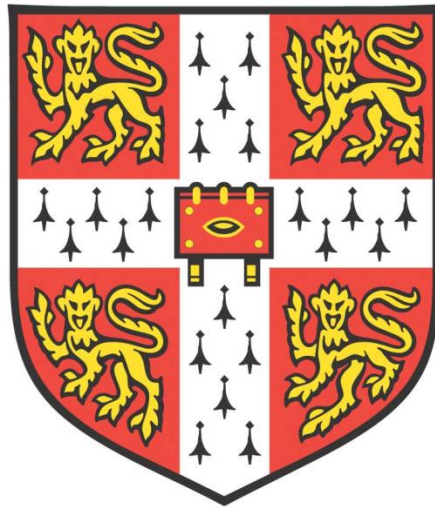


HELICAL RECONSTRUCTION IN RELION



Shaoda He

Hughes Hall

University of Cambridge

&

MRC Laboratory of Molecular Biology

This dissertation is submitted for the degree of Doctor of Philosophy

January 2018

DECLARATION

Name: Shaoda He

Title: Helical Reconstruction in RELION

The work described in this thesis was carried out by me in the MRC Laboratory of Molecular Biology between October 2014 and September 2017 under the supervision of Dr. Sjors Scheres. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

This thesis does not exceed 60,000 words in length and it contains less than 100 figures.

Shaoda He

ABSTRACT

Name: Shaoda He

Title: Helical Reconstruction in RELION

Helical assemblies of proteins are ubiquitous in nature and they perform vital functions in a wide range of organisms. The recent development of direct electron detectors and other imaging techniques in cryo-electron microscopy (cryo-EM) has opened new possibilities in solving helical structures at atomic resolution. Existing software packages for helical processing often require experience in tuning many *ad hoc* parameters to achieve optimal reconstruction results. REgularised LIkelihood Optimisation (RELION), an open-source single-particle analysis package, reduces the need for user expertise by the formulation of an empirical Bayesian framework, and has yielded some of the highest resolution density maps in recent years. Prior information about the helical assemblies can be conveniently incorporated into the statistical framework of RELION and thereby improves the helical reconstructions.

This PhD thesis describes the development of a helical processing computation workflow with reduced user intervention in RELION. Chapter 1 introduces the theoretical basis of cryo-EM data acquisition and single-particle data processing, the concepts of helical symmetry, and a previously described method for iterative real-space reconstruction of helical assemblies, to which the RELION implementation bears resemblance. Chapter 2 discusses multiple adaptations to RELION that are necessary for helical processing. Key elements include the imposition and local refinement of helical symmetry, masks on helical segments and references, expressions of angular and translational prior information, manual and automated segment picking as well as initial model generation for helices. Calculations have been performed on four test data sets showing that the developed methods in RELION yield results that are as good as or better than alternative approaches for the tests performed. Chapter 3 describes the same methodology adapted to helical sub-tomogram averaging in RELION. Chapter 4 introduces the local symmetry option developed for special types of filaments with pseudo-helical symmetry. The concept can be extended to general single-particle analysis as well. Chapter 5 describes four helical structures determined in collaboration with other research groups using helical RELION for data processing. Chapter 6 concludes the thesis with a brief summary and future prospects.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Sjors Scheres for his excellent day-to-day supervision of my research over the last three years and his valuable suggestions to the text. I am grateful to Nigel Unwin, my second supervisor, for the helpful discussions and his gracious support and guidance. Thanks are due also to Ben Luisi, my University supervisor, for his support in my first year assessment. I truly appreciate the examiners, Christopher Russo and Peter Rosenthal, for their great efforts in reviewing this thesis and presenting in my viva. I am deeply indebted to Jude Short, who taught me traditional methods in helical processing in great detail. I am also indebted to Andres Gonzalez, a Part II student who worked on testing helical sub-tomogram averaging in RELION under my day-to-day supervision in the spring of 2017. I am thankful to Rafael Fernandez for his help in using PyMol, Jake Grimmet and Toby Darling for maintaining the computer cluster. I have enjoyed the research collaborations with Frances Alvarez, Xian Deng, Anthony Fitzpatrick, Sam Lacey, Jan Löwe, Osman Salih and Peijun Zhang. I am also thankful to Tanmay Bharat, Tony Crowther, Linas Urnavicius for the discussions and Xiaochen Bai, Julian Conrad, Cornelius Gati and Thomas Martin as current and previous members of Sjors' research group for their company. In addition, I am grateful to Malcolm Morgan and Kayla Friedman for their online sharing of dissertation templates in Word format.

I am also grateful for the generous financial support from the LMB Scholarship, jointly funded by MRC Laboratory of Molecular Biology and Cambridge Commonwealth European and International Trust. Hughes Hall, my residential college, deserves my thanks for their help, support and advice mainly for accommodation, catering and administrative service. I am indebted as well to Sanctuary Housing at Addenbrooke's for their service of open-ended, affordable en-suite accommodation very close to the institute.

My gratitude rightly extends to family, Fengzhu Liu, Wei He and Xuan Li, for their love. I have really enjoyed my PhD research and my life at Cambridge.

CONTENTS

1 INTRODUCTION.....	1
1.1 STRUCTURAL BIOLOGY AND CRYO-ELECTRON MICROSCOPY	1
1.2 TRANSMISSION ELECTRON MICROSCOPE AND IMAGE FORMATION	2
1.3 SINGLE-PARTICLE ANALYSIS AND ELECTRON TOMOGRAPHY	7
1.4 THE RECONSTRUCTION ALGORITHM IN RELION	9
1.5 HELICAL ASSEMBLIES	15
1.6 HELICAL SYMMETRY.....	19
1.7 FOURIER-BESSEL METHOD FOR 3D HELICAL RECONSTRUCTION	20
1.8 ITERATIVE HELICAL REAL-SPACE RECONSTRUCTION ALGORITHM	23
2 IMPLEMENTED ALGORITHMS IN SINGLE-PARTICLE HELICAL RECONSTRUCTIONS	25
2.1 WORKFLOW OF HELICAL PROCESSING	25
2.2 IMPLEMENTED METHODS	28
2.2.1 <i>Definition of helical symmetry</i>	28
2.2.2 <i>Applying helical symmetry in Fourier-space</i>	30
2.2.3 <i>Applying helical symmetry in real-space</i>	32
2.2.4 <i>Local searches of helical symmetry in real-space</i>	33
2.2.5 <i>2D image masks and 3D reference masks</i>	35
2.2.6 <i>Incorporating prior information</i>	38
2.2.7 <i>Segment picking and extraction</i>	41
2.2.8 <i>Randomisation of image subsets</i>	43
2.2.9 <i>Initial model generation</i>	44
2.2.10 <i>Graphic user interface for helical reconstruction jobs</i>	46
2.2.11 <i>Simulations of 2D experimental segments</i>	46
2.3 INTRODUCTION TO THE TEST DATA SETS	48
2.4 RESULTS AND DISCUSSION	50
2.4.1 <i>Segment picking</i>	50
2.4.2 <i>2D classifications</i>	55
2.4.3 <i>Initial helical symmetry estimation and initial model generation</i>	57
2.4.4 <i>Pitfalls of imposing incorrect initial helical symmetry</i>	63
2.4.5 <i>Applying helical symmetry in Fourier-space</i>	67
2.4.6 <i>Inter-box distances and mask diameters</i>	68
2.4.7 <i>Applying helical symmetry in real-space</i>	71

2.4.8 Local searches of helical symmetry.....	71
2.4.9 Angular and translational priors	74
2.4.10 Prevention of over-fitting in 3D auto-refinements	78
2.4.11 High-resolution structures	82
3 IMPLEMENTED ALGORITHMS IN HELICAL SUB-TOMOGRAM AVERAGING	86
3.1 IMPLEMENTED METHODS	86
3.1.1 Manual-picking of sub-tomograms	86
3.1.2 3D image masks	87
3.1.3 Incorporating angular prior information.....	88
3.1.4 Simulations of 3D experimental sub-tomograms	89
3.2 INTRODUCTION TO THE TEST DATA SET	89
3.3 RESULTS AND DISCUSSION.....	90
4 LOCAL SYMMETRY.....	92
4.1 MICROTUBULES WITH SEAMS - BREAKING FROM NORMAL HELICAL SYMMETRY	92
4.2 IMPLEMENTED METHODS	96
4.2.1 Definition of local symmetry	96
4.2.2 Applying local symmetry	98
4.2.3 Local searches of local symmetry operators.....	100
4.2.4 Creating local symmetry masks and operators for a 13-microtubules with seam	101
4.2.5 Creating local symmetry masks and operators for arbitrary cryo-EM density maps.....	103
4.2.6 Verification of local symmetry operators.....	106
4.2.7 Enabling local symmetry option in 3D reconstructions.....	107
4.3 INTRODUCTION TO THE TEST DATA SET OF 13-MICROTUBULE WITH SEAM	107
4.4 RESULTS AND DISCUSSION.....	108
5 SOLVED STRUCTURES	111
5.1 BTUBABC MINI MICROTUBULE	113
5.2 MxB (MYXOVIRUS RESISTANCE PROTEINS).....	115
5.3 TYPE VI CONTRACTILE SHEATH TSSB1C1 FROM <i>PSEUDOMONAS AERUGINOSA</i>	117
5.4 TAU FILAMENTS	119
6 FUTURE PROSPECTS.....	121
7 REFERENCES.....	126
8 APPENDIX.....	137

LIST OF ABBREVIATIONS AND ACRONYMS

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
ϕ	phi or rotation, the first Euler angle
θ	theta or tilt, the second Euler angle
ψ	psi or in-plane rotation, the third Euler angle
σ	sigma, the standard deviation of a Gaussian distribution
μm	micrometre (10^{-6} metres)
$^{\circ}$	degree, ($\pi/180$ radians)
\pm	plus or minus
\leq	less than or equal to
\geq	greater than or equal to
\AA	Angstrom, (10^{-10} metres)
AChR	acetyl-choline receptor
AMPPNP	adenylyl-imidodiphosphate
Bsoft	Bernard's software package
CARD	caspase activation and recruitment domain
cc	cross-correlation
CCD	charge-coupled device
C_n	cyclic group of order n (point group symmetry)
COM	centre-of-mass
CPU	central processing unit
cryo-EM	cryo-electron microscopy OR electron cryo-microscopy
cryo-ET	cryo-electron tomography OR electron cryo-tomography

CTF	contrast transfer function
DED	direct electron detector
D _n	dihedral group of order <i>n</i> (point group symmetry)
DNA	deoxyribonucleic acid
DQE	detection quantum efficiency
e ⁻	electron
<i>E. coli</i>	<i>Escherichia coli</i>
EM	electron microscopy OR expectation-maximisation
EMAN	electron micrograph analysis
EMDB	electron microscopy data bank
EMPIAR	electron microscopy public image archive
FOM	figure-of-merit
FOV	field-of-view
FSC	Fourier-shell correlation
FREALIGN	Fourier reconstruction and alignment
FT	Fourier transform
GPU	graphics processing unit
GUI	graphic user interface
H	the number of asymmetrical units in 3D helical reconstructions
HIV	human immunodeficiency virus
ID	identification number
IFT	inverse Fourier transform
IHRSR	iterative helical real-space reconstruction
kV	kilovolt (1,000 V), a unit of electric potential
LMB	(MRC) Laboratory of Molecular Biology
MAVS	mitochondrial antiviral signalling proteins

MAP	maximum <i>a posteriori</i>
MBP	maltose binding protein
ML	maximum likelihood
MPI	message passing interface
MRC	(a file format of cryo-EM density maps developed by) Medical Research Council in the UK
Mx	myxovirus resistance protein
nm	nanometre (10^{-9} metres) also mark in text
NMR	nuclear magnetic resonance
<i>P. aeruginosa</i>	<i>Pseudomonas aeruginosa</i>
PDB	protein data bank
<i>P. dejongei</i>	<i>Prostheobacter dejongei</i>
PhD	Doctor of philosophy
PHF	paired helical (tau) filament
pm	picometer (10^{-12} metres) also mark in text
PSF	point spread function
RELION	regularised likelihood optimisation
RMS	root-mean-squared
SF	straight (tau) filament
SNR	signal-to-noise ratio
SPARX	Single Particle Analysis for Resolution eXtension
SPIDER	system for processing image data from electron microscopy and related fields
SPRING	single particle reconstruction from images of known geometry
STAR	self-defining text archive and retrieval (file format)
T6SS	(bacterial) type VI secretion system
TEM	transmission electron microscope

TMV	tobacco mosaic virus
UCSF	University of California, San Francisco
XIMDISP	X-window-based image display program
XMIPP	X-window-based microscopy image processing package

1 INTRODUCTION

1.1 Structural biology and cryo-electron microscopy

Structural biology seeks the atomic-level structures of biological macromolecules (mainly proteins and nucleic acids) and uses these structures to explain how they are related to their dynamics and functions. One of the most important aspects of structural biology is to visualise how the amino-acid sequences are folded into proteins with complicated conformations through model building (figure 1.1).

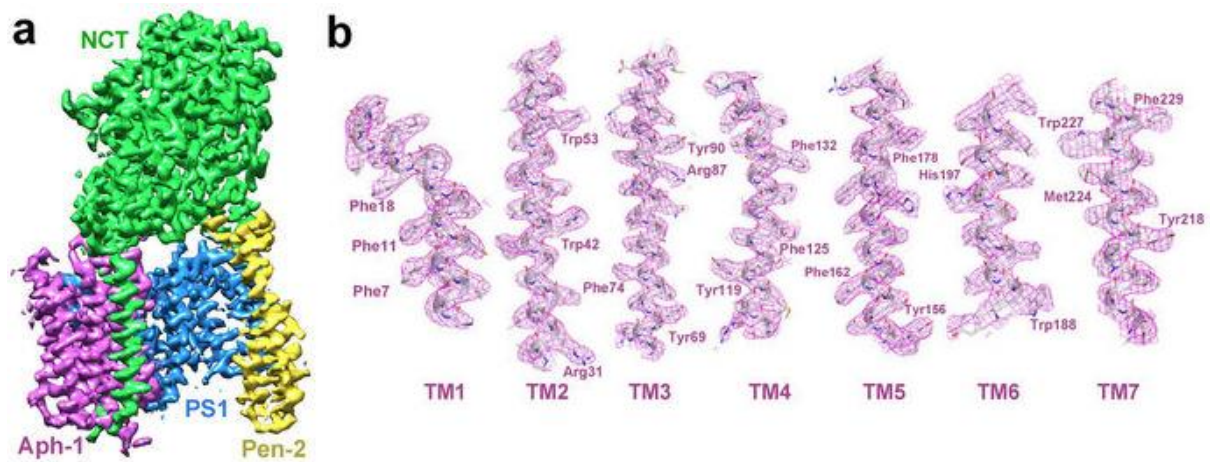


Figure 1.1. Published cryo-EM density map (a) and model building results (b) of some side chains in human γ -secretase (adapted from Bai et al., 2015).

Various techniques have been developed to determine the molecular structures, such as X-ray crystallography, nuclear magnetic resonance (NMR) and mass spectrometry. Particularly, X-ray crystallography has yielded many atomic resolution structures. Recent years have witnessed a growing

number of near-atomic resolution biomolecular structures solved by cryo-electron microscopy (cryo-EM) technique (figure 1.2, Patwardhan, 2017). An advantage of cryo-EM over traditional X-ray crystallography is that the former does not require the procedure of molecule crystallisation required by the latter. In addition, particles frozen in ice layers may adopt multiple conformations, which opens up possibilities in research of molecule dynamics. Cryo-EM is becoming increasingly popular and powerful for structural determination at atomic resolution with the rapid development of high-end microscopes, direct electron detectors, sample/grid preparation process, automated data collection, reconstruction algorithms and computing hardware.

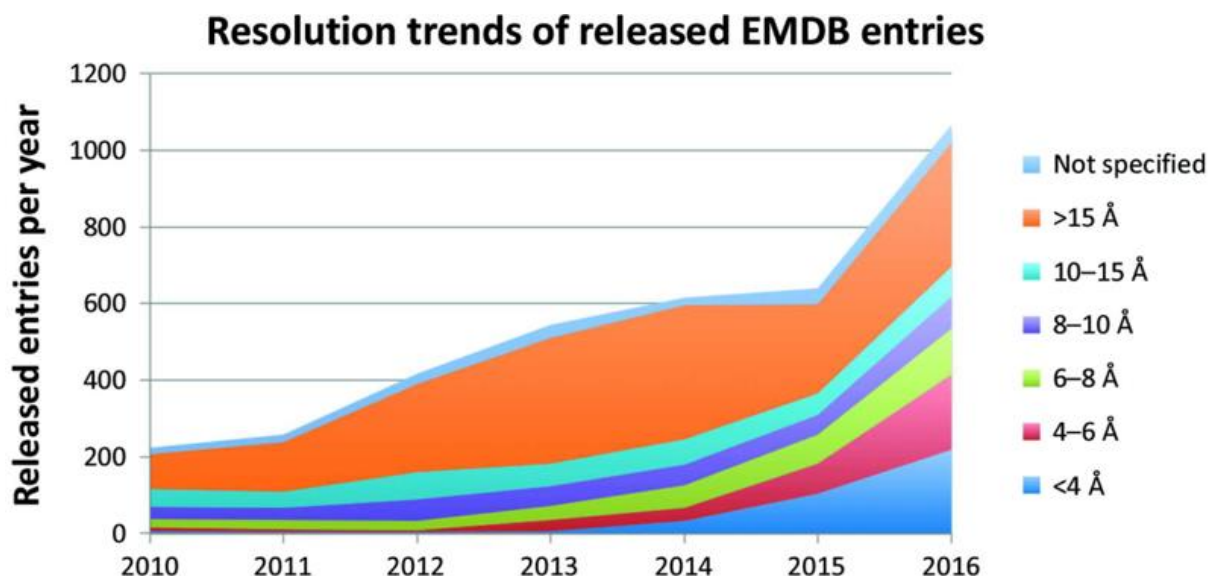


Figure 1.2. Resolution trends of released EMDb entries up to the year 2016. The stacked graph comprised of the number of released entries separated into different resolution bands (adapted from Patwardhan, 2017).

1.2 Transmission electron microscope and image formation

This section includes a brief description of the key elements in a typical transmission electron microscope (TEM) column (figure 1.3), starting from the electron source at the top all the way down to the detector at the bottom.

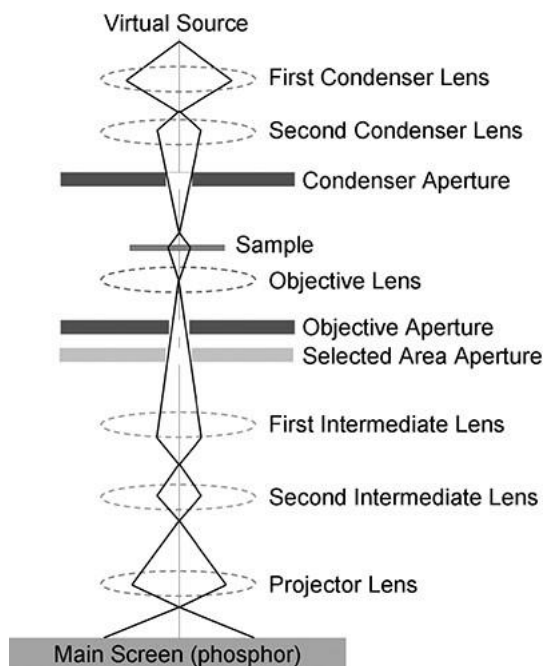


Figure 1.3. Schematic diagram of a transmission electron microscope (TEM) (adapted from Williams and Carter, 1996).

Macromolecules are so small that visible light, with wavelength $\sim 400\text{-}700$ nanometres (nm, 10^{-9} m), are not powerful enough to resolve the structural details of molecules. Instead, TEM uses high energy electrons ($\sim 80\text{-}300$ kV), with wavelengths at about 2-4 picometres (pm, 10^{-12} m). Modern electron microscopes generate electrons using field emission guns, which produce more coherent and monochromatic electron beams compared to thermal emission sources. The condenser lens system, which is located below the electron source, is designed to focus the high-energy coherent electron beam onto the specimen for better illumination.

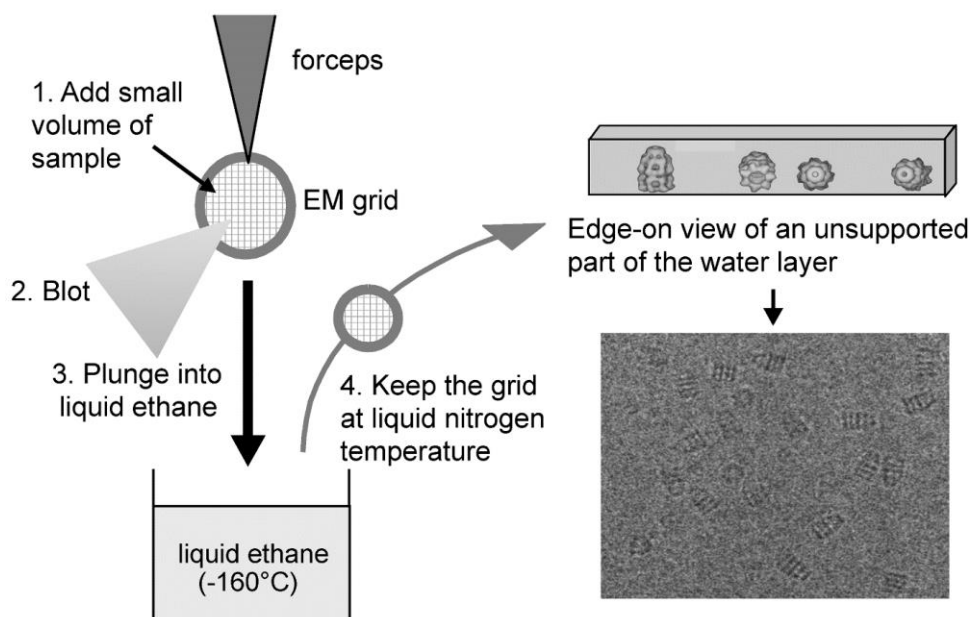


Figure 1.4. Schematic of plunge freezing and of a vitrified layer and an example cryo-EM micrograph (adapted from Orlova and Saibil, 2011).

The thin specimen is placed in vacuum column between the condenser system and objective lens. During specimen preparation, purified protein solutions are applied to grid support and plunge-frozen in liquid ethane (figure 1.4). The freezing process is performed extremely fast at $<-130^{\circ}\text{C}$ temperature to form thin vitreous (amorphous) ice layers instead of crystalline ice because the latter diffracts electrons strongly and degrades the image quality (Dubochet et al., 1988). The vitrified ice layer preserves many biomolecules which only exist under the action of forces which are acting in the liquid water, not in dry specimens (Dubochet et al., 1988). During imaging, electron doses are kept to a minimum ($\sim 20 \text{ e}/\text{\AA}^2$) to limit charging and radiation damage. Since biological specimen are thin and mainly consist of light atoms (H, C, N and O), most of the high-energy electrons ($\sim 80\%$) simply pass through the thin specimen unchanged. Elastically scattered electrons ($\sim 5\%$) interact with the atomic Coulomb potentials of the specimen and contribute to phase contrast in image formation. Inelastically scattered electrons ($\sim 15\%$) deposit energy, damage the biological specimen and produce noise in the collected images (Williams and Carter, 1996). Image contrast in typical cryo-EM, which consists of amplitude and phase contrast, is related to the scattered electrons. The amplitude contrast results from the electrons illuminating the specimen but not eventually arrived at the image detector. The incoming electrons can be diffracted by the atoms of the specimen, and the total scattering of an element of biological matter increases proportionally to the mass increment (Dubochet et al., 1988). The objective aperture allows the passing of unscattered electrons and those scattered within only a small angle. Therefore the excluded high-angle scattered electrons (and those diverted by the energy filter) result in the amplitude contrast. Phase contrast is generated when the remaining electrons are scattered elastically at different angles, experienced with location dependant phase shifts and interfered with the unscattered

beam. Biological specimens are also weak electron scatterers. In order to increase the contrast moderately and record the ‘invisible’ phase contrast as ‘visible’ amplitudes on the output images, cryo-EM imaging makes use of imperfections and special operating conditions of the objective lens, i.e. aberrations and defocus levels (Williams and Carter, 2009).

The electromagnetic objective lens forms the images but it is far from perfect due to various types of aberrations, such as spherical, chromatic aberrations and astigmatism. Spherical aberration means that the focus strength varies from the optical axis to the periphery of the objective lens. Chromatic aberration arises when lower-energy electrons are focused more strongly. Astigmatism is present when the lens is stronger in one direction compared to the perpendicular. Moreover, the objective lens is set underfocused in typical cryo-EM in order to generate sufficient image contrast. The effects of aberrations and defocus on the images formed can be described by the point spread function (PSF) in real-space, or equivalently by the contrast transfer function (CTF) in Fourier-space (Erickson and Klug, 1970). The CTFs describe the imaging properties of the objective lens systems. The spherical aberration gives rise to oscillating CTF signals at various spatial frequencies (figure 1.5) and must be estimated and corrected during data processing.

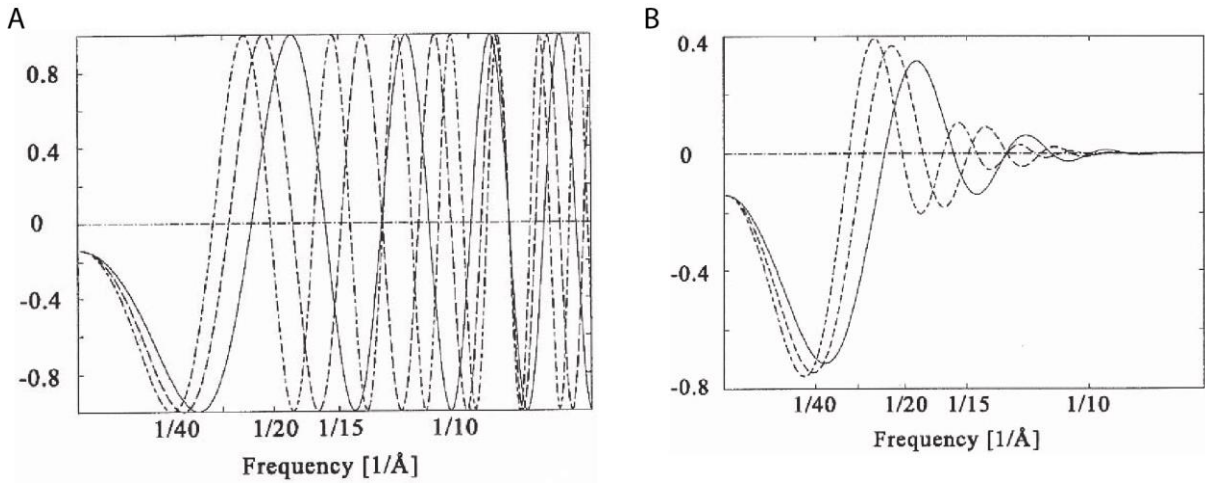


Figure 1.5. Examples of contrast transfer functions (CTF). (A) Theoretical CTF at three defocus values: 1.5 μm (—), 2.0 μm (---) and 2.5 μm (-.-). (B) With envelop functions effect (adapted from Penczek et al., 1997).

According to Rohou and Grigorieff 2015, for a 2D image, the contrast transform function (CTF) can be modelled in Fourier space as a 2D function of the frequency-dependent phase shift χ :

$$CTF(\lambda, \vec{g}, \Delta f, C_s, \Delta\varphi) = -\sqrt{1 - A^2} \sin[\chi(\lambda, \vec{g}, \Delta f, C_s, \Delta\varphi)] - A \cos[\chi(\lambda, \vec{g}, \Delta f, C_s, \Delta\varphi)]$$

A ($0 < A < 1$) is the fraction of amplitude contrast, arising for example from electrons scattered outside the objective aperture or removed by energy filtering. χ is a function of the electron wavelength, the length of the spatial frequency vector \vec{g} , the objective defocus Δf , the spherical aberration C_s , and the additional phase shift introduced by a phase plate $\Delta\varphi$ ($\Delta\varphi = 0$ if absent):

$$\chi(\lambda, \vec{g}, \Delta f, C_s, \Delta\varphi) = \pi\lambda|\vec{g}|^2 \left(\Delta f - \frac{1}{2}\lambda^2|\vec{g}|^2 C_s \right) + \Delta\varphi$$

The astigmatism of objective lens is incorporated in the defocus term Δf :

$$\Delta f = \frac{1}{2}[\Delta f_1 + \Delta f_2 + (\Delta f_1 - \Delta f_2) \cos(2[\alpha_g - \alpha_{ast}])]$$

Where Δf_1 and Δf_2 describe the defocus in perpendicular directions, α_{ast} is the angle between Δf_1 - Δf_2 axes and X-Y axes of the complex plane, and α_g is the angle between \vec{g} and the X axis (figure 1.6). The astigmatism is defined by Δf_1 , Δf_2 and α_{ast} .

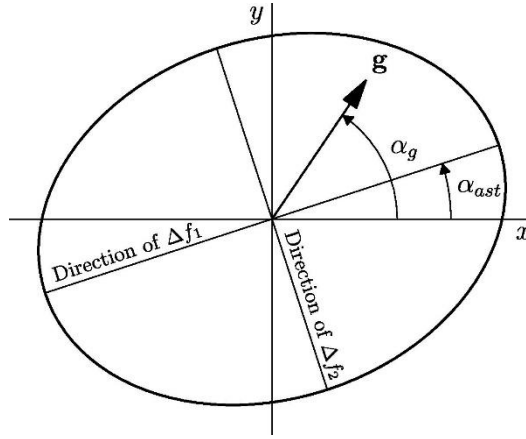


Figure 1.6. Two defocus values, Δf_1 and Δf_2 , and an angle, α_{ast} , define an astigmatic CTF. Δf is the effective defocus at an arbitrary point \vec{g} (scattering vector) in reciprocal space (adapted from Rohou and Grigorieff 2015).

The projector lens magnifies the image formed by the objective lens and projects the magnified results onto a detecting system, such as film, a CCD camera coupled to a phosphor screen or a direct electron detector (DED). The data collected for each field-of-view (FOV), a discretised grey-scale image of a magnified region, is called a micrograph. Until several years ago, photographic film was the most commonly used detector system for cryo-EM. The exposed film had to be taken out of the vacuum of the microscope and scanned for conversion to digital images. CCD detectors were also designed to create electron micrographs, which enabled software to acquire data automatically (Mastronarde, 2003). But the internal conversions (from electron to photon signals, and back again) of CCD-based electron detectors lower the detection quantum efficiency (DQE) (Dobbins et al., 1995), which is the ratio of output and input squared signal-to-noise ratios (SNRs) for all spatial frequencies. Recently, direct electron detectors (DEDs) have enhanced the image quality and contributed to the recent resolution revolution in cryo-EM. In the so-called integrating mode, electrons build up charges inside each pixel during the imaging process of each micrograph (Bichsel, 1988). However, electrons not only partially damage the specimen, but also cause the individual particles to move within their field-of-view (Brilot et al., 2012). Such beam-induced motions can be partially corrected during data processing if fractional

doses of electrons are recorded for every field-of-view during imaging (Scheres, 2014). In counting mode, latest electronics allow DEDs to detect individual electrons, leading to even higher DQEs and better images (McMullan et al., 2009; Bai et al., 2015).

The formed micrographs collected from the microscope contain numerous 2D images of protein particles enclosed in the thin ice layer. The 2D experimental image of a particle can be conceptually considered as an integral of Coulomb scattering potential of the specimen along the direction of the incoming electron beam, or an approximated 2D projection of the 3D particle (Vulović, 2014). Since the signal-to-noise ratios are extremely low due to limited electron doses, a large number ($>10^4$ or 10^5) of 2D experimental images of the same molecules are usually collected for single-particle data analysis.

1.3 Single-particle analysis and electron tomography

Images of individual particles can be extracted from the micrographs. The central slice theorem provides a theoretical foundation for recovering a 3D object according to its 2D projections from different directions (Midgley and Weyland, 2003). It states that the Fourier transform (FT) of a 2D projection corresponds to a slice through the centre of the Fourier transform of the 3D object, and the direction of the projection is orthogonal to this central slice. In cryo-EM single-particle analysis, a large number of 2D projections of individual 3D objects are imaged across different fields-of-view. The 2D projections can be inserted back into the 3D Fourier space correctly if their relative orientations are known. Subsequently, the 3D object can be recovered by the inverse Fourier transform (IFT), provided that sufficient orientations are sampled so that they leave no empty regions in the entire 3D Fourier space (figure 1.7).

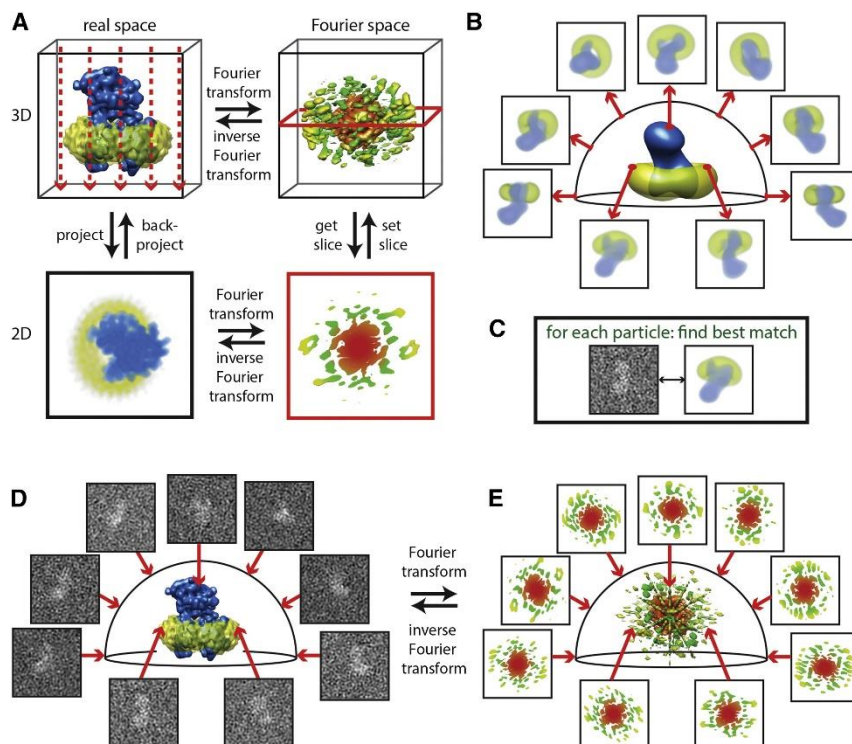


Figure 1.7. Basic concepts of 3D reconstruction in cryo-EM. Central slice theorem is described in (A). Reconstruction algorithm projects the low-passed 3D reference at various orientations (B). Experimental particles are matched with the generated projections (C) and their missing orientations are recovered. According to central slice theorem, 2D experimental images are inserted into 3D Fourier space (D) followed by a Fourier inversion (E) (adapted from Nogales and Scheres, 2015).

Cryo-electron tomography (cryo-ET) is another technique to reveal the structural details of a 3D object (Frank, 2013). Non-repetitive samples of unique 3D objects can be imaged and reconstructed using electron tomography whereas a large number of identical 3D structures must be collected and aligned for single-particle reconstructions. As is shown in figure 1.8, in tomography, a specimen with one protein particle in the centre is mounted onto a tilting stage and a series of images are collected at various tilt angles with respect to the electron beam direction. Tilt angles of the stage are limited to $\pm 70^\circ$ since the specimen becomes considerably thicker along the beam direction at high tilts and therefore blocks electrons for effective image formation. Similar to single-particle analysis, images collected from tilt series are 2D projections of the particle at different orientations so they can be inserted back into the 3D Fourier space according to the central slice theorem. The absence of high-tilt ($> \pm 70^\circ$) images unavoidably leaves missing wedges in Fourier-space reconstructed tomograms. Furthermore, if inside the tomogram there exist multiple copies of the same particles rather than one, they can be picked and extracted as smaller 3D boxes called sub-tomograms. Subsequently, 3D reconstructions of the particle could be performed through sub-tomogram averaging.

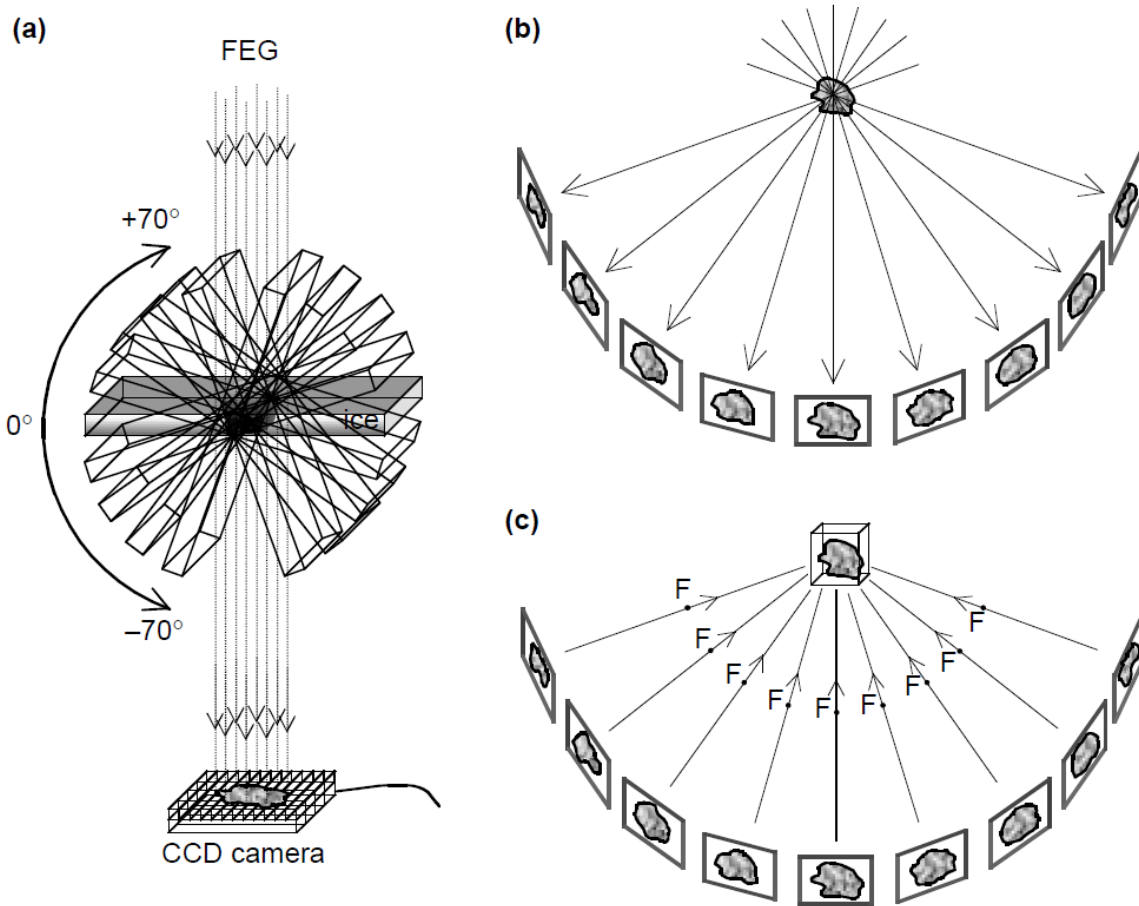


Figure 1.8. Data collection and tomogram reconstruction in cryo-ET. (a) Cryo-holder stage is tilted from -70° to $+70^\circ$ at equal intervals. (b) A series of tilted frames are recorded using a CCD camera. (c) Tomogram reconstruction using back-projection algorithm (adapted from Steven and Aeibi, 2003).

Cryo-ET is able to elucidate sub-cellular structures. It can provide images of structures *in vivo* within cells, which is a great advantage compared to single-particle cryo-EM and X-ray crystallography where samples have to be *in vitro* and in highly ordered crystalline states respectively. Although cryo-ET reconstructions often suffer from uncorrected fluid movements, high radiation damage, and extremely low image contrast in individual tilting frames, recent technical developments demonstrate that it is possible to reach near atomic resolution using cryo-ET and sub-tomogram averaging (Hagen et al., 2017).

1.4 The reconstruction algorithm in RELION

Many packages for cryo-EM single-particle analysis exist, such as SPIDER (Frank et al., 1996), IMAGIC (van Heel et al., 1996), IMOD (Kremer et al., 1996), XMIPP (Sorzano et al., 2004), SPARX (Hohn et al., 2007), EMAN2 (Tang et al., 2007), Bsoft (Heymann and Belnap, 2007), FREALIGN

(Lyumkis et al., 2013), etc. This PhD project is performed within the framework of RELION (Scheres, 2012).

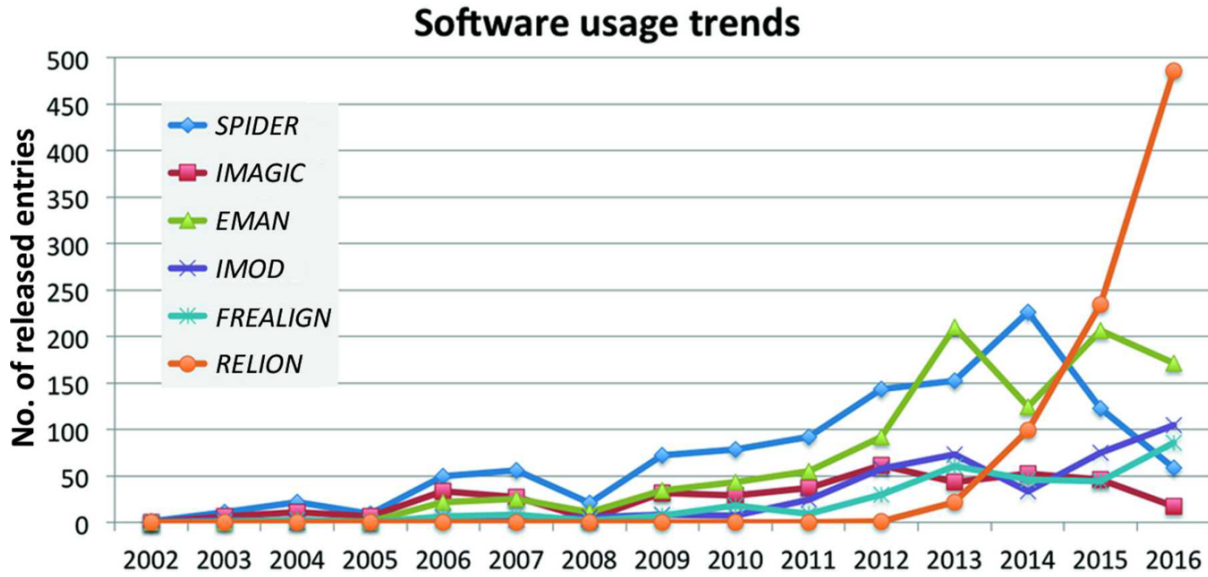


Figure 1.9. Usage trends for a selection of major EM software packages up to the year 2016. More than one package may have been used in the same experiment (adapted from Patwardhan, 2017).

RELION (REgularised Likelihood Optimisation) is a standalone C++ software package that provides a workflow for cryo-EM single-particle analysis and sub-tomogram averaging (Scheres, 2012). It employs an empirical Bayesian statistical framework that infers many model parameters from experimental data (figure 1.10). The framework therefore minimises user intervention in fine tuning *ad hoc* parameters and still delivers high-quality reconstructions with low risks of over-fitting. RELION has contributed to the resolution revolution of cryo-EM in recent years, yielding more high-resolution structures than any other existing package in the field (figure 1.9).

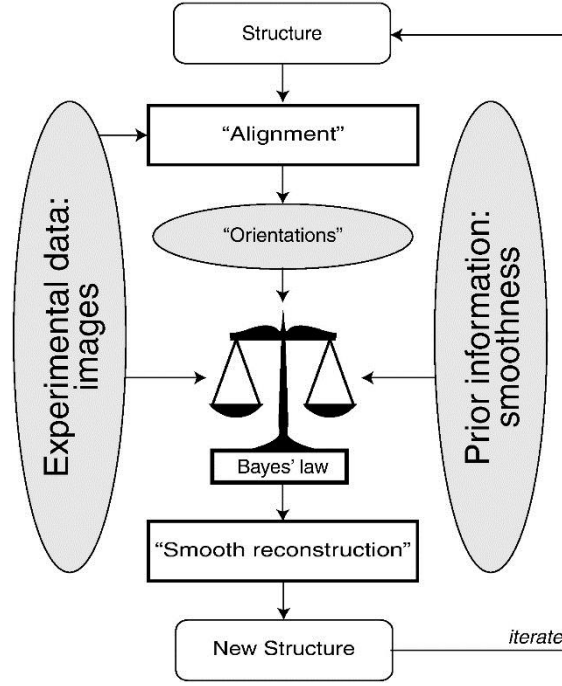


Figure 1.10. A schematic interpretation of RELION approach (adapted from Scheres, 2012).

