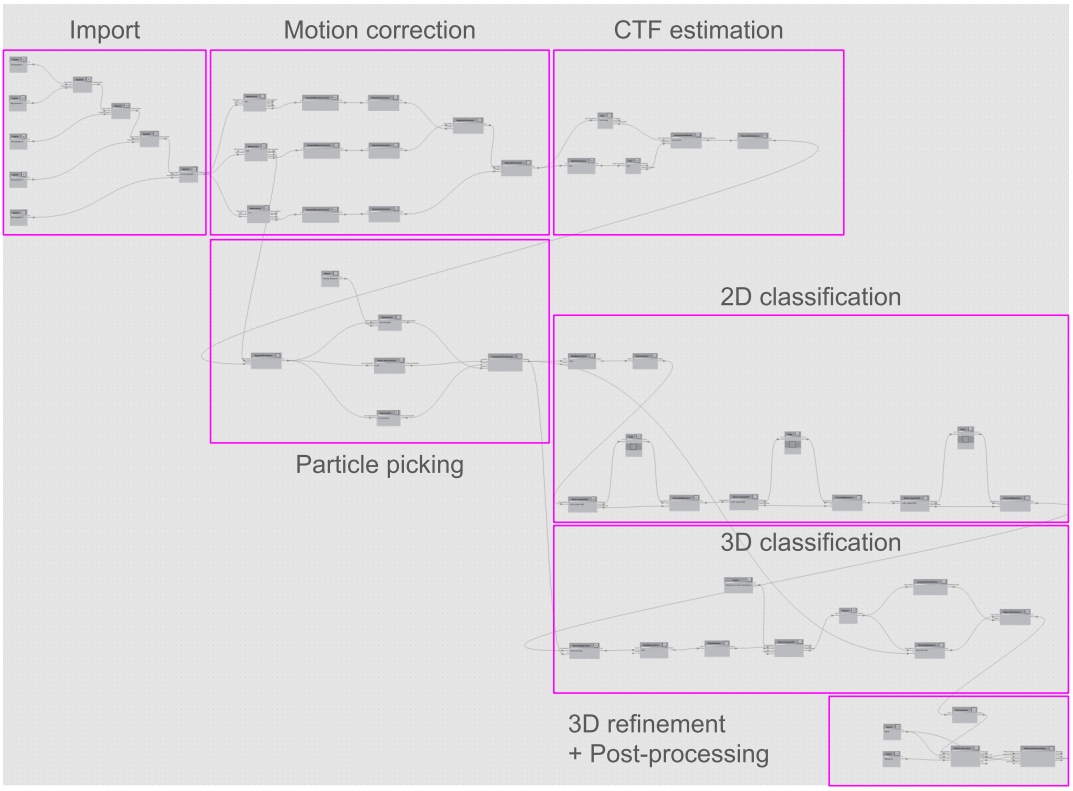


Figure B.3.: Overview of a workflow for image processing while collecting quality parameters in COW [61].

In figure B.3, an example of a basic cryo-EM image processing workflow with integrated quality parameter collection is shown. To re-build this workflow, a good starting point would be the RELION tutorial template provided in COW. The workflow starts with the import of the raw cryo-EM movies. Since the movie files are in five different folders, five *Import* logics are used. The movies are combined into a dataset using four consecutive Append logics.

The movies are then averaged using *MotionCor2*. Figure B.4 shows a close-up on the motion correction part of the workflow. The motion correction is carried out in three variants using 3x3, 5x5 and 7x7 patches per micrograph. In the downstream *AnalyzeMotionCorrection* logic, the motion-related quality parameters are computed and stored in the micrograph headers. The *RenameParameters* logic is used to add a suffix to the parameter names signifying the patch size used. The parameters from the