The image formation model in RELION is based on a weak-phase object approximation: an experimental 2D image is taken as the integral of the electron scattering potentials along the direction of projection. In Fourier-space, it assumes:

$$X_i = CTF_i \cdot P_i^\Phi V + N_i$$

The 3D reference (the underlying 3D signal) and the experimental 2D images have the same box size in real-space and they are transformed into V and X_i s in Fourier-space respectively. P_i^Φ is the operator which extracts the i th 2D noise-free slice $P_i^\Phi V$ according to the relative orientations Φ in Fourier-space. CTF_i is the 2D contrast transfer function of the i th experimental image X_i . Multiplication of CTF with the projection in Fourier space is equivalent to convolution with the imaging point spread function (PSF) in real-space. N_i is a 2D complex plane of Gaussian distributed noise (with each pixel filled with an independent random complex number drawn from a Gaussian distribution). The signal-to-noise levels are usually lower at higher spatial frequencies.

The goal of 3D structure determination can be seen as the inverse process of image formation, to find the best V given a set of experimental images X_i . For this purpose, RELION optimises (maximises) a single target function – the posterior probability $P(\theta|X, Y)$ of finding the model with parameter set θ given observed data X_i s and the prior information Y , according to Bayes' Law:

$$P(\theta|X, Y) \propto P(X|\theta, Y) \cdot P(\theta|Y)$$

Where the likelihood $P(X|\theta, Y)$ denotes the probability of the observed data given the model, prior $P(\theta|Y)$ quantifies the likeliness of the model given the prior. $P(X|\theta, Y)$ is called the maximum a

posteriori (MAP) estimate. The MAP estimator can be seen as the maximum likelihood (ML) estimator regularised by incorporated prior information Y . In RELION, Y expresses the expectation that the 3D signal is relatively smooth in real-space, i.e. the signal has limited power spectra in terms of intensities in Fourier-space. The prior is defined as:

$$P(\theta|Y) = \prod_{l=1}^L \frac{1}{2\pi\tau_l^2} \exp\left(\frac{|V_l|^2}{-2\tau_l^2}\right)$$

where V_l is the l th component of the 3D Fourier transform V of the underlying structure and τ_l^2 denotes the power spectra of signals of V_l . The electron scattering potentials of protein particles are continuous in real-space so the Fourier transform of such structures does not contain large coefficients at high spatial frequencies. Cryo-EM 3D reconstructions need such prior information because structures cannot be completely recovered from the noisy experimental images due to scarcity of information caused by extremely low signal-to-noise ratios at high spatial frequencies. Mathematically, this problem is considered ill-posed, where multiple models explain the experimental data equally well. Intuitively, the prior complements the scarce experimental data and thereby somewhat alleviates the ill-posed problem.

The likelihood (evidence) of the i th experimental image X_i is calculated by comparison with computer-generated projections of reference V given the image's resolution-dependent noise spectra σ_{ij}^2 , CTF information and relative orientations Φ .

$$P(X_i|\phi, \theta, Y) = \prod_{j=1}^J \frac{1}{2\pi\sigma_{ij}^2} \exp\left(\frac{|X_{ij} - CTF_{ij} \cdot P_{ij}^\Phi V|^2}{-2\sigma_{ij}^2}\right)$$

Where X_{ij} is the j th component of the 2D Fourier transform of the image X_i . The relative orientations Φ for all experimental 2D projections are not observed during the imaging process. RELION marginalises over all possible values of Φ to calculate the marginal likelihood of observing the i th image X_i :

$$P(X_i|\theta, Y) = \int_{all \phi'} P(X_i|\phi', \theta, Y) \cdot P(\phi'|\theta, Y) d\phi'$$

RELION incorporates MAP estimation into the expectation-maximisation (EM) algorithm (Dempster et al., 1977). This iterative algorithm performs two steps (E, M) alternately, recovering and making better estimations of missing orientations progressively given noisy images. In each iteration during the E-step, the algorithm calculates the posterior probability $\Gamma_{i\phi}$ of all orientations (Φ) given an experimental images X_i and the current model parameters $\theta^{(n)}$:

$$\Gamma_{i\phi}^{(n)} = \frac{P(X_i|\phi, \theta^{(n)}, Y) \cdot P(\phi|\theta^{(n)}, Y)}{\int_{all \phi'} P(X_i|\phi', \theta^{(n)}, Y) \cdot P(\phi'|\theta^{(n)}, Y) d\phi'}$$

Where n is the iteration number. The experimental images are contaminated with high level of noise due to strictly limited electron dose applied to the specimen to reduce radiation damage. Compared to approaches based on simple cross-correlation, marginalisation improves 3D reconstructions by giving more freedom for the orientations to change, in particular during the early stages of optimisation at low resolution, instead of fixing the orientations at some bad estimations too early and consequently deteriorating the 3D reference. Generally, the posterior distributions of orientations become sharper as reconstructions progress. Moreover, other hidden variables, such as the class assignment K for each image in classification tasks, are also marginalised together with relative orientations Φ .

During the maximisation (M) step, one maximises the lower-bound of the posterior distributions of all the experimental projections X_i s with prior information to produce an updated model for the next iteration. The model Θ includes the 3D Fourier transform V of the underlying structure, and the power spectra of both signal (τ^2) and noise (σ^2) of V (thus τ and σ enable the estimation of signal-to-noise ratios at all Fourier components).

$$V_l^{(n+1)} = \frac{\sum_{i=1}^N \int_{all \ \phi} \Gamma_{i\Phi}^{(n)} \sum_{j=1}^J (P_{ij}^\Phi)^T \frac{CTF_{ij} \cdot X_{ij}}{\sigma_{ij}^2(n)} d\phi}{\sum_{i=1}^N \int_{all \ \phi} \Gamma_{i\Phi}^{(n)} \sum_{j=1}^J (P_{ij}^\Phi)^T \frac{CTF_{ij}^2}{\sigma_{ij}^2(n)} d\phi + \frac{1}{\tau_l^2(n)}}$$

$$\tau_l^2(n+1) = \frac{1}{2} \left| V_l^{(n+1)} \right|^2, \quad \sigma_{ij}^2(n+1) = \frac{1}{2} \int_{all \ \phi} \Gamma_{i\Phi}^{(n)} \left| X_{ij} - CTF_{ij} \cdot P_{ij}^\Phi V_l^{(n)} \right|^2 d\phi$$

The updating scheme for V also behaves like a Wiener-filter (MacKay, 2003, pages 549-551), which corrects for the CTF and produces a low-pass filter that yields the least noisy map. The iterations are repeated until a user-specified number of iterations is reached or changes in both V and the most likely orientations become sufficiently small.

The expectation-maximisation (EM) algorithm has been implemented in multiple types of reconstructions. In 2D/3D classifications, the class tag K of each experimental particle is an additional hidden variable marginalised, together with the missing orientations Φ . Initially in reference-free 2D classifications, images are randomly distributed among a user-defined total number of classes K and initial 2D references for all classes are constructed through averaging from randomly oriented particles. MAP-EM is a local optimiser that can get stuck in local minima if the initial reference $V^{(0)}$ deviates too much from the true underlying structures. In 3D classification, a user-defined initial 3D reference is required for all classes since 3D reconstructions are more prone to convergence problems than 2D classifications. 3D initial models can be built from homologous structures, random conical tilt reconstructions (Radermacher et al., 1987), common line-based approaches (Van Heel, 1987), probabilistic projection matching (Penczek et al., 1994) and stochastic hill climbing (Elmlund et al.,

2013). More recently, stochastic gradient descent has been developed to ameliorate convergence problems with MAP estimation in cryo-EM single particle analysis (Punjani et al., 2017). In 3D auto-refinements, there is always only one class ($K = 1$) but the experimental data set is split randomly into two half-sets and reconstructed independently. Resolutions are estimated via the gold-standard FSC between the two halves (Scheres and Chen, 2012) and the two reconstructions are combined eventually to yield the final results. Moreover, optimal angular and translational sampling rates are estimated automatically across iterations of 3D auto-refinements according to the resolutions and the changes of hidden variables. The EM algorithm has also been adapted for 3D classifications/refinements of 3D sub-tomogram experimental data sets (Bharat et al., 2015).

Apart from the main reconstruction algorithm implemented in RELION for 2D and 3D reconstructions, it also comprises of a number of surrounding data processing tools, such as wrappers to other programs for motion correction and CTF estimation, programs for particle extraction and sorting, movie data processing, post-processing and local resolution estimation, along with a graphic user interface (GUI, figure 1.11) designed for setting up pipelining jobs (Fernandez-Leiro and Scheres, 2017). It thereby provides a general package for data-processing in cryo-EM single-particle analysis and sub-tomogram averaging.

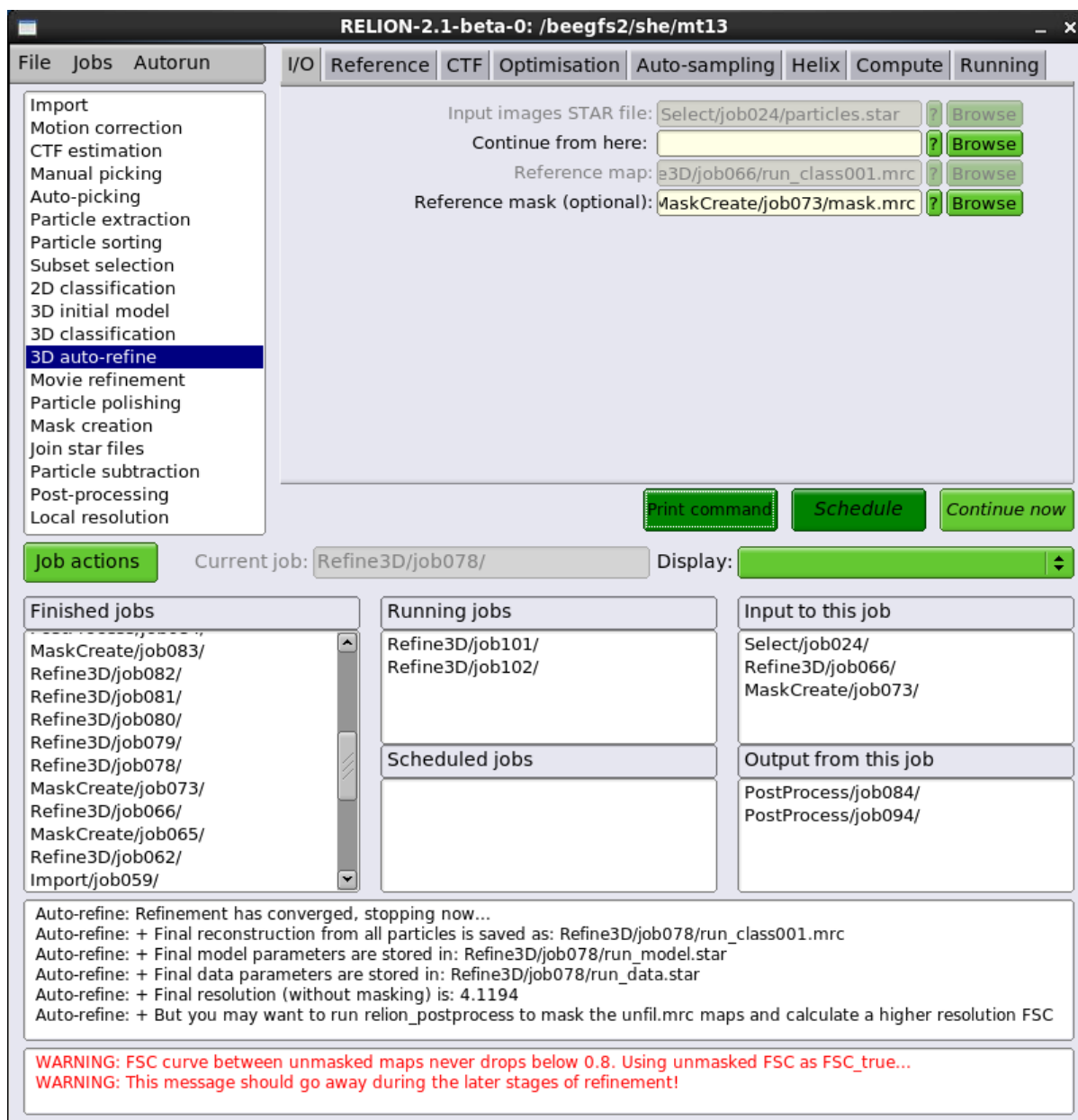


Figure 1.11. Graphic user interface (GUI) of RELION 2.1-beta. The upper half of the GUI consists of a job-type list and parameter tabs. Pipelining settings are shown in the middle. The bottom panels are for standard and error output of a selected job.

1.5 Helical assemblies

Helical assemblies of proteins are widely present and they perform vital functions in all types of organisms. For example, cytoskeleton filaments in archaea and eukaryotes are assembled into helical lattices with globular proteins (figure 1.12A, Holmes, 2009). Examples of extra-cellular helical fibres include flagella and pili, which provide structural foundations for bacterial movements (Diaz et al., 2010). Proteins such as RecA in bacteria (Hegner et al., 1999) or Rad51 in eukaryotes (Short et al.,

2016), which are related to DNA repair using homologous recombination, form into helical arrangements around nucleotide strands (figure 1.12D, Egelman, 2015). In addition, helical assemblies may also be involved in pathologies. For instance, sickle-cell haemoglobin (May and Heuhns, 1976) or amyloid formation (Sunde et al., 1997) occur when proteins undergo aberrant helical polymerization. Many viruses and bacteriophages either have helical coat proteins or helical components, such as nucleocapsids (Egelman, 1989) or tails (De Rosier and Klug, 1968; DeRosier and Moore, 1970). Apart from the naturally occurring helices, some membrane proteins can be induced to form helical lattices (curled 2D crystals) *in vitro* under certain conditions (Pomfret et al., 2007; Yamaguchi et al., 2010), for example, the acetyl-choline receptors (AChR) tubes (figure 1.12B, Unwin, 2017).

The ubiquity of helical assemblies might arise from a general packing rule. If any two molecules can form a specific packing interface which does not involve point group symmetry, a third molecule can be added so that interactions between the 1st and 2nd molecules are the same as the 2nd and 3rd. A helix can then be formed by repeated packing of new subunits along a single direction, i.e. the helical axis (Egelman, 2007). In addition, helices may also consist of multiple types of similarly shaped but different proteins (e.g. 13- and 4-microtubules, Nogales and Zhang, 2016; Deng et al., 2017) and other molecules might decorate the surfaces (both inner and outer) of helical lattices asymmetrically (e.g. figure 1.12C, doublet microtubules, Ichikawa, 2017).

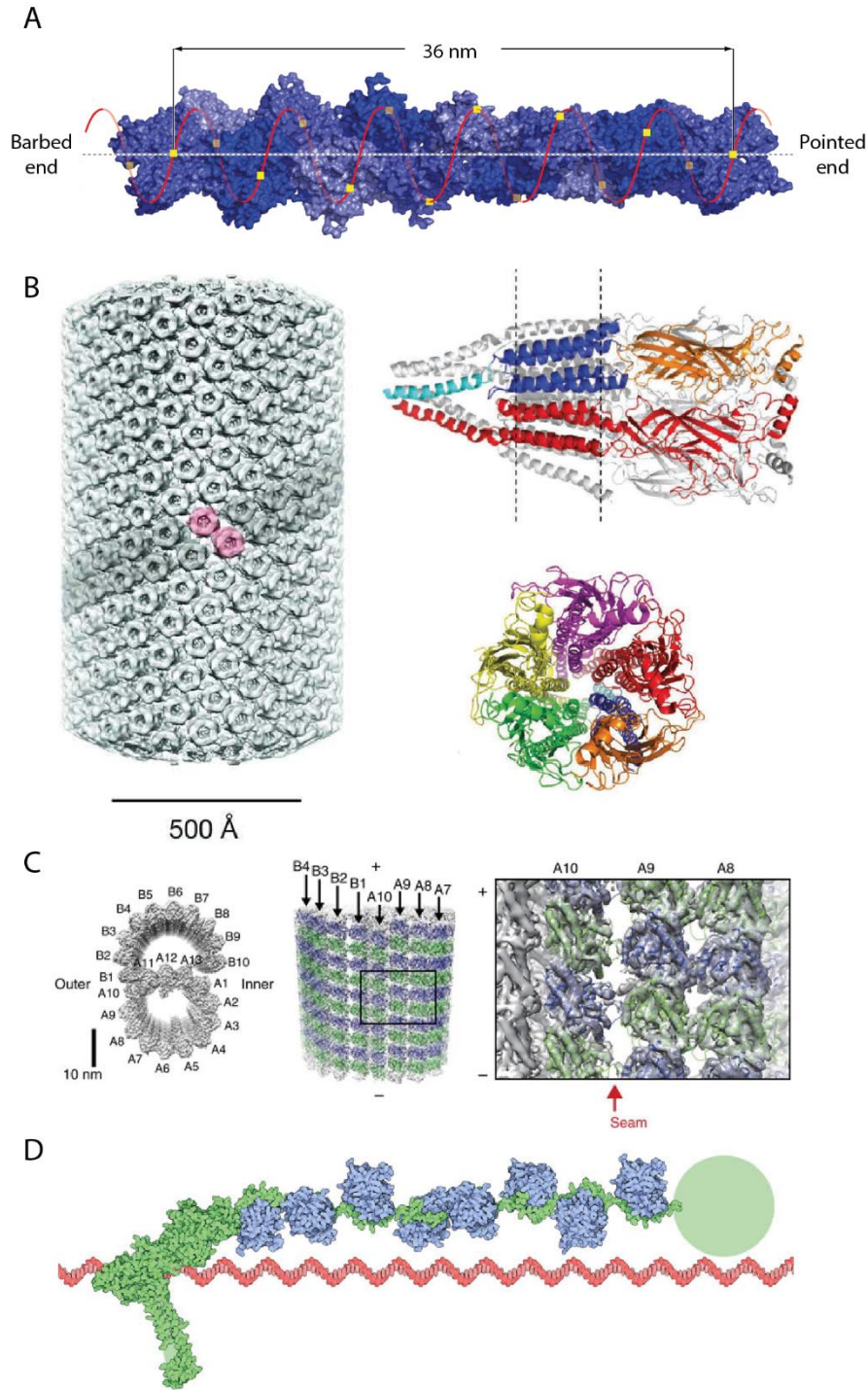


Figure 1.12. Examples of helical assemblies. (A) Cryo-EM density map of F-actin (adapted from Dominguez and Holmes, 2011). (B) Cryo-EM density map of AChR tube (adapted from Unwin, 2017) and PDB model of the constituent AChR subunits (side and top views) (adapted from Akabas, 2011). (C) Cryo-EM density map of doublet microtubule (side and top views, adapted from Ichikawa, 2017). (D) BRCA2 (green) assists with the assembly of the Rad51 (blue) helical filament around damaged DNA (red) (adapted from PDB-101: RecA and Rad51, <http://pdb101.rcsb.org/motm/172>).

The first biological molecular structure elucidated by 3D electron microscopy (3D-EM) was a helical object, the extended tail of T4 bacteriophage (figure 1.13, De Rosier and Klug, 1968; DeRosier and Moore, 1970; Zheng et al., 2017). Helical symmetry makes it an easy target compared to asymmetrical molecules since a single projection of a helical object contains multiple views of the structure with sufficient information for a complete 3D reconstruction. In principle, the total number of parameters required to be determined for a helical reconstruction is significantly reduced compared to analysis of single-particles with no or only simple symmetry. For example, the orientations needed to be recovered for single-particle images (projections) are all independent, whereas repeating asymmetrical subunits in a helical assembly have relative orientations fixed by helical symmetry. Thus, given the correct helical symmetry, the noise levels of input images can be brought down significantly by averaging a huge number of asymmetrical subunits during helical reconstructions. In addition, helical reconstructions rarely suffer from preferred orientation problem since the slowly twisting assemblies typically lie flat on the specimen support and hence represent a 360° range of views (except for the cases of pseudo-helices where helical twists = 0°). Currently, helical assemblies make up $\sim 7\%$ entries of all the determined structures in the Electron Microscopy Database (EMDB, Patwardhan, 2017).

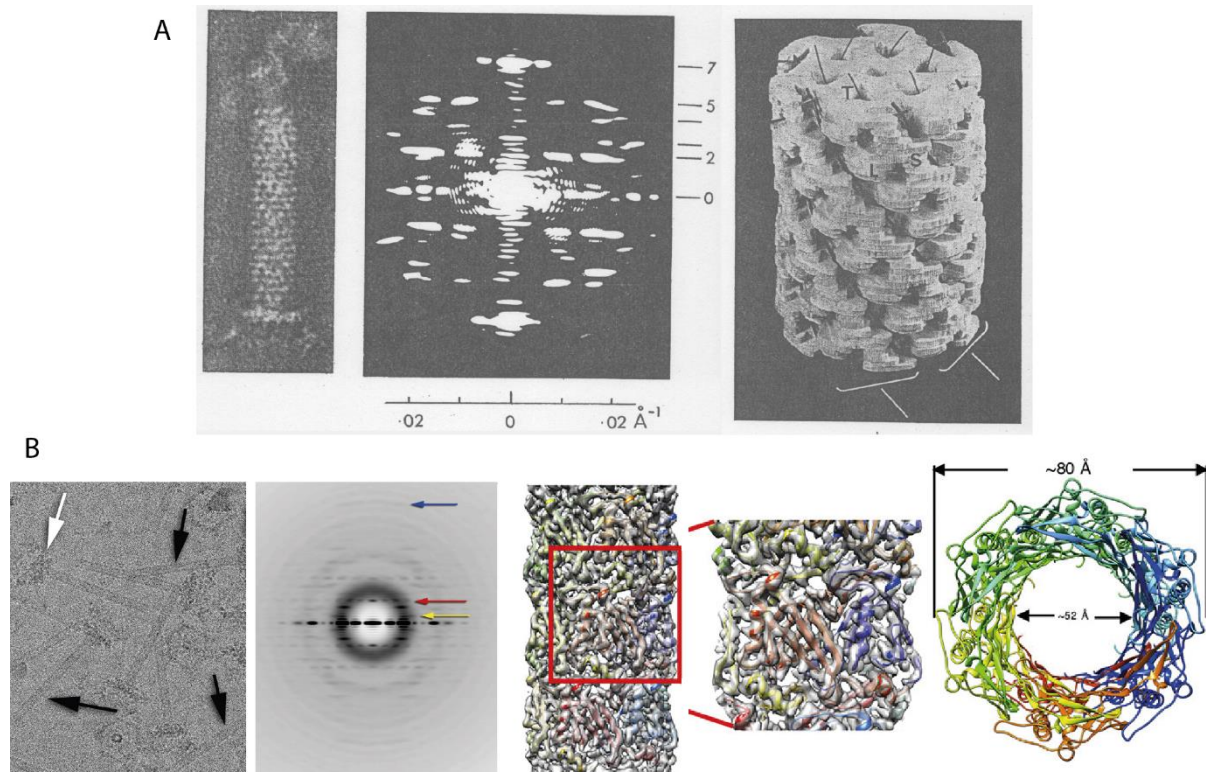


Figure 1.13. The cryo-EM structure of the extended tail of T4 bacteriophage. (A) 3D reconstruction of the extended tail of T4 bacteriophage at a resolution of ~ 35 Å by De Rosier and Klug, 1968. (From left to right) electron micrograph of bacteriophage T4 tail (negative stained with uranyl formate, 50,000 \times); optical diffraction pattern; side-view of 3D reconstructed model (tilted forward). (B) 3D reconstruction of the same target at 3.4 Å resolution by Zheng et al., 2017. (From left to right) a representative motion-corrected micrograph; averaged

diffraction pattern from 26,320 overlapping tube segments; side-views of the 3D reconstructed map; top-view of the atomic model.

1.6 Helical symmetry

There are many ways to describe the symmetry of helical assemblies. A schematic plot of the simplest 1-start (strand) helix is given in figure 1.14B. Helical twist and rise are the most widely used terms. Helical twist ($\Delta\phi$) is the azimuthal angle between neighbouring subunits. By convention, values are in degrees and plus and minus signs denotes right- and left-handedness respectively. Helical rise (Δz) is the translation along helical axis between two neighbouring subunits. The values of rise are in Angstroms and always positive.

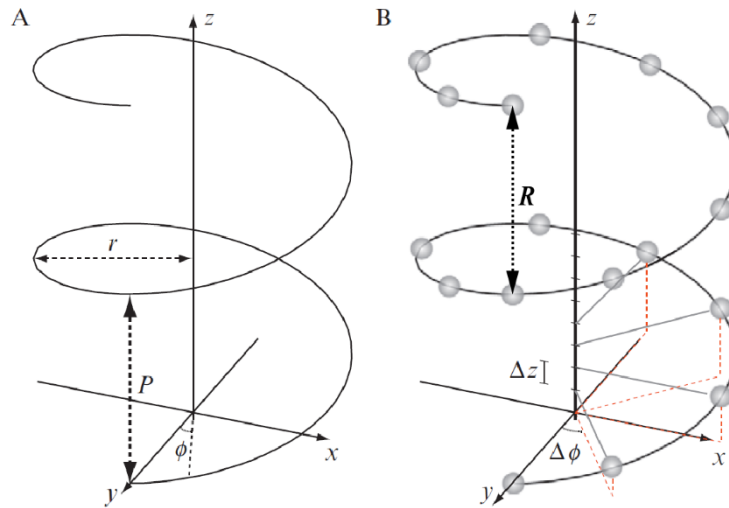


Figure 1.14. Helical parameters for a simple 1-start right-handed helix. (A) shows a continuous helical curve defined by radius (r) and pitch (P). (B) shows a 1-start helical assembly with twist $\Delta\phi = +45^\circ$ and rise Δz . Repeat distance of (B) is the same as its pitch since there are exactly $360^\circ / 45^\circ = 8$ subunits in a pitch. Generally, repeat (R) is $N \times \text{pitch}$ for a 1-start helix, where N is the smallest positive integer that makes $N \times 360^\circ / \text{twist}$ an integer (adapted from Diaz et al., 2010).

The formation of helices can be considered in at least two ways. Simple 1-start helices are formed by rotations and translations of one subunit onto all positions determined by helical twist and rise. Moreover, helices can be seen as curled up 2D lattices according to pairs of base vectors. Figure 1.15 shows such schematic plots of more complicated, n -start helices.

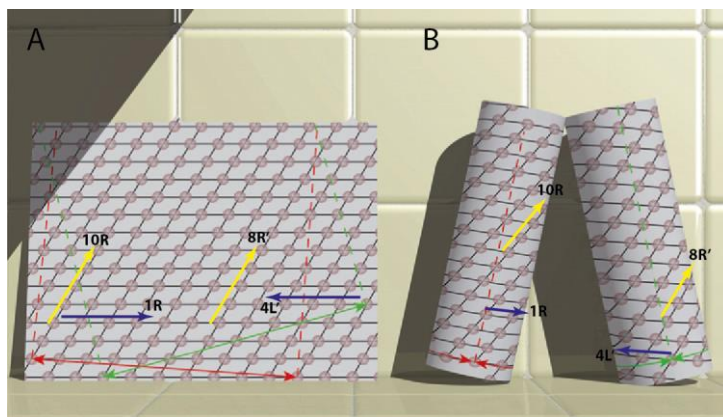


Figure 1.15. The formation of helices from a 2D surface lattice. Two pairs (red and green) of base vectors (solid and dashed) are shown in (A). Helices with different symmetries are formed in (B) although they come from the same surface lattice (A). Start number can be arbitrary for a complicated helical lattice like this. For example, right-handed 1- (blue) and 10-start (yellow) are prominent in helix on the left in (B). While left-handed 4-start (blue) and right-handed 8-start (yellow) seem prominent in helix on the right in (B) (adapted from Diaz et al., 2010).

Helical twist and rise are the same across all strands in an n -start helix. The simple 1-start helix shown in figure 1.14B consists of one strand without ambiguity. However, more complicated helical lattices in figure 1.15 may have strands (start numbers) defined arbitrarily and each start number leads to a set of different helical twist and rise. To avoid confusion, an approach is adopted to convert any n -start C_1 helix into an equivalent 1-start symmetry representation in section 2.2.1.

Two other terms, repeat and pitch, are also useful in explaining helical symmetry. Helical repeat (R in figure 1.14B) is the smallest distance from one subunit to the next which is exactly in register with the first one along the helical axis. According to this definition, an arbitrarily long helix can be seen as helical repeats piling up along the helical axis. For complicated helical lattices, helical repeat may span thousands of Angstroms. Helical pitch (P in figure 1.14A) is the smallest distance from one point to the next in register along the helical axis on its own helical strand. Pitch is defined on a continuous helical curve thus it is irrelevant to the actual positions of constituent protein subunits. Given the helical twist and rise on this strand, the pitch must be $\text{rise} \times 360^\circ / \text{twist}$. Consequently, helical repeat distance must be an integer multiple of the helical pitch.

1.7 Fourier-Bessel method for 3D helical reconstruction

The field of 3D reconstructions of helical assemblies has long been dominated by the Fourier-Bessel method, where the analysis of helical symmetry and calculation of helical density maps are both performed in Fourier-space. The initial mathematical analysis of the Fourier transform of helical objects was conducted by Cochran et al., 1952. Klug et al., 1958 further researched into the theory of diffraction

patterns of helical structures. The formation of helical diffraction patterns (figure 1.16f) can be generally explained as a convolution of two discrete functions (figure 1.16de) in Fourier-space.

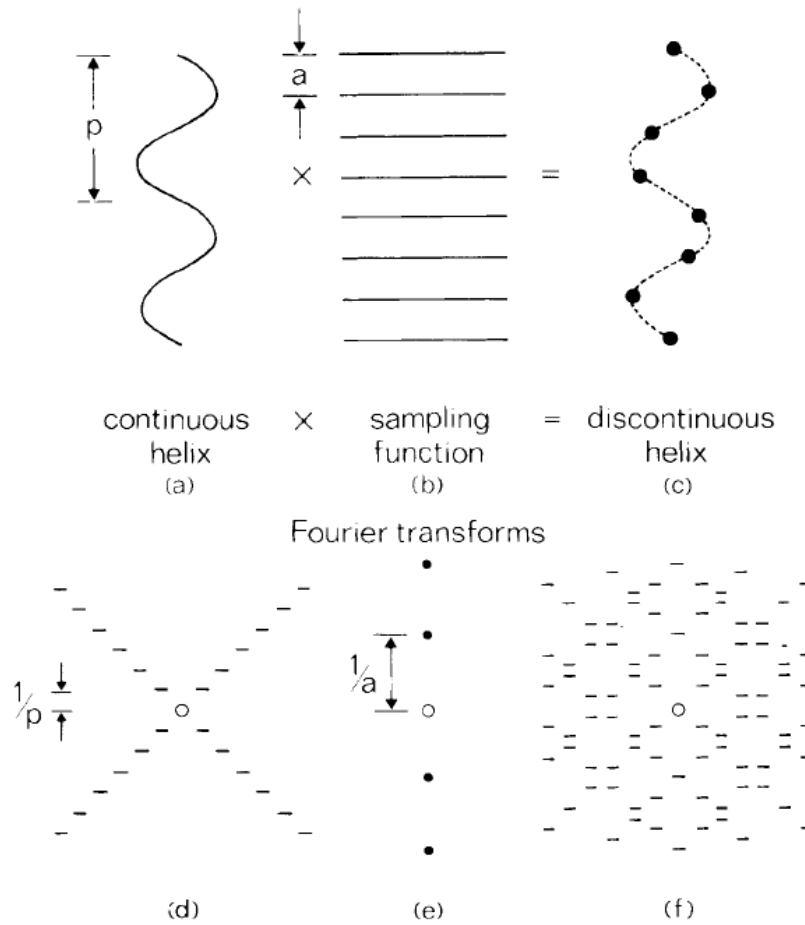


Figure 1.16. The schematic diagram of the formation of helical diffraction patterns. In real-space, a helical object consisting of protein subunits (c) can be conceptually considered as continuous helical densities (a) multiplied by an equally sampled function (b). Therefore, helical diffraction pattern (f) arises from a convolution of (d) and (e) which are the Fourier transforms of (a) and (b) respectively (adapted from Stewart, 1988).

The Fourier-Bessel method starts from the inspections of 2D Fourier-space diffraction patterns of helical filaments. Long filaments are often rotated, straightened and averaged together during data pre-processing in order to generate more discernible diffraction signals. As the formation of a helix can be conceptually understood as a curled up 2D surface lattice (figure 1.15), the effect of curling results in a convolution of a 2D surface lattice with the Bessel functions, a series of cylindrical harmonics, in Fourier space (the lower half of figure 1.16). Fourier transform of a 3D helical object shows radially oscillating amplitudes on some of the Z-axial planes. Thereby, according to the central slice theorem, these ring-like Fourier coefficients in 3D lead to symmetrical maxima across the meridian of every layer line in the Fourier transformations of 2D projections. First-order maxima on all discernible layer lines (except for the equator) are marked and separated into two symmetrical sets of points, corresponding to

the near and far side of the curled helical lattice in real-space (figure 1.17). Fourier-Bessel indexing assigns integers to first-order maxima according to their distance to the meridian, the layer lines they belong to and the actual width of the helical object in real-space (figure 1.18). Helical symmetry can be subsequently deduced, and a full 3D reconstruction of the object can be calculated by Fourier synthesis given the correct indexing scheme. The general procedure of Fourier-Bessel indexing and deductions of helical symmetry will be discussed in details in section 2.4.3.

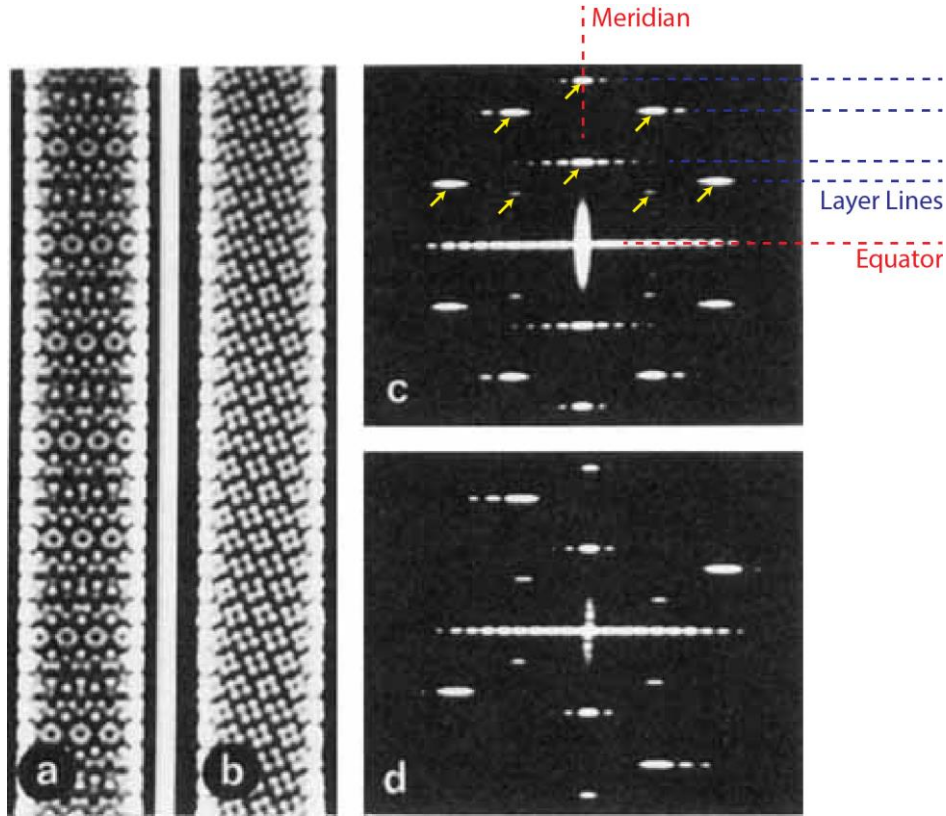


Figure 1.17. The helical diffraction pattern can be separated into two lattices, which corresponds to near and far sides of the helix in real-space. (a) shows a 2D projection of helical filament. (b) is the projection of only the near side. (c) and (d) are the Fourier transform amplitudes of (a) and (b) respectively. (c) can be decomposed into near and far side lattices - (d) and reflected (d) across the meridian. Points of first-order maximum on each layer line are marked with yellow arrows in the upper half of (c) (adapted from Stewart, 1988).

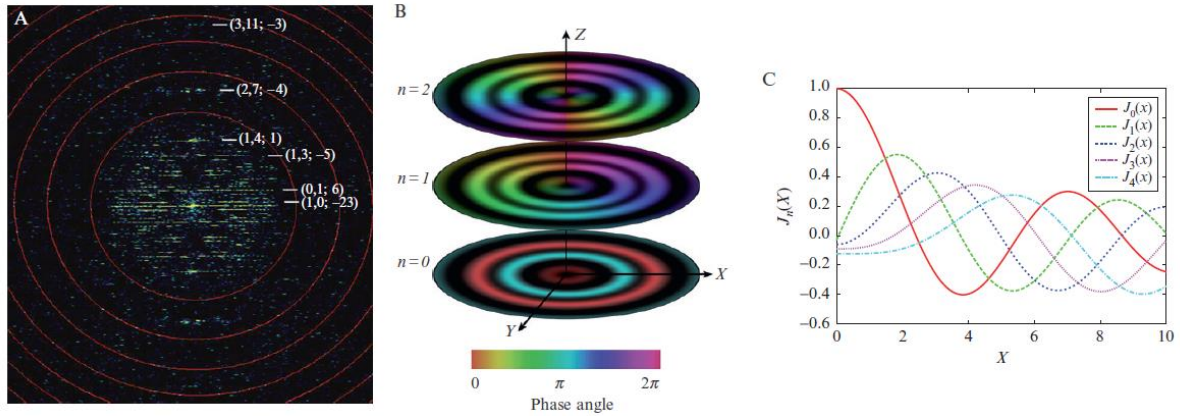


Figure 1.18. Bessel functions and Fourier transform of helical assemblies. (b) shows part of the Fourier transform phase information of a 3D helix on layer planes $n = 0, 1, 2$, where radially oscillating phases arising from Bessel functions are shown in rainbow colours. Diffraction pattern (a) can be seen as a 2D central slice from (b). (c) shows some of the amplitudes of Bessel functions of different orders (0-4), where the X positions of first-order maxima stretch out with increasing orders (adapted from Diaz et al., 2010).

The MRC image processing package developed at LMB contains several routines for Fourier-Bessel analysis (Crowther et al., 1996). The routines have also been adapted by other packages such as Brandeis Helical Package (Owen et al., 1996), Phoelix (Whittaker et al., 1995) and Ruby-Helix (Metlagel et al., 2007).

Fourier-Bessel analysis requires user expertise and experience, as the procedure of indexing is prone to errors, especially when the filaments are distorted by bending, tilting, skewing, flattening and other effects of short-range disorders. A mixture of slightly changing symmetry parameters along the filaments may further blur the diffraction patterns. Wrong indexing schemes lead to incorrect 3D reconstructions which might be hard to distinguish from the true structures at low resolutions. In addition, some helices do not show any separable layer lines in diffraction patterns, perhaps due to image quality, intrinsic properties of the surface lattice or extremely long repeat distances, making the Fourier-Bessel method infeasible to reveal their structures. Consequently, the Fourier-Bessel method yields the best results when the examined filaments have close-to-perfect helical symmetry. Despite of the drawbacks, Fourier-Bessel analysis is still considered as an important method for the initial estimation of helical symmetry prior to single-particle helical reconstruction.

1.8 Iterative helical real-space reconstruction algorithm

Beroukhi and Unwin, 1997 firstly introduced the idea of cutting helical filaments into short segments for alignments against a 3D reference during Fourier-Bessel reconstructions of AChR filaments. An alternative to Fourier-Bessel analysis, iterative helical real-space reconstruction (IHRSR) algorithm (Egelman, 2000), also adopts this idea of segmentation. In this robust approach, overlapping square

boxes are extracted at equal intervals along all helical filaments on the micrographs so that the same iterative algorithms in single-particle analysis can also be adapted to 3D helical reconstructions.

As is shown in figure 1.19, each iteration starts with projection matching, where side-view projections of an initial 3D reference are compared with overlapping 2D segments extracted along helical filaments from experimental micrographs by cross-correlations. After the optimal in-plane rotations, azimuthal angles and x, y shifts are found according to the highest cross-correlation, the aligned segments are back-projected into an asymmetric 3D volume. This algorithm is closely related to the EM reconstruction algorithm described in section 1.4, but without any statistical noise model or marginalisation of missing variables. The helical twist and rise are estimated locally by least-square fitting given their initial values, the 3D volume densities, user-defined search ranges and steps. The refined helical symmetry is subsequently imposed in real-space in order to generate a helically symmetric map as the reference for the next cycle. For favourable cases, correct structures can be reconstructed iteratively starting from featureless cylinders as initial references using this algorithm. This approach avoids some of the aforementioned difficulties in Fourier-Bessel analysis, reduces human intervention and allows the reconstructions of more flexible and bending helices by reducing the effects introduced by long-range disorders. This approach was first incorporated into the SPIDER package (Frank et al., 1996).

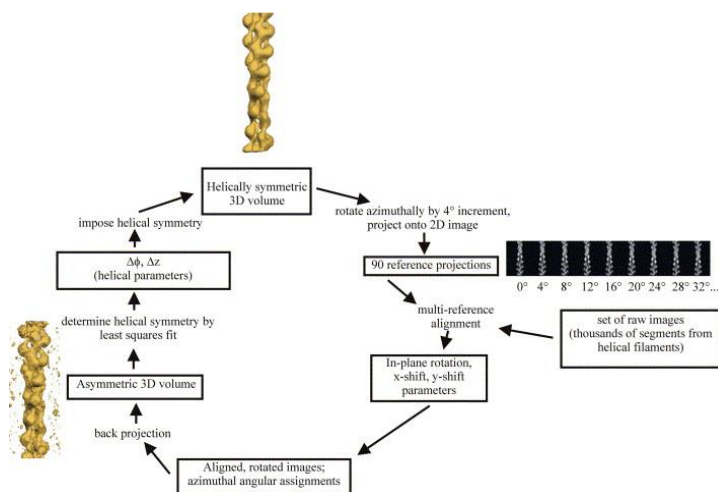


Figure 1.19. An iteration of the original IHRSR algorithm (adapted from Egelman, 2007).

More recent implementations of single-particle based approaches to helical processing also include FREALIGN (Lyumkis et al., 2013), SPARX (Behrmann et al., 2012), FREALIX (Rohou and Grigorieff, 2014) and SPRING (Desfosses et al., 2014). Single-particle based approaches have recently yielded multiple filament structures at near-atomic resolution (Lu et al., 2014; von der Ecken et al., 2015; Ge et al., 2015; Zhang et al., 2015).

2 IMPLEMENTED ALGORITHMS IN SINGLE- PARTICLE HELICAL RECONSTRUCTIONS

2.1 Workflow of helical processing

The general workflow of helical processing in RELION (figure 2.1) is adapted from that of the ordinary single-particle analysis. All the implemented methods are discussed in more details in the following sections of this chapter.

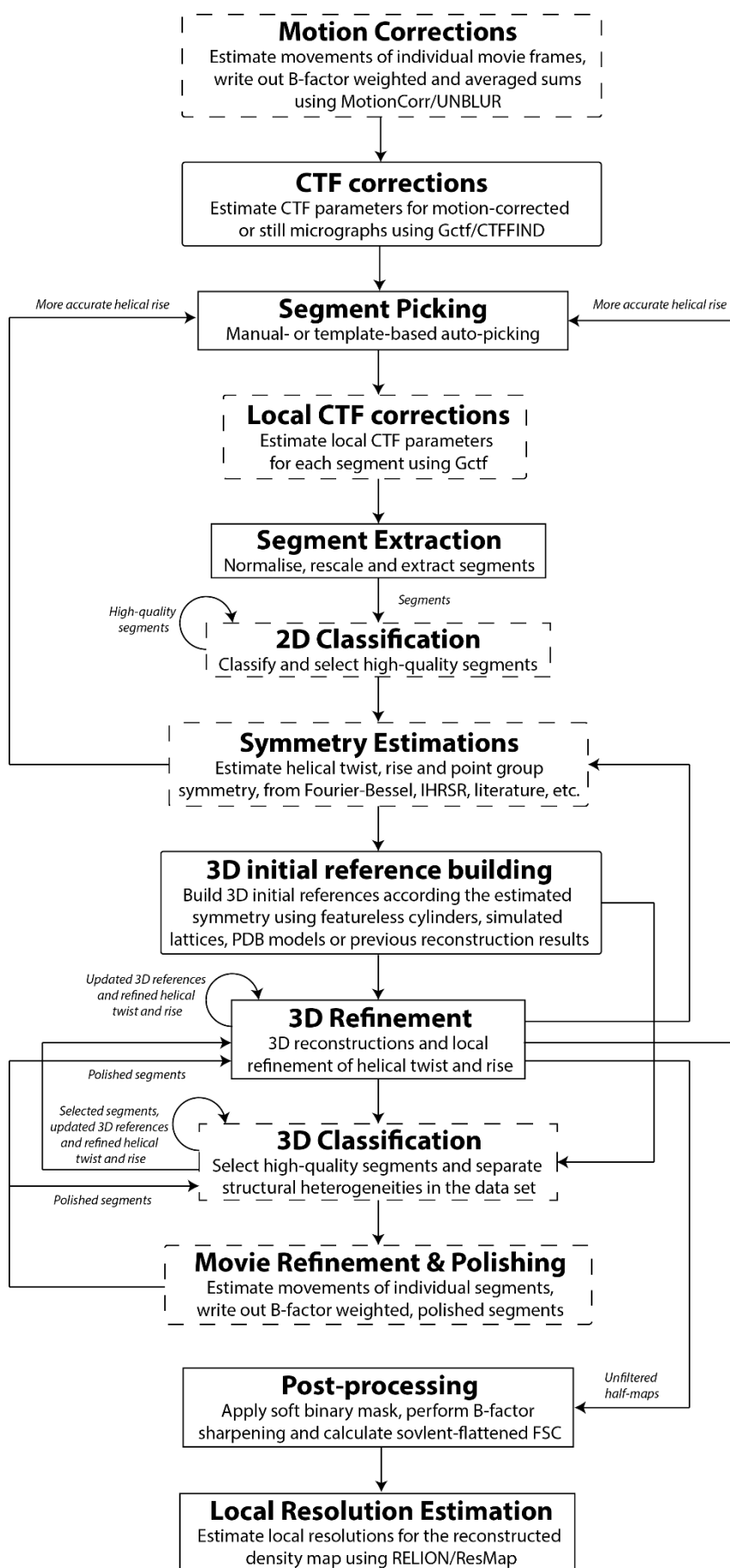


Figure 2.1. General helical processing workflow in RELION. Processing steps denoted with dashed boxes might not be necessary under certain circumstances. Arrows denote some possible sequences of steps for various types of data sets.

Usually, the first step of helical processing is to figure out the helical symmetry using whatever prior information, knowledge or measures possible. The workflow can still proceed starting with a featureless cylinder if the guess of symmetry cannot be made at all at the beginning (the method of which will be described in section 2.2.9). But the user is recommended to repeat part of the workflow once the symmetry and related settings are settled.

Much like in single-particle analysis, movie frames recorded on microscopes equipped with the latest detectors (DEDs) need to be corrected for beam-induced motions, realigned and averaged to produce still micrographs using RELION wrappers to UNBLUR (Grant and Grigorieff, 2015) or MOTIONCOR/2 (Li et al., 2013; Zheng et al., 2017). Averaged or directly collected still micrographs are then subject to CTF estimations with GCTF (Zhang, 2016) or CTFFIND3/4 (Rohou and Grigorieff, 2015). Based on the conventional CTF correction, Russo and Henderson (2018) has recently improved a single-sideband CTF correction algorithm to deal with the approximately Friedel-related diffraction patterns in underfocused images (also refer to section 6 and figure 6.1).

Then, coordinates are picked along helical filaments. In RELION, segment picking can be done either manually by specifying start and end points of filaments or semi-automatically using pre-made 2D class averages as templates. It is also possible to import filament/segment coordinates from other packages, such as e2HelixBoxer (Tang et al., 2007) and XIMDISP (Smith, 1999). Angular priors are determined from the orientations of the filaments specified on the micrographs and stored in the STAR format files (Hall, 1991) that also contain the picked coordinates.

Subsequently, normalised helical segments are extracted at all picked coordinates using a user-defined box size and inter-box distance. The latter is a user-defined multiple (i.e. the number of asymmetrical units, H , a positive integer) of the estimated helical rise.

2D classifications are often performed to separate false positives, e.g. bending segments or other abnormalities, from the high-quality segments in the picking results. It might also separate segments with different widths or diffraction patterns. Multiple rounds of 2D classifications are recommended if the picked data set is highly heterogeneous.

An initial 3D refinement should be set up with the cleaned data set. For some cases, high resolution structures can already be reconstructed at the first attempts. For others, the user might find out additional point group symmetry or the presence of hollow centres in the structures. New knowledge learnt can benefit further analysis with 3D classifications, which have the potential to separate and reconstruct helices of different properties.

If the input data are movies, movie refinement (Bai et al., 2013) and particle polishing (Scheres, 2014) enable tracking of the movement of individual helical segments on a series of movie frames. The above steps also estimate the radiation damage during the imaging of the movie stacks and write out B-factor sharpened segments as motion-corrected, weighted averages. These polished segments can be used as input for further 3D classification or refinement jobs.

Given the final results of refinement and a soft-edged binary mask as input, post-processing performs B-factor sharpening (Rosenthal and Henderson, 2003) and estimates the true overall resolution of the structure where noisy solvent parts are taken into account by Fourier-shell correlations between phase randomised and masked half-maps (Chen et al., 2013).

In the following sections, the most important implemented methods are described in details.

2.2 Implemented methods

2.2.1 Definition of helical symmetry

The helical axis is always the z axis of 3D density maps in RELION. To provide unique and unambiguous definitions for helices in general, the implementation in RELION makes use of two parameters, helical twist ($\Delta\phi$) and rise (Δz). Therefore, for every position (r, ϕ, z) in 3D real-space under cylindrical coordinate system, with the 3D volume V , radial distance r , azimuth angle ϕ , height z :

$$V(r, \phi, z) = V(r, \phi + k\Delta\phi, z + k\Delta z), -180^\circ < \Delta\phi < +180^\circ, \Delta z > 0, k \in \mathbb{N}$$

The above relationship holds when the object is perfectly helical and continuous (i.e. interpolation errors introduced by voxel samplings are ignored).

In addition to helical symmetry, point group symmetry may also exist in some helices. C_n -symmetry ($n \geq 2$) means the helix is n -fold rotationally symmetrical around the helical axis. C_n helices consist of n parallel strands, the subunits of which always appear at the same height z along the helical axis. D_n -symmetry ($n \geq 1$) is present when all the constituent subunits in a helix have dyad axes, which means the subunits are centrosymmetric (for example, the AChR subunits shown in figure 1.12B). This gives rise to an additional 2-fold rotational axis perpendicular to the helical axis for the structure as a whole. Thus, D_n helices lack polarities.

Due to distinct molecular shape and packing of subunits, some helices may be more conveniently interpreted as n -start structures even if C_n -symmetry is absent. In such cases, the n strands are both rotated and translated along the helical axis when assembled into helices. That is to say, for a given subunit, there exists no others located at the same height z along the z axis. RELION uses unique 1-start representations of such structures for consistency (figure 2.2). The conversion from n -start to 1-

start helical symmetry is possible since the differences in z-axial translations across all the n strands must be multiples of $\text{rise}_{n_start} / n$ in order to keep the helical lattice complete and intact.

$$\Delta z_1 = \frac{\Delta z_n}{n}, \quad \Delta \phi_1 = \begin{cases} \frac{\Delta \phi_n + 360^\circ}{n}, & \Delta \phi_n \geq 0 \\ \frac{\Delta \phi_n - 360^\circ}{n}, & \Delta \phi_n < 0 \end{cases}$$

Subscripts 1 and n denote helical parameters for transformed 1-start and original n -start representations respectively. If the converted $\Delta \phi_1$ is out of the range of valid twist values, i.e. $(-180^\circ, +180^\circ)$, multiples of 360° should be added or reduced. This conversion is mathematically correct, compatible with helices of various types and makes the settings in 3D helical reconstructions much easier. However, the new 1-start representation might change the handedness of the helical lattice. And the new representation does not necessarily have intuitive biological significance since neighbouring molecules on the strand might not have contacts to each other.

For example, there are two identical strands in the paired helical filaments (PHF) of tau (figure 5.4). In every strand, C-shaped subunits closely interact and are assembled according to the helical twist and rise of $\Delta \phi_2 = -1.2^\circ$ and $\Delta z_2 = 4.72 \text{ \AA}$ respectively. To transform such 2-start symmetry into 1-start representation, the new rise must be $\Delta z_1 = 4.72 \text{ \AA} / 2 = 2.36 \text{ \AA}$, and the new twist is calculated as $\Delta \phi_1 = (-1.2^\circ - 360^\circ) / 2 = -180.6^\circ$ since each of the single strands is left-handed ($\Delta \phi_2 < 0^\circ$). Apparently, a twist of -180.6° is equivalent to $+179.4^\circ$ (because $-180.6^\circ + 360^\circ = +179.4^\circ$). The helical symmetry $(+179.4^\circ, 2.36 \text{ \AA})$ is used in the final 3D reconstruction of PHF tau (table 5.1 and section 5.4).

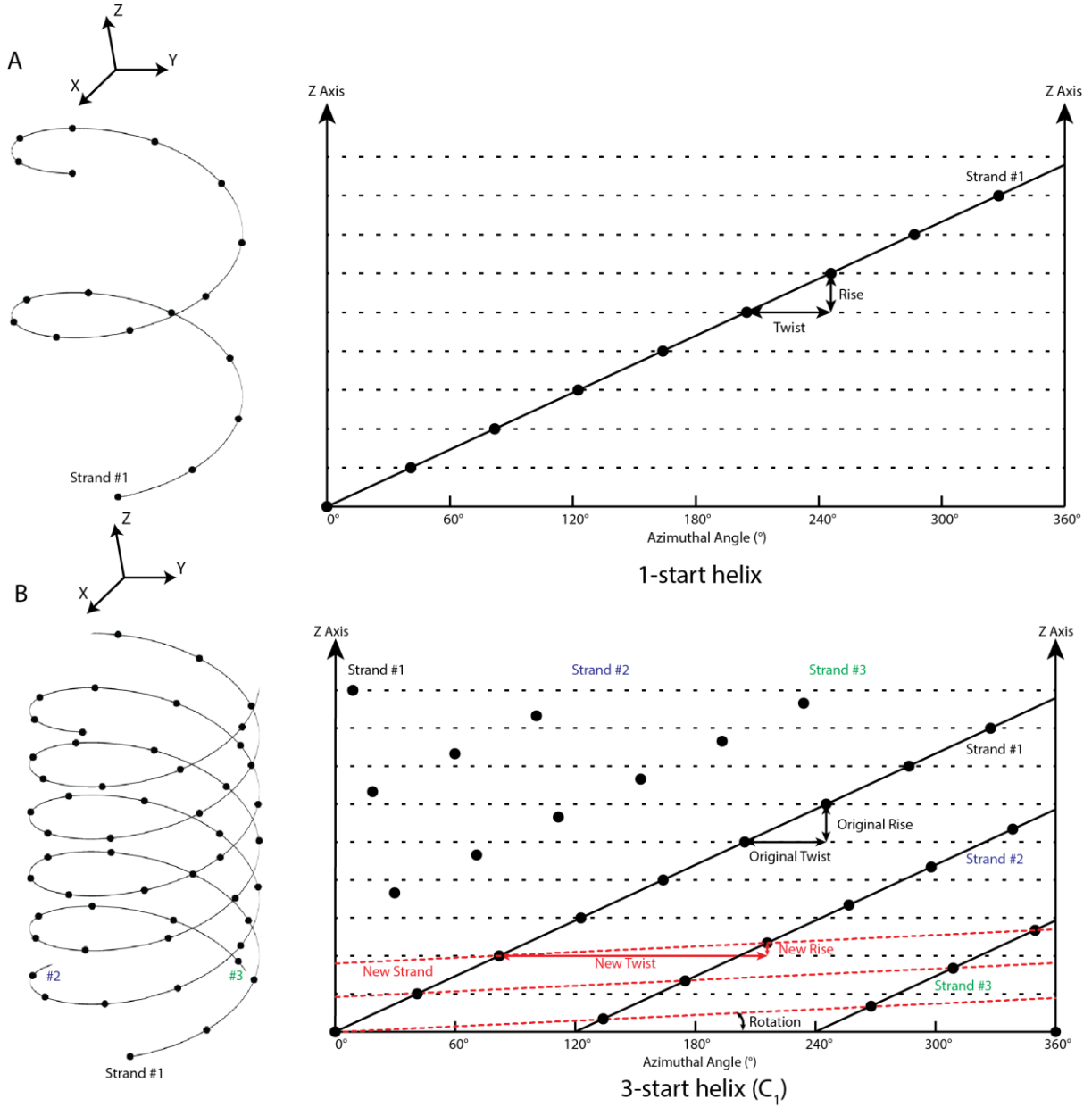


Figure 2.2. Definition of helical symmetry for n -start helices. (A) Example of a 1-start helix with no point group symmetry (C_1). The 3D helical structure on the left can be ‘unfolded’ into a 2D plot of the z -height against the azimuthal angle ϕ on the right. The dots represent the centres of the subunits on the 2D surface lattice. (B) Example of a 3-start helix (with no point group symmetry), where instead of using the original twist and rise of the individual three strands (in black), RELION expresses the helical symmetry using one third of the original rise (in red) (adapted from He and Scheres, 2017).

2.2.2 Applying helical symmetry in Fourier-space

As is described in section 1.4, RELION makes use of the estimated signal-to-noise ratios (SNRs) at all Fourier frequencies to update the reconstructed volume V in each iteration. In order to employ this empirical Bayesian approach, the algorithm requires the number of asymmetric units (H , a positive

integer) for the reconstructions. In ordinary single-particle analysis, each experimental 2D image (2D projection) contains the information of only one asymmetric unit if point group symmetry is absent from the structure. In helical reconstructions of RELION, the inter-box distance, the distance between centres of adjacent helical segments extracted, determines the number of asymmetrical units H in each new segment. Therefore, the typically used inter-box distances are H times the helical rise. The user needs to provide H in each segment for subsequent 3D reconstruction jobs. However, if there is no reliable estimation of helical rise at the start of a project, an educated guess, or simply $\sim 10\%$ of the box size, might be used as the inter-box distance for the segment extraction. In such cases, H is temporarily set to one. Once the helical twist and rise values become clear, re-extractions of all segments with a more reasonable inter-box distance and H are necessary for better reconstructions.

For $H = 1$, helical reconstructions are the same as ordinary single-particle reconstructions without any point group symmetry. For $H \geq 2$, the algorithm could insert every experimental 2D projection H times into Fourier space, each time rotated and translated by $h \times \text{twist}$ and $h \times \text{rise}$ respectively ($h = 0, 1, \dots, H - 1$). The translations in real-space are actually implemented as phase shifts to all coefficients in Fourier space. Other packages, such as FREALIGN (Lyumkis et al., 2013), have already adopted this concept of Fourier-space symmetrisation. However, the implementation in RELION firstly inserts each original 2D projection only once without any transformations ($h = 0$), and then it rotates and translates the asymmetric 3D volume F_0 as a whole, making the solution computationally cheaper (figure 2.3B):

$$F_{sym} = \frac{\sum_{i=0}^{H-1} T^{i \cdot \Delta} \cdot R^{i \cdot \phi} \cdot F_0}{H}, (H \geq 2, H \in \mathbb{N}^+)$$

Where F_0 denotes the accumulated sum of all the aligned experimental 2D projections in Fourier space, F_{sym} the symmetrised map, and R^ϕ , T^Δ the rotational and translational operators. FREALIX also makes use of a similar implementation (Rohou and Grigorieff, 2014).

In addition, translations of Fourier volumes make coefficients on top of the particle boxes wrapped around to the bottom and vice versa. These artefacts mess up densities at both ends of helices. Thereby, only the central z slices are kept intact after Fourier space symmetrisation (figure 2.3D). This effect imposes a strict upper limit on the inter-box distance at 50% the box size.

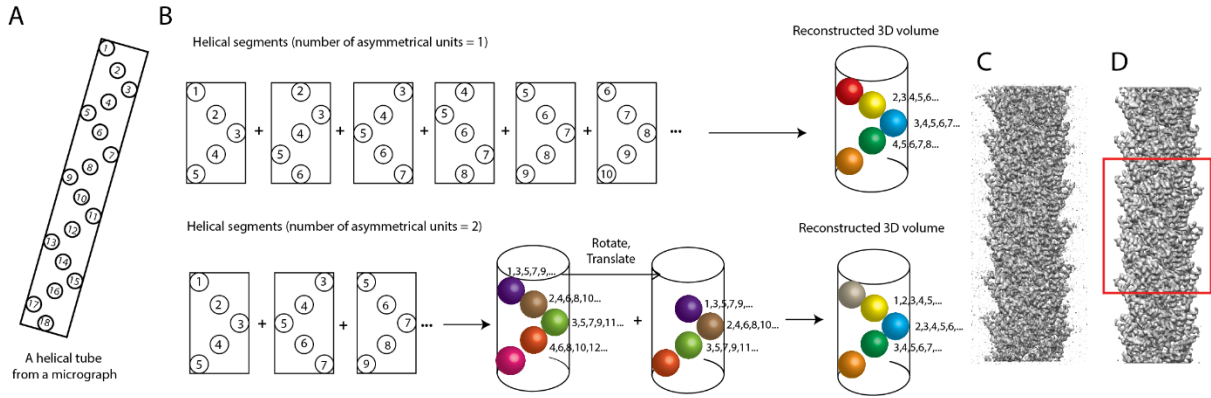


Figure 2.3. Imposition of helical symmetry in Fourier space. (A) A schematic representation of a helical filament with 18 asymmetrical units. (B) Extracted segments with $H = 1$ (top) or 2 (bottom) per box lead to similar reconstructed volumes after symmetrisation in Fourier-space. Fourier-space symmetrisation is not necessary if the inter-box distance is one helical rise ($H = 1$). Otherwise, the reconstructed volume from the helical segments is rotated and translated according to the helical symmetry, and summed up in Fourier space. (C) The reconstruction of MAVS from manually picked filaments using $H = 6$. Without Fourier-space symmetrisation, the 6 asymmetrical units in each segment box are not averaged. (D) The MAVS reconstruction after Fourier-space symmetrisation. The central z sections along the helical axis (within the red box) are averaged over all available asymmetrical units. At the top and the bottom of the helix, blurring is caused by artefacts (adapted from He and Scheres, 2017).

For a given data set, larger values of H result in fewer extracted helical segments, which speeds up computation. However, a fixed set of micrographs holds invariant amount of information regardless of what inter-box distances are used. Theoretically, different values of H should give rise to identical signal and noise power spectra in 3D helical reconstructions after Fourier-space symmetrisation.

Fourier-space symmetrisation is applied in every iteration of 3D classification and refinement jobs. The method is also implemented in a standalone program ‘relion_reconstruct’. The following command symmetrises a TMV reconstruction in Fourier-space using $H = 30$:

```
relion_reconstruct --i TMV.mrc --nr_helical_asu 30 --helical_rise 1.408 --
helical_twist 22.03 --angpix 1.126 --o TMVsym.mrc
```

2.2.3 Applying helical symmetry in real-space

There exist other factors blurring the density maps of helical structures in addition to the artefacts introduced by Fourier-space symmetrisation. For example, although cut into segments, helices are still not perfectly straight even within short spanning ranges. Similar to springs, helices can be locally stretched and pressed, resulting in slightly varying twists and rises (and this is not considered as a mixture of completely different helical lattices). Other distortions include shears or flattening of the helical tubes, which are especially prominent with wide helical tubes. Inaccuracies of alignments also blur the whole reconstructions. Most of the factors mentioned above have more smearing effects to the

top and bottom densities of the 3D reconstructions compared to the central z slices. Thus, it improves the quality of reconstructions to calculate a real-space symmetrised map based on the densities only within a number of central z sections:

$$V_{sym}(r, \phi, z) = \overline{V_0(r, \phi + i\Delta\phi, z + i\Delta z)} = \sum_{i: -\frac{pD}{2} \leq z + i\Delta z \leq \frac{pD}{2}} V_0(r, \phi + i\Delta\phi, z + i\Delta z) / n(i), (i \in \mathbb{N})$$

Where V_0 is the previous 3D volume symmetrised in Fourier-space according to the number of asymmetrical units (H), and V_{sym} is the real-space symmetrised result. A user-defined percentage value $p(\%)$ and the box dimension (D) together specify that the central $-pD/2 \sim +pD/2$ z slices are used for real-space symmetrisation. $n(i)$ is the number of voxels within the central slices that are used to calculate the averaged (symmetrised) voxel value at 3D cylindrical coordinate (r, ϕ, z) . The values of i , making $z + i\Delta z$ located inside the central $-pD/2 \sim +pD/2$ z slices, may differ across voxels at various z heights in the 3D volume. The equation is equivalent to but computationally cheaper than the operations of symmetrising the central slices according to the helical parameters and then spanning the perfect central densities to the full length of box size.

Real-space symmetrisation is applied at every iteration in 3D classification and refinement jobs with one exception. Since this symmetrisation has masking effects in real-space, it is not applied to the last iteration of 3D refinement jobs where the unfiltered half-maps are compared for gold-standard Fourier-shell correlations (FSC). The method is also implemented in a standalone program ‘relion_helix_toolbox’. The following command symmetrises a TMV reconstruction in real-space:

```
relion_helix_toolbox --impose --i TMV.mrc --o TMVsym.mrc --
cyl_outer_diameter 200 --angpix 1.126 --rise 1.408 --twist 22.03 --
z_percentage 0.3
```

2.2.4 Local searches of helical symmetry in real-space

The initial helical twist and rise provided by the user, estimated using prior information or Fourier-Bessel method, might not be accurate enough for high resolution reconstructions. The implementation in RELION optionally performs local searches of helical symmetry in real-space before imposition of the symmetry in real-space.

The algorithm performs 2D grid searches and starts with user-defined initial search ranges of helical twist and rise along with their customised searching steps (optional). Lower and upper bounds of ranges can be set to the same value to enable 1D searches of either twist or rise when appropriate. This could be useful as many helices have twist values that are much more stable than rise, since a small change in twist might distort the helical lattice remarkably while a small change in rise has limited pulling and stretching effects along z axis only. If searching steps are not defined by the user (left as 0 or negative),

the algorithm will set $<1^\circ$ and $<1\%$ as the step sizes for helical twist and rise respectively. The step sizes should also satisfy the condition that ≥ 5 and ≤ 1000 samplings are to be searched for either parameter. This initial sampling scheme is designed for effective searches at limited computation costs for general cases.

For each sampled symmetry, the algorithm treats the density map as if it conforms to the symmetry being tested and calculates the variance between averaged central z sections within a user-defined percentage of the box length (p%). Theoretically, the error should be 0 when the tested symmetry fits the map perfectly if interpolation errors are ignored. The sampled symmetry with the smallest variance is selected as the best candidate for this batch of samplings.

Once the best symmetry is found given the initial samplings, the algorithm decreases both searching step sizes of twist and rise by half. Centred at the most possible symmetry estimated from the last attempt, the new searching ranges are set to $\pm 2 \times$ the new steps, allowing $\geq 5 \times 5$ samplings in total. The searching ranges are shrunk shorter progressively in further iterations. The algorithm stops and returns a refined symmetry when both the step sizes of twist and rise are smaller than the floating-point number accuracy (i.e. $\sim 10^{-5}$ of a floating-point number).

The algorithm tries to find the best possible symmetry within the user-defined ranges. However, if in any iteration the best symmetry sits at the lower or upper limits given by the user, it is highly likely that the true symmetry might be completely different from the initial estimation. In such cases, the algorithm outputs a warning message. This out-of-range warning could also appear in early stages of 3D refinement jobs, where resolutions are still low and the user has specified tightly confined search ranges of helical symmetry. In such cases, it is expected that the warnings should disappear during the later stages of reconstructions.

Helical symmetry is searched independently in each class in 3D classifications. The searches are also independent in either of the two half-maps in 3D refinement but the averaged symmetry is used as the new estimation for the next iteration. However, the searches are not conducted in the first iterations of 3D classification and refinement jobs for stability.

This algorithm is different from the original IHRSR algorithm (Egelman, 2000) in two aspects. Firstly, it performs 2D-grid searches if both ranges of twist and rise are provided by the user, while IHRSR always searches rise and twist in 1D alternately. Secondly, the implementation in RELION does not make any assumptions on the landscape where the local minimum of variance lies and samplings steps are adaptively decreased. This is in contrast to IHRSR, which keeps sampling steps fixed and estimates the minimum variance and best symmetry by fitting a quadratic function of calculated variances with respect to all the sampled symmetry.

The method to locally refine helical symmetry is also implemented in a standalone program ‘relion_helix_toolbox’. The following command searches helical symmetry for TMV within ranges of $+21^\circ$ - $+23^\circ$ and 1.3 \AA - 1.5 \AA for twist and rise respectively and writes out the final results to the screen:

```
relion_helix_toolbox --search --i TMV.mrc --cyl_outer_diameter 200 --angpix
1.126 --rise_min 1.3 --rise_max 1.5 --twist_min 21 --twist_max 23 --
z_percentage 0.3
```

2.2.5 2D image masks and 3D reference masks

Prior to the introduction of helical masks, it is necessary to explain the definition of coordinate system in cryo-EM data processing using RELION (figure 2.4). RELION adopts the convention of right-handed, Cartesian coordinates with right-handed rotations designated as positive (Heymann et al., 2005). The rotation centre of a 2D or 3D input image is the geometric centre. For example, for a 2D image with dimensions of X_{dim} and Y_{dim} pixels, the rotation centre is set to the position $(\lfloor \frac{X_{dim}}{2} \rfloor, \lfloor \frac{Y_{dim}}{2} \rfloor)$ on the 2D grey-scale pixel grid, with (0, 0) representing the 1st pixel on the upper left. For most cases, all the 2 or 3 dimensions are equal and in even numbers. The translational offsets (estimated by the classification or refinement algorithms described in section 1.4) of individual input images are applied before the rotations. The definition of rotations, or the three Euler angles (ϕ , θ , ψ), is shown in figure 2.4 (Heymann et al., 2005). Since the helical axis is always the z axis of the density maps of 3D references in RELION (section 2.2.1), for each input 2D or 3D image of a helical segment, the 1st angle ϕ ($-180^\circ < \phi < +180^\circ$) denotes the rotation around the helical axis, the 2nd angle ($\theta - 90^\circ$) ($0^\circ < \theta < 180^\circ$) denotes the out-of-plane tilt (i.e. $\theta = 0^\circ$, 90° , 180° means the top view, the side view with no tilting and the bottom view of the helix respectively), and the 3rd Euler angle ψ ($-180^\circ < \psi < +180^\circ$) denotes the in-plane rotation (for all 3 rotations, the angles are with respect to the 3D reference).

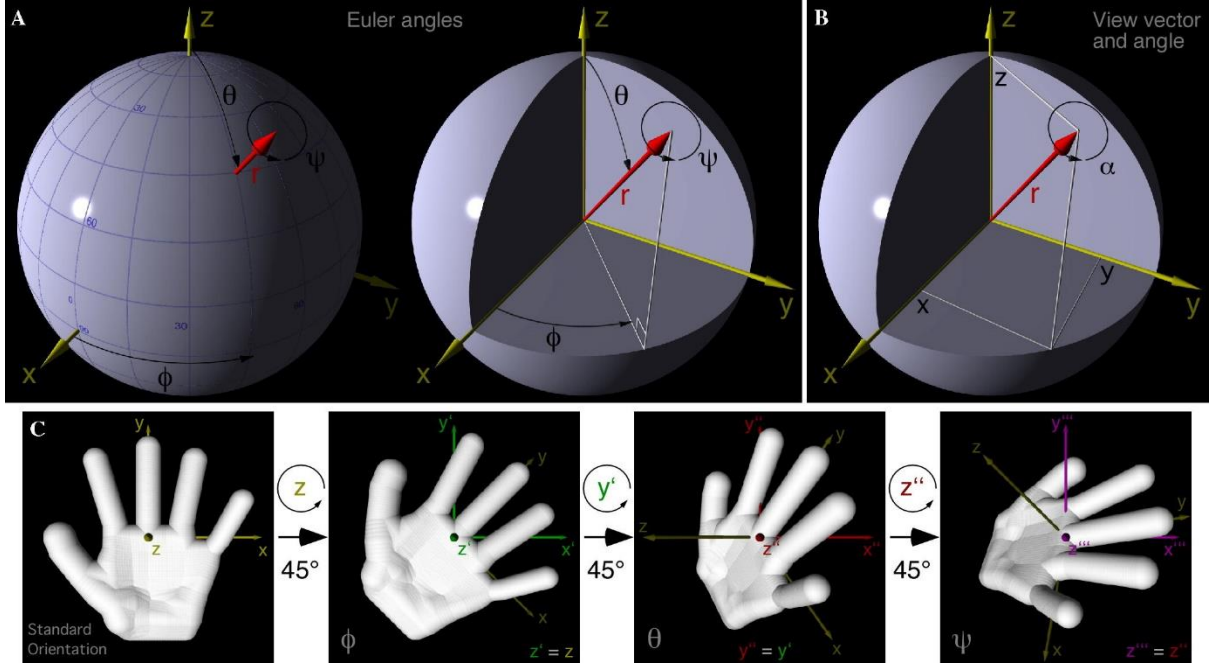


Figure 2.4. The definition of orientations. **(A) Euler angles.** The position of any vector beginning at the origin can be described by the angles ϕ and θ (in this example, $\phi = \theta = 45^\circ$). The angle ϕ is the angle measured in the anti-clockwise direction between a projection of the vector \mathbf{r} onto the x - y plane and the positive x -axis. (ϕ is analogous to longitude.) The angle θ is the angle between the vector and the positive z -axis. (θ is analogous to latitude.) Rotation about the vector is described by the angle ψ (analogous to compass heading). **(B) View vector and angle.** The vector \mathbf{r} is the same as the corresponding vector in panel A. \mathbf{r} is described by coordinates $\{x, y, z\}$ (labeled in black) in each of the three Cartesian axes x, y, z . The view angle (α) is a rotation about \mathbf{r} . **(C) Rotation matrices for Euler angles.** A model of a human left hand is used to demonstrate the three angles. As applied in rotation matrices, the convention involves three right-handed rotations about successive orthogonal axes: first, rotation about z by ϕ ; second, rotation about y' by θ ; and third, rotation about z'' by ψ . The standard axes $\{x, y, z\}$ are shown in yellow in each frame. The prime, double-prime, and triple-prime axes are shown in green, red, and magenta, respectively. Note, when the axis of rotation is pointing at the viewer, the coordinate system is rotated anti-clockwise and the object clockwise (adapted from Heymann et al., 2005).

Helical masks are implemented both for 3D reconstructions and 2D/3D experimental segments.

Referential masks for 3D reconstructions largely remove background noise in the defined solvent regions (figure 2.5B). Structures and masks of helices both span along the entire z axis of 3D segment boxes. A cylindrical mask (M_{ref}) is multiplied with the real-space reconstruction map after the maximisation step, resulting in a better 3D reference for the next iteration. The user-defined parameter R_o is the width of the cylindrical mask and an optional parameter R_i can also be specified if the helical structure contains a hollow centre.

$$M_{ref}(r, \phi, z) = \begin{cases} 1, & 0 \leq R_i < r < R_o, r^2 + z^2 < R^2 \\ 0, & \text{otherwise} \end{cases}$$

R denotes the radius of the 3D spherical mask which is usually set to 40% - 45% of the segment box sizes for 3D helical reconstructions. Evidently, the condition $R > R_o > R_i$ must be satisfied. In addition, 3-5 pixels of cosine soft-edges with values ranging 0-1 are added around the binary cylindrical mask.

The significance of masks for 2D/3D experimental segments is twofold. Firstly, they are used to determine the mean and standard deviation of background for segment normalisation. And they wipe out noise at solvent regions and reduce the effects from neighbouring segments on alignments during the expectation step in the EM algorithm. Unlike circular masks in single-particle analysis, masks around 2D/3D experimental helical segments are no longer rotationally invariant. These masks are rotated according to the spanning directions of helices (i.e. angular priors) estimated through manual or template-based auto-picking (figure 2.5A). For each 2D helical segment, the in-plane rotation angle (ψ) is arbitrary, and the tilt prior (θ) is always set to 90° (the assumption of perfect side views).

For each pixel of a 2D segment, its x,y coordinates are rotated back to the same 2D Cartesian coordinate system according to angular priors of the segment and the distance from the central axis determine the value in the binary mask:

$$A_{2 \times 2} = \begin{pmatrix} \cos \psi & \sin \psi \\ -\sin \psi & \cos \psi \end{pmatrix}, M_{exp}(x, y) = \begin{cases} 1, & A_{2 \times 2}^T \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x' \\ y' \end{pmatrix}, |y'| < R_o \text{ and } x'^2 + y'^2 < R^2 \\ 0, & \text{otherwise} \end{cases}$$

Cosine soft-edges are also applied around the binary masks of the experimental segments.

Although helical segments always extend across experimental images, circular or spherical components of masks are still necessary in removing artefacts at box corners. If signals at the edges and corners of the particle boxes are not masked away, rotations or projections of images performed in Fourier-space may cause artefacts in real-space as is shown in figure 2.5C-F. In addition, the presence of circular components in the masks also enables fair comparisons between experimental images and reference projections of the same segment lengths and shapes.

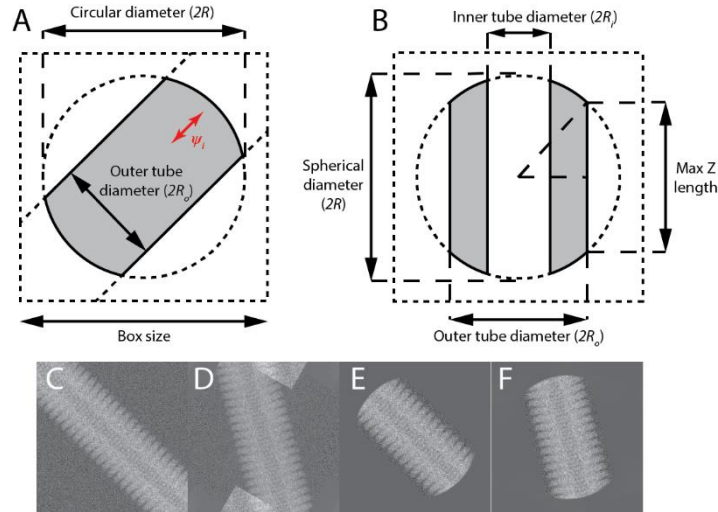


Figure 2.5. Masks for 2D experimental images (A) and 3D referential maps (B) and the reduction of real-space artefacts. Areas of a filament inside the mask are shown in grey, areas outside the mask in white. In panel (A), the helical axis runs at 45° with respect to the horizontal direction, which represents an arbitrary in-plane rotation angle ψ . The helical (z) axis runs parallel to the vertical direction in (B). Panels (C)-(F) illustrate the need for the circular (or spherical) component of these masks. Rotation of an unmasked segment (C) leads to artefacts (top and bottom in D), due to wrapping effects caused by rotations in the Fourier domain. The application of a circular mask to the original image (E) allows rotation without artefacts (F) (adapted from He and Scheres, 2017).

2.2.6 Incorporating prior information

As a statistical data processing framework, RELION has the ability to incorporate prior information from experimental data sets to improve the quality of 3D reconstructions. Helical segments come with abundant prior knowledge both in orientations and translations, which can be exploited to perform more accurate alignments with reduced computation costs.

Unlike single-particle analysis, simple assumptions on particle orientations can be made for 2D helical segments. The initial estimation of such orientation priors are retrieved from manual- or auto-picking of helical segments (explained in the next section). Firstly, most helical filaments take approximately horizontal positions in thin ice layers. This information is expressed as Gaussian prior probability functions initially centred at 90° for tilt angles (θ), with a user-defined standard deviation to account for different degrees of tilting within a specific data set (figure 2.6B). A standard deviation of 5° (--sigma_tilt 5) means that the program will search for $\pm 15^\circ$ (3σ) around the tilt prior of each segment in the first few iterations. The tilt priors are re-centered to the best angles found in the last iteration, unless '--keep_tilt_prior_fixed' is provided to keep tilt priors fixed always at 90° in the initial global angular searching steps. Secondly, directions of filaments on 2D micrographs retrieved from manual or auto-picking can be expressed by bimodal Gaussian priors of in-plane rotations ψ (figure 2.6C). For helices without point group symmetry or with C_n symmetry only, these directional priors have to be bimodal

because the polarity of each filament relative to the 3D reference is unknown before alignment. The polarities are expected to be resolved during the 3D reconstructions. However, directional priors for D_n helices can be set as unimodal since these helices lack polarities. An initial standard deviation (`--sigma_psi X`) is required from the user as well to account for the accuracies of in-plane angle (ψ) estimations from manual or auto-picking. As the same in single-particle analysis, the standard deviations of θ and ψ and their searching steps are both decreased automatically as alignments proceed to finer stages.

In addition, there is additional prior knowledge for C_n helices that all segments coming from the same filament must conform to the same polarity. In the implementation, the two opposing components of bimodal Gaussian priors have the same weights (50% vs 50%) before the start of a 3D reconstruction. As global angular searches progress, segments coming from the same filament may accumulate higher probability for one of the two polarities. Assume that for a continuous filament with N segments, $x^{(i)}$ segments are found pointing 'upward' with respect to the reference in the i th iteration. Then the weights of 'upward' and 'downward' components of bimodal priors in $(i+1)$ th iteration are:

$$w_{up}^{(i+1)} = \frac{x^{(i)} + 1}{N + 2}, w_{down}^{(i+1)} = \frac{N - x^{(i)} + 1}{N + 2}, w_{up}^{(i+1)} + w_{down}^{(i+1)} = 1, N \in \mathbb{N}^+; x^{(i)} \in \mathbb{N}, 0 \leq x^{(i)} \leq N.$$

Additional smoothing factors '1's are appended to the numbers of segments $x^{(i)}$ and $N - x^{(i)}$ so that the priors are always bimodal, although highly biased, even if (almost) all segments along a filament are found to be pointing into the same direction ($x^{(i)} = 0$ or $x^{(i)} = N$). This approach is useful especially for short filaments so that the segments will not get stuck in wrong unimodal polarity assignments at initial stages of 3D reconstructions. Bimodal priors are only implemented during global angular searches. When 3D refinements proceed into local searching steps, such directional priors are locked to the component with higher weight (either $w_{up} = 1$ or $w_{down} = 1$), therefore no longer bimodal. The above solution for polarity correction is enabled by the additional option '`--helical_sigma_distance X`', which will be discussed at the end of this section. By default, the option is disabled and fixed 50%-50% unbiased bimodal priors are applied to all segments during global angular searches, and the polarities are fixed thereafter according to the best ψ angles found.

Theoretically, the first Euler angles, $\text{rot}(\varphi)$, should have fixed increments of $H \times \text{twist}$ for consecutive segments along a perfect filament. However in reality, rot angles (φ) are heavily influenced by translational offsets and flexibility of helical segments. Moreover, since RELION only extract segments at integer coordinates to avoid additional interpolation, the exact inter-box distances between neighbouring segments may deviate slightly (< 1 pixel) from the user-defined value. rot priors (φ) are therefore not implemented currently because expressing this information is too complicated and it is not expected to bring evident improvements to helical reconstructions.

For both single-particle and helical reconstructions, there are inevitable translational offsets between the centre of particles/segments and the actual centres of image boxes introduced by particle/segment picking. In single-particle analysis, RELION estimates standard deviations of isotropic Gaussian offset priors on image x, y axes from the experimental 2D data during iterative 3D reconstructions. In helical processing, translations on the helical direction are treated differently from translations perpendicular to the helical axis. In-plane rotation (ψ) prior defines the direction of the helical axis of each segment and the prior distribution of offset along that direction is a top-hat function with non-zero values within range $[-\text{rise} / 2, +\text{rise} / 2]$ around the image centre (figure 2.6A). This strictly confined distribution prevents segments from being translated on top of the next asymmetric unit, which would result in duplicating the same signal in neighbouring overlapping segments. For the perpendicular direction (one for 2D segments), the standard deviations of their Gaussian offset priors are firstly given by the user and subsequently re-evaluated at every iteration in the same way as in single-particle analysis. The translational searching steps are usually the same or comparable along both helical and orthogonal directions. However, for helices with tiny rise values, the top-hat priors become very narrow. In these cases, at least 5 samplings are imposed along the helical axis in order to enable sufficient translations in that direction.

The translation along the helical axis is coupled with the rot angle (ϕ), the rotation of a segment around its axis with respect to the 3D reference. A translational step of ΔX Angstroms is equivalent to $\text{twist} \times \Delta X / \text{rise}$ degrees of change in rot angle (ϕ). For simplicity, the current implementation still treats rot angles (ϕ) as free parameters, and its local angular searching ranges (3σ) should be at least $\pm 0.5 \times \text{twist} \times \Delta X / \text{rise}$ around the centre of its prior estimated from last iteration to guarantee sufficient orientation searches.

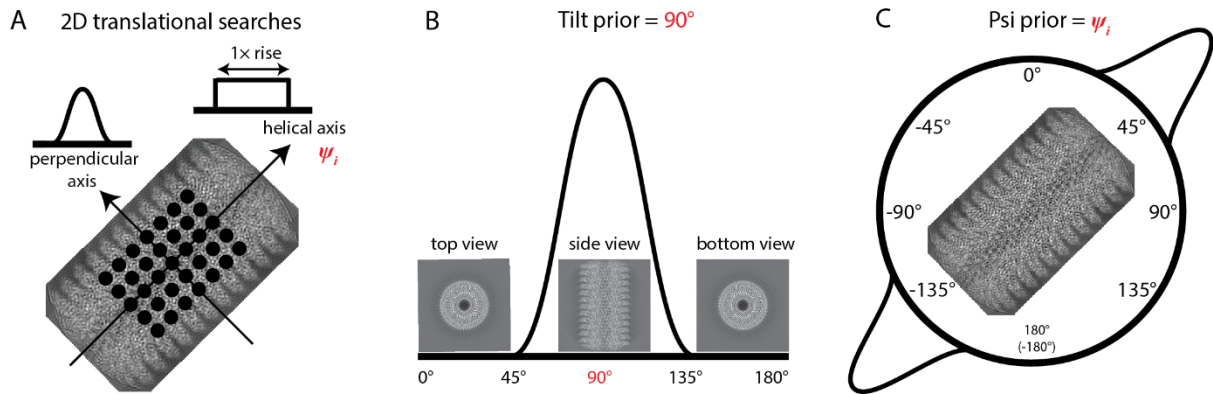


Figure 2.6. Priors on the orientation parameters. (A) A top-hat prior on the in-plane translations along the helical axis, and a Gaussian prior on the in-plane translations perpendicular to the helical axis. (B) A mono-modal Gaussian prior on the tilt angle ($\theta = 90^\circ$), which describes the out-of-plane rocking of filaments in the ice layer. (C) A bimodal Gaussian directional prior on the ψ angle, which describes the in-plane rotation of the filament (adapted from He and Scheres, 2017).

A major difference between ordinary single-particle and helical reconstruction is that overlapping helical segments along the same filament are not completely independent. Their images share overlapping regions, possess the same polarity and neighbouring segments should only have gradually changing orientation and translations due to continuity of the helical filaments. The first two properties have already been dealt with using the concept of asymmetrical units (H) and bimodal in-plane rotation (ψ) priors. As for the last one, some unexpected factors, such as strong contamination signals or otherwise irregular pixel values, may cause a few segments to be wrongly assigned with very different orientations and translations from their neighbours during alignment. Therefore, local averaging of segments has been implemented as an additional option in 3D reconstructions. For both tilt (θ) and in-plane (ψ) angles, the implementation centres priors of the outliers on the best possible orientations averaged according to the orientations of their neighbouring segments within a short distance given by the user through option ‘--helical_sigma_distance X’. X means that all the neighbouring segments along the same helical filament within a distance of X times the segments box size are taken into account for calculating the averaged θ and ψ orientation priors. A Gaussian function is applied to lower weights of segments lying closer to the edges of averaging window. If the orientation alignment of a segment deviates more than 3σ (--sigma_tilt or --sigma_psi) of either its θ or ψ , this segment is considered as an outlier and its orientation priors will be replaced by the locally averaged θ and ψ . Local averaging of segments effectively corrects outliers of wrong alignments.

2.2.7 Segment picking and extraction

The routines for manual and semi-automated picking have also been modified for 2D projection images of helical segments.

Manual picking of 2D helical segments requires the user to specify start and end points of helical filaments through subsequent left-clicks of mouse on every 2D micrograph displayed in RELION (see section 2.4.1 and figure 2.10AB). It assumes that the picked filaments are straight. The picking results are saved into STAR files (Hall, 1991) containing columns of x, y coordinates. During the extraction step, coordinates and priors on the orientations of the picked segments are written into the final particle STAR file along with all the cropped 2D image boxes. Additional settings of helical rise and the number of asymmetrical units ($H \geq 1$) need to be provided by the user to calculate the inter-box distance $\Delta L = H \times \text{rise} / \text{pixel_size}$.