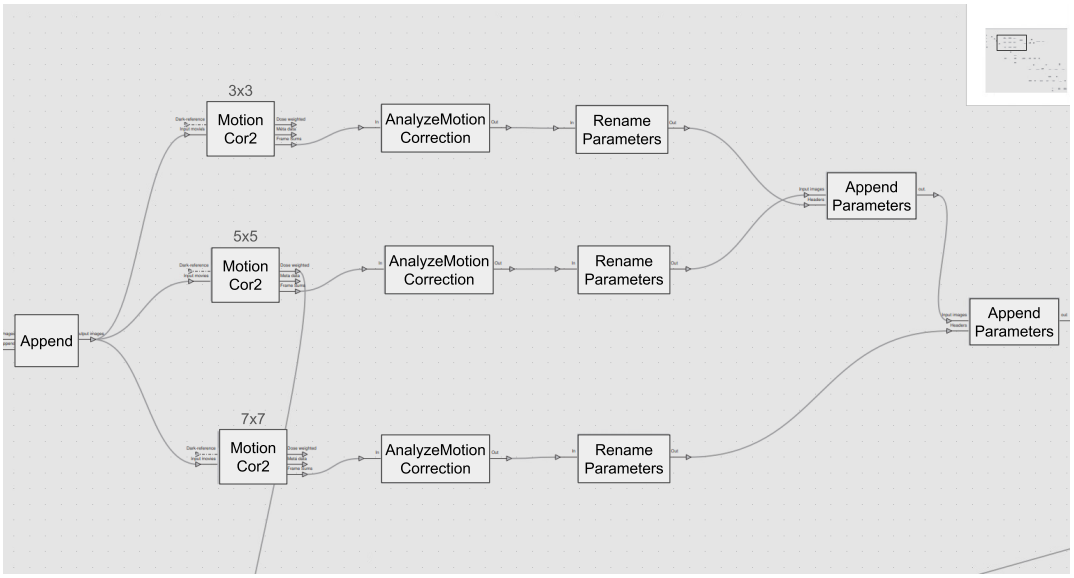


Figure B.4.: Close-up on the *Motion correction* part of figure B.3.

3x3 and 7x7 motion correction and parameter calculation are then appended to the headers of the summed micrographs created by using 5x5 patches. These micrographs now contain the quality parameters from all three motion correction variants.

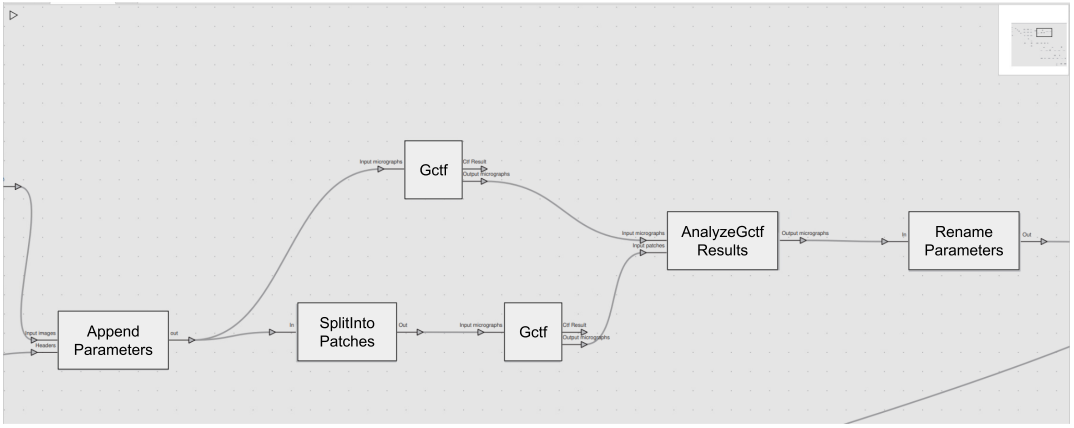


Figure B.5.: Close-up on the *CTF estimation* part of figure B.3.

The micrographs are forwarded to CTF estimation, which is shown magnified in figure B.5. The summed micrographs are divided into 5x5 patches in the *SplitIntoPatches* logic. The CTF estimation is then carried out with the *GCTF* wrapper logic on both the full micrographs and the micrograph patches. The CTF quality parameters are then calculated from micrograph and patch results in the *AnalyzeGCTFResults* logic.

The micrographs keep the estimated defocus parameters from the full frame estimate. The quality parameters get a 5x5 suffix in the *RenameParameters* logic.