For example, if a filament starts at x, y positions (x_s, y_s) and ends at (x_e, y_e) , the inter-box distance, i th segment coordinates (x_i, y_i) and their tilt (θ) and in-plane rotation (ψ) priors are calculated as:

$$\text{For all possible } i, \Delta l = \frac{H \times \text{rise}}{\text{pixel_size}}, \theta = 90^\circ, \psi = \tan^{-1} \frac{y_e - y_s}{x_e - x_s},$$

$$\begin{aligned} x_i &= x_s + i\Delta l \times \cos \theta \\ y_i &= y_s + i\Delta l \times \sin \theta, x_e \neq x_s. \end{aligned}$$

Each 2D segment is marked with an identification number (ID) for the filament it belongs to, and its position on that filament track. All the information mentioned above is used to generate rotational and translational priors described in the previous section. Alternatively, the implementation also supports other formats of input filament coordinates, such as box files from EMAN2 helix_boxer (Tang et al., 2007) and plain text files containing even number of entries of x, y coordinates in the first two columns.

Manual picking of filaments is a formidable task for projects with thousands of 2D micrographs. In order to speed up the process, the original template-based auto-picking algorithm in RELION (Scheres, 2015) has been adapted for helices on 2D micrographs (figure 2.7). Auto-picking yields better detections of segments for non-straight filaments. Auto-picking requires high-quality 2D references of target helices as input. They can be imported from elsewhere or generated through 2D classifications of a small number of manually-picked filaments. Figure-of-merit calculation, during which the similarities are compared between the arbitrarily rotated references and all local regions of micrographs, remains the same as in the original algorithm for single-particle analysis. New features of the adaptation cover filament detections and cross-over removal. In addition to the inter-box distance, the user must also provide a binary threshold, helical width and the maximum curvature of filaments (0-1, defined as the fraction of curvature of a 2D circular mask). Firstly, the algorithm filters figure-of-merit peaks with intensities higher than the threshold. Then it starts from the highest peak to fit helical tracks on the figure-of-merit map, taking into account the adjacent peaks within half of the template box size and the maximum curvature allowed. The algorithm iteratively searches for possible tracks of filaments along both directions until no new peaks can be detected on either ends. Afterwards, the algorithm repeats the above procedure of filament detections starting from the next highest peak which is not covered by any of the previously fitted filament tracks within helical widths. Segment coordinates are then calculated based on the user-defined inter-box distance. In particular, coordinates are discarded if they lie at cross-overs of overlapping filament widths, or at the edges and corners of micrographs. Prior information, such as angular priors ($\theta = 90^\circ$, ψ) and relative positions on filament tracks, is also estimated and written into the output STAR file along with the segmented coordinates.

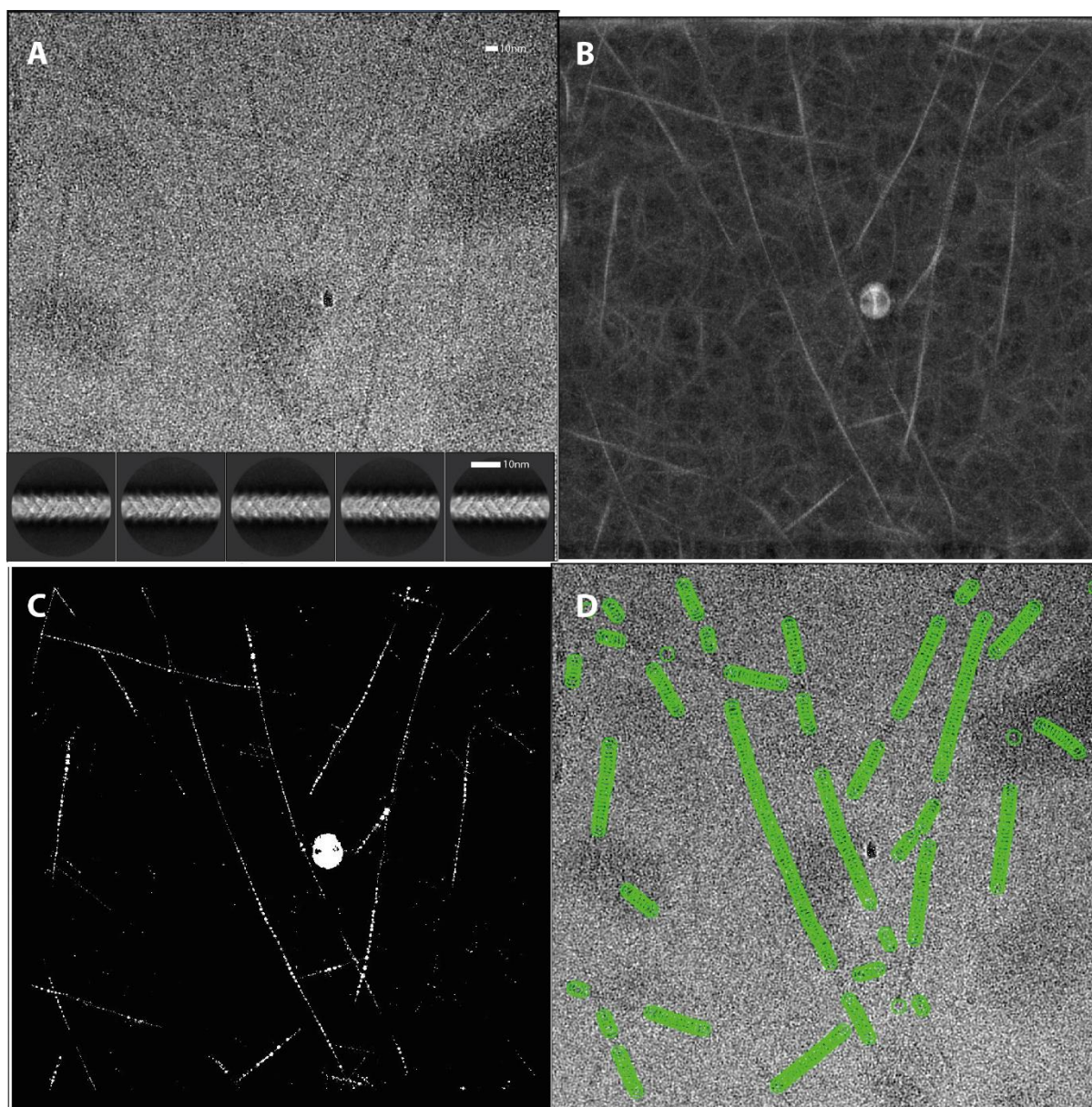


Figure 2.7. Auto-picking filaments on a 2D micrograph. (A) A representative micrograph low-pass filtered to 20 Å and the five 2D class templates. (B) Figure-of-merit map generated after comparing 2D templates with the micrograph (A). (C) Filtered figure-of-merit map with filament tracks fitted to remaining peaks. (D) Final picked coordinates (each denoted by a green circle) at an inter-box distance of 50 Å.

2.2.8 Randomisation of image subsets

In 3D auto-refinement jobs, input particles are randomly separated into two half data sets. Reconstructions are then performed independently to prevent over-fitting to prevent the build-up of high-frequency noise. For single-particle analysis, each particle is assigned to one of the half-sets using a uniform random number generator. However, this approach is problematic for helical segments, which always have >50% of overlap with their directly adjacent neighbours. Overlapping regions distributed

into different halves may introduce redundant correlations between the two half-sets, which could lead to over-fitting and spurious resolution estimates. Therefore, filament-wise randomised subsets can be a simple solution for this problem. Every helical segment is tagged with an ID denoting the filament it belongs to on a certain micrograph during extraction. The modified method traverses through the list of all segments, count the total number of filaments (N) found on all micrographs and the number of segments each filament contains (S_i , $1 \leq i \leq N$). Then it randomises the sequence of N filaments using Knuth's shuffle (or Fisher-Yates shuffle, Knuth, 1997) so that any of the $N!$ permutations can occur with equal probability. The shuffled filaments are subsequently collected sequentially into two half-sets, with balance maintained for the numbers of segments across the two half-sets.

```

For I <- 1 to N,
    Sequence[i] <- i
For I <- 1 to N,                                     ### Knuth's shuffle
    J <- uniform_random_integers(i~N);
    If I != J
        Swap(Sequence[i], Sequence[j]);
half1_size <- 0, half2_size <- 0
For I <- 1 to N,                                     ### Two halves collect filaments
    If half1_size < half2_size
        Half1.addFilament(Sequence[i]);
        half1_size += Sequence[i].s;
    Else
        Half2.addFilament(Sequence[i]);
        half2_size += Sequence[i].s;

```

Code 2.1. Pseudo-code for randomisations of subsets in helical 3D refinements.

2.2.9 Initial model generation

The MAP-EM estimator in RELION is a local optimiser. Therefore, the initial reference must lie within the convergence radius of the algorithm in order to find the correct reconstruction. Various options, that are generally useful to construct initial models, have been implemented, which are described in more details below. Apart from the following methods, the user may also choose previously refined density maps or sub-tomogram averaging results as 3D references if appropriate.

2.2.9.1 Featureless cylinders

Featureless cylinders are unbiased initial models that are also non-polar, and they are often a good choice for simple helices. A cylindrical reference can be generated using the following command:

```
relion_helix_toolbox --cylinder --o out.mrc --boxdim 300 --
cyl_inner_diameter 50 --cyl_outer_diameter 200 --angpix 1.34 --
sphere_percentage 0.9 --width 5
```

In this example, the program writes out a cylinder (out.mrc) with voxel values set to one within the inner and outer diameter (80 Å and 200 Å). Cosine soft-edges (5 pixels) are applied around its side surfaces as well as a spherical mask (90% of the box size) at the periphery. The outer diameter should be set to approximately the actual width of filaments estimated from micrographs or 2D class averages.

2.2.9.2 Simulated lattices

Simulated helical lattices are better initial models for complicated symmetries compared to featureless cylinders as they provide helical lattice information for the alignment. Such references may be built if the helical symmetry can be reliably estimated based on prior knowledge or clues from the Fourier-Bessel method. Each subunit, with its voxel values all set to one, is drawn as a small cosine soft-edged sphere centred at the positions specified by the helical and rotational (C_n) symmetry. The references also conform to D_n symmetry since all constituent subunits (spheres) are rotationally invariant. Ideally, the diameters of spheres should be large enough to simulate the surface contacts of neighbouring subunits.

```
relion_helix_toolbox --simulate_helix --o out.mrc --subunit_diameter 30 --
cyl_outer_diameter 200 --angpix 1.126 --rise 1.408 --twist 22.03 --boxdim
300 --sym_Cn 1 --sphere_percentage 0.9 --width 5
```

2.2.9.3 Converted density maps from PDB models

Initial references can also be constructed from crystal structures of the subunits (or their homologues) in the helical assemblies. These structures might be readily available in PDB formats, fitted into density maps which have been reconstructed previously. For example, if a TMV subunit in PDB format has the same orientation as any of the subunits on a TMV helix, the following command can make 100 copies of the input PDB model and assemble them into a helix according to the correct helical symmetry of TMV:

```
relion_helix_toolbox --pdb_helix --i in.pdb --o out.pdb --rise 1.408 --twist
22.03 --nr_subunits 100
```

The helix spans along the z axis of the PDB coordinate system with subunits positioned on xy planes at a radius calculated internally given the input PDB model. If the user wants a modified radius (filament width) at which the subunits are positioned, ‘--center_pdb’ should be provided along with the helical tube diameter ‘--cyl_outer_diameter X’ in Angstroms.

The output PDB assembly needs to be converted to MRC densities, and the density map must be further applied with a spherical mask at 80%-90% of the box size to avoid artefacts in Fourier-space due to the presence of densities at box edges (section 2.2.5). For example, the following two commands write out a 3D reference low-pass filtered to 10 Å in a 300-pixel box:

```
e2pdb2mrc.py out.pdb out.mrc -apix 1.34 -box 300 -res 10
relion_helix_toolbox --spherical_mask --i out.mrc --o masked.mrc --
sphere_percentage 0.9
```

The three types of models explained above do not reflect the true structural densities of the data sets so they are not on the absolute grey-scales. In addition, none of them is affected by CTFs.

2.2.10 Graphic user interface for helical reconstruction jobs

The graphical user interface (GUI) in RELION provides a user-friendly way of setting up jobs, superseding the usage of command lines and scripts. Specific tabs on the GUI with helical settings have been implemented for the auto-picking, particle extraction, 2D/3D classification, 3D refinement (figure 2.8), particle polishing and mask-creation jobs types. The implementation in RELION requires fewer parameters provided by the user than other existing helical packages since the statistical framework of RELION learns many parameters automatically from the experimental data sets (section 1.4).

Figure 2.8. ‘Helix’ tab in 3D classification/refinement job types.

2.2.11 Simulations of 2D experimental segments

Simulated image data is useful to detect bugs, evaluate performance and elucidate concepts of the implemented algorithms.

To start with, a 3D density map of a perfect helix must be provided as the reference with known helical symmetry. It can be either a high-resolution reconstruction result or a density map converted from a PDB helix (section 2.2.9.3). In order to simulate helical segments with box sizes of D pixels, the helix must span along the 3D reference with a box size of at least $1.5 \times D$. Otherwise some of the simulated segments might not span the full length of the particle boxes (for example, the segments along the diagonal directions). The following commands enlarge the box size (300 pixels) of the original TMV reference ‘TMVref300.mrc’ and impose the helical symmetry of TMV to produce a longer reference ‘TMVref450.mrc’ (450 pixels):

```
relion_image_handler --i TMVref300.mrc --new_box 450 --o tmp.mrc

relion_helix_toolbox --impose --i tmp.mrc --o TMVref450.mrc --
cyl_outer_diameter 200 --angpix 1.126 --rise 1.408 --twist 22.03 --
z_percentage 0.3 --sphere_percentage 0.9
```

A 3D spherical mask of $\leq 90\%$ the box size should also be applied to the 3D reference to avoid artefacts introduced by the image transformations afterwards (section 2.2.5).

Then the following command writes out several STAR files with angular and translational priors (‘segments_info.star’ and ‘segments_helical_priors.star’) and an executable shell script (‘./simulate-2d-segments.sh’).

```
relion_helix_toolbox --simulate_segments --i TMVref450.mrc --o segments.star
--boxdim 300 --nr_subunits 6000 --nr_asu 30 --nr_tubes 5 --twist 22.03 --
rise 1.41 --cyl_outer_diameter 200 --angpix 1.126 --bimodal --sigma_tilt 3
--sigma_psi 3 --sigma_offset 3 --white_noise 5 --random_seed 1400014000
```

It means that the user wants to simulate 6,000 TMV subunits on 5 helical filaments with a number of asymmetrical units (H) set to 30 using box sizes of 300 pixels. The final simulated data will consist of $6,000 / 30 = 200$ 2D segments with $200 / 5 = 40$ segments on each filament. Sigmas of tilts (θ , centred at 90°), in-plane rotations (ψ) and offsets lead to a certain degree of randomness imposed on the angular and translational priors. The random seed needs to be modified if different results in multiple runs are required. The value of white noise decides how much Gaussian random noise is to be added to the simulated segments in real-space and it needs to be tuned for each 3D reference. Running the shell script ‘./simulate-2d-segments.sh’ generated by the above command writes out all the simulated 2D segment images.

The shell script projects the 3D reference according to the angular and translational priors in the previously written in ‘segments_info.star’ file. For each simulated filament, the constituent segments have their in-plane rotations (ψ) at the same random value and rot angles (ϕ) increased at a fixed increment of $H \times \text{twist}$ degrees. Tilts (θ) are all initially centred at 90° . All angular and translational

priors are subsequently perturbed by small random numbers coming from uniform distributions around 0 with user-defined widths (σ), although the simulated segments are all perfectly straight. However, the simulated segments are perfectly straight with fixed helical symmetry without any heterogeneities, which does not exactly reflect the properties of true helical filaments collected from experimental micrographs or tomograms. The output file ‘segments_helical_priors.star’ can be used as input for 3D helical reconstructions in RELION, as it removes rot angles (ϕ) and sets all translational priors to zeros since such information cannot be estimated during the picking of true experimental data sets.

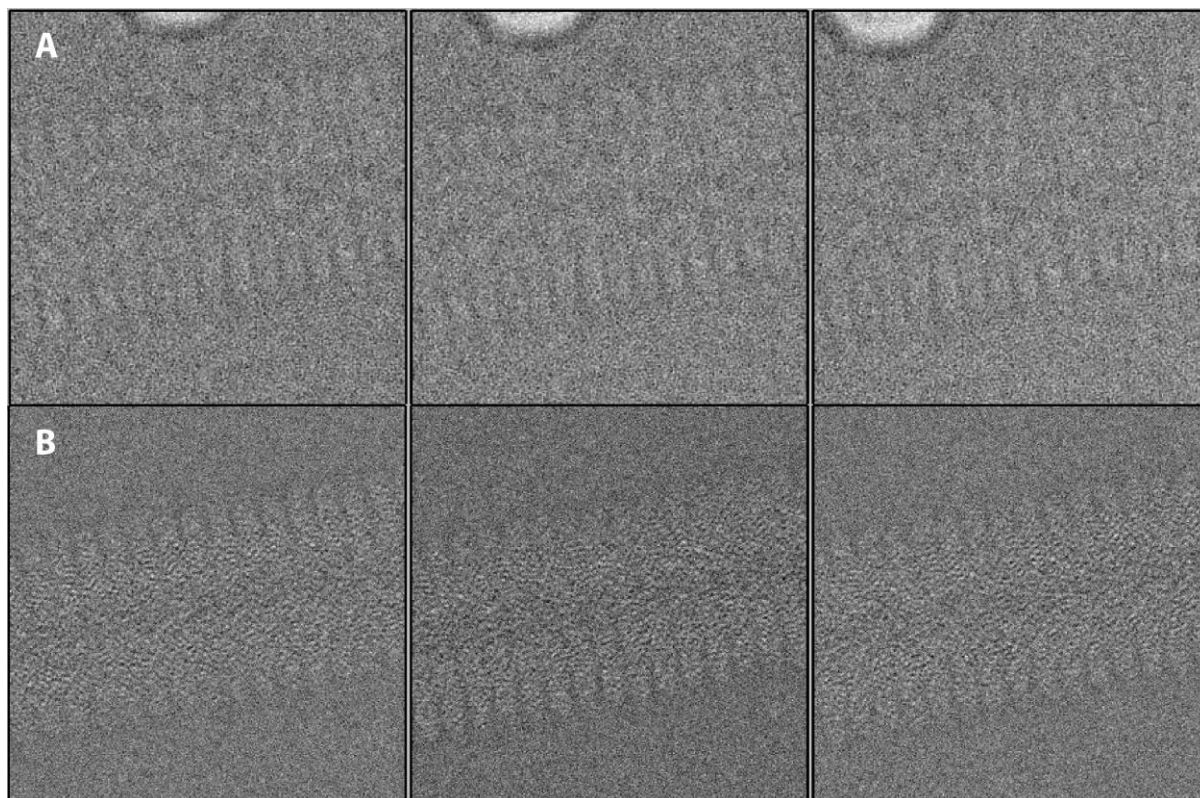


Figure 2.9. Comparison between true and simulated 2D experimental images. (A) 3 TMV segments extracted from EMPIAR-10020 (Fromm et al., 2015). **(B)** 3 simulated TMV segments generated from 3D reference EMDB-2833 (Fromm et al., 2015) with a white noise level set to 20.

2.3 Introduction to the test data sets

Four cryo-EM data sets are used repeatedly in this section in order to demonstrate the significance and benefits of all the implemented methods explained above. Original collected movies/micrographs are available from EMPIAR (<https://www.ebi.ac.uk/pdbe/emdb/empiar/>, Iudin et al., 2016) for three data sets: tobacco mosaic virus TMV (Fromm et al., 2015; EMPIAR-10020), contractile sheath (formed by VipA/VipB proteins) of the bacterial type-VI secretion system (Kudryashev et al., 2015; EMPIAR-10019), prion-like aggregates of mitochondrial antiviral signalling proteins mediated by N-terminal caspase activation and recruitment domain MAVS/CARD (Xu et al., 2015; EMPIAR-10031). The

fourth data set, bacterial actin homolog MamK (Löwe et al., 2016), was collected in-house by Jan Lowe's group at LMB.

As can be seen in Table 2.1, these four data sets of helices are different in many aspects. They were collected from various microscopes and direct-electron detectors. The helices conform to various combinations of helical and point group symmetry parameters. Moreover, the filaments are different in widths and flexibilities as well. TMV has been a widely popular test case for helical processing packages (Desfosses et al., 2014; Rohou and Grigorieff, 2014). TMV filaments are highly straight and regular, with some end-to-end joints (red arrows in figure 2.10A) connecting individual filaments into longer ones. With higher degree of symmetry, VipA/VipB are straight and regular as well but they tend to form shorter ones compared to TMV. MAVS/CARD and MamK helices are more bending with thinner filament widths. Their flexibilities could be partly explained by simpler helical and point group symmetries, which result in fewer contacts and therefore less cohesions among neighbouring subunits on their lattices. In addition, the MamK test data set here is refined to a lower resolution (3.9 Å) compared to the published results (3.6 Å, Löwe et al., 2016) because it consists of only a fraction of movies available (260 out of 2,086). Löwe et al. refined the 3.6 Å structure using the implemented workflow of RELION-2 helical processing described in this thesis.

	TMV	VipA/VipB	MAVS/CARD	MamK
Microscope	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios	FEI Polara 300
Detector	Gatan K2	Gatan K2	FEI Falcon-II	FEI Falcon-III
Pixel size (Å)	1.126	1.00	1.05	1.34
Movies collected?	Yes	No	Yes	Yes*
Nr. Movies/micrographs	14	77	512	260 (2086) ¹
Nr. frames per movie	22	-	16	(46) ¹
Accumulated dose (e⁻/Å²)	~43	~30	~35	~40
Defocus range (-,µm)	1.5~2.3	0.5~1.5	2.5~4.0	1.1~3.0
EMPIAR entry	10020	10019	10031	-
EMDB entry	2833	2699	6428	4062
EMDB software	SPRING	IHRSR/SPIDER	RELION-1.2 (custom)	RELION-2 α
EMDB resolution (Å)	4.0	3.5	4.2	3.6

PDB entry	4UDV	3J9G	3J6J	5LJW, 5LJV
	RELION-2 processing of test data sets			
Manual/auto-picked?	Manual	Manual / Auto ²	Auto	Auto
Auto-picking threshold	-	0.10	0.50	0.15
Auto-picking maximum curvature	-	0.05	0.30	0.30
Box size (pixels)	300	450	300	280
Inter-box distance (Å)	42.2	43.6	30.4	52.2
Nr. asym. units/box, H	30	2	6	1
Nr. segments (picked)	1,761	11,114	227,176	118,911
Nr. segments (refined)	1,761	9,480	133,237	110,055
Filament widths (Å)	180	100~280 ³	105	195
Point group symmetry	C ₁	C ₆	C ₁	C ₂
Helical rise, refined (Å)	1.41	21.79	5.07	52.11
Helical twist, refined (°)	+22.03	+29.41	-101.2	+23.75
B-factor (Å²)	-83	-91	-135	-119
Final resolution (Å)	3.9	3.3	4.0	3.9

¹ The full data set of MamK consists of 1,655 movies with 46 frames each. The test data set of MamK consists of only 260 motion-corrected micrographs.

² Both manual- and auto-picking are performed on VipA/VipB data set (section 2.4.1).

³ Inner and outer diameters of the hollow tubes.

Table 2.1: Key characteristics of four test data sets used in section 2.4.

2.4 Results and Discussion

2.4.1 Segment picking

Straight filaments, TMV and VipA/VipB, can be manual-picked effectively (figure 2.10AB). Especially, TMV must only be picked manually since template-based auto-picking cannot identify end-

to-end joints (red arrows in figure 2.10A) which connect the termini of two individual filaments. Figure 2.10CD shows the representative auto-picking results of bending filaments, MAVS/CARD and MamK.

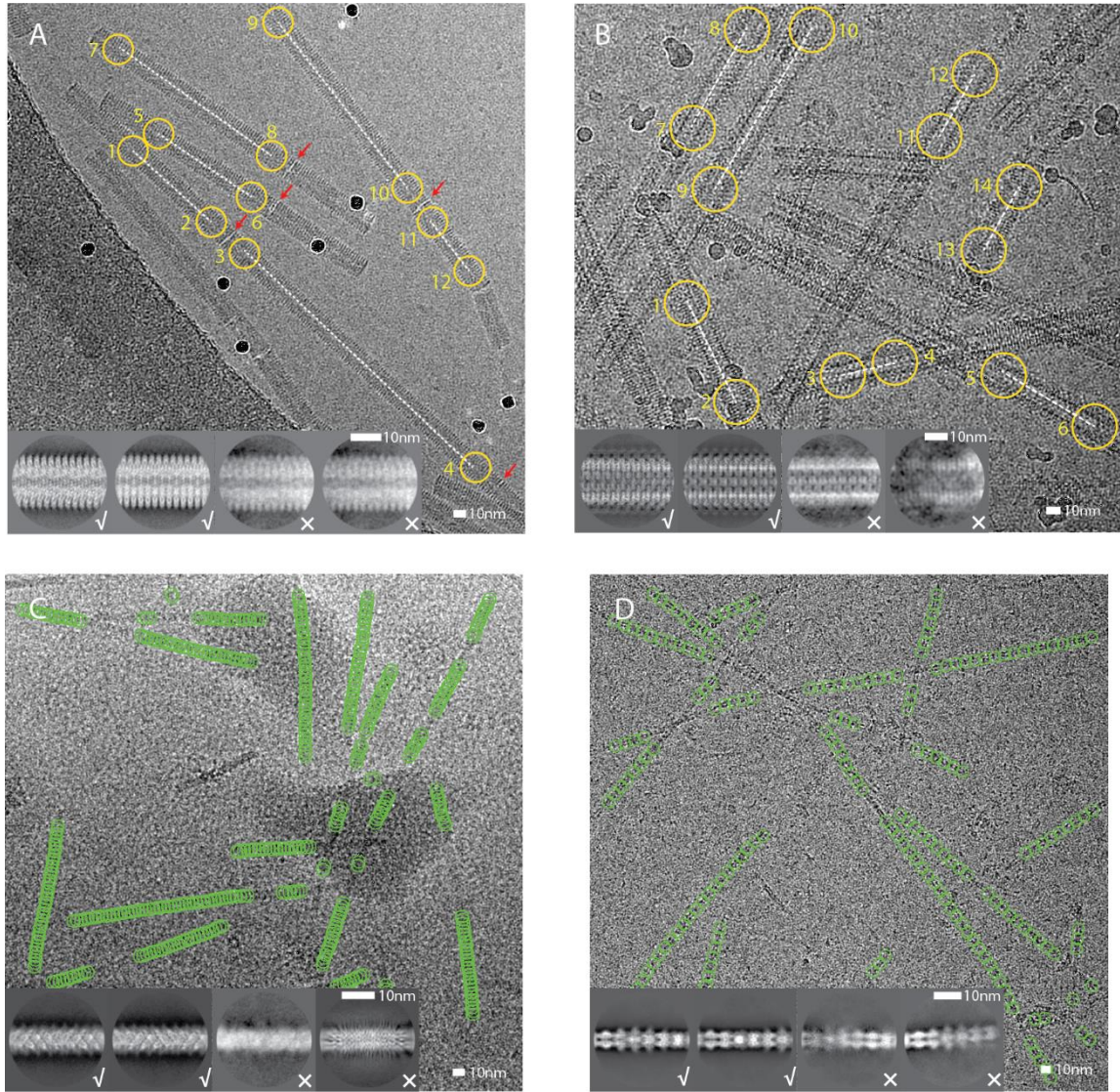


Figure 2.10. Representative micrographs, particle picking and 2D classification results for the four test cases. (A) Manually-picked coordinates (yellow) at the starts and ends of the TMV filaments. Non-helical end-to-end joints are indicated with red arrows. (B) Manually-picked coordinates (yellow) for a VipA/VipB micrograph. (C) Auto-picked coordinates (green circles) for a MAVS/CARD micrograph. (D) Auto-picked coordinates (green circles) for a MamK micrograph. For each test case, four 2D class averages are shown as insets. Good class averages are indicated with a tick and suboptimal ones with a cross (adapted from He & Scheres, 2017).

Auto-picking of helical segments is similar to that of the ordinary single-particles, but with more user-defined parameters required. The first step is the choice of 2D class averages as templates. In general, the selected templates should be clear in details and among the classes of best-quality. The box size of such templates is recommended to be 1-1.5 \times of the box size of segments used in the actual 3D

reconstructions. Smaller box sizes with insufficient signals might lead to sub-optimal calculations of figure-of-merit values between the micrographs and the templates, while very large box sizes increase computation costs and exclude more segments to be picked near the edges of the micrographs. Normally, 3-5 classes are sufficient for auto-picking of helices. However, more representative templates may need to be collected if there are heterogeneities present in the 2D classification results (e.g. varying helical diameters, evident different structural details and mixtures of helical symmetries).

Once the 2D templates are chosen, 5-10 representative micrographs need to be selected in order to figure out the best settings of auto-picking parameters. They should cover the full range of defocus values and include micrographs with both densely and scarcely populated filaments, large contamination signals and good/bad imaging conditions. The tuned parameters (introduced in the following paragraphs) should have overall satisfactory performance on the representative set. The overall recall rate of helical segments matters more than accuracies since the non-helical segments are expected to be separated out after 2D classifications performed on the whole picked data set.

Some of the auto-picking parameters can be determined without much tuning. Low-pass filters are usually set to 20Å and high-pass filters are not usually needed. Angular samplings are set to $\leq 3^\circ$ because helical segments are long in shapes and span across the full lengths of the boxes. The option, maximum standard deviations of noise, is usually set to -1 (disabled) since the process of fitting helical tracks ignores many of the small contamination signals (e.g. round-shaped ice and carbon particles). Shrink factors < 0.5 are not recommended since heavily binned micrographs could blur helical tracks. Helical widths can be measured from the 2D classes and the product of number of asymmetrical units and the helical rise is the actual inter-box distance used in picking. If the helical symmetry is not yet clear, the product can be set to $\sim 10\%$ of the segment box size used in subsequent 3D reconstructions. Maximum curvatures differ according to the flexibilities of filaments, with 0.05-0.10 suitable for straight filaments (such as VipA/VipB) and 0.20-0.40 for more bending ones (MamK and MAVS/CARD). Minimum lengths of tracks are used to exclude very short filaments in case that the algorithm picks too many false-positives, non-helical artefacts.

Given a set of 2D templates with fixed box sizes, the mask diameter may have profound influences on the auto-picking results. The mask diameter must be strictly smaller than the template box sizes. Theoretically, larger values lead to higher figure-of-merit values when the templates match the helical filaments on the micrographs. However, a very large diameter might include more short-range disorders of the filaments on the micrographs and the signals at the periphery of the diameter might give worse correlations with the templates. Especially for flexible helices, inappropriately large masked regions on the micrographs cover more bending parts of the filaments which become less comparable with the straight 2D class averages. Thus, the template box size and the structural flexibilities together impose an upper limit to the mask diameter. On the other hand, the mask diameter should be large enough for

the calculated figure-of-merit values to distinguish filamentous regions from the background noise effectively. The template box size, the 2D selected templates and the mask diameter must be determined before comparisons of different picking thresholds can be performed.

The picking threshold is the most important user-defined parameter in helical auto-picking algorithm. As is explained in section 2.2.7, figure-of-merit maps (*CCF.spi, figure 2.6B; *PLOT.spi, figure 2.6C) corresponding to the 5-10 representative micrographs should be written out initially and their pixel values along the helical filaments can be examined using ‘relion_display’ in order to find a range of reasonable picking threshold values for testing. The user need to find the best threshold within the range by comparing the total numbers of segments picked. The optimal choice is the largest threshold which gives high recall rate of helical segments with most of the long filaments picked continuously. Given a fixed inter-box distance and the manually picked coordinates along straight filaments on the representative micrographs, the recall rate is calculated as the number of auto-picked segments along these filaments divided by the number of manually picked ones. Eventually, the optimal settings are used to auto-pick the full data set.

For VipA/VipB and MAVS/CARD filaments, comparisons have been done to show the influence of different picking methods on the final resolutions of reconstructions (table 2.2). VipA/VipB were manual-picked by the author and the start-end coordinates of MAVS/CARD filaments are readily available online in EMAN2 helix_boxer format (Tang et al., 2007) through accession number EMPIAR-10031 (Xu et al., 2015). Auto-picking of both data sets was performed starting from 2D class averages of a small subset of manually picked segments (for the parameters used, please refer to entries ‘auto-picking thresholds’ and ‘auto-picking maximum curvatures’ in table 2.1). The results of 2D classification and 3D reconstruction will be discussed in details in the following sections.

	Segments picked	Segments kept	Final resolution
VipA/VipB (manual)	10,830	10,211	3.3 Å
VipA/VipB (auto)	11,114	9,480	3.3 Å
MAVS/CARD (manual)	34,397	34,397	4.3 Å
MAVS/CARD (auto)	227,176	133,237	4.1 Å

Table 2.2. Comparison of picking methods performed on VipA/VipB and MAVS/CARD filaments. Total number of segments after picking and after segment selections based on high-quality 2D class averages are shown, along with final map resolutions after 3D refinement and post-processing (adapted from He & Scheres, 2017).

As is shown in table 2.2, template-based auto-picking demonstrates performance at least as satisfactory as careful manual-picking. Particularly, template-based picking captured a great numbers of segments

at bending parts of MAVS/CARD filaments, which result in better resolution of its final map. Therefore, auto-picking is more beneficial to long and bending filaments. Moreover, auto-picking may greatly accelerate initial steps of helical processing. According to the performance with test data sets (table 2.3), it is expected that auto-picking jobs running on GPU cards can be finished within a couple of hours given thousands of micrographs as input. The general GPU acceleration of RELION, implemented by Kimanius et al., 2016, is applicable to auto-picking, 2D classification, 3D classification and refinement job types for both single-particle and helical processing.

	Size of micrographs (pixels)	# of micrographs	Size of templates (pixels)	# of templates	# of GPU cards	Time (approx.)
VipA/VipB	3838 × 3710	77	600 × 600	8	1	1 h 30 m
MAVS/CARD	4096 × 4096	512	300 × 300	5	1	2 h 35 m
MamK	4096 × 4096	260	238 × 238	5	1	55 m

Table 2.3. Processing time of auto-picking jobs performed on Nvidia GTX 1080 Ti GPU cards. Angular samplings are set to 3° for all test cases.

The 2D helical segments can be picked once the coordinates are determined through manual or auto-picking. Empirically, common values of box sizes are at least 1.5~2x the filament widths or 200 pixels, whichever is larger. On one hand, a very large box size is usually unnecessary and makes the subsequent classification or reconstruction jobs computationally expensive, and it risks including more bending or distorted parts of the helices across a long distance in the segment boxes and therefore worsens the reconstructions. In addition, it is not guaranteed that the helical symmetry is perfectly uniform and the deviations may add up across a long distance, especially for the bending filaments. On the other hand, a too small box size may truncate the signals at the periphery of the filaments which are not clearly visible on the raw micrographs (with 80%~90% soft-edged segment masks applied in the reconstructions, refer to section 2.2.5). Furthermore, the images are not sufficiently sampled in Fourier space in this case and, along with the lack of signals, lead to inaccurate results in orientation and translational searches. Another important factor that sets the lower bound of the appropriate box size is the effect of CTF delocalisation. Twin images can be observed in bright-field defocused phase-contrast images in cryo-EM, as the pseudo-Friedel-related Fourier components scattered from a specimen particle are physically separated along trajectories that follow their respective Bragg scattering angles and are rendered separate on the micrograph (Russo and Henderson, 2018. Refer to section 6 and figure 6.1). In the other way, it can be seen as the information contained at a certain location in the diffraction pattern in Fourier-space can be delocalised to the neighbouring twin images close to the centred one in

the real space. The centres of the two components are separated by a distance of $2\Delta F\lambda/d$, with ΔF denoting the defocus, λ the electron wavelength, and d the resolution in units of distance. For example, the distance is ~ 100 Å for the information at 4 Å resolution if the defocus is $\sim 1\mu\text{m}$ using a 300 keV microscope ($\lambda = 0.02$ Å). Therefore, the box size should be at least the filament width plus 100 Å in order to preserve the structural information available at 4 Å, which is a common final resolution for many helical structures resolved in the recent years.

2.4.2 2D classifications

Reference-free 2D classifications serve at least three major purposes in helical processing: generating 2D templates for auto-picking using a small number of manually picked segments as input, generating segment averages with better signals for Fourier-Bessel analysis (section 1.7 and 2.4.3) and separating poor-quality segments from the good ones suitable for 3D reconstructions. Examples of optimal and poor 2D class averages are shown in figure 2.10A-D (insets). Sub-optimal segments and class averages might result from contaminations, extremely distorted filaments, coordinates picked at filament ends or false-positives of picking.

Classifications of 2D helical segments make use of in-plane angular (ψ) priors. The algorithm only searches around the priors within a small range specified by the user ($\pm 6^\circ$ by default), which speeds up computation greatly. Optimal angular search ranges depend on the accuracies of segment picking and the degrees of bending of the filaments. The in-plane angular (ψ) priors are bimodal for C_n and unimodal for D_n helices (section 2.2.6). 2D classifications of helical segments generally require finer angular samplings compared to those of single particles because the signal of helices span across the full diameter of the 2D circular masks. When angular samplings are not sufficiently small, blurring effect can be seen at the periphery of the circular masks. The user may find out the upper limit of samplings in the output file. It is calculated using current resolution estimates (X , in Angstroms) and the diameter of the circular/spherical masks (D , in Angstroms).

$$\Delta\theta_{max} = \arctan \frac{2 \times X}{D}$$

0.5° or 1° are typical values used for helical 2D classifications. The translational search range is ± 5 pixels by default and should be increased when the segment box size is large (≥ 500 pixels) or the picked coordinates are not accurate.

2D classifications are more likely to over-fit if the user uses tiny inter-box distances ($< 5\%$ the box size). Signs of over-fitting in this step include streaky densities in solvent areas, concentrated classes with similar appearances (sometimes with non-protein shapes), lots of empty classes unexpectedly, ever increasing estimated resolutions which are unreasonable for the scale of the data set, etc. One explanation might be that tiny inter-box distances combined with translational searches duplicate some

projections among different classes and create redundant correlations as each segment is free to be considered for any arbitrary class in 2D/3D classifications (over-fitting prevention using randomised subsets are only enabled in 3D auto-refinements, explained in section 2.4.10). Apparently, one solution is holding on to the convention of using 5%-30% as the inter-box distance (which will be explained in details in section 2.4.6). Alternatively as a general way of over-fitting prevention, limiting the expectation (E) step to a moderate resolution (5 Å - 10 Å) could also alleviate the problem (i.e. providing ‘--strict_highres_exp X’ as an additional option).

2D classifications may be the most time-consuming steps in the whole workflow of helical processing if hundreds of thousands of helical segments have been extracted for analysis. In order to get rid of apparently abnormal segments and false-positives, it is suggested to divide any large data set into a few ~100,000-segment subsets and perform 2D classifications on these subsets with class numbers (K) of ~100 independently. For example, the following command divides the input particle STAR file ‘in.star’ into 4 parts.

```
relicon_helix_toolbox --divide --i in.star --nr_outfiles 4
```

The subsets can be imported into the workflow again. After 2D classifications and particle selections for all subsets, the full data set can be merged using ‘combine STAR file’ job-type in the GUI. Not only will this measure save a lot of computation resources, but also enable finer 2D classifications of the helical segments.

The key settings and processing time of 2D classification jobs are listed in table 2.4 for the four test cases.

	Size of segments (pixels)	# of segments	# of 2D classes	Angular range & sampling	Trans range & sampling	CPU cores (# of MPI X threads)	Time (approx.)
TMV	300 × 300	1,761	20	±6° / 1°	±5 / 1 pix	15×8=120	10 m
VipA/VipB	450 × 450	11,114	50	±6° / 0.5°	±5 / 1 pix	48×4=192	3 h 10 m
MAVS/CARD	300 × 300	227,176	100	±6° / 1°	±5 / 1 pix	90×8=720	10 h
MamK	280 × 280	118,911	100	±6° / 1°	±5 / 1 pix	60×4=240	12 h 50 m

Table 2.4. Processing time of 2D classification jobs performed on Intel Xeon E5-2643 (3.4~3.5 GHz) and E5-2667 (2.9 GHz) CPU clusters. Angular searches were bimodal for all test cases. None of the data sets was divided into subsets.

2.4.3 Initial helical symmetry estimation and initial model generation

There is not a one-size-fit-all method of initial model generation which suits every case of helical reconstructions. Apart from cylindrical, simulated and PDB models mentioned in section 2.2.9, other options include Fourier-Bessel reconstruction, sub-tomogram averaging or previous single-particle reconstruction results from similar structures.

Featureless cylindrical references are feasible for both the MAVS/CARD and the MamK data sets. In the first round, 3D refinement started from a cylinder in each case using additional option ‘--ignore_helical_symmetry’ to make use of angular prior information retrieved from segment auto-picking while no helical symmetry was imposed (the inter-box distances are set to 10% the box sizes for both cases). The reconstructions yielded density maps with apparent helical contours (figure XAB, middle). The user is expected to give some rough estimations of the size of subunits and possible ranges of helical twist and rise by looking at the densities. Then, helical symmetry can be searched through ‘--search’ option in ‘relion_helix_toolbox’. The searching results were -101.1° , 5.150 \AA for MAVS/CARD and $+18.29^\circ$, 55.24 \AA for MamK. The maps from the first round were symmetrised according to the searched symmetry and used as initial references for the next 3D helical refinement with symmetry imposed and locally searched at every iteration. The final symmetry parameters were -101.2° , 5.071 \AA for MAVS/CARD and $+23.57^\circ$, 52.11 \AA for MamK. These final results revealed the correct structures (figure 2.10AB, right).

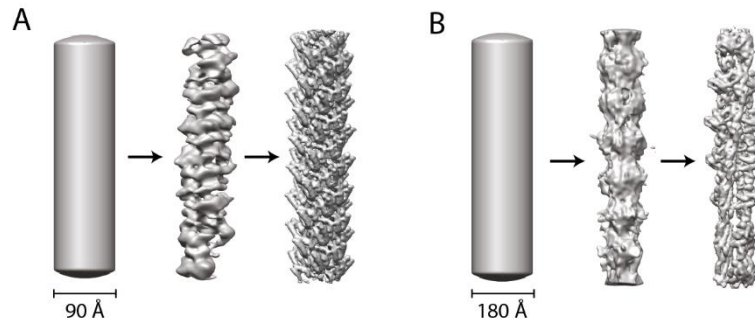


Figure 2.11. Initial model generation for MAVS/CARD (A) and MamK (B). In each panel, the featureless cylinder (the width of which is estimated from the 2D class averages) that was used as an initial model is shown on the left. The result from a 3D refinement without imposing any symmetry is shown in the middle. The result (unfiltered reconstructed maps) from a subsequent 3D refinement with local searches of helical symmetry is shown on the right.

Featureless cylinders do not work for TMV and VipA/VipB filaments. The reconstruction results were extremely noisy without any structural features (figure 2.12). The claimed resolutions were 33.8 \AA and 26.5 \AA for TMV and VipA/VipB respectively after 3D refinements and such resolutions were even worse than the low-pass filters applied to the featureless cylindrical references (20 \AA). For wide and complicated helical lattices like TMV and VipA/VipB, the initial references must provide information

of the positions of protein subunits on the helical lattices and the correct helical symmetry should be applied during 3D reconstructions to keep the refined structure within the radius of convergence of the algorithm. Therefore, their diffraction patterns should be analysed to estimate their helical twist and rise values using Fourier-Bessel method (section 1.7). Fourier transforms can be performed on the best-quality 2D class averages of helical segments (figure 2.13A, 2.14A and 2.15A) or some cropped filaments on raw micrographs. The 2D class averages contain less noise and they generate clear diffraction patterns with more discernible layer lines. The 2D classes may as well give a clue of the presence of various helical symmetries and widths in the data set. The Fourier-Bessel analysis of 2D class averages is more applicable for bending, short, thin or long-repeat filaments. However, 2D classification might average segments with different helical symmetries into the same classes if there exists a high degree of heterogeneity in the data set. It is expected that the diversity of helical symmetry can be observed more often in thick and hollow filaments with complicated induced helical lattices. In such cases, it is recommended to perform Fourier-Bessel analysis on a careful selection of cropped and straightened filaments instead of 2D averages.

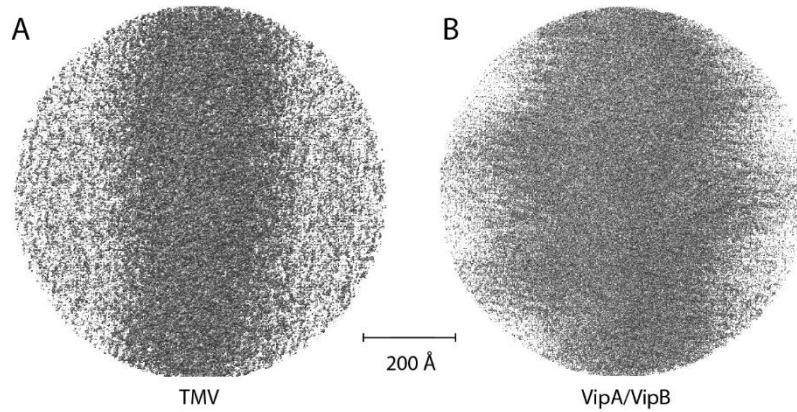


Figure 2.12. The 3D refinement results (side views of the unfiltered maps) of TMV (A) and VipA/VipB (B) without any helical symmetry applied using featureless cylinders low-pass filtered to 20 Å as the initial models.

As the first step in the Fourier-Bessel method, helical repeat distance (ρ , in Angstroms) can be estimated after identifying the equator and the first layer line (blue dashed lines on figure 2.13B, 2.14B and 2.15B) on the diffraction pattern:

$$\rho = \frac{\text{pixel_size} \times B}{\Delta L}$$

Where ΔL is the number of reciprocal pixels counted between the equator and the first layer line. B denotes the box size of the diffraction pattern in pixels. Then, the number of subunits in each repeat needs to be discovered. All the discernible first maxima (small red and green boxes in figure 2.13B, 2.14B and 2.15B) should be marked on every layer line and fitted into a 2D lattice (red dashed lines in

figure 2.13B, 2.14B and 2.15B) which traverses through half of the maxima, while the reflected lattice (green dashed lines in figure 2.13B, 2.14B and 2.15B) across the meridian links the other half symmetrically. These two lattices in reciprocal space represent the near and far sides of the helical tube in real-space. It is sufficient to assign integer Bessel orders to the first maxima using only one of the lattices, which is spanned by two unit vectors (red arrows in figure 2.13B, 2.14B and 2.15B). The positions and the Bessel orders of all the first maxima must be the linear combinations of those of the unit vectors. To calculate the Bessel orders for maxima on the ends of unit vectors, coefficient C should be firstly figured out using the reciprocal distance from the meridian to the first maxima (ΔD , in reciprocal pixels) along the layer lines and the real-space filament diameters (d , in Angstroms) observed in 2D class averages.

$$C = \frac{\pi \times d \times \Delta D}{pixel_size \times B}$$

The Bessel orders of the first maxima can be estimated from table 2.5 according to C . The values of C for the maxima located to the left of meridian are negative.

n	0	1	2	3	4	5	6	7	8	9	10
 C 	0	1.84	3.05	4.20	5.32	6.42	7.50	8.58	9.65	10.71	11.77
n	11	12	13	14	15	16	17	18	19	20	...
 C 	12.83	13.88	14.93	15.98	17.02	18.06	19.10	20.14	21.18	22.22	...

Table 2.5. A table of first maxima (n) of the Bessel functions $J_n(x)$ for $n = 0$ to $n = 20$ (Andrews, 1981).

In reality, the diffraction patterns might be affected by tilts and real-space distortions, leading to multiple choices of n which seem equally likely. Empirically, high Bessel orders (≥ 10 , maxima further away from the meridian) are widely considered as less reliable than the lower ones. A small error introduced by the estimated filament diameter (d) may lead to large errors especially for high Bessel orders. However, only one unique solution is correct and the user is encouraged to make more attempts to find it out, taking all information on the diffraction patterns into consideration. The lattice geometry along with the Bessel orders calculated unravels the real-space lattice pattern, from which helical twist and rise can be read out (figure 2.13CD, 2.14CD and 2.15CD).

For example, TMV filaments are approximately 180 Å wide (on micrographs with ~1.126 Å pixel size). In the diffraction of TMV (figure 2.13B), first maxima $\Delta D_1 = 11$ (on layer line 1) and $\Delta D_3 = -1$ (on layer line 3) are marked as ends of unit vectors with Bessel orders $n = 16$ ($C = 18.3$) and $n = -1$ ($C = -1.67$) assigned respectively. Thereby, the vector pair can be marked with coordinates $(-1, 3)$ and $(+16, 1)$ in Fourier-space (figure 2.13C). This vector pair are converted into real-space vectors (figure 2.13D)

by a 90° rotation and a mirror operation. Equivalently, flipping the x, y components of the coordinates in Fourier-space gives the real-space coordinates (3, -1) and (1, 16). The new vector pair spans a complete real-space lattice in figure 2.13D. The only strand in this real-space lattice intersects with the azimuthal axis and the z-axis at point (49, 0) and (0, 49) respectively. This lattice determination leads to a full repeat of 3 turns ($16 + 1/3 \times 3 = 49$ subunits), an equivalent of 1.45 Å helical rise. The helical twist can be calculated as 22.04° (360° divided by $16 + 1/3$ subunits in a turn, or a pitch). The helical twist (30°) and rise (22.4 Å) of VipA/VipB filaments can be derived in a similar way (figure 2.14).

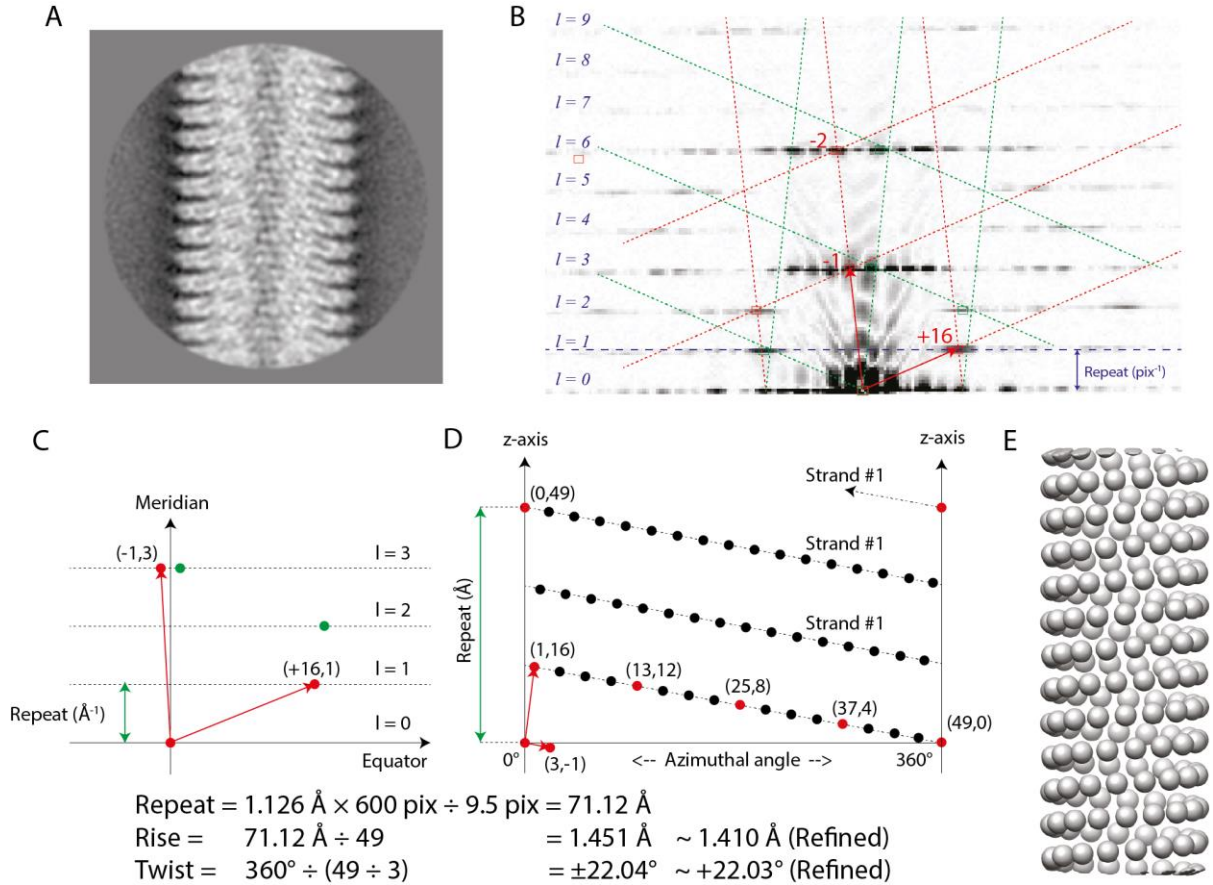


Figure 2.13. Helical lattice determination and initial model generation of TMV. (A) A selected 2D class average. (B) The amplitudes of Fourier-space diffraction pattern of (A). The layer lines (l) are numbered and the vector pair that constructs the helical lattice is indicated with red arrows. The Bessel orders of this vector pair is -1 and +16. A helical lattice and its mirror counterpart are shown with red and green dashed lines respectively. (C) The schematic plot of Fourier-space lattice. (D) The schematic plot of real-space lattice. The base vectors in (C) are converted into a pair of real-space vectors in (D) by a 90° rotation and a mirror operation. The real-space vectors span the whole real-space lattice D. The reciprocal repeat distance (green arrow in C) is the distance in real-space where the lattice folds back on itself on the z-axis (green arrow in D). The helical rise is then calculated by dividing the repeat distance by the number of lattice points in a full repeat (49 for TMV). The helical twist is calculated by dividing 360° by the number of lattice points in a single turn. (E) The corresponding initial spheres-only model (adapted from He & Scheres, 2017).

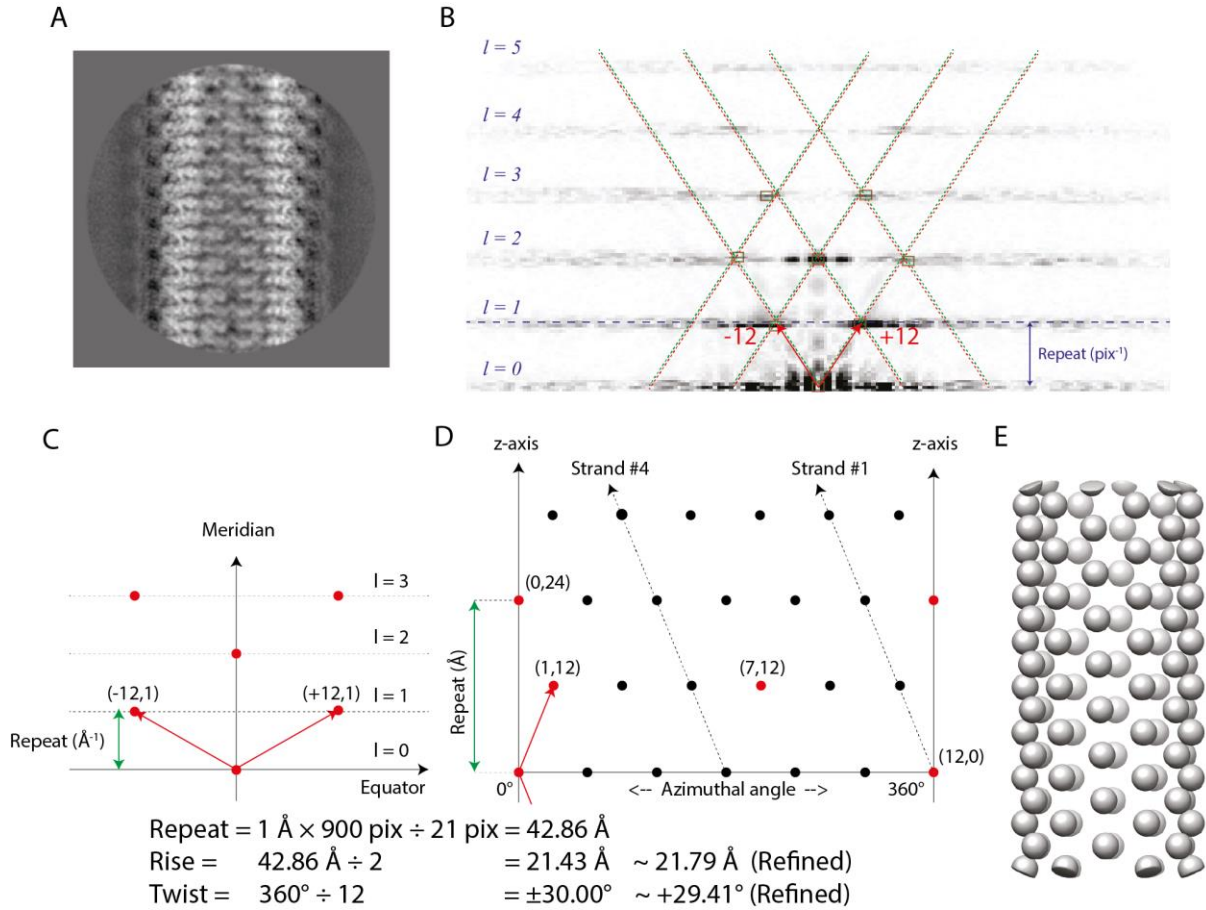


Figure 2.14. Helical lattice determination and initial model generation of VipA/VipB. Refer to figure 2.12 for the explanations of each panel. As is shown in panel (D), VipA/VipB conforms to C_6 point group symmetry and there are 2 subunits (lattice points) in a full repeat on every strand (green arrow in D) (adapted from He & Scheres, 2017).

Simulated lattices are then constructed using ‘relion_helix_toolbox --simulate_helix’ as described in section 2.2.9.2. The diameter of artificial spheres should be large enough (80 Å for both TMV and VipA/VipB) so that their peripheral soft-edges touch with their direct neighbours on the simulated lattice. Reconstructions are subsequently performed using the references with helical symmetry imposed and locally refined at every iteration. The final, refined symmetry parameters are $+22.03^\circ$, 1.410 Å for TMV, and $+29.41^\circ$, 21.79 Å for VipA/VipB and the refinements converges to correct near-atomic resolution structures. In addition, TMV simulated lattices might occasionally show unsatisfactory convergence behaviour. A possible explanation could be that the long-shaped TMV subunits are less similar to the shapes of simulated spheres.

Empirically, very simple helices, such as MamK filaments, often do not generate suitable diffraction patterns for Bessel order assignments and subsequent symmetry estimations (figure 2.16). But the user might escape the trouble of Fourier-Bessel by refining these structures from cylinders, without any prior knowledge of helical symmetry (figure 2.11). On the contrary, complicated helical lattices, such as

TMV and VipA/VipB filaments, contain richer information in their diffractions which makes symmetry determination from Fourier-Bessel analysis rather straightforward. And these helices, prominent as curled 2D surface lattices, hardly yield anything reasonable if reconstructed using cylinders as initial references. Therefore, it seems that in many situations there could be a way to start a helical reconstruction, as long as the suitable method is chosen. MAVS/CARD refinement can be initialised in both ways as its symmetry has moderate complexity (figure 2.11A and 2.15).

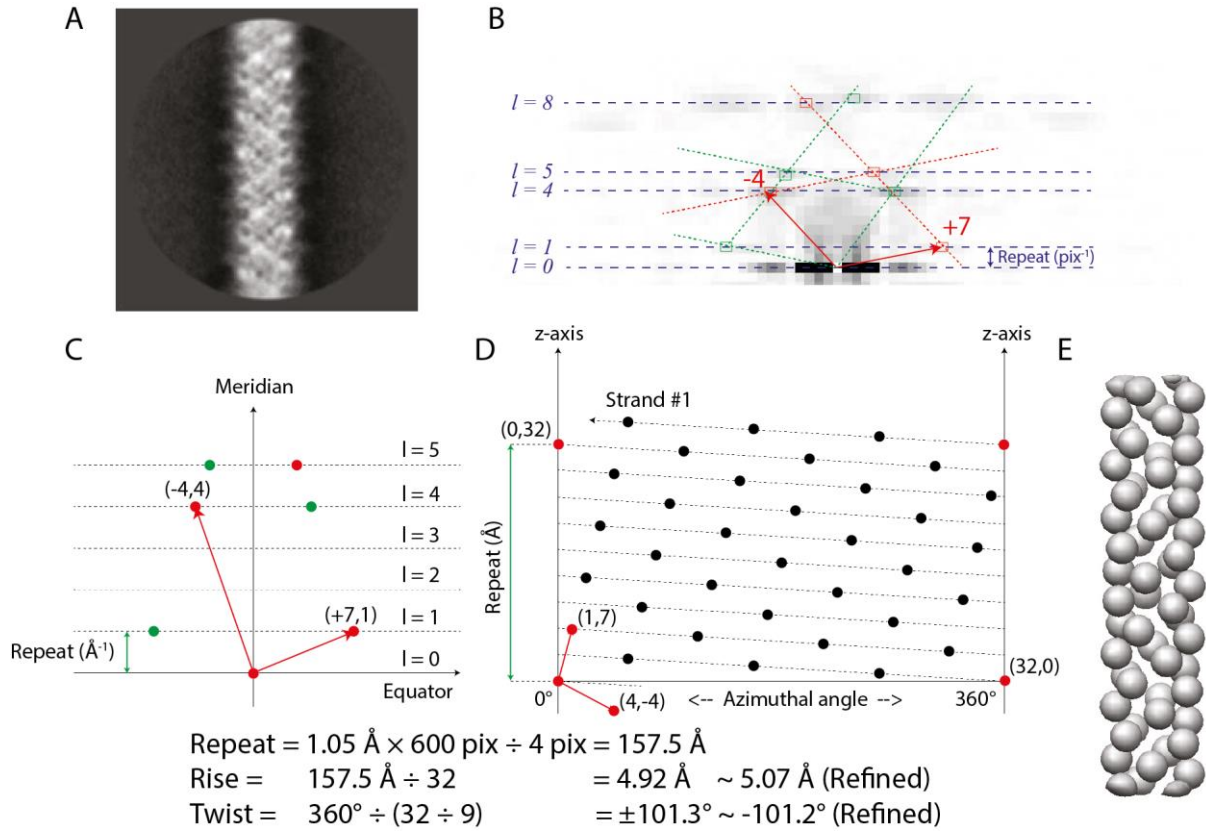


Figure 2.15. Helical lattice determination of MAVS/CARD. Refer to figure 2.12 for the explanations of each panel. As is shown in panel (D), MAVS/CARD has no point group symmetry and there are 32 subunits (lattice points) in a full repeat if the helix is considered as 1-start (green arrow in D).

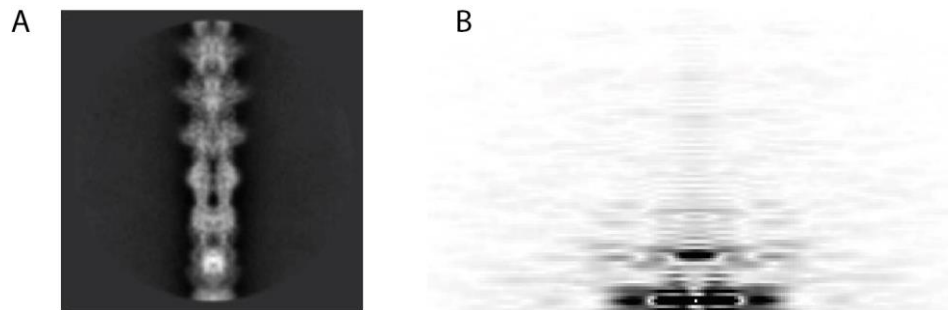


Figure 2.16. A representative 2D class average (A) and the amplitudes of Fourier-space diffraction pattern (B) of MamK. Layer lines and positions of first maxima are not discernible on B.

Alternatively, references from PDB models require prior knowledge of the subunit (or homolog) structure and reliable estimates of helical symmetry. For the published work of MamK filament, Löwe et al., 2016 placed actin-like crystal structures onto the helical lattice using the method described in section 2.2.9.3.

Low-pass filters should still be applied to all helical references, but usually not as stringent as the ones in single-particle analysis. Although filtered to prevent model bias, the references must provide basic information of the helical lattices (i.e. positions of constituent subunits). Empirically, this information still exists if the references are low-pass filtered to $20\text{ \AA} - 30\text{ \AA}$, but it is hardly preserved using the values of $50\text{ \AA} - 60\text{ \AA}$ which are common for single-particle reconstructions. Less stringent values are suggested for complicated helical lattices with smaller protein subunits.

Starting from simulated cylinders or lattices as the initial references, the final distributions of the tilt angles (θ) of all the experimental images in some 3D-refinements have been found bimodal (e.g. at both $\sim 70^\circ$ and $\sim 110^\circ$), not mono-modal around 90° as normally expected. Such results should be considered sub-optimal unless there are specific reasons why the data sets should conform to bimodal tilt (θ) priors. However, the users might not notice since the refined maps are interpretable and the final claimed resolutions are still at sub-nanometre levels. The split distribution of tilts could be explained by the accumulated alignment errors introduced by the simulated initial references. During the first few iterations, the tilt angles (θ) deviate from 90° as the experimental images are compared with the projections of the references which are slightly different from the target structures. The centres of tilt priors are always reset to the tilts estimated in the previous iteration. Therefore, tilt angles might move away from 90° step by step and get stuck in incorrect values when the search ranges are too small to include the true values of tilts. The additional option ‘--keep_tilt_prior_fixed’ may alleviate this problem by fixing the tilt prior centres at 90° in global angular searching steps. The tilt priors are re-centred only when the reconstructions proceed into local searching steps.

For some other cases, it is necessary to run additional rounds of refinement to reach the highest resolution possible since the initial references might be slightly deviated from the radius of convergence, due to the differences in shapes between real proteins and the simulated spheres/cylinders. Even if the first-round reconstruction converges to the correct symmetry and reveals contours of proteins, the subsequent runs may reach higher resolutions, yield better FSC curves and more details of the structures.

2.4.4 Pitfalls of imposing incorrect initial helical symmetry

Initial models and symmetry should be carefully considered so that they sit within range of convergence radius of RELION optimisation algorithm, which is a local optimiser (section 1.4). This statement is also true for other helical processing workflows that are based on IHRSR-like projection matching

(section 1.8). Incorrect structures, albeit with rather high claimed resolutions, have been reconstructed starting from references with poorly estimated helical twist and rise (reviewed by Desfosses et al., 2014; Egelman, 2010, 2014). Therefore, appropriate measures of determining helical symmetry and building initial models are perhaps the most crucial steps in successful helical processing.

To demonstrate the danger of applying incorrect initial helical symmetry, a wrong symmetry ($+64.9^\circ$, 3.63 \AA , C_1) was deliberately imposed on VipA/VipB, instead of the correct symmetry ($+29.41^\circ$, 21.79 \AA , C_6 point group). The wrong symmetry was calculated if the correct helical symmetry on any of the 6 rotationally symmetric strands was converted to a 1-start C_1 representation for the actual 6-start C_6 structure ($\Delta\phi_{\text{wrong}} = (360^\circ + 29.41^\circ) / 6 = +64.9^\circ$, $\Delta z_{\text{wrong}} = 21.79 \text{ \AA} / 6 = 3.63 \text{ \AA}$, section 2.2.1). The initial references were constructed using simulated lattices with corresponding symmetries. Surprisingly, the 3D refinement with such initial settings still converged within 25 iterations (figure 2.16B). The refined helical parameters were $+65.07^\circ$ and 3.76 \AA , lying around their initial values and well included in the specified searching ranges ($+60.0^\circ - +70.0^\circ$, $3.20 \text{ \AA} - 4.00 \text{ \AA}$). Post-processing yielded solvent-corrected resolution of 4.5 \AA (figure 2.17C, black). If one pretends that the correct symmetry is unknown, the above indicators suggest that the result could still be a reasonable near-atomic structure. However, on the density map figure 2.17B, one cannot find any clues of protein-like features, such as α -helices or β -sheets which should be clearly observed under the claimed resolution. Its side-view 2D projections (figure 2.17F, lower half) have very much different appearances with streaky densities in solvent areas compared to those of the correct map (figure 2.17F, upper half) and the 2D class averages of segments (figure 2.17E). In addition, its FSC curve (figure 2.17C, black) is not smooth anywhere, which rarely happens in perfect reconstructions.

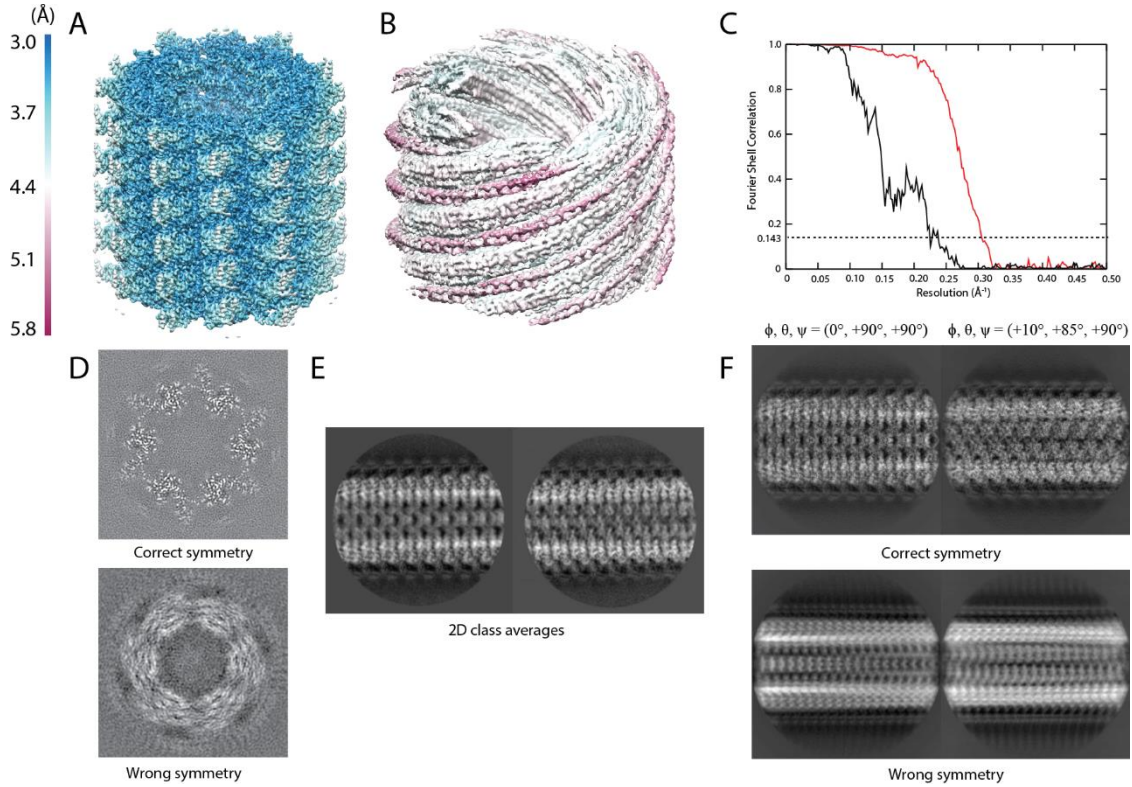


Figure 2.17. Effects of imposing the correct and wrong helical symmetries for the VipA/VipB data set. (A) Reconstruction with the correct helical twist and rise imposed ($+29.41^\circ$, 21.79 \AA , C_6). (B) Reconstruction with the wrong helical symmetry imposed ($+64.9^\circ$, 3.63 \AA , C_1). The maps in (A) and (B) are coloured according to local-resolution estimates (from 3.0 \AA in blue to 5.8 \AA in maroon). (C) FSC curves between two independently refined half-sets for the refinement with the correct helical symmetry (red) and the wrong helical symmetry (black). (D) Central z slices through the unsharpened map from the refinement with the correct and wrong helical symmetry. (E) Two typical 2D class averages of VipA/VipB segments. (F) The comparison of 2D projections of correct and wrong density maps at Euler angles $(\phi, \theta, \psi) = (0^\circ, +90^\circ, +90^\circ)$ and $(+10^\circ, +85^\circ, +90^\circ)$. The 2D projections were generated using the standalone executable ‘--relion_project’ (adapted from He & Scheres, 2017).

In reality, such mistakes may happen unnoticed. The authors of the original MAVS/CARD paper (Xu et al., 2014) firstly made errors in estimations of helical symmetry (pointed out by Egelman, 2014). They had tried hard to interpret a wrong density map which was not at all similar to the corrected results (Xu et al., 2015). In addition, a near-atomic nominal resolution in 3D refinements is not a guarantee for success either, since FSC only measures consistency between the data half-sets. Incorrectness is consistent in both half-maps imposed with similarly wrong symmetry. Incorrect reconstructions are unavoidable in RELION with initially wrong symmetry since both MAP-EM and searches of helical parameters are local optimisers. Other widely used data processing packages, such as IHRSR (Egelman, 2000) in SPIDER (Frank et al., 1996), SPARX (Behrmann et al., 2012) and FREALIX (Rohou and Grigorieff, 2014), all suffer from the same problem. Exhaustive searches of all possible helical symmetries (e.g. SPRING, Desfosses et al., 2014) might cover the true twist and rise but extra costs

include much more expensive computations and more free parameters to tune in 3D reconstructions. Eventually, it is still required to manually find the correct symmetry (if it has been considered) through inspections of a series of independent 3D reconstruction results. Tilt-pair analysis (Henderson et al., 2011) serves as a robust validation method for both single-particle and helical reconstructions. Should it have been performed for the structures mentioned at the beginning of the paragraph, the incorrectly reconstructed results could have been avoided from being published.

The most convenient way (if possible) to validate a helical reconstruction is to examine protein-like features in the final density map. For maps with high claimed resolutions, this can be easily done by identifying clearly separated α -helices, β -strands and amino-acid side chains. The density map must lead to a fitted, stereo-chemically sound atomic model. Such criteria are widely accepted since the whole cryo-EM community are aiming at more near-atomic resolution structures in the recent years. Summarised below are some useful tips which might rule out most possibilities of imposing wrong symmetry, which can also be used as evidence (but not sound proof) of the correctness of medium to low resolution reconstructions.

Firstly, one should attempt to build initial models with both cylinders and Fourier-Bessel methods if there are still doubts about the estimated helical symmetry. If the helical lattice is relatively simple, cylinders can lead to unbiased reconstructions from which the true helical parameters can be determined by eye inspections along with searching (e.g. MAVS/CARD and MamK, figure 2.11). On the other hand, Fourier-Bessel indexing is more feasible for complicated lattices and require users' expertise. The indexing process generally leads to better understanding of the helices and thus helps to avoid mistakes. Secondly, helical symmetry must be terribly wrong or at least badly optimised (for a specific class) if the twist and rise converge to the edges of searching ranges during 3D reconstructions. This happened when the wrong symmetry of MAVS/CARD, $(+53.6^\circ, 16.8 \text{ \AA}, C_3)$ published by Xu et al., 2014, was applied in helical reconstructions of RELION. Thirdly, the side-view 2D projections of the refined map should match with the 2D class averages of raw segments. This provides evidence of consistency between the reconstructed volume and the input images. In addition, doubts of wrong symmetry should also be raised if either angular distributions (alignments) of segments (recorded in .bild files) or the post-processing FSC curves appear abnormally spiky.

The handedness of the helix in the final density map can be incorrect as the information of chirality is not recovered from the single-particle reconstructions of 2D input images. The handedness of the final map might be fixed by that of the initial reference. In this case, there is no need to perform the reconstruction from start. The final 3D map can be simply flipped across the helical axis (z) in order to obtain the true conformation. There are several ways to discover the handedness with cryo-EM/ET imaging and data processing if no prior knowledge is available. Firstly, the chirality information is preserved in cryo-ET sub-tomogram averaging, which sometimes used in preliminary experiments

before cryo-EM analyses of helices are performed. Secondly, the handedness can be figured out by cryo-EM imaging of a tilted stage. Collected from a stage tilted at specific small angles, the 2D images of helices may show characteristic left-right, serrated smooth asymmetry across the outer surfaces of helices. The handedness can be determined based on the tilt directions and angles, along with the sense of asymmetry observed from the images. Alternatively, simulated projections of both left and right-handed helical lattices are generated at a series of tilt angles and compared with the experimentally tilted images in order to determine the handedness (Finch, 1972). In addition, if the resolution of the final map is high enough to resolve the densities of α -helices, the handedness can be confirmed as correct if these secondary structures appear as right-handed spirals.

2.4.5 Applying helical symmetry in Fourier-space

Simulated data sets of TMV were used to illustrate the significance of applying Fourier-space helical symmetry in 3D reconstructions. Using the method introduced in section 2.2.9.3, the 3D reference of TMV was constructed in a 300-pixel particle box using PDB entry 4UDV (Fromm et al., 2015). Then two simulated data sets of 15,000 TMV subunits on 10 filaments were generated using 5 and 30 as the numbers of asymmetrical units (H) respectively, resulting in two particle stacks containing 3,000 and 500 segments. For simplicity, all tilt angles (θ) of these segments were set to 90° . In-plane rotations (ψ) were randomised for the 10 filaments and the Gaussian white noise level was set to 20.

Three experiments of 3D reconstructions were performed using combinations of different H and the simulated segment stacks (figure 2.18). As the first steps of the reconstructions, the experimental segments were inserted back into 3D according to the central-slice theorem (figure 1.7). This could be done using the above data sets because the alignment information was known during segment simulations. Thus, the simulated 3D reconstructions are reference-free and non-iterative. Afterwards, helical symmetrisation was applied to the reconstructions in Fourier-space according to H. The following command generated the reconstruction ‘reconA.mrc’ in figure 2.18B using ‘3000segments_H5.star’ as the STAR file of simulated segments. Reconstructions in figures 2.18CD were generated similarly.

```
relion_reconstruct --i 3000segments_H5.star --o reconA.mrc --nr_helical_asu
5 --helical_rise 1.408 --helical_twist 22.03 --angpix 1.062
```

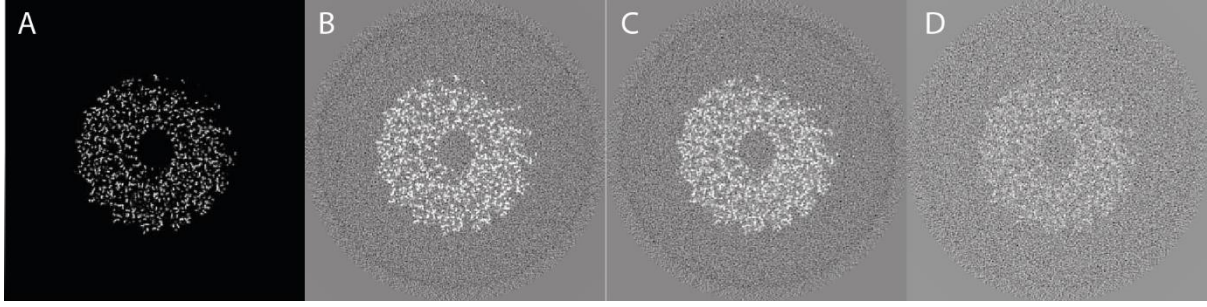


Figure 2.18. Applying Fourier-space helical symmetry using different numbers of asymmetrical units (H) in the simulated TMV data set consisting of 15,000 subunits. The central z slices of the original and averaged half-maps are shown for all test cases of simulated 3D reconstructions. **(A)** The original density map converted from PDB entry 4UDV for data simulations. **(B)** Data set simulated with $H = 5$ (3,000 segments) and correctly refined with $H = 5$ ($5 \times 3,000 = 15,000$ subunits averaged). **(C)** Data set simulated with $H = 30$ (500 segments) and correctly refined with $H = 30$ ($30 \times 500 = 15,000$ subunits averaged). **(D)** Data set simulated with $H = 30$ (500 segments) and incorrectly refined with $H = 5$ ($5 \times 500 = 2,500$ subunits averaged).

The reconstructed maps were symmetrised in Fourier-space H times according to the helical twist and rise values when $H \geq 2$ (section 2.2.2). As is seen in figure 2.18B&C, the signal-to-noise levels are correctly estimated when the H used in segment simulations (extractions) matches with the H used in simulated 3D reconstructions. However, noise levels are not properly averaged down if the H in simulated 3D reconstructions is considerably lower than the H used in segment simulations (figure 2.18D). Therefore, Fourier-space symmetrisation is crucial for the correct estimations of signal-to-noise ratios (SNRs) at all spatial frequencies.

2.4.6 Inter-box distances and mask diameters

Cutting overlapping segments gives rise to an additional free parameter, the inter-box distance (or equivalently the number of asymmetrical units H in each new segment, given a known helical rise). Since helices span across segment boxes, the determination of mask diameters is not straightforward compared to the reconstructions in single-particle analysis, where the sizes of circular (spherical) masks are set to allow some extra spaces around the particle densities. To investigate the effects of these two user-defined parameters, tests were performed on both TMV and MAVS/CARD data sets. On one hand, 3D refinements had fixed circular mask diameters with varying inter-box distance according to different number of asymmetrical units provided in segment extractions (figure 2.19AB). On the other hand, more refinements were performed with fixed inter-box distance but varying mask diameters (figure 2.19CD).

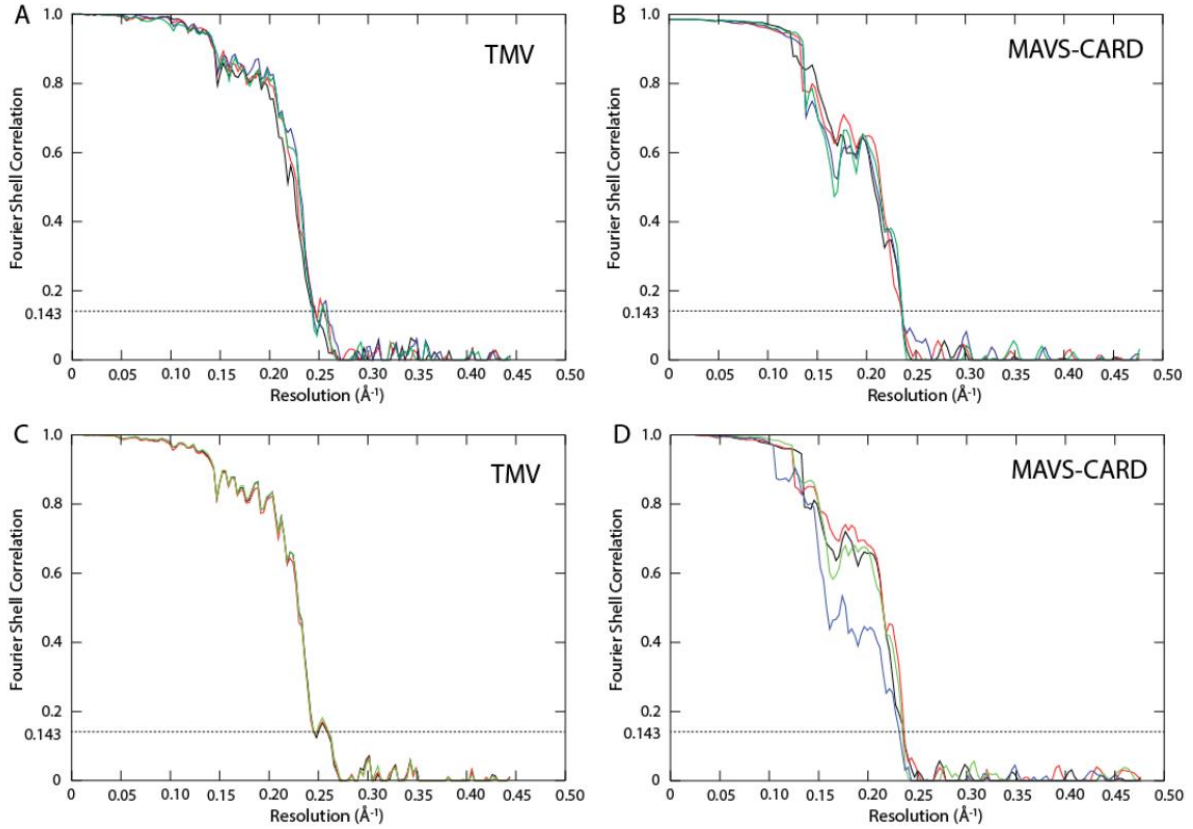


Figure 2.19. Refinements with varying inter-box distances. (A) FSC curves for TMV refinements with inter-box distances corresponding to 5 (black), 15 (red) and 30 (blue) and 75 (green) new asymmetrical units per segment, corresponding to 2.1 %, 6.3%, 12.5% and 31.3% of the box size, respectively. The rise of TMV is 1.41 Å and the box size is 338 Å. The mask diameter in these refinements was fixed at 90% of the box size. (B) FSC curves for MAVS/CARD refinements with inter-box distances corresponding to 2 (black), 8 (red), 14 (blue), and 20 (green) new asymmetrical units per segment, corresponding to 3.2%, 12.9%, 22.5% and 32.2% of the box size, respectively. The rise of MAVS/CARD is 5.07 Å and the box size is 315 Å. The mask diameter in these refinements was fixed at 90% of the box size. (C) FSC curves for TMV refinements with mask diameters corresponding to 90% (black), 85% (red) and 80% (green) of the box size. The inter-box distance in these refinements was fixed to yield 30 new asymmetrical units per segment. (D) FSC curves for MAVS/CARD refinements with mask diameters corresponding to 90% (black), 80% (red), 70% (green) and 60% (blue) of the box size. The inter-box distance in these refinements was fixed to yield 8 new asymmetrical units per segment (adapted from He & Scheres, 2017).

A smaller inter-box distance yields more helical segments but the total amount information contained in the extracted segments remains invariant. More segments per unit-length on filaments allow more orientation and translational parameters to be assigned to the segments during alignment. This higher degree of freedom might benefit flexible or bending filaments. Moreover, smaller diameters of circular (and spherical) masks in 2D (and 3D) reduce correlations of alignments among neighbouring overlapping segments. When the filaments are highly bent, a smaller diameter masks out more signals

near box edges, cutting the segments into shorter and straighter ones, which might improve alignments. Thus, for TMV filaments which are almost perfectly straight, qualities of refinements remained largely the same no matter how these two parameters change (2%–30% inter-box distances, 80%–90% circular masks). For bending MAVS/CARD filaments, it is expected that shorter inter-box distance and smaller masks could describe their flexibilities more accurately. However, experiments showed that large inter-box distances (up to 30% the box size) were still equally satisfactory even for bending filaments of MAVS/CARD. It can be explained that the majority of highly curved MAVS/CARD segments are firstly excluded by auto-picking algorithm (section 2.2.7) and subsequently discarded after selections of good-quality segments according to 2D class averages. The effects of various mask diameters agreed with predictions on MAVS/CARD segments. A mask of 80% box size was marginally better compared to 70% or 90%. But it does not necessarily mean that smaller values are always more advantageous. The resolution of MAVS/CARD reconstruction was evidently decreased with a 60% circular mask since the signals remained were insufficient for accurate alignments.

The results above imply the following empirical requirements on the initial settings of segment box size, inter-box distance and circular mask diameter for high-quality reconstructions. The box size must be large enough to accommodate helical segments and their circular masks so it is usually set to at least $1.5\times$ the filament widths. A bigger box size leads to an advantage in finer samplings of resolution shells in Fourier-space but the maximum limit of box size depends on total amount of memory available on the CPU or GPU cluster. Its minimum should still ensure accurate alignments of segments (together with a suitable mask diameter). Circular mask diameter must be greater than actual filament widths and less than $\sim 90\%$ the box size to effectively prevent real-space artefacts explained in section 2.2.5. Smaller masks (or box sizes) might be more appropriate for flexible helices. As for the inter-box distance, it hardly affects the final resolutions of reconstructions of straight filaments with perfectly stable symmetries therefore higher values are recommended to reduce the number of segments extracted to gain linear speed-up of computations. Practically, its maximum should be around 30% the box size, as a number of central z sections (in this case, $1 - 30\% \times 2 = 40\%$) are required for local searches and impositions of helical symmetry in real-space (section 2.2.3 and 2.2.4). Inter-box distances of 5%-10% lead to more segments and their alignment parameters for flexible or straight filaments with varying (or mixed) symmetries. Scaling the mask diameter may tune the dependencies of alignment parameters among neighbouring segments. Since all segments are extracted at (rounded) integer coordinates on micrographs in RELION, tiny inter-box distances (e.g. $\leq 2\%$) are not recommended since it might make interpolation errors more evident.

2.4.7 Applying helical symmetry in real-space

The importance of helical symmetrisation in real-space is demonstrated by 3D refinements of MamK helix in figure 2.20. These helical segments were picked at an inter-box distance of 1 helical rise (~ 52.2 Å, $H = 1$) therefore the Fourier space symmetrisation was not applicable. Two pairs of reconstructions were performed using the cylindrical and proper MamK helical references respectively. All the references were low-pass filtered to 30 Å. For 3D reconstructions starting from either type of initial references, real-space helical symmetry improved the refinement results in terms of both claimed resolutions and resolved details on the density maps. In particular, the real-space helical symmetry contributed to the reconstruction from the featureless cylinder considerably (red and black curves in figure 2.20A). In addition, the application of real-space symmetry compensated for the lack of information in cylindrical starting model and delivered similar results (black curve in figure 2.20A) compared to the refinement from the helical reference of MamK (blue curve in figure 2.20A). The application of helical symmetry in real-space improves the quality of density maps and yields better convergence.