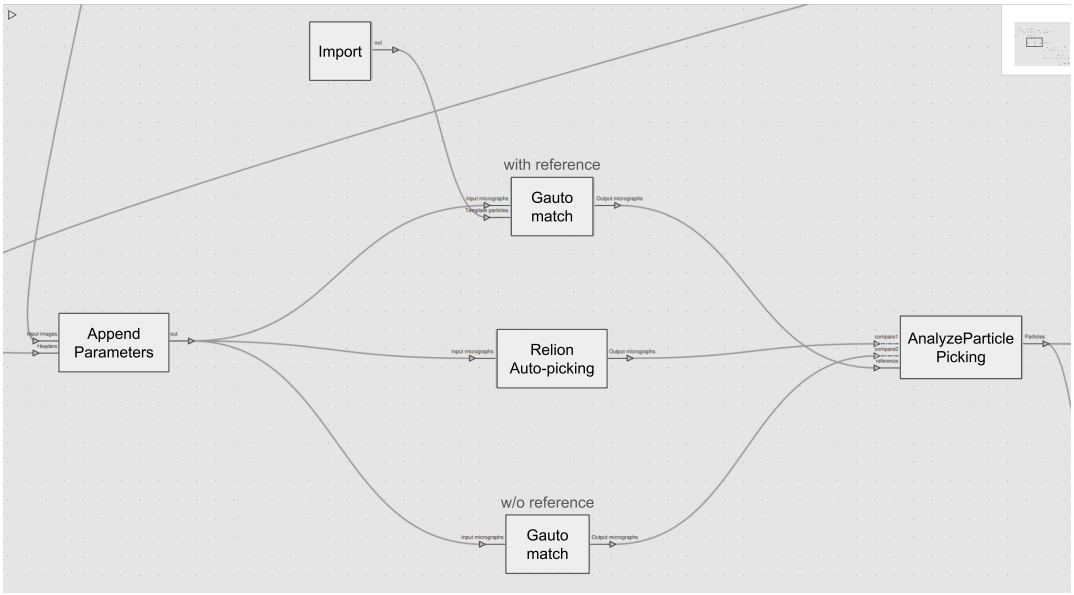


Figure B.6.: Close-up on the *Particle picking* part of figure B.3.

The next workflow step is particle picking (see figure B.6). To maximize image quality, the dose-weighted frame sums from the 5x5 motion correction are used. The quality parameters from motion correction and CTF estimation as well as the full micrograph CTF defocus parameters are thus appended to the dose-weighted frame sums. Three different picking strategies are applied to the micrographs: Template-based *Gautomatch*, template-free *Gautomatch* and template-free *LoG RELION Auto-picking*. For the template-based picking, reference images are imported. All three picking logics write the particle coordinates in the respective micrograph headers. The micrographs with the picking coordinates are then forwarded to the *AnalyzeParticlePicking* logic. This logic extracts the particle images using the coordinates from the *reference* input and calculates the PickingCounter parameter by checking whether the particles are contained in the two *compare* inputs. The logic also transfers all quality parameters from the micrograph headers to the respective particle headers.

The processing pipeline of the extracted particles consisting of 2D classification, 3D classification and 3D refinement is magnified in figure B.7. For the 2D classification, particle images are first binned to a larger pixel size using the *BoxManipulation* logic and then normalized. The particles are then subjected to three rounds of RELION