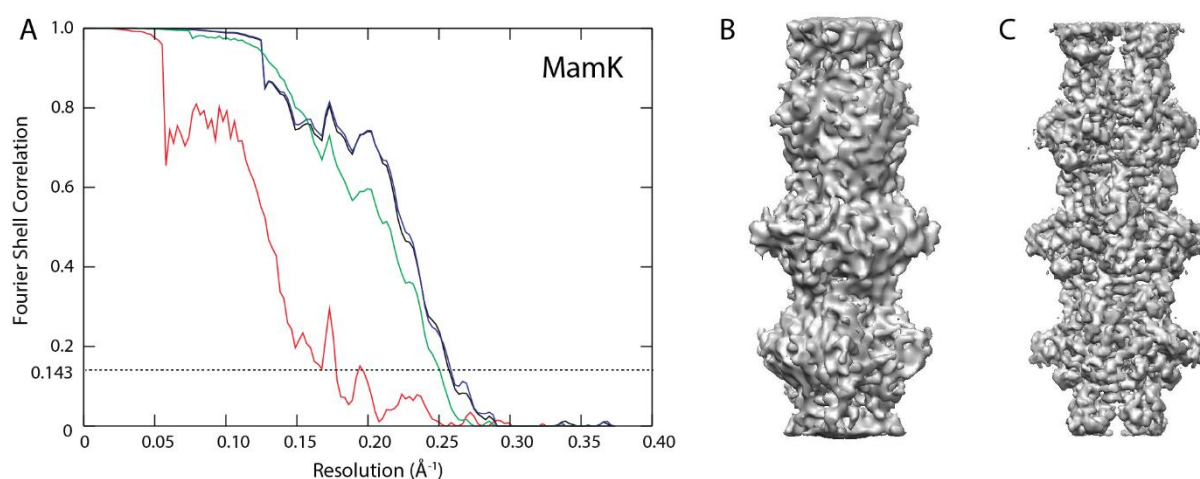


Figure 2.20. The importance of real-space helical symmetry in 3D refinements of MamK. The post-processed FSC curves (A) show the refinements from the featureless cylinders (180 Å wide) ended up at 5.7 Å (red) and 3.9 Å (black) without or with real-space helical symmetry applied. The respective side views of these two density maps are shown in panels (B, 3.9 Å) and (C, 5.7 Å). The refinements from MamK helical references ended up at 4.1 Å (green) and 3.9 Å (blue) without or with real-space helical symmetry applied.

2.4.8 Local searches of helical symmetry

As discussed in sections 2.4.3, initial helical symmetry and 3D references must lie within radius of convergence of the optimisation algorithm in order to yield correct helical reconstructions. Helical twist and rise can be locally optimised in real-space at the end of every iteration in 3D reconstructions. Refinements with different initial twist and rise values (table 2.6) were performed on MamK to demonstrate the robustness of such local optimisations. All the jobs started from the same 180 Å wide

featureless cylinder low-pass filtered to 30 Å. Convergence criteria here, which remain the same as in the single-particle reconstructions in RELION, consider current resolution estimates, accuracies and accumulated changes of orientations and translations. Fortunately, these criteria also guarantee convergence of helical symmetries for all the tested refinements, no matter whether the optimised symmetries were correct or not (figure 2.21A&B). Out of all combinations of twist and rise tested for MamK, initial values within ranges of $+19^\circ - +25^\circ$ and $42 \text{ Å} - 62 \text{ Å}$ were observed leading to correctly converged reconstructions. Incorrectly refined symmetries could be identified by weird densities, low resolution estimates and twist/rise values lying at the edges of searching ranges. Moreover, the side-view projections of incorrect structures are strikingly different from any 2D class averages of helical segments. Interestingly, some initial symmetries, far from the perfect values ($+23.77^\circ$, 52.12 Å), could still be optimised correctly but the resolution estimates of such cases (e.g. $+17^\circ$, 37 Å ; $+17^\circ$, 42 Å and $+21^\circ$, 67 Å) were slightly worse than the optimal results. It could be explained that in 3D refinements, exhaustive orientation searches had been stopped too early when the helical symmetries were still experiencing large changes to reach the correct values. One solution is to keep angular searches global for finer angular steps ($\leq 0.9^\circ$) in 3D auto-refinements. Single-class 3D classifications, with exhaustive searches always to be performed for an arbitrary number of iterations, could be another approach to approximate the correct symmetry parameters. Or generally, if the initial helical twist and rise values seem considerably different from correct ones, additional rounds of 3D refinement should be performed starting from the symmetry and the structure refined from the previous round.

Twist → Rise ↓	17° [14-26°]	19° [16-26°]	21° [18-26°]	23.746° [21-26°]	25° [21-29°]	27° [21-31°]	29° [21-33°]
37 Å [32-57 Å]	23.63° 52.24 Å 6.47 Å *	23.73° 52.13 Å 5.44 Å	23.72° 52.14 Å 5.44 Å	23.58° 52.28 Å 6.70 Å *	24.23° 52.14 Å 6.05 Å *	23.60° 52.16 Å 5.44 Å	28.54° * 37.03 Å * 8.93 Å *
42 Å [37-57 Å]	23.18° 52.26 Å 6.36 Å *	23.73° 52.11 Å 4.69 Å	23.64° 52.09 Å 4.75 Å	23.74° 52.12 Å 5.36 Å	20.97° * 46.09 Å * 6.70 Å *	23.69° 52.13 Å 4.81 Å	31.96° * 51.02 Å * 8.53 Å *
47 Å [42-57 Å]	23.71° 52.10 Å 4.75 Å	23.73° 52.10 Å 4.81 Å	23.72° 52.10 Å 4.63 Å	23.73° 52.12 Å 4.63 Å	23.73° 52.10 Å 4.63 Å	23.73° 52.11 Å 4.75 Å	32.31° * 52.26 Å * 7.36 Å *

52.15 Å [47-57 Å]	23.65°	23.73°	23.76°	23.77°	23.74°	23.73°	23.71°
	52.19 Å	52.11 Å	52.12 Å	52.12 Å	52.12 Å	52.12 Å	52.13 Å
	6.25 Å *	4.81 Å	4.63 Å	4.69 Å	4.63 Å	4.63 Å	4.63 Å
57 Å [47-62 Å]	23.69°	23.72°	23.73°	23.75°	23.74°	23.71°	23.73°
	52.13 Å	52.11 Å	52.16 Å	52.12 Å	52.12 Å	52.12 Å	52.12 Å
	4.75 Å	4.69 Å	4.75 Å	4.63 Å	4.69 Å	4.63 Å	4.63 Å
62 Å [47-67 Å]	16.00° *	23.73°	23.69°	23.73°	23.73°	23.76°	23.74°
	53.14 Å	52.18 Å	52.61 Å	52.13 Å	52.12 Å	52.14 Å	52.18 Å
	7.36 Å *	4.87 Å	6.15 Å	4.63 Å	4.81 Å	4.75 Å	5.36 Å
67 Å [47-72 Å]	13.59° *	18.29° *	22.82°	23.14°	22.98°	24.78° *	31.24° *
	60.39 Å *	59.40 Å *	53.91 Å	53.34 Å	55.35 Å *	63.00 Å *	62.99 Å *
	7.98 Å *	8.34 Å *	7.66 Å *	6.82 Å *	7.66 Å *	8.34 Å *	7.98 Å *

Table 2.6. Local optimisation of helical parameters for MamK. Starting from different initial estimates for the helical twist (top row, with search ranges between square brackets) and rise (first column, with search ranges between square brackets), the final refined values for the helical twist (top), rise (middle) and resolution (bottom) are shown. Asterisks and red colours are used to indicate values that deviate more than 1° or 1 Å from the correct ones, which are shown in bold and green in the middle (adapted from He & Scheres, 2017).

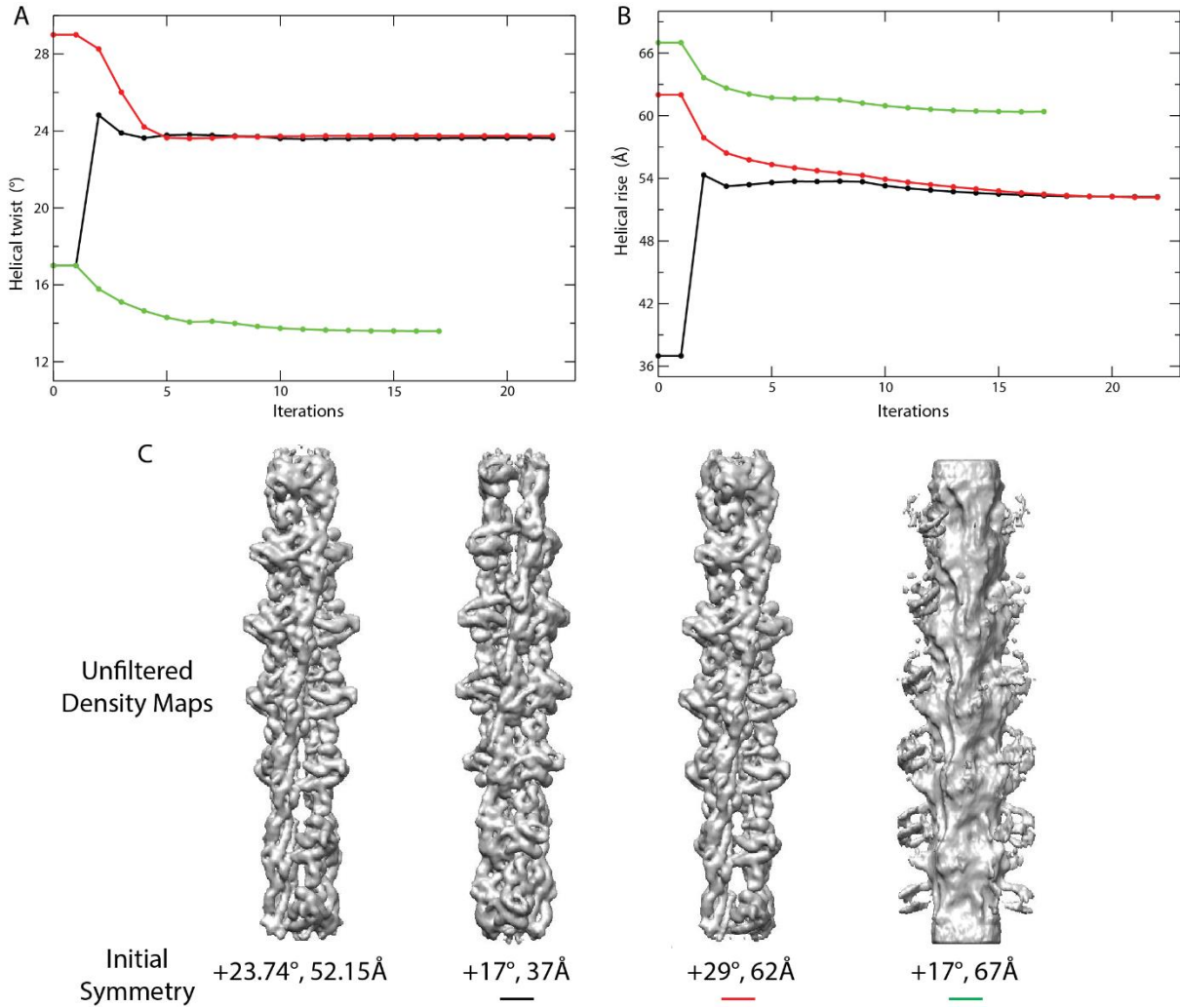


Figure 2.21. Convergence of helical parameters within 3D auto-refinement runs. Plots of the refined helical twist (A) and rise (B) for each iteration of three different 3D auto-refinement runs: starting from initial twist and rise settings of: +17° and 37 Å (black); +29° and 62 Å (red); and +17° and 67 Å (green). The first two refinements converged onto the correct values, whereas the last did not. (C) shows the side views of the unfiltered reconstructed maps starting from the aforementioned combinations of twist and rise values. Note that the 2nd side view from the left is flipped upside-down with respect to the 1st (adapted from He & Scheres, 2017).

2.4.9 Angular and translational priors

In the original IHRSR algorithm (Egelman, 2000) incorporated into SPIDER (Frank et al., 1996), the orientations and translational priors are expressed as user-defined searching ranges. These ranges are equivalent to top-hat shaped priors in the statistical learning framework of RELION. That means, orientations and translations within the ranges have weights all set to 1 while those outside the specified ranges have weights of 0. Theoretically, reconstructions can be improved if the prior shapes match better with the actual distributions of the underlying parameters. The improvements can be more prominent when the input images have higher noise levels. RELION makes use of priors with shapes

of Gaussian distributions which are expected to yield better results compared to top-hat priors. In addition, the statistical framework has the potential to update the optimal search ranges according to the data set. The update scheme in RELION decreases the angular searching ranges and re-estimates the widths of translational priors at every iteration in 3D auto-refinement.

Comparisons between Gaussian and top-hat priors were performed on MamK and VipA/VipB filaments. The widths of top-hat priors were set to $\pm 3\sigma$ of the corresponding Gaussian priors at all times. Figure 2.22 shows the FSC curves of these 3D refinements after post-processing. For the VipA/VipB data, Gaussian priors made notable improvements over top-hat ones. However, both priors led to similar results for MamK filaments. The flexibilities of MamK bending filaments seem to be the main obstacle to getting a higher-quality map.

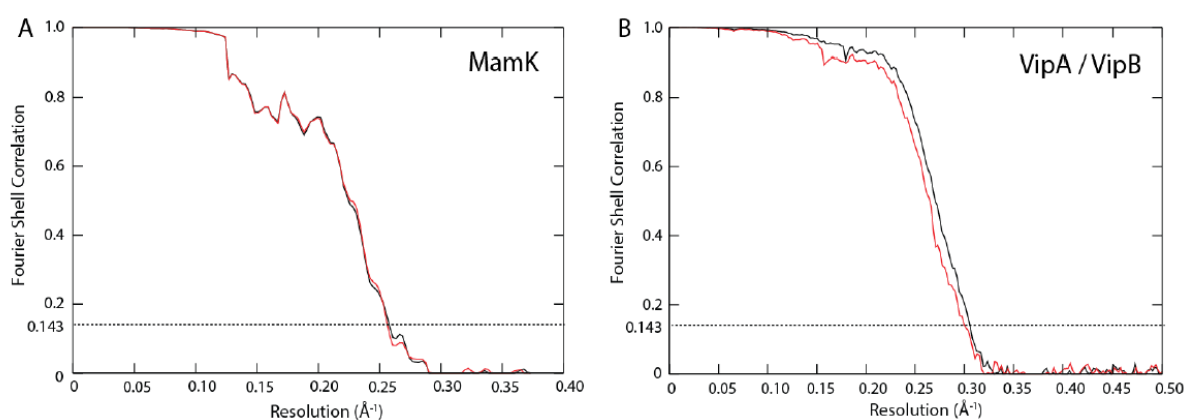


Figure 2.22. Comparison of different priors. FSC curves of post-processed maps for MamK (A) and VipA/VipB (B), using either Gaussian priors (black) or top-hat priors (red) in the refinement (adapted from He & Scheres, 2017).

As discussed in section 2.2.6, in-plane angular (ψ) priors are bimodal for all helical segments if the target structure has no or only C_n point group symmetry. In such cases, the helical segments coming from the same filament should adopt the same polarity. If the additional option ‘--helical_sigma_distance X’ is enabled in 3D reconstructions, the weights of the two components of the bimodal ψ priors are assigned to each segment according to the number of segments conforming to either of the polarities on the given filament. Here, a 3D refinement of MamK was performed to estimate how many segments adopted the majority polarity using the bimodal ψ priors. Total numbers of segments with the majority polarity are listed in table 2.7 and 2.8 for each iteration of global angular searching in MamK and TMV refinements. About 2/3 of all MamK segments were assigned with the majority polarity of their corresponding filaments after the first iteration. The percentage increased to 93.5% before the refinement proceeded into local angular searching steps. The outliers, i.e. the rest 6.5% segments, were subsequently flipped to the majority polarity. An experiment was also performed to compare the MamK refinements with and without such polarity correction. Figure 2.23A shows that

the flipping slightly improved the final post-processed resolution of MamK by a few Fourier-shells in FSC curves.

# of iterations	# segments with majority polarity	# segments with opposite polarity	% of outliers
1	88,541	44,696	33.5%
2	98,545	34,692	26.0%
3	115,333	17,904	22.4%
4	123,850	9,387	7.0%
5	123,832	9,405	7.1%
6	123,814	9,423	7.1%
7	124,533	8,704 (flipped)	6.5% (flipped)
8+ (Local searches)	133,237 (total)	-	-

Table 2.7. The total numbers of MamK segments with majority and the opposite polarities in each iteration of global angular searching in 3D refinements.

The polarity statistics was also analysed for TMV refinements (Table 4.8). Compared to MamK, the polarities of TMV segments were much less ambiguous and only ~0.2% of all segments were considered as outliers when the refinement went into local angular searches. Thus, flipping the polarities of the outliers hardly improved the final reconstructions of TMV (figure 2.23B). The results could be explained by the richer information contained in the TMV helical lattice and the less flexibilities of TMV filaments.

# of iterations	# segments with majority polarity	# segments with opposite polarity	% of outliers
1	1,684	77	4.4%
2	1,756	5	0.3%
3	1,753	8	0.4%
4	1,759	2	0.1%
5	1,758	3	0.2%
6	1,758	3 (flipped)	0.2% (flipped)
7+ (Local searches)	1,761 (total)	-	-

Table 2.8. The total numbers of TMV segments with majority and the opposite polarities in each iteration of global angular searching in 3D refinements.

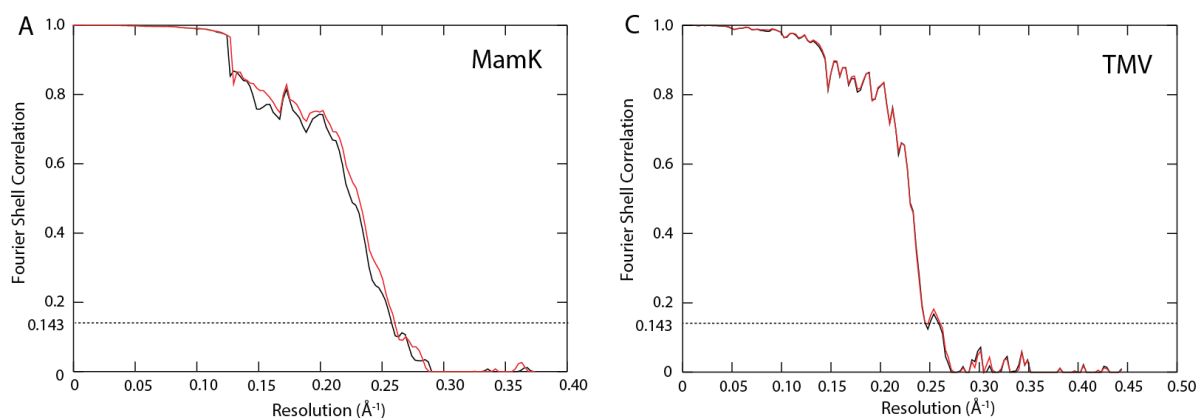


Figure 2.23. The comparisons of MamK (A) and TMV (B) refinements with and without polarity corrections. For both panels, red curves show the 3D refinements with bimodal ψ prior corrections switched on by the additional option ‘--helical_sigma_distance 2’ (explained in section 2.2.6) and with polarities of the outliers flipped in the last iteration of global angular searches (table 2.7 and 2.8). Black curves show the 3D refinements with fixed 50%-50% unbiased bimodal ψ priors during all iterations of global angular searches.

Therefore, the additional option ‘--helical_sigma_distance X’ (section 2.2.6) is recommended for bending and less rigid filaments for possible polarity corrections. In addition, polarity statistics (table 2.7 and 2.8) can provide a clue whether the data set conforms to D_n symmetry (without polarity). In such cases, distributions of opposite polarities are expected to be around 50%-50% with no accumulations in either side during all iterations of global alignment.

In fact, the ground truth of the polarity of an experimental segment is never known. If only several segments have been extracted from a filament, there is a possibility that the algorithm assigns wrong polarities to the majority, flips the outliers with correctly assigned polarities and makes the reconstruction even worse, especially when the filaments are flexible and short, recorded on images with higher noise levels. In order to make polarity correction works, it is advised to pick as many longer filaments as possible, which lowers the joint probability of making such mistakes. Otherwise, the naive, 50%-50% fixed bimodal ψ priors should be adopted.

2.4.10 Prevention of over-fitting in 3D auto-refinements

In general, over-fitting in cryo-EM means that some occurrences of pure noise are treated as signals of molecular features during data processing. When over-fitting happens, signal-to-noise ratios (SNRs) are incorrectly estimated to higher values at some spatial frequencies and the claimed final resolution of the density maps might be exaggerated. In 3D reconstructions, the orientation (ϕ , θ , ψ) and translational (Δx , Δy) parameters of the experimental 2D particles are iteratively refined through comparisons with the gradually improved 3D reference model ($V^{(n)}$, section 1.4). The raw 2D images contain more random noise than feature signals especially at higher spatial frequencies. For this reason, the parameters of 2D images may slightly biased to fit noise in the reference as features. The iterative process accumulates such false features in the updated models, leading to over-fitting in the final reconstruction (Chen et al., 2013).

Several existing measures have been taken to limit over-fitting in single particle cryo-EM reconstructions. Stringent low-pass filters (e.g. 50 Å – 60 Å for ordinary single particles) are applied to get rid of high frequency data in the initial references and alleviate model bias (where the parameters are biased to fit existing features in the initial reference but not generating true features contained in the experimental data set, Henderson, 2013). In 3D auto-refinements, the 2D images are divided into two random halves, and the reconstructions are performed independently. Fourier-shell correlations (FSC) and current resolution estimates are calculated between the two half-maps at the end of each iteration. In addition, if an atomic model is built according to the reconstructed density map, a Fourier-shell correlation (FSC) between the model and the map could further validate the signal-to-noise ratios (SNRs) and the overall resolution estimates.

In helical reconstruction, overlapping segment boxes are cut off along helical filaments from 2D micrographs. Two specific measures are taken in order to prevent overfitting from overlapping signals in 3D auto-refinements: randomly dividing the data set into halves filament-wise (section 2.2.8), and limiting the translational priors within a single rise along the helical axis of every segment (without accumulating the translations during the iterative reconstruction process, section 2.2.6 and 2.4.9). The latter is also adopted for all 3D reconstructions (including both 3D classifications and auto-refinements).

Four tests of 3D refinements were performed on TMV data set to evaluate these two measures (figure 2.24). The refinement with both measures adopted (black line) led to a final resolution of 4.1 Å after post-processing, which agreed with the visible features of the density map, and the FSC curve dropped close to zero beyond 3.8 Å. This resolution was lower than the one reported in figure 2.28 and table 2.1 since non-polished segments were used in all the four tests in this section. The refinement with neither safeguard measures (green) caused severe over-fitting indicated by a heavily inflated FSC curve. Here, division of half-sets was performed segment-wise and the translational priors were isotropic in 2D, both of which were the same as in ordinary single-particle analysis. In this case, neighbouring segments along the same filament are separated into different half-sets, which introduces redundant correlations between the same parts of experimental images. Furthermore, the translations along the helical axis accumulate during the isotropic 2D translational searches in the iterative refinement, therefore some segments may be translated so much that they can superimpose onto the neighbouring segments in the other half-set. These two factors severely inflate the (green) FSC curve all the way to Nyquist frequency. The third refinement (blue line), with filament-wise randomisation and isotropic priors, demonstrated no signs of over-fitting. However, it had a lower claimed resolution compared to the first test (black line), probably due to worse translational priors. The last experiment (red line) adopted segment-wise division with helical translational priors, where slight over-fitting effects could be observed. The inter-box distance used in all the four cases was 42.2 Å (37.5 pixels, $H = 30$). It is expected that the effects of over-fitting could be more evident if smaller inter-box distances were used in the segment extraction.

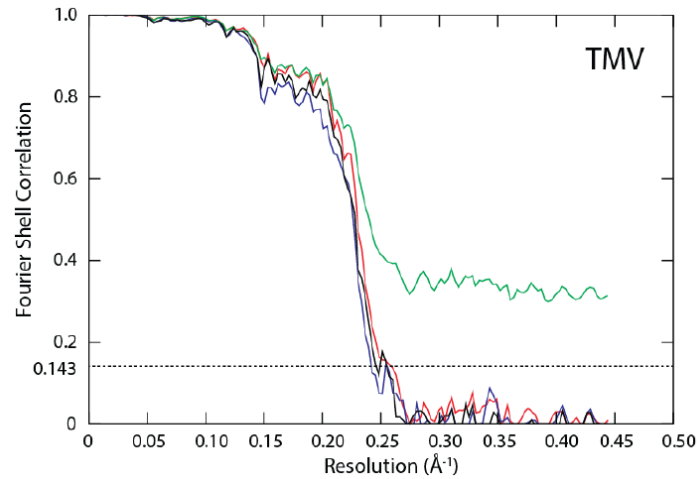


Figure 2.24. Prevention of overfitting in 3D auto-refinement by random half-sets. The FSC curve for TMV refinements with: filament-based half-sets with a top-hat prior on the translation along the helical axis (black); filament-based half-sets without a translational prior (blue); segment-based half-sets with a translational prior (red); and segment-based half-sets without a translational prior (green) (adapted from He & Scheres, 2017).

Thus, filament-wise randomisation alone can prevent iterative over-fitting effects. The two measures combined deliver the best results as expected theoretically. Therefore, the helical 3D refinement switches on both options by default. However, ‘--ignore_helical_symmetry’ option prevents the use of anisotropic translational priors. This option is provided by the user when the helical symmetry (rise in particular) is unknown or not at all reliable, therefore it is impossible to construct such priors. This could only happen at the initial stages of a helical processing project.

A special case, ParM actin-like filament, is present here to show that the intrinsic structural factors of helices might as well lead to over-fitting. As is explained in section 1.7, helical diffraction patterns contain discrete Fourier amplitudes due to the periodicity of helical structures along z axis in real-space. Together with structural factors of the helices, the FSC curves calculated between half-maps in 3D refinements may contain sharp peaks at certain resolutions. The ParM actin-like filament consists of two anti-parallel strands and the whole structure has a twist of $+165.0^\circ$ and a rise of 24.7 Å. The full data set of ParM+AMPPNP containing 4,552 still micrographs was collected on an FEI Krios microscope with a K2 direct electron detector by Bharat et al., 2015. To illustrate the concept, the final PDB model of the ParM helix (5AEY) was firstly converted to a density map. The map was then duplicated and added with two sets of white Gaussian noise at the same level of 5 in real-space. 90% spherical masks were also applied before calculating the FSC curves between this pair of simulated maps of ParM helices:

```
e2pdb2mrc.py 20merParM.pdb 20merParM.mrc --apix 1.07 --res 3 --box 350

relion_project --3d_rot --i 20merParM.mrc --o sub1-n5.mrc --add_noise --
white_noise 5 --random_seed 1444400000

relion_helix_toolbox --spherical_mask --i sub1-n5.mrc --o sub1-n5m.mrc --
spherical_percentage 0.9

relion_project --3d_rot --i 20merParM.mrc --o sub2-n5.mrc --add_noise --
white_noise 5 --random_seed 1444499999

relion_helix_toolbox --spherical_mask --i sub2-n5.mrc --o sub2-n5m.mrc --
spherical_percentage 0.9

relion_image_handler --i sub1-n5m.mrc --fsc sub2-n5m.mrc --angpix 1.07 >
fsc.txt
```

The FSC curve is shown in blue in figure 2.25A. A sharp peak can be seen clearly at 4.9 Å. Small peaks can also be identified at around the multiples of ~5 Å (i.e. 10.1 Å, 15.0 Å). Interestingly, lots of α -helices can be seen in real-space with the grooves coincidentally positioned along the direction of z axis at a spacing of ~5 Å (figure 2.26B). This special feature of the ParM helix leads to higher amplitudes on the corresponding layer lines in the diffraction pattern. The multiples of ~5 Å spacings along z

direction in real-space can be seen as anchor points and, to a large extent, fix the results of orientation and translational searches of the 2D experimental images at an early stage of 3D refinement. Therefore, the artefacts due to over-fitting build up in the course of the iterative reconstruction, which might be the main reason why the FSC curve has bumps (especially at 4.9 Å) as the input 2D images contain rather high levels of noise. 3D refinements of ParM were performed with 826,969 segments extracted using $H = 2$ and box sizes of 350 pixels (375 Å). Normal settings in the helical processing workflow were adopted initially. The reconstruction ended up at 4.5 Å (figure 2.26A&C) with a huge peak at 4.9 Å in FSC curve (red curves in figure 2.25A&B). As is seen in figure 2.26A&C, there are spiky sectional artefacts at about 4.6 pixels (pixel size = 1.07 Å) apart in layers all the way along z axis. It is an evidence that the refinement over-estimated the signal-to-noise ratio (SNR) at the spatial frequency of 4.9 Å due to the intrinsic structure of the ParM helix. To overcome the problem, the highest resolution was limited to 5.1 Å for the expectation step ('--strict_highres_exp 5.1') in the 3D refinement in order to reduce over-fitting at the cost of slightly less accurate alignments. As a result, the spuriously high FSC values at 4.9 Å was dampened (black curves in figure 2.25A&B) and the post-processed density map (figure 2.26B&D) did not contain real-space artefacts.

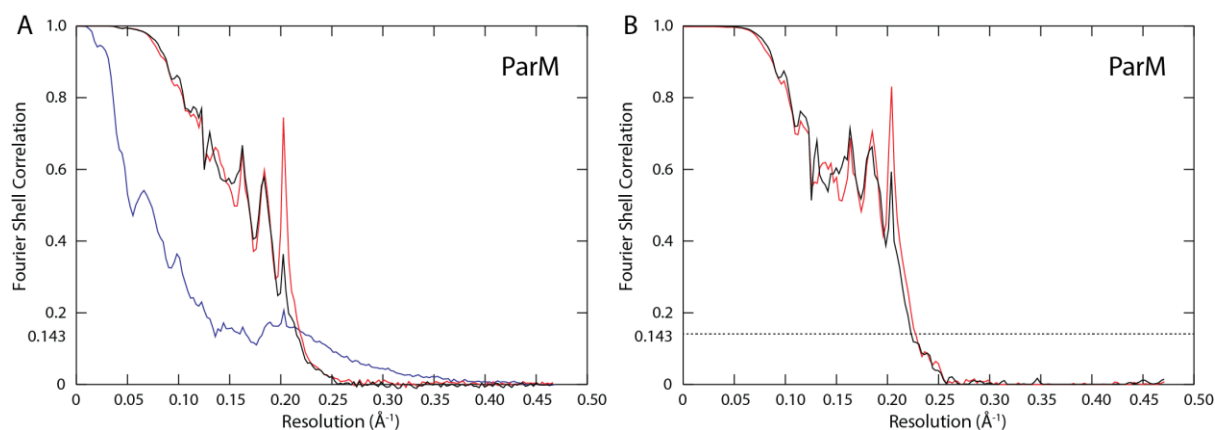


Figure 2.25. Prevention of over-fitting in 3D-refinements of ParM. (A) Blue: the FSC curve between two density maps converted from the same PDB helix model of ParM and added with white Gaussian noise at the same level of 5 in real-space. A sharp peak can be seen at the spatial frequency of 4.9 Å (0.20 Å^{-1}). Red: FSC curve between half-maps after 3D refinements of ParM data set without such over-fitting prevention measure. Black: FSC curve between half-maps after 3D refinements of ParM data set with resolutions limited to 5.1 Å for alignments in the expectation steps. (B) Red: FSC curve of (A) red after post-processing with a claimed resolution of 4.46 Å. Black: FSC curve of (A) black after post-processing with a claimed resolution of 4.51 Å. The sharp peaks at 4.9 Å (0.20 Å^{-1}) are abnormally strengthened in both red curves but appropriately reduced in both black curves.

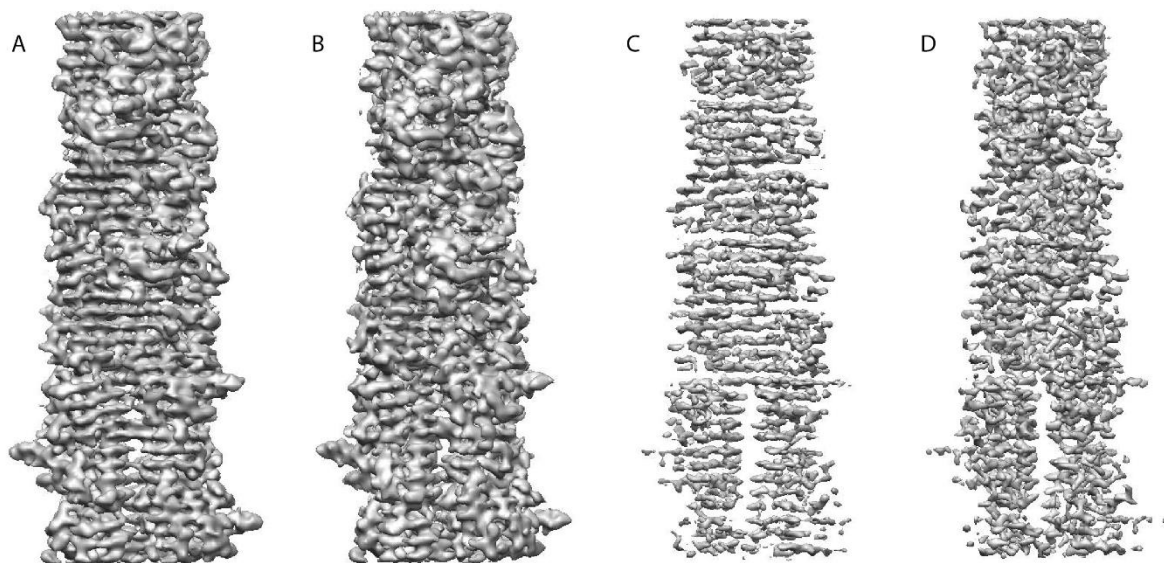


Figure 2.26. Side views of the ParM density maps after post-processing. Filtered and displayed at a contour level of 0.05 in UCSF Chimera (Pettersen et al., 2004), (A) and (B) show the results reconstructed without and with the over-fitting prevention measure respectively. (C) and (D) are the same pair of results but filtered at a contour level of 0.13. Artefacts at 4.9 Å spacing along vertical axis (Z) can be clearly seen in (A) and (C) but not in (B) and (D).

2.4.11 High-resolution structures

The workflow introduced in section 2.1 was used to process the four test data sets listed in table 2.1. For TMV and MAVS/CARD data sets, a first-round 3D refinement was conducted using motion-corrected micrographs. Movie refinements were done subsequently to track the movements of individual segments in every movie stack (Bai et al., 2013). Particle polishing estimated the B-factor weighting for each individual movie frame (figure 2.27) and wrote out polished segments (Scheres, 2014). The polished segments were then subject to the final round of 3D refinement using the reconstructed map calculated previously as the initial reference. The processing time of the final round of 3D refinements is listed in table 2.9. VipA/VipB filaments were recorded on still micrographs so no movie processing was performed. A subset of motion-corrected micrographs (260) of MamK filaments was processed as a test case because the large scale of full data set (2,086 micrographs with 600,000+ segments) made it inconvenient for testing purpose. Löwe et al. 2016 have already run the particle polishing of the complete MamK data with this helical processing workflow in RELION-2.0. In the hope of even better-quality subsets of segments, 3D classifications of polished segments were attempted without further alignment changes permitted. However, no subsets of segments in any of the four cases led to improved reconstructions. Post-processing steps sharpened the density maps coming from 3D refinements by automatically estimated B-factors (Rosenthal and Henderson, 2003) and calculated corrected FSC curves between half-maps according to the masked, protein densities through phase-

randomisation (Chen et al., 2013). FSC = 0.143 criterion was adopted to estimate the final overall resolutions (Scheres and Chen, 2012). Heat-maps of local resolutions were also generated with a built-in function in RELION, using a small sphere as a sliding mask to estimate resolution at every region of the map locally by phase-randomisation as well (Fernandez-Leiro and Scheres, 2017). Figure 2.28 shows the final density maps (with local resolution estimates), solvent-corrected FSC curves, representative central z slices, and fitted atomic models of all the four test cases. For TMV, VipA/VipB and MAVS/CARD filaments, representative side chain densities were compared between results from RELION and from other packages (available on EMDB) in the last columns of figure 2.28. The results from implemented helical workflow of RELION are better than or at least the same as the results from other packages in terms of the final resolutions and the quality of protein densities. Key parameters used are summarised in table 2.1 and 2.9.

	Size of segments (pixels)	# of segments	Tilt (θ) & psi (ϕ) ranges	Angular sampling	Trans range & sampling	CPU cores (# of MPI X threads)	Time (approx.)
TMV	300 × 300	1,761	±15°, ±5°	1.8° / 0.9°	±10 / 1 pix	5×24=120	2h 30m
VipA/VipB	450 × 450	9,480	±15°, ±5°	1.8° / 0.9°	±10 / 1 pix	6×24=144	13h 30m
MAVS/CARD	300 × 300	133,237	±15°, ±15°	3.7° / 1.8°	±5 / 1 pix	15×24=360	17h 30m
MamK	280 × 280	110,055	±15°, ±15°	3.7° / 1.8°	±5 / 1 pix	10×24=240	15h

Table 2.9. Processing time of the final 3D auto-refinement jobs performed on Intel Xeon E5-2643 (3.4-3.5 GHz) and E5-2667 (2.9 GHz) CPU clusters. Initial in-plane (ψ) angular searches were bimodal for all test cases. For the ‘angular sampling’ column, the entry ‘1.8° / 0.9°’ means the refinement started with an angular sampling of 1.8° for global searches and proceeded into local searches of all three Euler angles (ϕ , θ , ψ) when the angular sampling was automatically decreased to 0.9° or smaller. Larger initial angular samplings might not be sufficient for the optimal reconstruction results. Similarly, translational ranges and samplings shown in this table are for the initial, global alignment steps only. Additional referential masks, covering the full-length helical densities, were also adopted to calculate solvent-corrected FSC curves at the end of each iteration.

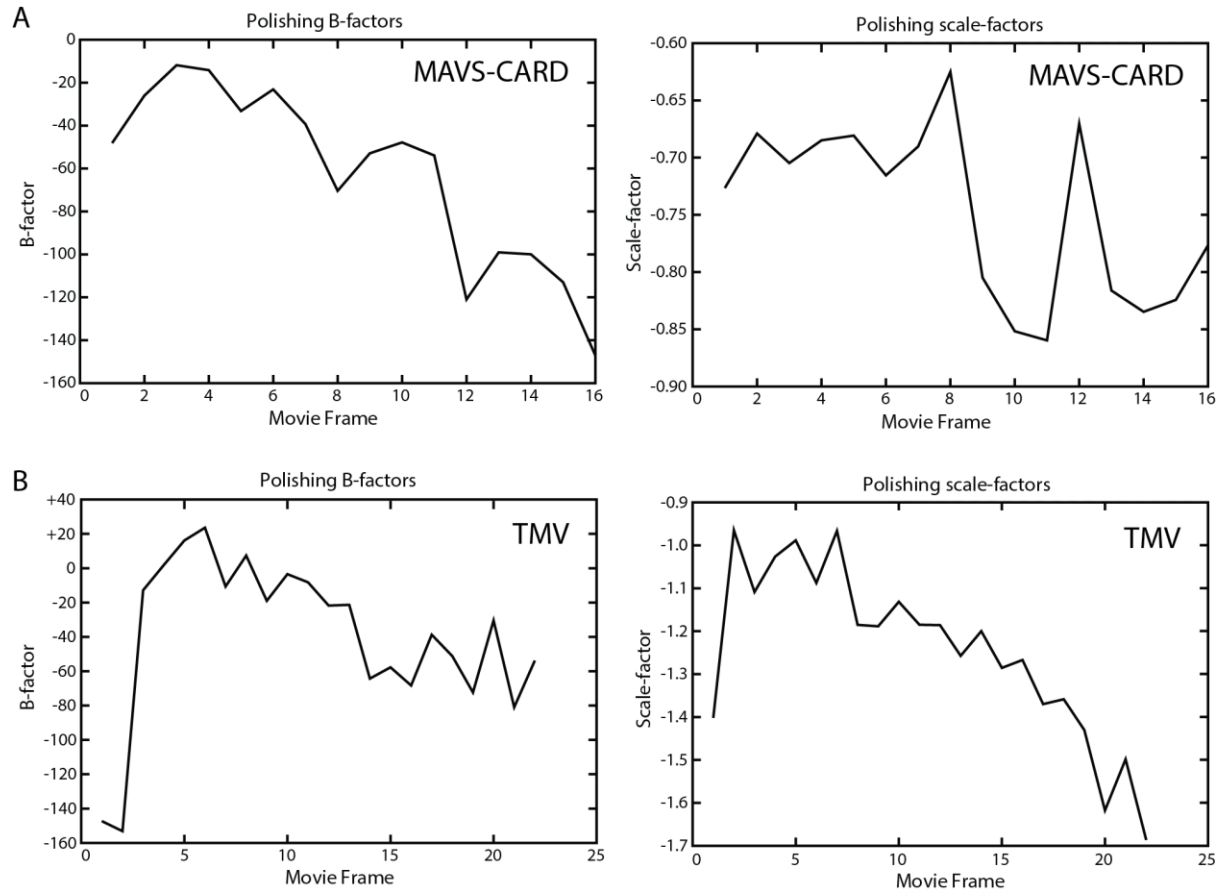


Figure 2.27. Particle polishing B-factor and scale-factor plots of movie data sets (A, TMV, 16 frames each; B, MAVS-CARD, 22 frames each).

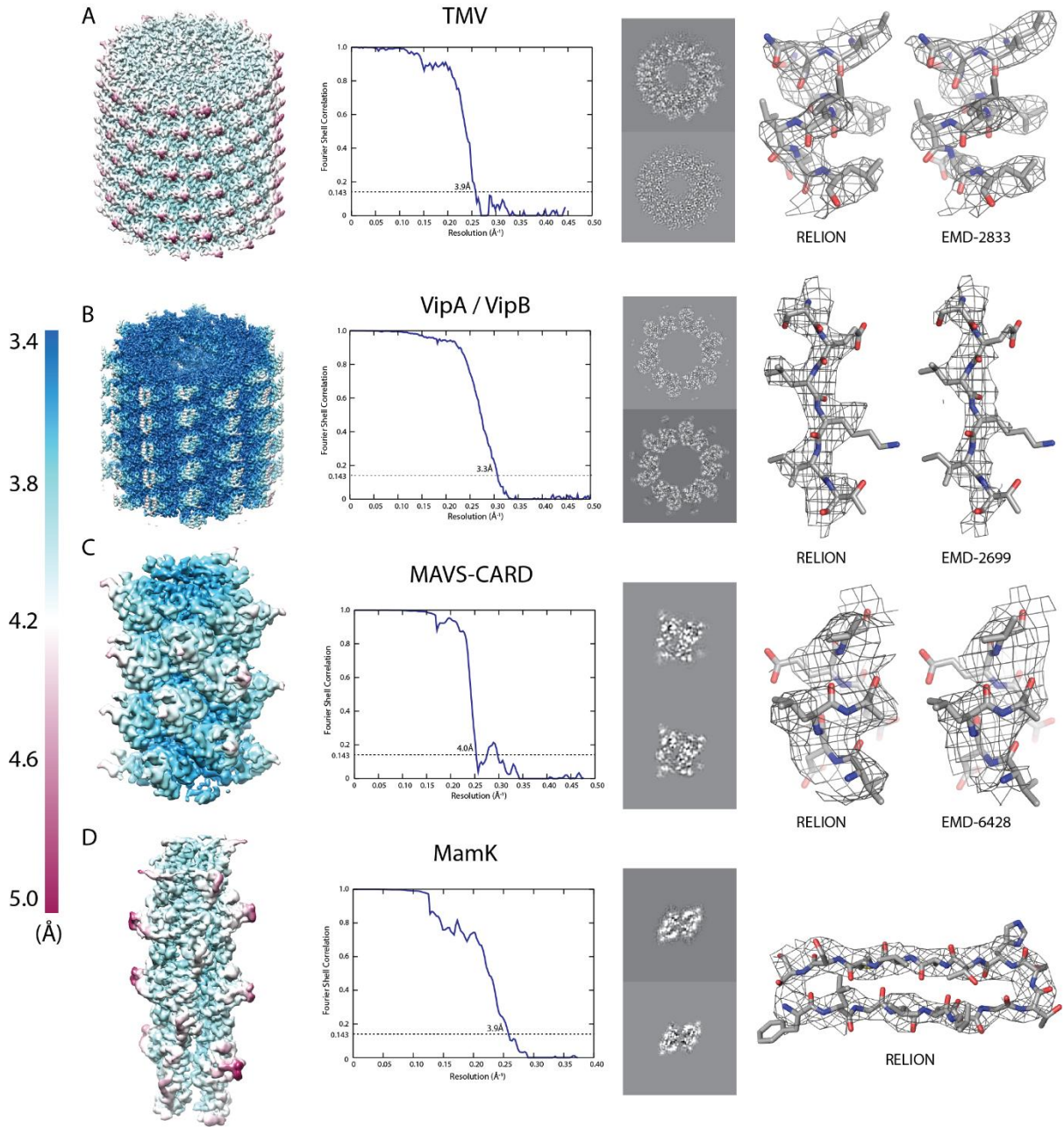


Figure 2.28. Reconstructions for the four test cases: TMV (A), VipA/VipB (B), MAVS/CARD (C) and MamK (D). A view of the reconstructed helix coloured according to local resolution estimates (from 3.5 Å in blue to 5 Å in maroon) is shown on the left. The FSC curve between the two independently refined half-sets is shown in the middle. A detail of the structure with the corresponding atomic model is shown on the right. For TMV, VipA/VipB and MAVS/CARD, the same details from the EMD maps are also shown (adapted from He & Scheres, 2017).

3 IMPLEMENTED

ALGORITHMS IN HELICAL

SUB-TOMOGRAM

AVERAGING

3.1 Implemented methods

Adaptations of single-particle helical reconstructions have been implemented for helical sub-tomograms. The main difference is that sub-tomogram averaging has input particles in 3D and each particle corresponds to a distinct 3D CTF model. All the implemented methods discussed in section 2.2 are re-used in helical sub-tomogram averaging, with the following adaptations for 3D images.

3.1.1 Manual-picking of sub-tomograms

Manual picking of 3D helical sub-tomograms currently needs to be performed through other packages since currently RELION does not have a built-in display for 3D tomograms. Additional support has been developed for IMOD (Kremer et al., 1996) where the user scrolls through continuous z slices of a 3D tomogram and clicks along every 3D filament. Here, the assumption that manually picked filaments are straight does not hold any more. By clicking, the user can specify the spanning direction of each 3D filament which might take arbitrary shapes and lengths. IMOD writes out a .MOD binary file for each picked filament and 'relion_helix_toolbox --interpo' can then generate 3D segment coordinates based on the inter-box distance (i.e, a user-defined helical rise, '--rise' and the number of asymmetrical units H, '--nr_asu'):

```
relion_helix_toolbox --interpo --i_root Btub_tomol --o_root _interpo --
nr_asu 1 --rise 52.77 --angpix 2.18 --boxdim 200 --bin 1
```


'--i_root' and '--o_root' denote the root-names of input MOD and output STAR files respectively. And '--bin' denotes the binning factor (a positive integer) used when reconstructing the 3D tomogram (1 = no binning). Internally, the program interpolates smooth 3D curves for all filaments using user-specified 3D segment coordinates as anchor points. The curves are subsequently segmented into individual coordinates for 3D segment extractions. Angular priors of helical sub-tomograms are calculated differently compared to manually picked 2D segments. Tilt priors (θ) are not necessarily 90° since the 3D sub-tomogram segments may adopt arbitrary views in the reconstructed tomograms. Filaments are segmented along smooth curves of interpolated high-order polynomials instead of simple straight lines so segments on a filament can have varying tilt (θ) and in-plane rotation (ψ) priors. The algorithm moves a point slowly in 3D from the start of a curve and records its i th position (x_i, y_i, z_i) when it has moved by the i th multiple of inter-box distance. The slightly perturbed position ($x_i + \Delta X, y_i + \Delta Y, z_i + \Delta Z$) along the smooth curve (C) is also estimated in order to calculate θ, ψ priors for the i th segment:

$$\theta_i = \cos^{-1} \frac{\Delta z_i}{\sqrt{(\Delta x_i)^2 + (\Delta y_i)^2 + (\Delta z_i)^2}}, \psi_i = \tan^{-1} \frac{\Delta y_i}{\Delta x_i},$$

$$(x_i, y_i, z_i) \in C, (x_i + \Delta x_i, y_i + \Delta y_i, z_i + \Delta z_i) \in C, \Delta x_i \rightarrow 0^+$$

Auto-picking of 3D sub-tomograms has not been implemented since the 2D auto-picking algorithm described in section 2.2.7 will be computationally expensive if performed on large 3D tomograms. Moreover, manual-picking of 3D sub-tomograms is typically affordable due to limited numbers of tomograms collected (<100).

After the extraction of sub-tomograms, distinct 3D CTF models must be calculated for every 3D segment box before 3D reconstructions. For each tomogram, 2D CTFs are estimated for the 2D images in the tilt series. Based on the central slice theorem (section 1.3), the 2D CTFs are subsequently placed back into 3D Fourier space according to their tilt angles. The 3D CTF models for the sub-tomograms are constructed given their positions in the tomogram (Bharat et al., 2015). Note that changing the inter-box distance in manual-picking will cost a lot more computation in helical sub-tomogram averaging since the 3D CTF models need to be re-calculated for every different set of picked 3D coordinates.

3.1.2 3D image masks

Masks are 3D for references and experimental segments in helical sub-tomogram averaging. The cylindrical masks (M_{ref}) for references remain the same as they are in single-particle reconstructions (section 2.2.5) since the reconstructed helices still span along the z axis. The 3D segment masks are rotated according to the spanning directions of helices estimated through manual-picking. As is explained above, the ψ and θ angles of every helical sub-tomogram are calculated through

interpolations. For each pixel of a given 3D sub-tomogram, its x, y, z coordinates are rotated back to the same 3D Cartesian coordinate system according to the angular priors ψ, θ . The distance of the point from the helical axis determines its value in the binary mask M_{exp} :

$$A_{3 \times 3} = \begin{pmatrix} \cos \psi \cos \theta & \sin \psi & -\cos \psi \sin \theta \\ -\sin \psi \cos \theta & \cos \psi & \sin \psi \sin \theta \\ \sin \theta & 0 & \cos \theta \end{pmatrix}$$

$$M_{exp}(x, y, z) = \begin{cases} 1, & A_{3 \times 3}^T \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}, x'^2 + y'^2 < R_o^2 \text{ and } x'^2 + y'^2 + z'^2 < R^2 \\ 0, & \text{otherwise} \end{cases}$$

$A_{3 \times 3}$ is the 3D rotational matrix with the first Euler angle (ϕ) always set to 0° . Cosine soft-edges are also applied around the 3D binary masks of the experimental segments. This M_{exp} is a 3D extension of 2D M_{exp} explained in section 2.2.5.

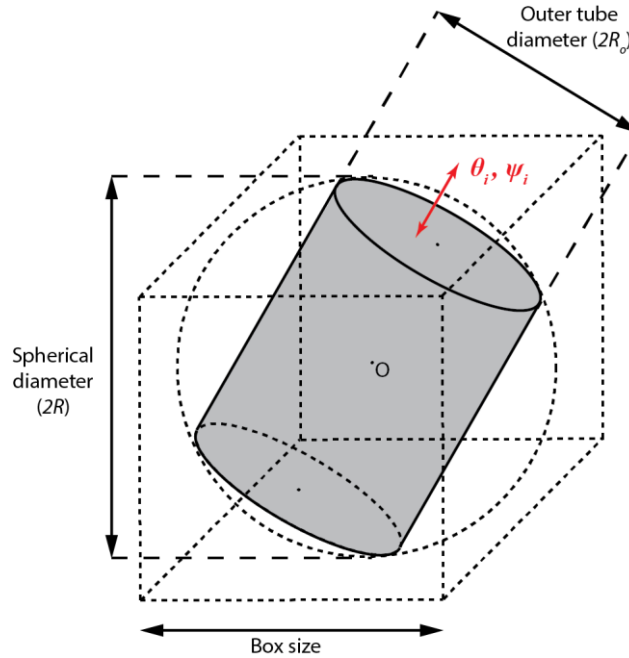


Figure 3.1. Mask for an experimental 3D helical sub-tomogram, given the angular priors ψ, θ , the spherical and outer tube diameters ($2R$ and $2R_o$). This figure is a 3D extension of the 2D segment mask in figure 2.4A.

3.1.3 Incorporating angular prior information

The angular priors of 3D helical sub-tomograms remain the same as those of 2D segments, except that the tilt priors (θ) are not always centred at 90° but specifically estimated through manual-picking. However, the translational priors has two components in the directions perpendicular to the helical axis to accommodate 3D data sets (figure 3.2). For each sub-tomogram, the prior distribution of offset, along the direction defined by the in-plane rotation (ψ) prior, is still a top-hat function with non-zero values

within range $[-\text{rise} / 2, +\text{rise} / 2]$ around the image centre to limit over-fitting (section 2.2.6 and 2.4.10). The standard deviation of the 2D Gaussian offset priors is defined on the 2D plane perpendicular to the spanning direction of the helical sub-tomogram.

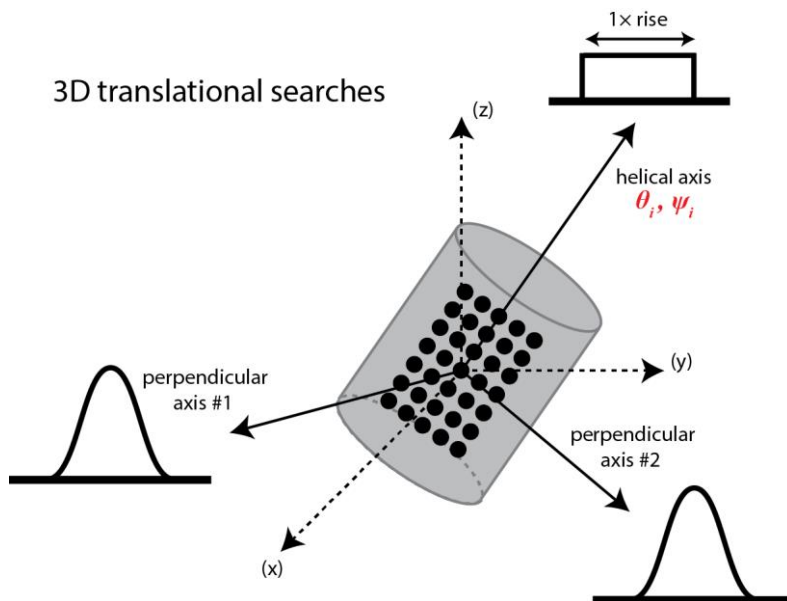


Figure 3.2. Priors on the 3D translational parameters for helical sub-tomogram averaging. A top-hat prior on the translation along the helical axis and a 2D Gaussian prior on the two translations perpendicular to the helical axis, given the angular priors ψ , θ and the helical rise. This figure is a 3D extension of the 2D translational parameters in figure 2.5A.

3.1.4 Simulations of 3D experimental sub-tomograms

3.1.4 Simulation of 3D experimental data

To simulate 3D sub-tomograms, 3D rotations of the perfect 3D helical reference should be performed rather than the 2D projections described in section 2.2.11. The additional option ‘--3d_tomo’ must be provided in the command starting with ‘relion_helix_toolbox --simulate_segments’, whereas all the other settings remain unchanged. For simulated sub-tomograms, θ and ψ are randomised at any possible values within 0° - 180° and -180° - $+180^\circ$ respectively. The shell script ‘./simulate-3d-subtomograms.sh’ generated by the above command writes out all the sub-tomograms.

3.2 Introduction to the test data set

A small cryo-ET data set of BtubAB is used in order to test helical sub-tomogram averaging. The data set was collected by Deng et al., 2017 on an FEI Titan Krios microscope with a K2 detector with a magnification of $57,000\times$ (2.38 \AA per pixel). It consists of 6 tilted series at 1° or 3° interval within ranges of -60° to $+60^\circ$. A total number of 4,333 segments are extracted using an inter-box distance of

80 Å. The filament has $\alpha\beta$ subunits in special configuration (the schematic plot is shown in figure 4.2A). However, the two subunit types cannot be distinguished under the low resolutions of sub-tomogram averaging. All subunits treated as the same type, the structure can be seen as a perfect helix without point group symmetry and the helical twist and rise are -90.6° and 9.96 Å respectively (see section 4.1 for detailed explanation). The segment boxes have 3D dimensions of 100 pixels (4.76 Å per pixel) using a binning factor of 2 in order to reduce computation costs. The results of this data set were initially used to determine the number of proto-filaments in the structure (Deng et al., 2017) and have not been published.

3.3 Results and discussion

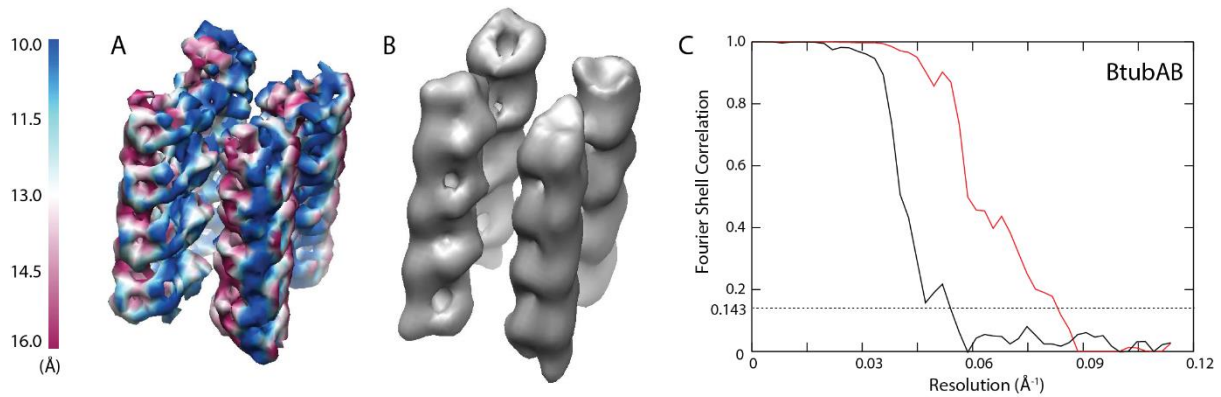


Figure 3.3 Results of helical sub-tomogram reconstructions of BtubAB filaments. (A) Post-processed density map after helical sub-tomogram averaging with local resolutions estimated within the range of 10 Å – 16 Å. (B) Post-processed density map after the ordinary single-particle sub-tomogram averaging. Both A and B maps are sharpened with a B-factor of -1000 Å^2 during post-processing. (C) Comparisons of FSC curves of the helical sub-tomogram reconstruction (red line, $\sim 12.5 \text{ Å}$ at 0.143 resolution cut-off) and the single-particle sub-tomogram averaging (black line, $\sim 18.9 \text{ Å}$ at 0.143 resolution cut-off) using the same cryo-ET BtubAB data set described in section 3.2.

The results of helical sub-tomogram reconstruction of BtubAB filaments are shown in figure 3.3. Andres Gonzalez, a Part II student in Biochemistry department of University of Cambridge, collaborated with me on testing the methods implemented in section 3.1 in the spring of 2017. The initial reference was a simple average of all the extracted sub-tomograms according to their angular priors. The 3D refinements were run on a local machine with four NVIDIA Titan X GPU cards and lasted less than 2 hours. The subunits can be clearly seen on the four-stranded structures although it is not expected to distinguish the two types of subunits at this resolution (worse than 10 Å). The helical sub-tomogram reconstruction made use of the helical symmetry (-90.6° and 9.96 Å), the prior information available in the extracted data set and outperformed single-particle sub-tomogram averaging in terms of final resolutions and structural details.

4 LOCAL SYMMETRY

4.1 Microtubules with seams - Breaking from normal helical symmetry

As is mentioned in the general introduction to helical structures (section 1.4), microtubules are filamentous complexes found in eukaryotic cells. They provide the structural basis for intracellular transport (Vale, 1987), nucleic acid and cell division (Nigg, 2001), movements of cilia and flagella (Silflow and Lefebvre, 2001), etc. Their building blocks are tubulin heterodimers, the $\alpha\beta$ subunits of which are alternately assembled into linear proto-filaments (figure 4.1). Multiple proto-filaments are translated and rotated, forming tubular microtubule walls (12-, 13-, 14- proto-filaments are among the most common cases).

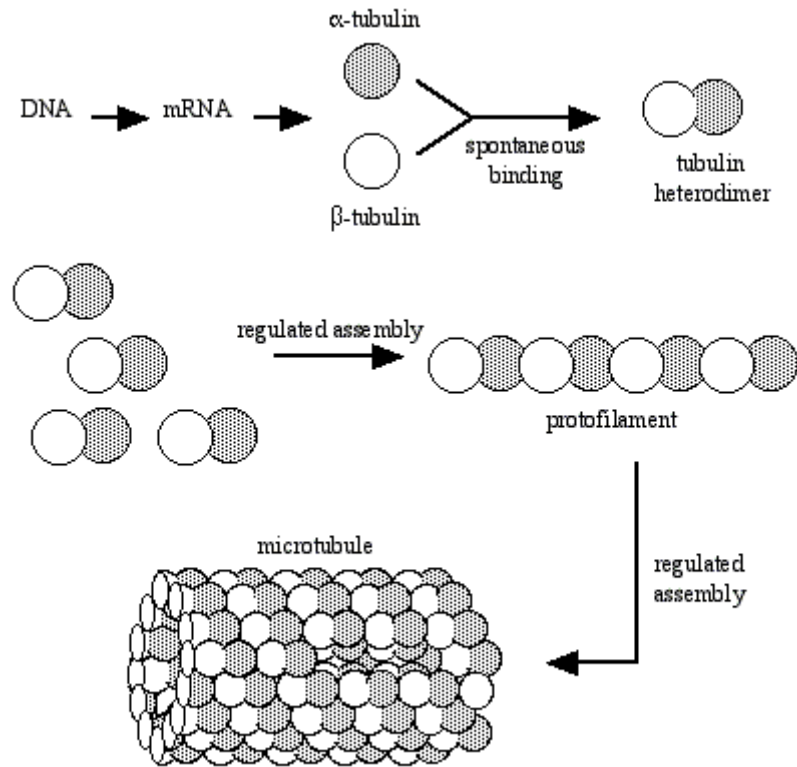


Figure 4.1 α , β -tubulins can be assembled into microtubules. All eukaryotic cells produce tubulin proteins. α , β -tubulins spontaneously bind one another to form heterodimers. When intracellular conditions favour assembly, tubulin heterodimers assemble into linear proto-filaments and then into microtubules (adapted from <http://www.ruf.rice.edu/~bioslabs/studies/invertebrates/microtubules.html>).

Breaking from normal helical symmetry can be observed when $\alpha\beta$ -tubulins make specific lateral contacts and are assembled into unique patterns of helical lattices. In general, structural details of $\alpha\beta$ -tubulins in the heterodimers are only slightly different (mainly at peripheral densities) and stability of the lattice as a whole is still largely preserved even if there is a change of tubulin configurations. When curled up to form 3D microtubule walls, lattices with only lateral contacts being $\alpha\alpha$ and $\beta\beta$ form normal helices (figure 4.1, bottom), the symmetry of which as a whole can be perfectly defined by considering each pair of adjacent $\alpha\beta$ tubulins along proto-filaments as a bigger subunit. On the contrary, in addition to $\alpha\alpha$ and $\beta\beta$ lateral contacts, $\alpha\beta$ and $\beta\alpha$ lateral contacts might as well be present in a neighbouring pair of proto-filaments (proto-filaments #1 and #4 in figure 4.2A, #1 and #13 in figure 4.2B). By convention, the surface between this pair of proto-filaments is called the seam (black vertical lines in figure 4.2). The seam breaks the overall helical symmetry, giving rise to strikingly different chemical environments along the seam compared to other parts of homogeneous surfaces of the microtubule walls.

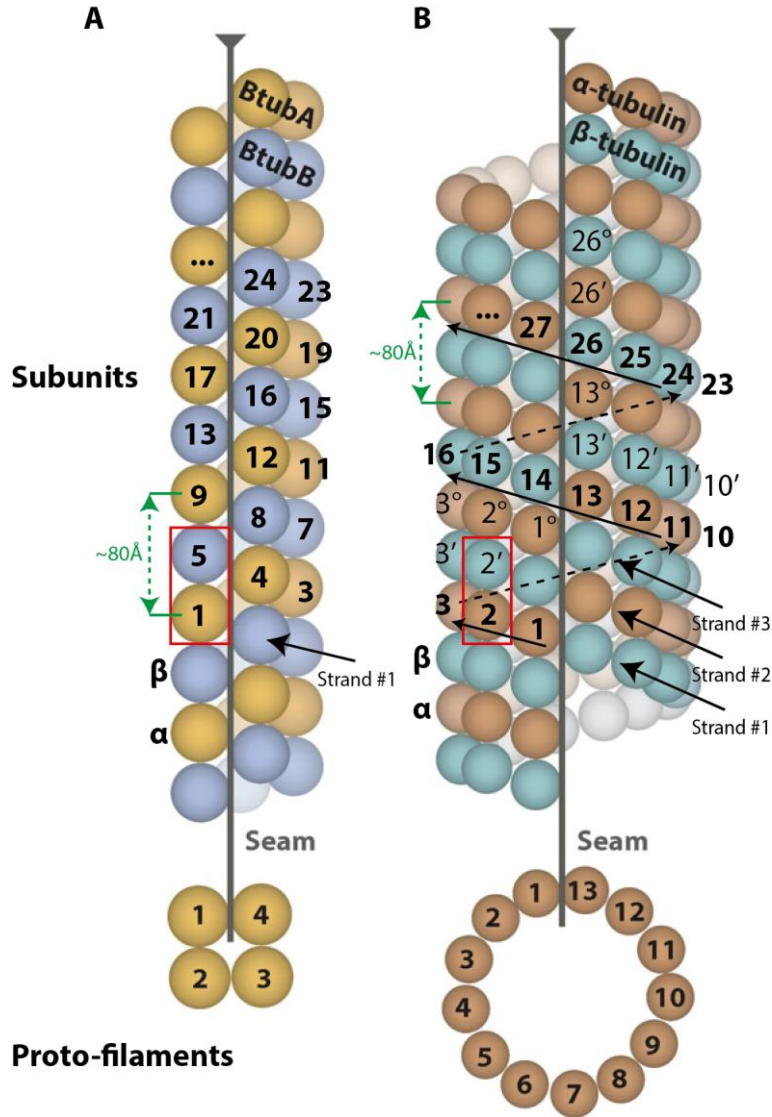


Figure 4.2. Schematic plots showing 4- and 13- microtubules with seams. Here, both the 4- and 13-microtubules have $\alpha\beta$ and $\beta\alpha$ lateral contacts between the pair of proto-filaments closest to the seam, in addition to the common $\alpha\alpha$ and $\beta\beta$ lateral contacts in the 13-microtubule without the seam shown at the bottom of figure 4.1. In reality, the proto-filaments and the seams are not necessarily positioned strictly parallel to the helical axes (adapted from Deng et al., 2017).

The first step to examine the symmetry of such microtubules with seams is to ignore the differences between the subunit types and figure out the symmetry as if they are normal helices consisting of all-the-same tubulins. There are true structures analogous to the schematic plots shown in figure 4.2. For example, the BtubABC microtubule (section 5.1) corresponding to the 4-microtubule in figure 4.2A, can be seen as a 1-start left-handed helix with a twist of -90.7° and a rise of 9.91 \AA under such assumption. The 13-microtubule (section 4.4) corresponding to figure 4.2B, can be seen as a 3-start (i.e. 3-strand) left-handed helix with a twist of -27.7° and a rise of 9.46 \AA on each strand. Such 3-start

symmetry can be converted to 1-start symmetry of $(-27.7^\circ - 360^\circ) / 3 = -129.2^\circ$ and $9.46 / 3 = 3.15 \text{ \AA}$ according to the definitions in section 2.2.1.

To consider the actual break of symmetry, $\alpha\beta$ subunit pairs should be seen as the smallest, indivisible building blocks of the whole structures instead of the individual tubulins. One possible way could be considering one adjacent pair of subunits α and β on a single proto-filament as the smallest asymmetrical unit (for example, $1\alpha\&5\beta$ (left) or $2\alpha\&2'\beta$ (right), the red rectangles in figure 4.2). Thus, the new overall helical symmetry can only be applied along the directions of the proto-filaments. For figure 4.2A, such new helical twist and rise are $-90.7^\circ \times 8 + 720^\circ = -5.60^\circ$ and $9.91 \text{ \AA} \times 8 = 79.3 \text{ \AA}$. The above calculations are based on the symmetry given in the last paragraph, since $4 \times 2 = 8$ subunits (the product of the numbers of proto-filaments and tubulin types) are traversed when one α (1α) is moved to the next α (9α) along the proto-filament #1 (figure 4.2A). Equivalently, this new helical symmetry alone treats 8 consecutive subunits 1α - 8β , 9α - 16β , ... as the building blocks of the whole 4-proto-filament. Such helical symmetry was used in BtubABC reconstructions by Deng et al., 2017 (also in section 5.1). Similarly, the new helical twist and rise are $(-27.7^\circ \times 26 + 360^\circ) / 3 = 0^\circ$ and $9.46 \text{ \AA} \times 26 / 3 = 82.0 \text{ \AA}$ for figure 4.2B. Here, $13 \times 2 = 26$ subunits are traversed when 1α is moved to 27α along the proto-filament #1 and the strand #1 (figure 4.2B). Both the new twist and rise are divided by 3 since the previous symmetry is given on one strand of the 3-start structure. The new symmetry defines the relationship between adjacent asymmetrical units on every proto-filament in figure 4.2B, for example, $(2\alpha, 2'\beta)$ and $(2^\circ\alpha, 15\beta)$ on proto-filament #2.

The above helical symmetry considers the alternating $\alpha\beta$ -tubulin pairs along each proto-filament, but the tubulins across multiple proto-filament are not yet averaged. Positioned uniformly around the z (helical) axis and with relative z translations, all the proto-filaments consist of the same information of alternating $\alpha\beta$ pairs. Neither point group symmetry nor overall helical symmetry is able to represent such structural self-similarities. Therefore, local symmetry operators need to be introduced to define the relative positions of the proto-filaments in the density map, and a mask should also be provided to cover any one of the proto-filaments in order to identify its region in the density map of the whole microtubule. For example, a mask around the whole proto-filament #1 can be defined for the 4-microtubule in figure 4.2A, and 3 operators should transform $(1\alpha, 5\beta, 9\alpha, \dots)$ tubulins into the positions of $(2\alpha, 6\beta, 10\alpha, \dots)$; $(3\alpha, 7\beta, 11\alpha, \dots)$; $(4\alpha, 8\beta, 12\alpha, \dots)$ tubulin groups along proto-filament #2, #3, #4 respectively. Thereby, the 3 operators need to represent rotations of $N \times -90.7^\circ$ around the z axis and translations of $N \times 9.91 \text{ \AA}$ ($N = 1, 2, 3$) along the z axis. For figure 4.2B, one mask and 12 operators are required in a similar way. The 12 operators can be rotations of $N \times -27.7^\circ$ around the z axis and translations of $N \times 9.46 \text{ \AA}$ ($N = 1, 2, \dots, 12$) along the z axis, if the mask is provided for the proto-filament #1. The helical symmetry along each proto-filament together with the local symmetry operators

defining the relations between proto-filaments describe the full symmetry of the microtubules with seams and average the smallest asymmetrical units (building blocks, $\alpha\beta$ -tubulin pairs) properly.

Image processing procedures, such as Ruby-helix (Metlagel et al., 2007) and SPIDER scripts written by Fourniol et al., 2010, have been designed to tackle the problem of seams specifically. The concept of local symmetry introduced above is not only applicable to microtubules, but also ordinary protein structures containing identical regions in single-particle reconstructions as well. For example, a molecule may contain a tetramer binding to another protein resulting in an asymmetrical complex, the symmetry of which cannot be described by overall point groups. Such structures call for arbitrary, local symmetry operators to express the relative positions of identical regions in order to average the densities locally for higher quality maps. Inspired by non-crystal symmetry methodology in X-ray crystallography (Kleywegt, 1996), built-in functions in RELION have been developed to search for and apply local symmetry for cryo-EM density maps.

4.2 Implemented methods

4.2.1 Definition of local symmetry

For each set of rigid and identical regions related by local symmetry, one soft-edged binary mask (M_i) is needed, covering any one of the asymmetrical subunits of the local symmetry. All the other regions are assumed to be identical in densities and their positions and orientations relative to the masked region are defined by local symmetry operators. Therefore, a mask defines the shape of a set of rigid and identical asymmetrical subunits. It is also used to collect the real-space densities of all the asymmetrical subunits in the set for the subsequent real-space symmetrisation step once the correct local symmetry operators are determined. The symmetrised densities for a set of identical asymmetrical subunits are firstly calculated and written into the masked region, and then copied into other regions in this set according to the local symmetry operators.

For example, two masks (M_{tri} , M_{tetra}) are required for a complex that consists of a trimer of one protein that is bound to a tetramer of another protein. One (M_{tri}) covers a monomer of the trimer (along with two operators) and the other (M_{tetra}) covers a monomer of the tetramer (along with three operators). The implementation in RELION supports the definitions of an arbitrary number of local symmetry sets and an arbitrary number of operators within each set. The masks and the original map, in MRC format, should all have the same cubic dimensions ($D \times D \times D$, in pixels) and the definition of origin at x, y, z coordinates (0, 0, 0) in their MRC headers (Cheng et al., 2015).

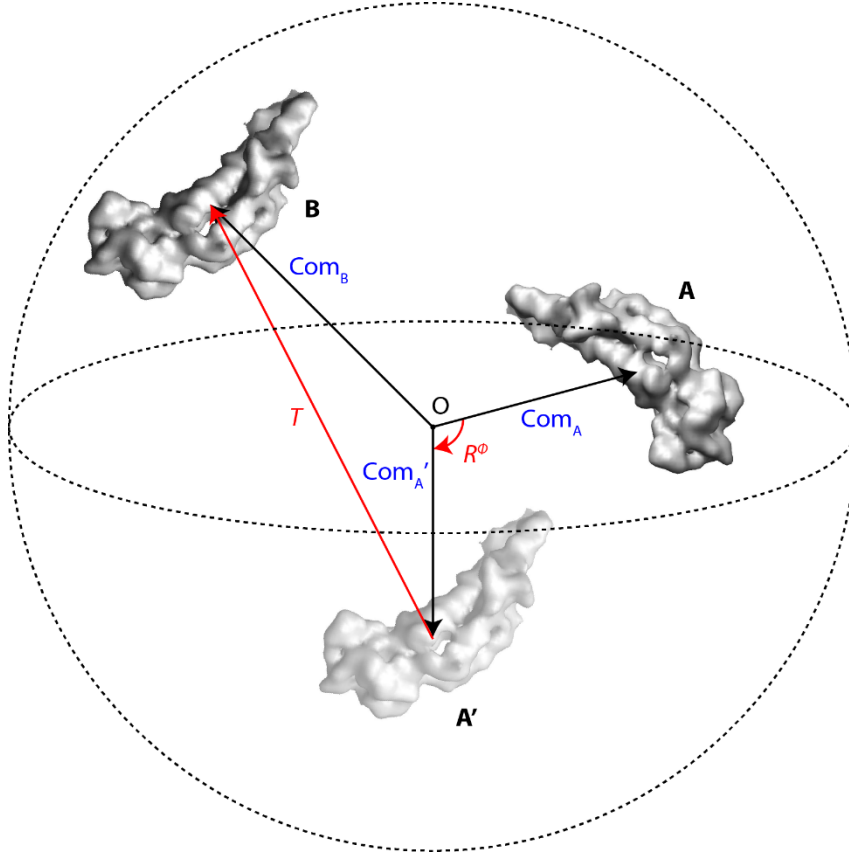


Figure 4.3. Definition of local symmetry. The box contains two identical regions A and B, the centre-of-mass coordinates of which are Com_A and Com_B (blue). The origin of the particle box is O, which is also the centre of any rotations. The rotational and translational operators, R^ϕ and T (red), transform the densities of A on top of B. $\vec{T} = \vec{\text{Com}_B} - R^\phi * \vec{\text{Com}_A}$. Given a soft-edged binary mask around A, the only local symmetry operator of this map is defined as $Op = [R^\phi, T]$. In such way, the identical regions A and B are related by local symmetry.

A transformation operator contains 6 parameters, three rotations ϕ, θ, ψ (following the definition of Euler angles in RELION) in degrees and three translations $\Delta x, \Delta y, \Delta z$ in Angstroms. The masked region (region A in figure 4.3) must be located within the maximum sphere (dashed circles) inside the 3D particle box ($D \times D \times D$, in pixels). The pair of identical pieces of densities (A and B) corresponds to a unique operator $Op = [\phi, \theta, \psi, \Delta x, \Delta y, \Delta z]$, which firstly rotates the masked region A to A' around the centre O of the 3D particle box, and then translates the rotated densities A' to fit onto the symmetry-related region B. The following equations show how an operator associates two identical regions in a cryo-EM density map (ignoring interpolation artefacts). x_m, y_m, z_m denote 3D coordinates in volume V . Note that an operator Op is divided into two parts, a rotational matrix ($R^{3 \times 3}$) and a translational vector ($T^{3 \times 1}$) in the following equations.

$$Op = (\phi, \theta, \psi, \Delta x, \Delta y, \Delta z), \quad \begin{array}{l} -180^\circ \leq \phi < 180^\circ \\ 0^\circ \leq \theta < 180^\circ \\ -180^\circ \leq \psi < 180^\circ \end{array}, \quad -\infty < \Delta x, \Delta y, \Delta z < +\infty$$

$$R = \begin{pmatrix} \cos \psi \cos \theta \cos \varphi - \sin \psi \sin \varphi & \cos \psi \cos \theta \sin \varphi + \sin \psi \cos \varphi & -\cos \psi \sin \theta \\ -\sin \psi \cos \theta \cos \varphi - \cos \psi \sin \varphi & -\sin \psi \cos \theta \sin \varphi + \cos \psi \cos \varphi & \sin \psi \sin \theta \\ \sin \theta \cos \varphi & \sin \theta \sin \varphi & \cos \theta \end{pmatrix}.$$

$$T = \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} / pixel_size.$$

$$\begin{pmatrix} x'_m \\ y'_m \\ z'_m \end{pmatrix} = R * \begin{pmatrix} x_m \\ y_m \\ z_m \end{pmatrix} + T.$$

$$\forall x_m, y_m, z_m \in \left\{ -\frac{D}{2}, -\frac{D}{2} + 1, \dots, \frac{D}{2} - 1 \right\} \text{ s. t. } M(x_m, y_m, z_m) = 1.$$

$$V(x'_m, y'_m, z'_m) = V(x_m, y_m, z_m), \quad \text{if } -\frac{D}{2} \leq x'_m, y'_m, z'_m \leq \frac{D}{2} - 1.$$

4.2.2 Applying local symmetry

Given accurate masks and local symmetry operators, the implemented algorithm symmetrises every set of similar regions in real-space. Suppose that there exist N sets of regions each with K_i ($i = 1, 2, \dots, N$, $K_i \geq 2$, $K_i \in \mathbb{N}^+$) similar regions need to be symmetrised separately (in the simplest example shown in figure 4.3, $N = 1$ and $K_1 = 2$). Within each set, a mask (M_i , $i = 1, 2, \dots, N$) covering one region is provided along with a list of $K_i - 1$ symmetry operators ($Op_{ij} = [\varphi_{ij}, \theta_{ij}, \psi_{ij}, \Delta x_{ij}, \Delta y_{ij}, \Delta z_{ij}]$, $i = 1, 2, \dots, N$, $j = 1, 2, \dots, K_i - 1$). Each operator Op_{ij} is decomposed into rotational ($R_{ij}^{3 \times 3}$) and translational ($T_{ij}^{3 \times 1}$) parts internally in the program. Volume operators Op_{ij} work in ‘no wrap’ mode so that they might move some voxels outside the dimensions of the original particle box ($D \times D \times D$, in pixels) and the information contained there is lost in the transformed map, rather than coming (wrapped) into the particle box from the opposite sides. The voxels in the transformed map are set to 0 if they are not mapped inversely to the inside of the original map by the operator.

Firstly, the algorithm generates averaged maps for each set of regions. In the following equation, V_i holds averaged densities inside the mask from the i th set of identical regions (R_{ij}^{-1} and T_{ij}^{-1} denote inverse rotation and translations in opposite directions of R_{ij} and T_{ij} respectively, $*$ denotes that a transform operation to its left is applied onto the 3D volume to its right, and \circ denotes element-wise product of 3D volumes with the same dimensions, $D \times D \times D$):

$$V_i = M_i \circ \frac{V_{ori} + \sum_{j=1}^{K_i-1} (R_{ij}^{-1} * T_{ij}^{-1} * V_{ori})}{K_i}$$

The averaged densities within the mask are copied into all the other $K_i - 1$ regions for every set. Volume S contains symmetrised regions at all positions defined by the list of masks and their corresponding operators:

$$S = \sum_{i=1}^N \left(V_i + \sum_{j=1}^{K_i-1} T_{ij} * R_{ij} * V_i \right)$$

A weighted sum (W) of masks at all original and transformed positions is also calculated. The voxel values in W can be larger than 1 if the rotated and translated regions ($T_{ij} * R_{ij} * M_i$) partly overlap with the originals (M_i).

$$W = \sum_{i=1}^N \left(M_i + \sum_{j=1}^{K_i-1} T_{ij} * R_{ij} * M_i \right)$$

In order to take care of the smoothness of symmetrised density map given soft-edged masks as input, binarised W_{bin} records the weight of symmetrised densities against the background (original) densities at each 3D coordinate.

$$W_{bin}(x_i, y_i, z_i) = \begin{cases} 0, & W(x_i, y_i, z_i) = 0 \\ \max(1, W(x_i, y_i, z_i)), & W(x_i, y_i, z_i) > 0 \end{cases}$$

$$\text{for } \forall x_i, y_i, z_i \in \left\{ -\frac{D}{2}, -\frac{D}{2} + 1, \dots, \frac{D}{2} - 1 \right\}$$

The perfectly symmetrised densities S/W are eventually combined with the original density map V_{ori} :

$$V_{sym} = \frac{S}{W} \circ W_{bin} + V_{ori} \circ (1_{D \times D \times D} - W_{bin})$$

where S/W means the element-wise division of S by W , $1_{D \times D \times D}$ denotes a matrix filled with all 1s, the size of which, $D \times D \times D$ in pixels, is equal to that of V_{ori}). The weighted sum V_{sym} is calculated here to apply smooth density transitions where the peripheries of symmetrised regions meet the original map in order to minimise real-space artefacts.

Practically, a STAR file (Hall, 1991) is used to communicate the local symmetry information to 3D reconstruction jobs in RELION. Every entry in this file has a 6-parameter operator Op_{ij} plus the mask filename of M_i . If there exist multiple sets of identical regions ($i > 1$), different mask filenames along with their corresponding operators need to be specified. Examples of local symmetry STAR files can be found in section 4.2.4 and 4.2.5. Soft-edged binary masks (M_i) must be in MRC format with all the voxel values within range $[0, 1]$ inclusive. In the MRC header (Cheng et al., 2015), the coordinate of origin must be the geometric centre of the 3D volume. For example, given a 3D volume of size $200 \times 200 \times 200$ in pixels, the x, y, z coordinates can be any integers within range $[-100, 99]$ inclusive and the

coordinate of origin in the MRC header must be set to $(x, y, z) = (0, 0, 0)$ for a local symmetry mask. The positive densities $(0, 1]$ of the mask are filled into relative positions in the 3D data arrays of the MRC files, corresponding to the regions of interest in the density map of the molecular complex. Note that masks produced by some other programs might always have $(0, 1]$ positive values written around the geometric centres of the 3D data arrays, with the coordinates of origin set to specific values other than $(0, 0, 0)$ in the MRC headers to record the relative positions of the masks with respect to the corresponding regions in the molecular density map. Such masks must be converted to masks with origins set to $(0, 0, 0)$ in order to comply with the local symmetry definition in RELION. In a word, the binary $[0, 1]$ densities of masks should be in the absolute positions in their MRC format maps.

4.2.3 Local searches of local symmetry operators

To improve the accuracies of estimated local symmetry operators, an option ‘--search’ was implemented in the standalone executable ‘relion_local_symmetry’. The brute-force search algorithm tests all possible combinations of local symmetry operators within the ranges specified by the user and returns the operator with the lowest root-mean-squared (RMS) error. RMS errors are calculated between the masked region and all possible destinations of transformations specified by the operators.

Suppose that a mask M is provided around a specific region in map V_{ori} . For all $Op_i = [\varphi_i, \theta_i, \psi_i, \Delta x_i, \Delta y_i, \Delta z_i]$, $i = 1, 2, \dots, N$, or equivalently, all combinations of their rotational and translational parts $R_i^{3 \times 3}$ and $T_i^{3 \times 1}$ to be tested, calculate:

$$RMS_{min} = \underset{i}{\operatorname{argmin}} \sqrt{\frac{\sum_x \sum_y \sum_z M \circ (V_{ori} - R_i^{-1} * T_i^{-1} * V_{ori})}{\sum_x \sum_y \sum_z M}}$$

$$\text{for } \forall x, y, z \in \left\{ -\frac{D}{2}, -\frac{D}{2} + 1, \dots, \frac{D}{2} - 1 \right\} \text{ s. t. } 0 < M(x, y, z) \leq 1$$

Op_i , which minimises the above argument, is the refined operator.

A local symmetry operator can be refined successfully given reasonable searching ranges and sufficiently fine steps of these 6 free parameters. Sometimes optimised operators appear unstable, moving around within small ranges (e.g. ± 1 pixels) due to real-space interpolation errors and slight structural variations across regions on the unsymmetrised original map. The implementation is also parallelised using MPI to accelerate the searches. The usage of ‘relion_local_symmetry’ program will be explained in section 4.2.4 and 4.2.5.

4.2.4 Creating local symmetry masks and operators for a 13-microtubules with seam

This section uses the 13-microtubule in figure 4.2B as an example, the symmetry of which has been explained in full details in section 4.1, and the helices always span along z axes of the particle boxes according to the definition in RELION helical processing (section 2.2.1).

It is recommended to reconstruct the 13-microtubule firstly as if all the tubulins are identical, without considering the presence of seam. For example, the overall helical symmetry of -27.7° and 9.46 \AA can be used to reconstruct a density map where the differences between $\alpha\beta$ tubulins are ignored. Such initial density maps are useful for mask generation, which is a prerequisite step for applying local symmetry in RELION.

As the next step, if the initial density map (from other sources) already reveals the seam, the PDB models of α and β tubulins can be docked into the proto-filament #7 which is opposite to the seam. Whereas starting from a map where all tubulins are the same, the PDB models of α and β tubulins can be docked into an arbitrary proto-filament and such proto-filament is defined as #7 for the calculation of local symmetry operators explained in the following paragraphs. It is worth mentioning that the subunits close to the top and bottom of the particle box (within a distance of $6 \times 9.46 \text{ \AA} = 56.76 \text{ \AA}$, the largest `_rlnOriginZ` absolute value in the STAR file 'filament_round001.star' below) should not be included in the PDB model in order to prevent real-space artefacts due to large translations along z when applying the local symmetry. The fitted PDB models are subsequently converted to an MRC format map, containing densities of the central parts of the proto-filament #7. It is also important to check whether the converted MRC densities can be superimposed with the original map in UCSF Chimera (Pettersen et al., 2004), and that both the original and converted maps have the same map origin defined as (0, 0, 0). If not, fit and resample the densities onto the original using the command line in UCSF Chimera and save the resampled copy map '#3':

```
vop resample #1 onGrid #2
```

where map '#1' and '#2' denote volumes of converted and original densities respectively.