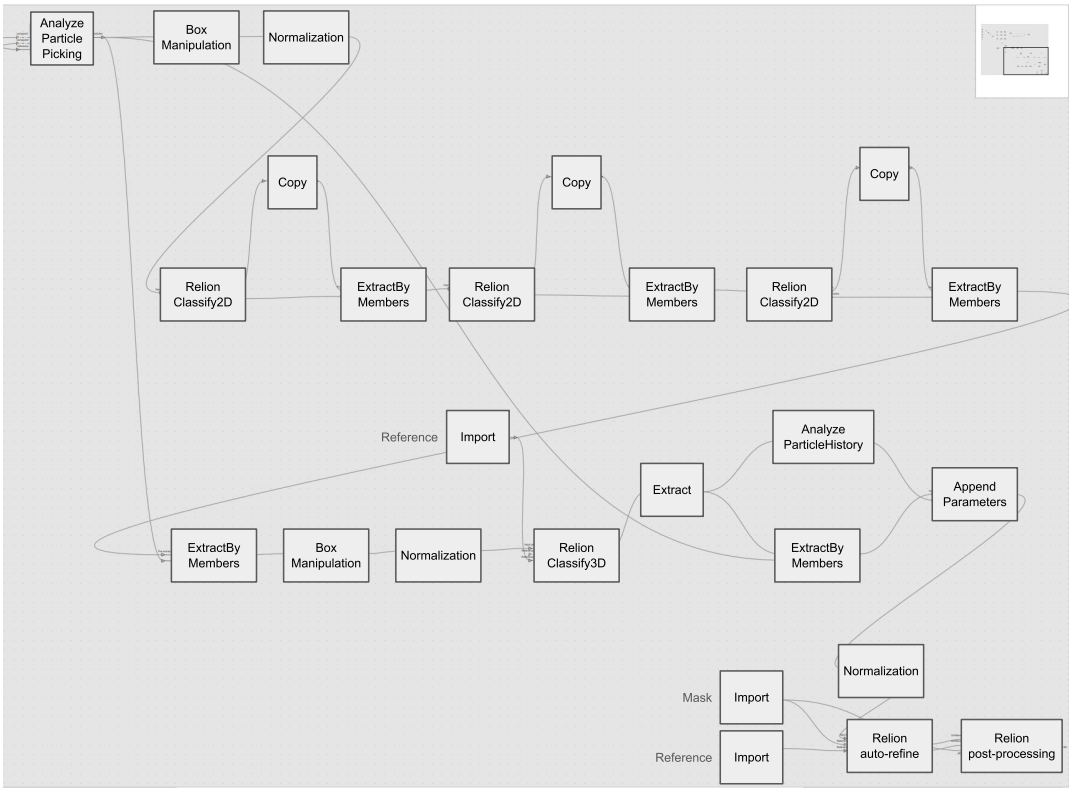


Figure B.7.: Close-up on the particle processing pipeline consisting of the *2D classification*, *3D classification* and *3D refinement + Post-processing* parts of figure B.3.

2D classification. After each round, the good classes are selected in the *Copy* logic by the user and the corresponding particles extracted in the *ExtractByMembers* logic. For the 3D classification, the *ExtractByMembers* logic is used – this time on the image IDs – to extract the particles selected during 2D classification among the unbinned images. The selected original images are then binned for 3D classification (usually by a smaller factor than for 2D classification) and normalized. Afterwards, the RELION 3D classification algorithm is applied using the *Store history* option to store shift, rotation and class assignments over the classification iterations for each particle. The required reference map is imported. After visual inspection of the 3D classification results, the user can select particles by providing the IDs of the good classes in the *Extract* logic. Only the particles which have these numbers as their *clusterMember* entry will be kept. For the selected particles, the 3D classification-related quality parameters are computed based on the shift, rotation and class history in the *An-*

alyzeParticleHistory logic. Finally, the *ExtractByMembers* logic is used to extract the particles remaining after 3D classification among the unbinned images. The 3D classification quality parameters are added to the headers of these particles with the *AppendParameters* logic. The images are normalized before reconstructing a baseline map with RELION auto-refine. The reference and mask used for the refinement are imported. After refinement, a RELION post-processing run is carried out to compute the FSC resolution.

B.4. COW: Orientation consistency

The workflow for computing the orientation consistency parameter is available as a template in COW. In this template, orientations are determined in three different ways for each particle and the average pairwise distance between the orientations is calculated. Figure B.8 shows a schematic illustration of the template workflow.

The workflow starts with the *Import* of the experimental images and the reference map. CTF correction is carried out on the images in the *ApplyCTFParameters* logic. It is assumed that the images already carry the estimated CTF parameters in their headers. The images as well as the reference are coarsened by the same amount in the *BoxManipulation* logic. Afterwards, 2D projections are created based on the reference in the *Projection* logic. Two sets of projections are created: one coarse set of projections to be used as an anchor set for the angular reconstitution (lower logic) and a finer one to be used as reference images for the alignment (upper logic). The experimental images are aligned to the finer projection set in the *Alignment* logic. The *Alignment* logic outputs the alignment parameters which can be applied to the images in the *AlignmentApply* logic. This means that the images are rotated by the in-plane rotation angle *eulerAlpha* so that they fit the assigned reference projection and *eulerBeta* and *eulerGamma* in the image header are copied from the reference projection. In addition, the determined x-/y-shifts from the alignment are applied. The aligned images are subjected to MSA-based classification in the *Classification* logic and summed up to a user-defined number of class average images. These class sums are aligned to the same reference projections as before and the alignment parameters are applied. Concurrently, the class sums are subjected to orientation determination in the *AngularReconstitution* logic. This logic is in principle reference-free, but very few projections are required as an anchor set to fix the global orientation of the protein of interest in the 3D coordinate system. The coarse projections are therefore converted to sinograms and then inserted into *AngularReconstitution* logic as the anchor set. At this point, the orientation has been determined three times for each image: Once directly on the particle images and twice on derived class sums. In the *MapEulerAngles* logic, the orientations determined on class sums are mapped back to the particles the classes originated from. Each particle image then contains three sets of Euler angles, one from direct alignment, one from class sum alignment and one from class sum angular reconstitution. In the *AnalyzeClassificationConsistency* logic,

pairwise distances between the orientations are computed as well as the average of the pairwise distances. These parameters are appended to the original images with the *AppendParameters* logic. The *CompareParticleCoords* logic can be used to check whether the images after and before the workflow are in the same order and no images are missing by comparing cropping coordinates and micrograph names at each input position. Finally, filtering of the particles can be carried out for example by setting a threshold for the orientation distance parameter in the *Extract* logic or by using the *DirectionalFiltering* logic.

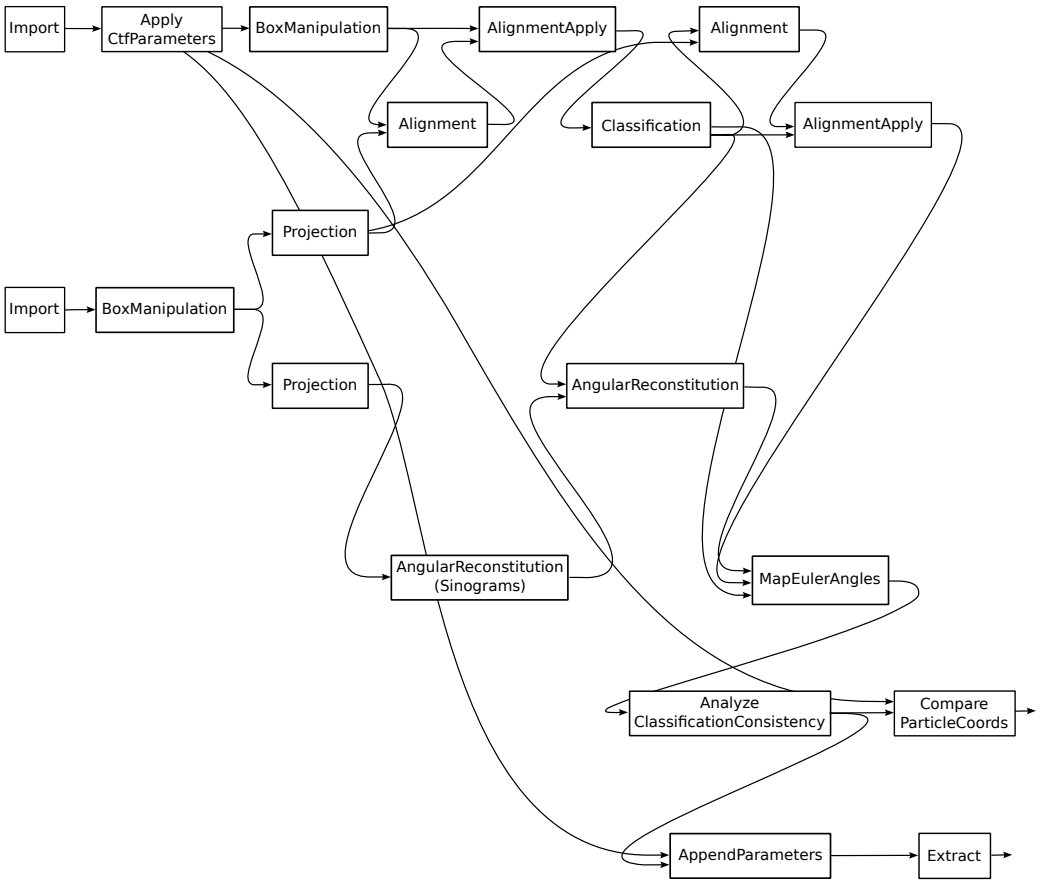


Figure B.8.: Illustration of the orientation consistency workflow template in COW [61].

B.5. COW: Directional filtering

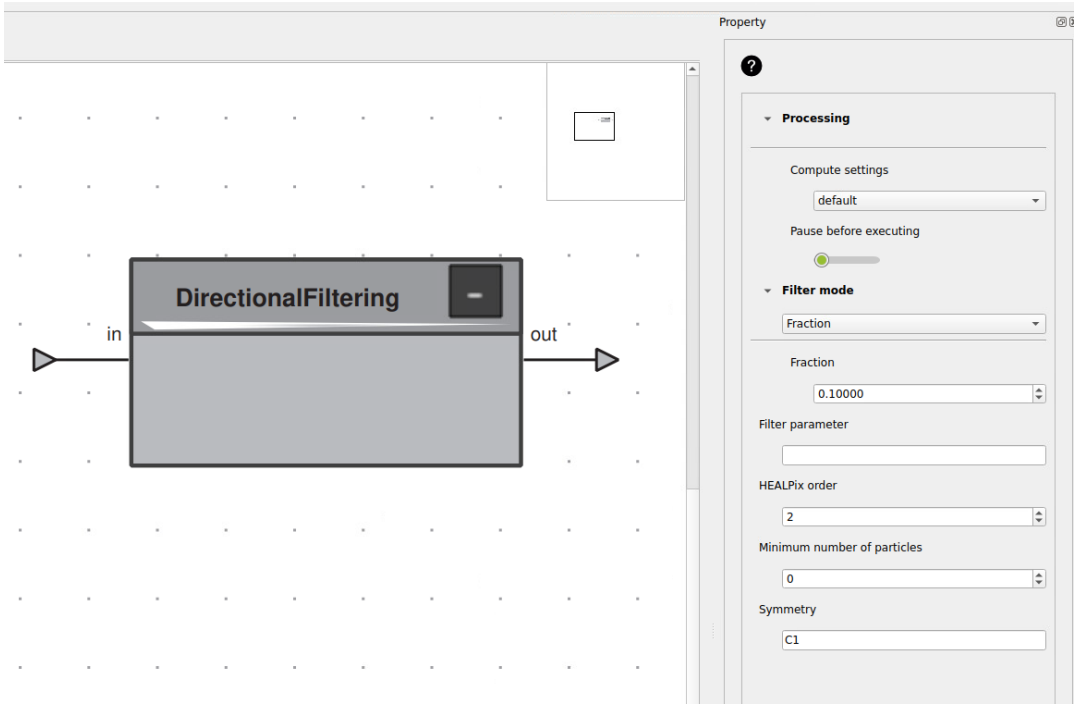


Figure B.9.: Interface of the *Directional Filtering* logic in COW [61].

The *DirectionalFiltering* algorithm is available as a logic in COW. The interface of the logic is relatively simple. There is a single input for the unfiltered images and the filtered image set is provided in the output. Next to the standard *Processing* options, which are the same for all logics, the input parameters include the *Filter mode*, which can either be to filter a user-defined *Fraction* of particles per bin or to filter boxplot outliers from each bin. In addition, the *Filter parameter* to be used for filtering needs to be defined as well as the fineness of the binning defined by the HEALPix order (default is 2). The user can set a *Minimum number of particles* to always be kept per bin to protect underrepresented orientations. If applicable, the structural symmetry can be defined. This should only be done, if the orientations of the input particles are within the asymmetric unit of the given symmetry group as the HEALPix reference orientations are only created in this area. The *DirectionalFiltering* logic is versatile and can be run with any numerical parameter.

C. Publications & Presentations

The work in this thesis was presented at the following occasions:

- **A. T. Cavasin**, M. Luettich, H. Stark, M. Kolbe and M. Rarey, *Metadata-based analysis of image quality for single-particle cryo-EM*, CDCS Opening Symposium, Hamburg, 2022 (Poster)

- **A. T. Cavasin**, M. Luettich, H. Stark, M. Kolbe and M. Rarey, *Metadata-based image sorting for single-particle cryo-EM*, EMBO Practical Course: Advances in cryo-electron microscopy and 3D image processing, Heidelberg, 2022 (Student talk)

The following presentations were given in the context of this thesis:

- **A. T. Cavasin**, *Theoretical background and algorithms for 3D reconstruction in single-particle cryo-EM*, CSSB CryoEM Course, Hamburg, 2021 & 2022 (Educational talk)

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Hamburg, den 24.03.2025

Anna Theresa Cavasin