Afterwards, a soft-edged binary mask ('1fil-mask.mrc') around the converted densities on proto-filament #7 can be generated by 'relion_mask_create' executable with a proper binary threshold.

```
relion_mask_create --i converted_resampled.mrc --ini_threshold 0.01 --  
extend_inimask 0 --width_soft_edge 3 --o 1fil-mask.mrc
```

As explained in section 4.1, the relative rotations and translations between proto-filaments are multiples of $(N \times) -27.7^\circ$ and 9.46 \AA respectively. The mask is provided as if the masked densities are on the proto-filament #7, thereby the possible integer values of N are -1 to -6 and +1 to +6 (inclusive), so that the proto-filament #7 can be transformed and superimposed onto proto-filament #8 to #13 and #1 to #6

(inclusive) in real-space. In this case, $N = 1$ and $K_l = 13$. Thus, the 12 local symmetry operators are written into the following STAR file 'filament_round001.star'. Only ψ (`_rlnAnglePsi`, in degrees) and ΔZ (`_rlnOriginZ`, in Angstroms) values need to be calculated whereas all the other values are left as zeros.

```
data_
loop_
  _rlnMaskName #1
  _rlnAngleRot #2
  _rlnAngleTilt #3
  _rlnAnglePsi #4
  _rlnOriginX #5
  _rlnOriginY #6
  _rlnOriginZ #7
1fil-mask.mrc      0.00      0.00     -27.69      0.00      0.00      -9.46
1fil-mask.mrc      0.00      0.00     -55.38      0.00      0.00     -18.92
1fil-mask.mrc      0.00      0.00     -83.08      0.00      0.00     -28.38
1fil-mask.mrc      0.00      0.00    -110.77      0.00      0.00     -37.84
1fil-mask.mrc      0.00      0.00    -138.46      0.00      0.00     -47.30
1fil-mask.mrc      0.00      0.00    -166.15      0.00      0.00     -56.76
1fil-mask.mrc      0.00      0.00      27.69      0.00      0.00      9.46
1fil-mask.mrc      0.00      0.00      55.38      0.00      0.00     18.92
1fil-mask.mrc      0.00      0.00      83.08      0.00      0.00     28.38
1fil-mask.mrc      0.00      0.00     110.77      0.00      0.00     37.84
1fil-mask.mrc      0.00      0.00     138.46      0.00      0.00     47.30
1fil-mask.mrc      0.00      0.00     166.15      0.00      0.00     56.76
```

If the helical twist and rise of the full 2D surface lattice are not accurate enough, ψ and ΔZ values can be locally optimised for all the manually calculated symmetry operators, by using the following command. Further details of local searches of operators will be given in the next section.

```
mpirun -n 30 relion_localsym_mpi --search --i_map input.mrc --i_mask_info
filament-round001.star --o_mask_info filament-round002.star --angpix 1.34 (-
-psi_range 1.0 --ang_step 0.1) --offset_z_range 2.0 --offset_step 0.1
```


In the actual 3D reconstructions, the helical twist and rise along every proto-filament, 0° and 82.0 \AA , can be used as the initial helical settings, together with the list of operators given above. The option to apply local symmetry ('--local_symmetry') in 3D reconstructions will be introduced in section 4.2.7.

4.2.5 Creating local symmetry masks and operators for arbitrary cryo-EM density maps

In ordinary single particle reconstructions, regions with similar densities can appear anywhere in the 3D particle box. Local symmetry operators have to be searched without that much prior information available in microtubules.

The following implemented algorithm calculates the centre-of-mass (COM) coordinates for all input masks in the first round of searches when initial estimations of local symmetry parameters do not exist. For example, for the AB dimer in figure 4.3, centre-of-mass coordinates of subunits A and B are calculated as $\overrightarrow{Com_A}$ and $\overrightarrow{Com_B}$. Regional boxes, slightly larger than the masks of A and B (i.e. all regions of interest), are cropped around their centre-of-mass coordinates in order to save time for subsequent parameter searches. All the masked densities are repositioned at the centres of cropped boxes, which enables the separation of orientation from translational searches. Three Euler angles ϕ , θ , ψ of all orientations are sampled and searched according to the user-defined angular ranges and step size. Given a rotational operator $R^{3X3} = (\phi, \theta, \psi)$, the translational vector T^{3X1} can be calculated as

$$\vec{T} = \overrightarrow{Com_B} - R * \overrightarrow{Com_A}$$

The translational vector T^{3X1} is subject to further local refinements. Afterwards, the local symmetry operator estimated in this pass is $Op = (R, T)$. Subsequent rounds of local searches are basically done in the same way, given the operators found previously and smaller ranges and sampling rates defined by the user. In principle, what the local searches do are to further refine the orientation angles and centre-of-mass coordinates of each region.

In the following, a protocol is described that serves as a general approach of finding local symmetry operators for single-particle reconstructions.

To start with, the user needs a density map and masks for all regions of interest. As an example, 3 plus 4 masks are needed for a complex with a trimer and a tetramer ($N = 2$, $K_1 = 3$, $K_2 = 4$). The density map can either come from a high-resolution reconstruction or from PDB atomic models that are fitted into such a map and then converted to densities ('partall_sums.mrc'). Fitted PDB structures provide more accountable information across various regions of interest and often lead to more accurate and stable local symmetry operators than the reconstructed cryo-EM density map itself, especially when some regions of the map are more blurred compared to their symmetry-related counterparts. The same PDB model should be fitted to each set of identical regions. A total of $K_1 + K_2 = 7$ properly centred, soft-

edged binary masks ('Trimer_mask1/2/3.mrc' and 'Tetramer_mask1/2/3/4.mrc') are then generated for all regions of interest as a start using 'relion_mask_create'.

```
relion_mask_create --i Trimer_pdb1.mrc --ini_threshold 0.01 --extend_inimask
0 --width_soft_edge 3 --o Trimer_mask1.mrc
```

...

A STAR file 'init_masklist.star' with all mask filenames should be generated as the only input text file, with related regions ($K_1 + K_2 = 7$) marked using positive integer IDs.

```
data_
loop_
_rlnMaskName #1
_rlnAreaId #2
Trimer_mask1.mrc 1
Trimer_mask2.mrc 1
Trimer_mask3.mrc 1
Tetramer_mask1.mrc 2
Tetramer_mask2.mrc 2
Tetramer_mask3.mrc 2
Tetramer_mask4.mrc 2
```

The above STAR file specifies 'Trimer_mask1/2/3.mrc' as the masks for all monomers of the trimer (ID = 1), and 'Tetramer_mask1/2/3/4.mrc' as the masks for all monomers of the tetramer (ID = 2). The next step is to perform initial global searches of the corresponding local symmetry operators.

```
mpirun -n 40 relion_localsym_mpi --search --i_map partall_sums.mrc --
i_op_mask_info init_masklist.star --o_mask_info maskinfo_iter000.star --
angpix 1.34 --ang_step 5
```

The above command performs global orientation searches on three Euler angles ϕ , θ , ψ at an angular sampling rate of 5° using the reconstructed (or PDB converted) density map 'partall_sums.mrc' (with a pixel size of 1.34 Å) and masks specified by 'init_masklist.star'. MPI parallelisation distributes the workload equally onto 40 CPU cores for acceleration. A non-parallel version is also implemented in 'relion_localsym'. For global searches, an angular step coarser than 5° might result in worse results with poorer convergence behaviour in the following local searching steps, and a much finer sampling may take too much computation time. Binning the cropped boxes by a factor of 2 (additional option '-bin 2') will also lead to better efficiency. The results are written into the local symmetry STAR file

‘maskinfo_iter000.star’. All the Euler angles ($_rlnAngle^*$) are in degrees and translations ($_rlnOrigin^?$) in Angstroms.

```
data_
loop_
_rlnMaskName #1
_rlnAngleRot #2
_rlnAngleTilt #3
_rlnAnglePsi #4
_rlnOriginX #5
_rlnOriginY #6
_rlnOriginZ #7

Trimer_mask1.mrc      -70.00   45.00   70.00   29.71   -7.65   -84.00
Trimer_mask1.mrc      155.00   30.00  175.00  -48.97  -32.08  121.36
Tetramer_mask1.mrc    130.00   70.00  -50.00  177.25   43.81 -131.91
Tetramer_mask1.mrc    -70.00   55.00   55.00   17.07   -9.14  -89.64
Tetramer_mask1.mrc    110.00   30.00 -110.00 -33.46  -61.76  139.06
```

This is the exact format of a local symmetry description file for the purpose of 3D reconstructions in RELION. Note that only one mask is used for each set of regions (‘Trimer_mask1.mrc’ for the trimer and ‘Tetramer_mask1.mrc’ for the tetramer) in this output file and the number of operators for trimer and tetramer are $K_1 - 1 = 2$ and $K_2 - 1 = 3$. The centre-of-mass positions of other masks (*_mask2/3/4.mrc) can be deducted by the $2 + 3 = 5$ operators, and thereby not kept in this STAR file. The above 5 operators need to be improved step by step with finer samplings in local searches:

```
mpirun -n 40 `which relion_localsym_mpi` --search --i_map partall_sums.mrc
--i_mask_info maskinfo_iter000.star --o_mask_info maskinfo_iter001.star --
angpix 1.34 --ang_range 2 --ang_step 0.5 --offset_range 2 --offset_step 0.5
```

The same density map ‘partall_sums.mrc’ must be used throughout the whole process of finding the operators and the filenames and densities of masks must also remain unchanged as well. Both the input density map and the translational operators in the local symmetry STAR file should also agree with the same pixel size given by ‘--angpix X’. The above command generates a list of refined operators ‘maskinfo_iter001.star’ from input ‘maskinfo_iter000.star’, using angular and translations searches of every 0.5° within $\pm 2^\circ$ and every 0.5 \AA within $\pm 2 \text{ \AA}$ respectively. The translational searches (of the center-of-mass coordinates) are not expected to require large searching ranges (‘--offset_range X’, X is

in Angstroms and usually an equivalence of $\leq 3 \times$ pixel size) as long as the input masks and densities are similar in every set of identical regions. The input STAR files can always be split into subsets of operators if any set of regions need different searching strategies from others, and combined before the actual 3D reconstructions.

It is recommended to perform multiple rounds of searches with gradually decreased searching ranges and steps. A general searching scheme in table 4.1 is designed so that each round lasts for ~2 minutes with ~100,000 samplings for each operator using a cropped box size of <100 pixels on a 30-core CPU machine (Intel Xeon E5-2643, 3.4-3.5 GHz):

Round	Angular ranges \pm ($^{\circ}$)	Angular samplings ($^{\circ}$)	Translational ranges \pm (pix)	Translational samplings (pix)
1	(360, all angles)	5	--	--
2	10	1	--	--
3	2	0.5	2	0.5
4	2	0.1	--	--
5	0.5	0.1	0.5	0.1

Table 4.1. An example of searching scheme for local symmetry operators in single particle analysis.

There is no need to search with very fine samplings ($<0.1^{\circ}$ or <0.1 pixels) as real-space interpolation errors become the dominant factor influencing the RMS errors. The optimal searching schemes may vary for different structures. For some structures, it might be worthwhile using the same settings several times until the refined values are all stabilised. Proceed the searches for all operators until all of them converge to required precisions (usually $\sim 0.1^{\circ}$ and ~ 0.1 pixels). Note that operators with low tilts ($< +10^{\circ}$) are often not stable since the ϕ (`_rlnAngleRot`) and ψ (`_rlnAnglePsi`) angles are inseparable at low θ ($\leq 30^{\circ}$, `_rlnAngleTilt`) values. For such cases, the searching ranges of θ and either ϕ or ψ can be set to zero to limit the orientation refinements to only one Euler angle for stability.

4.2.6 Verification of local symmetry operators

One way to verify the optimised operators is to apply the local symmetry to a previously reconstructed density map ('unsym.mrc') and inspect the symmetrised densities ('sym.mrc'). The symmetrised regions on the output map should be at least as clear as the unsymmetrised regions on the input map if the correct local symmetry operators and masks ('mask_and_operators.star', similar to 'maskinfo_iter000.star' in the above section 4.2.5) are provided.

```
relion_localsym --apply --angpix 1.34 --i_map unsym.mrc --i_mask_info
mask_and_operators.star --o_map sym.mrc
```

Verification can also be performed by replicating the masks according to the symmetry operators. This method might be beneficial for flexible complexes where some regions appear more blurry than their counterparts and applying local symmetry to the original map makes all regions vague:

```
relion_localsym --duplicate ('--i_map unsym.mrc') --angpix 1.34 --
i_mask_info mask_and_operators.star --o_map replicated.mrc
```

Masks (or masked densities) in the input STAR file are copied, rotated and repositioned onto every centre-of-mass coordinates specified by the local symmetry operators accordingly. Original density map ('unsym.mrc') is optional for this command. If the original map is provided, masked densities, instead of the masks in the local symmetry description file ('mask_and_operators.star'), are subject to the transformations described above. The output map ('replicated.mrc') can be displayed with the original density map ('unsym.mrc') simultaneously in UCSF Chimera (Pettersen et al., 2004). The contours of all masks (or masked densities) should enclose (or fit into) the original protein densities well if the operators are correct.

4.2.7 Enabling local symmetry option in 3D reconstructions

Once the correct local symmetry operators have been obtained, they can be used inside any 3D classification/refinement jobs using the option '--local_symmetry mask_and_operators.star'. Imposition of local symmetry has masking effects in real-space so it is always disabled in the final iterations of 3D auto-refinement jobs where unfiltered half-maps are compared for Fourier-shell correlations. The same sets of operators can also be applied to the post-processed maps. Unlike helical symmetry, local auto-refinement of local symmetry operators is not implemented during 3D reconstructions since there are 6 free parameters in every operator and the operators might diverge given a heavily low-pass filtered density map as the 3D initial reference.

4.3 Introduction to the test data set of 13-microtubule with seam

A small cryo-EM data set of 13-microtubules with seams is used in order to test local symmetry option implemented in RELION. The data set was collected by Sam Lacey in Andrew Carter's group on FEI Polara 300 microscope with FEI Falcon-II detector with an actual pixel size of 1.38 Å. The data set, consisting of 1,263 movie stacks each with 30 frames, was previously motion-corrected by MOTIONCOR2 (Zheng et al., 2017). A total number of 187,737 segment coordinates were auto-picked and extracted using a picking threshold of 0.20, maximum curvature of 0.3 and an inter-box distance of 82 Å (~15% of 552 Å or 400-pixel box size). 176,853 segments were kept after 2D classifications.

4.4 Results and discussion

3D auto-refinements and subsequent post-processing were performed with and without local symmetry operators. For both reconstructions, initial helical twist = 0° and rise = 82.0 Å (equivalent to $H = 1$) were applied on every proto-filament. The inner and outer diameters were set to 100 Å and 400 Å respectively. The initial 3D model was a density map of 13-microtubule with all identical tubulins low-pass filtered to 30 Å (figure 4.4A). In order to apply the local symmetry, a soft-edged binary mask ‘1fil-mask.mrc’ was made around the central 4 tubulins (two pairs of $\alpha\beta$ tubulins) along the proto-filament opposite to the seam (#7 proto-filament marked in figure 4.4C, right). The STAR file containing the local symmetry operators used here was the same as the file ‘filament_round001.star’ in section 4.2.4. After 3D auto-refinements, the twist and rise were refined to $+0.1^\circ$ and 83.1 Å. Multiple views of the post-processed density maps are listed in figure 4.4CDE. Although the post-processed FSC curves are similar with claimed resolution at 4.2 Å, the reconstruction with local symmetry has resolved clearer peripheral densities for tubulins close to the seam (figure 4.4CDE, right) while such densities are totally blurred out in the other reconstruction (figure 4.4CDE, left). However, the peripheral densities close to the seam are still slightly dampened even if with local symmetry applied since there are a mixture of various types of microtubules (e.g. 12-, 14-microtubules and those without seams) in the data set. This test demonstrates that local symmetry option in RELION is useful for reconstructions of microtubules with seams. This resolution (4.2 Å) achieved using RELION is better than similar microtubules resolved (5-9 Å) with alternative software packages (Sindelar and Downing, 2007; Atherton et al., 2014; Shang et al., 2014).

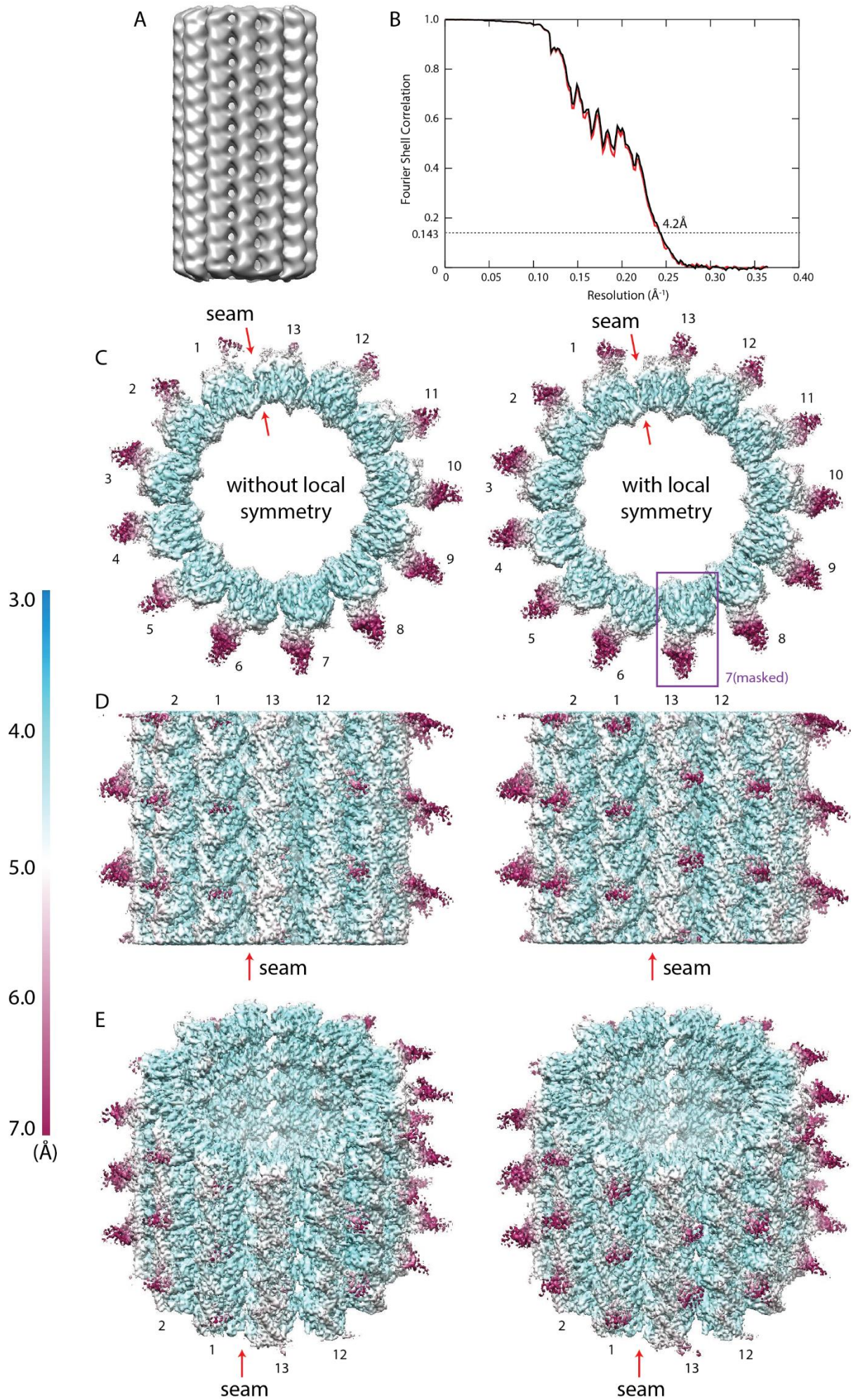


Figure 4.4. Results of applying local symmetry to a 13-microtubule in 3D auto-refinements. (A) Low-pass filtered to 30 Å, an initial 3D model of 13-microtubule with all identical tubulins (ignoring the seams) is used in both reconstructions. (B) Post-processed FSC curves of 13-microtubule auto-refinements with (red) and without (black) local symmetry operators. (C, D, E) bottom, side and tilted views of post-processed maps with local resolution estimates, panels on the left/right are the results without/with local symmetry operators. The positions of the seam (lateral $\alpha\beta$ tubulin contacts) are marked with vertical red lines. The 13 proto-filaments are numbered in the same way as in figure 4.2B.

5 SOLVED STRUCTURES

This chapter briefly describes cryo-EM helical structures from four research papers. These data sets have all been processed with the helical reconstruction workflow in RELION 2.0, either by myself or some of the co-authors independently. Through our collaborations, these projects have revealed exciting structural details and provided better understanding of some important biological processes. The related research papers are Deng et al., 2017; Salih et al., 2017; Fitzpatrick et al., 2017 and Alvarez et al., 2017. Key information of the data sets is summarised in Table 5.1.

	TssB,C/T6SS	BtubABC	MxB	Tau (PHF)	Tau (SF)
Microscope	FEI Titan Krios	FEI Polara 300	FEI Polara 300	FEI Titan Krios	
Detector	FEI Falcon-II	Gatan K2	FEI Falcon-II	Gatan K2	
Pixel size (Å)	1.34	1.34	1.147	1.04	
Movies collected?	Yes	Yes	Yes	Yes	
Nr. Movies/micrographs	1,405	6,105	630	>2,000	
Nr. frames per movie	19	30	7	50	
Accumulated dose (e ⁻ /Å ²)	~30	~40	~40	~60	
Defocus range (-,µm)	1.0-2.5	1.5-3.0	1.5-3.5	1.0-3.0	
	RELION-2 processing of test data sets				
Manual/auto-picked?	Auto	Auto	Auto	Manual	
Auto-picking threshold	0.0	0.2	0.0	-	
Auto-picking maximum curvature	0.05	0.20	0.10	-	
Box size (pixels)	350	280	450	270	270
Inter-box distance (Å)	33.6	80.0	49.4	28.0	28.0

Nr. asym. units/box, H	10	1 ²	6	12	6
Nr. segments (picked)	107,157	345,777	51,553	214,757	84,701
Nr. segments (refined)	71,264	257,656	44,955	23,086	24,689
Filament widths (Å)	100-285 ⁴	150	55-275 ⁴	~100	~100
Point group symmetry	C ₁	C ₁	C ₁	C ₁	C ₁
Helical rise, refined (Å)	3.359	79.31 ²	8.237	2.36	4.74
Helical twist, refined (°)	+55.40	-5.54 ²	+58.4	+179.4	-1.05
B-factor (Å²)	-76	-92	-163	-105	-114
Final resolution (Å)	3.3	3.6	4.6	3.4 ⁵	3.4 ⁵
EMDB entries	3600 ¹	3726	8577 ¹	3741	3743
PDB entries	5N8N ¹	5O01, 5O09 ³	5UOT ¹	5O3L	5O3T

¹ Structures which might not be publicly accessible at the time of thesis submission (October 2017).

² These parameters correspond to the helical symmetry on every proto-filament in the BtubAB(C) structure.

³ The entries 5O01 and 5O09 correspond to the atomic models of BtubC and BtuBAB respectively.

⁴ Inner and outer diameters of the hollow tubes.

⁵ The resolution estimates of full-length tau PHF and SF structures.

Table 5.1. Key characteristics of the four projects.

5.1 BtubABC mini microtubule

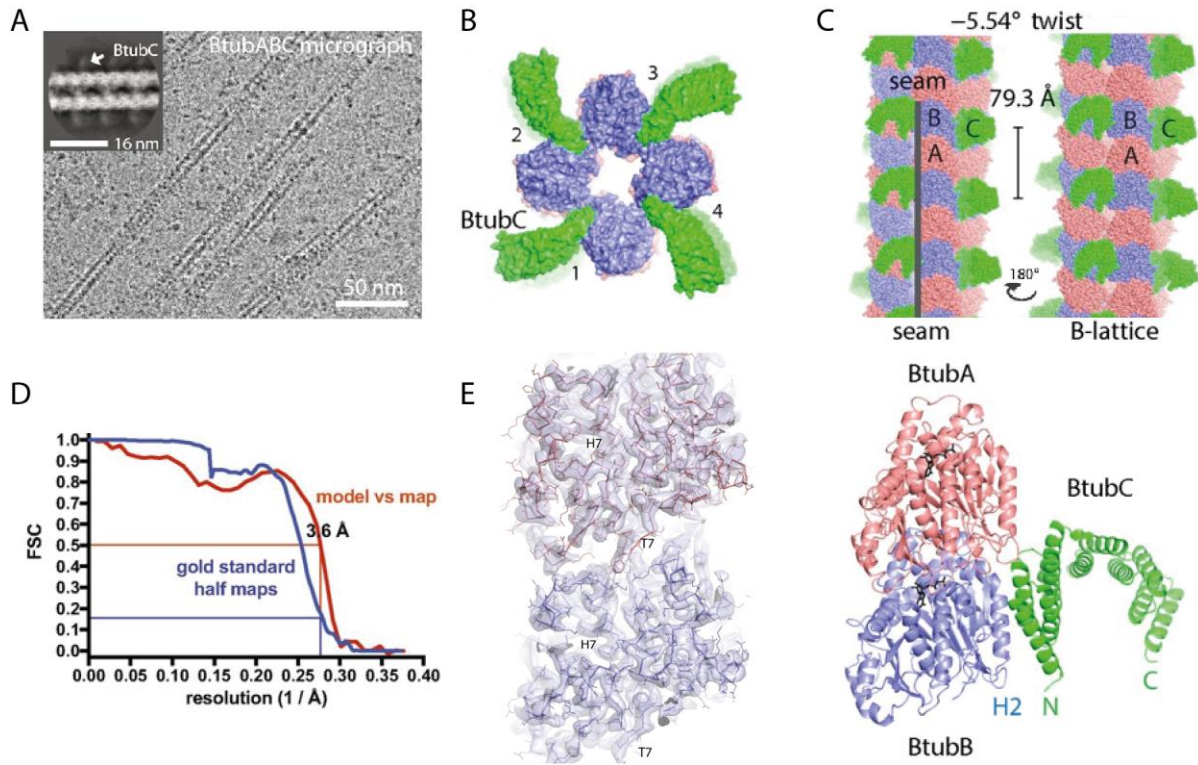


Figure 5.1. Cryo-EM data analysis of BtubABC mini microtubule. (A) A representative cryo-EM micrograph of *in vitro* polymerized BtubABC and a 2D class average showing the binding of BtubC at every ~80 Å. (B) Top-view along the filament axis showing the 4-protofilaments in purple and the protruding BtubC subunits in green. (C, left) The side-view highlighting the seam showing BtubA-BtubB lateral contacts. (C, right) Another side-view (opposite to the seam) by a rotation of 180° showing BtubA-BtubA and BtubB-BtubB lateral contacts. (D) The FSC curve between cryo-EM density half-maps after post-processing and the model-vs-map FSC curve. (E) Cryo-EM densities and fitted atomic models of BtubA and BtubB. (F) The atomic models of BtubA, BtubB and BtubC subunits (all the panels in this figure are adapted from Deng et al., 2017).

As explained in section 4.1, hollow tubes of microtubules can be assembled by $\alpha\beta$ -tubulin heterodimers, which were previously thought to be present only in eukaryotic cells. However, certain bacterial *Prostheco bacter* species contain BtubAB proteins that can also form hollow tubes. Deng et al., 2017 produced unmodified *P. dejongeii* BtubAB proteins in *E. coli* and long, straight 4-protofilament tubes were observed by cryo-ET when GTP was added to BtubAB. I mainly provided suggestions on the analysis of helical symmetry for this project. The authors firstly performed cryo-ET sub-tomogram averaging of BtubAB filaments and manually derived pseudo-helical parameters (twist = -90° , rise = 10 Å) through inspection of the map. These parameters averaged BtubA and BtubB together and ignored the presence of a seam. Helical reconstructions of cryo-EM BtubAB data set refined the pseudo-helical parameters to -90.7° , 10.4 Å and yielded a final map at 4.2 Å resolution. Since BtubA and BtubB are extremely similar in structural details (figure 5.1EF), a new feature was required to mark the ~80 Å

heterodimer repeat on each proto-filament in order to distinguish BtubA from BtubB during 3D reconstructions. The authors then expressed the third gene *bklc* in the *btub* gene cluster and found that *P. dejongei* BtubC binds to BtubAB filaments in a stoichiometric fashion (figure 5.1ABC). For this BtubABC data set, the initial helical twist and rise were set to -5.6° ($-0.7^\circ \times 4 \times 2$) and 80.0 \AA ($10 \text{ \AA} \times 4 \times 2$) respectively to describe the symmetry on each individual proto-filament only. Thereby, BtubA and BtubB subunits could be separated, details of the seam were preserved but the four identical proto-filaments (figure 5.1B, marked by 1, 2, 3 and 4) were not averaged (also see section 4.1). This solution was adopted as a work-around because local symmetry option in RELION had not been implemented at that time in 2016. Helical parameters of each BtubABC proto-filament were refined to -5.54° and 79.31 \AA . The overall resolution of the post-processed map was 3.6 \AA (figure 5.1D). The densities of BtubC were not clearly resolved by cryo-EM data processing since BtubC proteins bound to the filaments at specific locations but with random orientations. The crystal structure of BtubC (PDB: 5O01) was independently solved by X-ray crystallography. The project shows that, compared to eukaryotic microtubules, BtubABC mini microtubules are much smaller in diameter and the number of constituent proto-filaments. But they still share common features with ordinary microtubules such as alternating subunits, a seam that breaks the overall helical lattice, and dynamic instability (treadmilling). The study sheds light on the evolution of microtubule structures and provides tools for investigation of filament dynamics (Deng et al., 2017).

5.2 MxB (myxovirus resistance proteins)

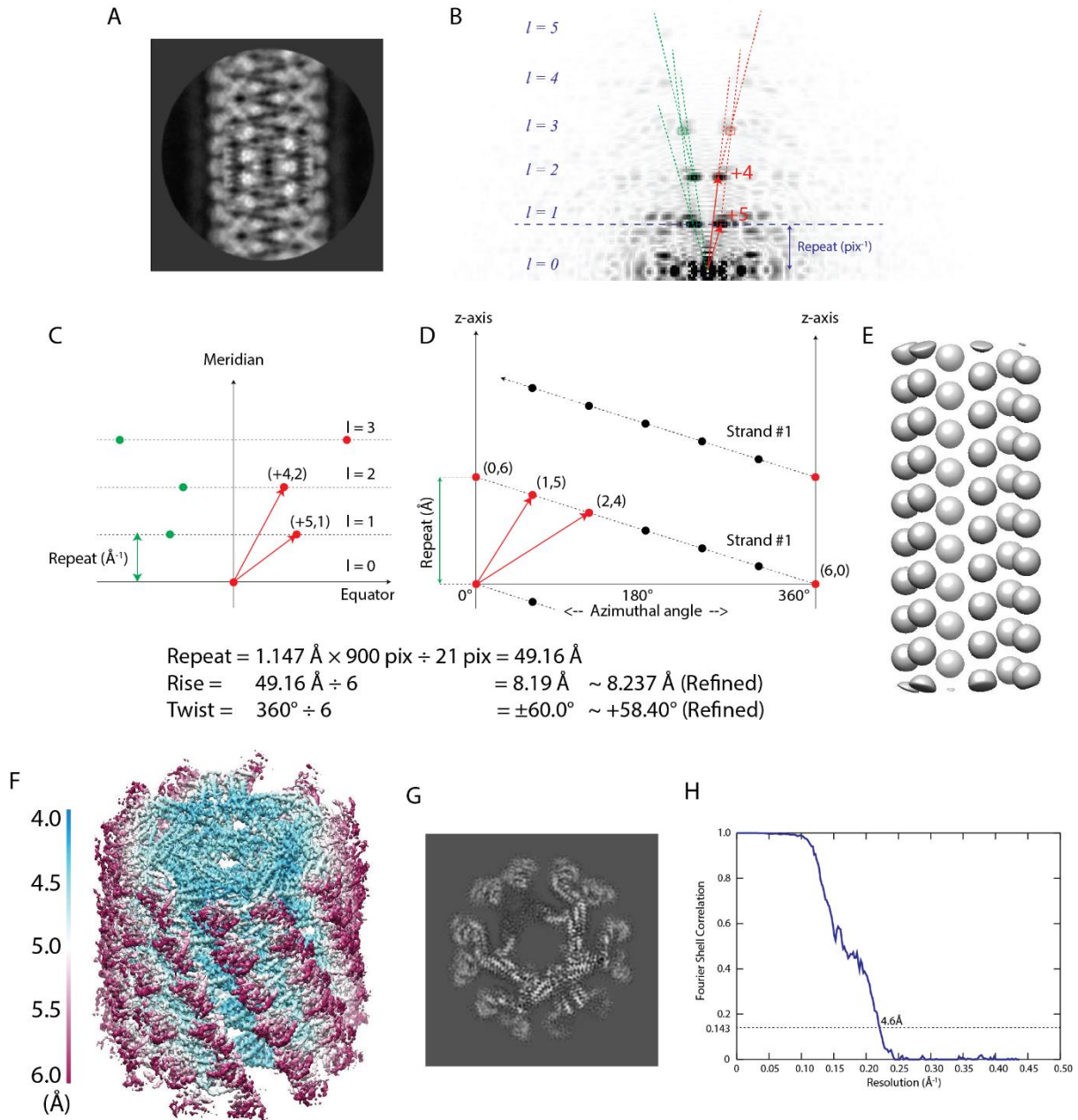


Figure 5.2. Cryo-EM data analysis of MxB (myxovirus resistance proteins). (A) A representative 2D class average. (B) The diffraction pattern of (A) and the Fourier-space helical lattice. (C) The indexing scheme of (B). (D) The real-space helical lattice deduced from (C). (E) The side view of the simulated initial 3D reference. (F) The local resolution heat-map of the post-processed density map. (G) The central z section of the post-processed density map. (H) The post-processing FSC curve.

Mx (myxovirus resistance) proteins are important restriction factors in the interferon response against viruses. In mammals, the two isoforms are classified as Mx1 (MxA)-like and Mx2 (MxB)-like proteins. Human dynamin-like Mx2/MxB has shown strong inhibition of HIV-1 replication. The antiviral

activities of MxB depend on both the N-terminal region and oligomerisation, and their structural details were previously not clear due to difficulties in sample preparations. Frances et al., 2017 expressed and purified MBP-fused full-length wild-type MxB and the protein oligomers self-assembled into helical lattices. I processed this cryo-EM data set using helical reconstruction workflow implemented in RELION. As is seen from the 2D class averages (figure 5.2A), the MxB helices have an outer diameter of ~ 275 Å with some faint unresolved densities extended to ~ 360 Å. Fourier-Bessel analysis has revealed that the helical structure is 1-start and contains ~ 6 subunits within each ~ 50 Å pitch (figure 5.2BCD). The data processing followed the ordinary workflow explained in section 2.1. The 3D-refinement started from the estimated symmetry ($+60.0^\circ$ and 8.19 Å) and a simulated lattice low-pass filtered to 30 Å (figure 5.2E). The final resolution was 4.6 Å after post-processing (figure 5.2H) and the helical twist and rise were refined to $+58.4^\circ$ and 8.237 Å respectively. This is the first near-atomic resolution structure of the mammalian dynamins. The authors has identified novel interfaces for MxB oligomerisation which are crucial to its assembly and anti-HIV-1 functions through inspection of the density map and additional mutational analyses (Frances et al., 2017).

5.3 Type VI Contractile Sheath TssB1C1 from *Pseudomonas aeruginosa*

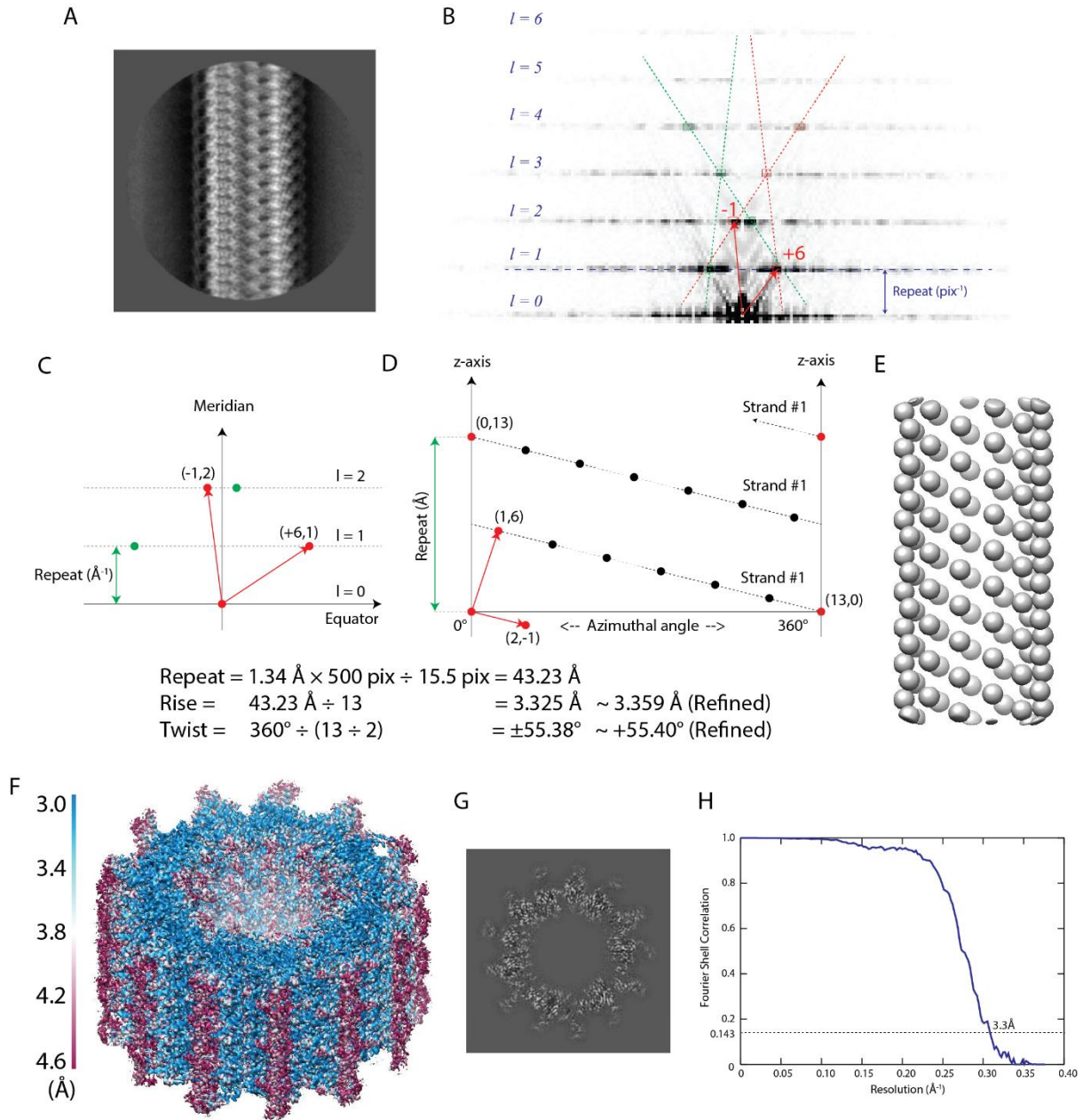


Figure 5.3. Cryo-EM data analysis of the Type VI Contractile Sheath TssB1C1 from *P. aeruginosa*. (A) A representative 2D class average. (B) The diffraction pattern of (A) and the Fourier-space helical lattice. (C) The indexing scheme of (B). (D) The real-space helical lattice deduced from (C). (E) The side view of simulated initial 3D reference. (F) The local resolution heat-map of the post-processed density map. (G) The central Z section of the post-processed density map. (H) The post-processing FSC curve.

The type VI secretion system (T6SS) is widely present in Gram-negative bacteria and it is an injection machine that utilises the contraction of a sheath structure to deliver toxins or effectors into bacteria or

eukaryotic cells. The sheath in T6SS is a tubular structure made of two proteins TssB (VipA/IglA) and TssC (VipB/IglB). The paper by Salih et al., 2017 presents the group-3 TssB1C1 sheath encoded by the H1-T6SS gene cluster (TssB1C1) from *P. aeruginosa*. I was in charge of the whole cryo-EM data processing workflow of this project. The co-authors initially applied helical twist and rise which conform to 12-fold with C_6 point group symmetry since all known contractile systems display such symmetry in native conditions (for example, the VipA/VipB structure used as a test data set in section 2.4, Kudryashev et al., 2015). However, both the asymmetric 2D class averages (figure 5.3A) and the Fourier-space helical lattices (figure 5.3B) indicate that the symmetry is not likely to be 12-fold or rotationally symmetrical around the helical axis. The 2D class densities (figure 5.3A) are apparently asymmetrical across the meridian. Furthermore, no maxima are found along the meridian of the diffraction pattern (figure 5.3B), which suggests that the structure lacks any rotational C_n ($n \geq 2$) symmetry. The Fourier-Bessel analysis shows that its helical twist and rise are $\sim +55.4^\circ$ and ~ 3.3 Å respectively, leading to a 13-fold symmetry structure with exactly two pitches in every helical repeat (figure 5.3CD). This estimated symmetry also agrees with the evidence that some of the short helical filaments display 13-fold cogwheel-like top-views on the motion-corrected micrographs. Thus, Salih et al. assumed that the sample preparation must have induced the structures into a predominant 13-fold symmetry. Once the helical symmetry was correctly resolved, the data processing followed the ordinary workflow explained in section 2.1. The 3D-refinement started from the estimated symmetry ($+55.4^\circ$ and 3.3 Å) and a simulated lattice low-pass filtered to 20 Å (figure 5.3E). The final resolution was 3.3 Å after post-processing (figure 5.3H) and the helical twist and rise were refined to $+55.40^\circ$ and 3.359 Å respectively. The claimed overall resolution of TssB1C1 sheath is the highest among all the helical data sets that I have processed with RELION during my PhD. Based on the resolved cryo-EM structure, the authors has proposed an extended spring-type mechanism of the T6SS sheath contraction. It has also been discovered that this system is highly conserved in mechanism and functions compared to other contractile assemblies, such as T4 bacteriophages and R-type pyocins (Salih et al, 2017).

5.4 Tau filaments

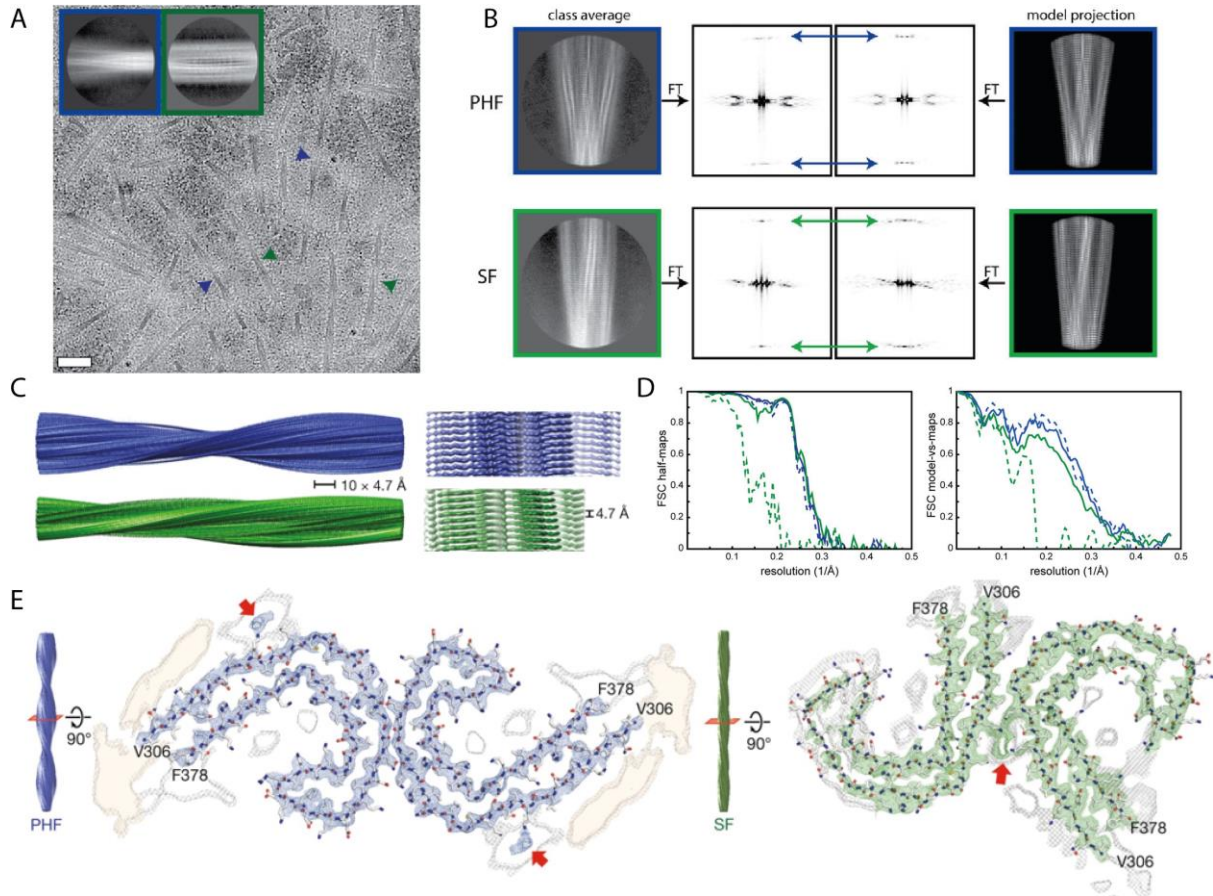


Figure 5.4. Cryo-EM structures of tau filaments from Alzheimer's disease. (A) A representative cryo-EM micrograph and 2D class averages (blue insets for PHFs, green insets for SFs). (B) Power spectra of 2D class averages and the corresponding side view projections ($\theta = 90^\circ$) of the atomic models. (C) The side views of the reconstructed cryo-EM density maps of PHFs and SFs. (D) FSC curves between two independently refined half-maps for the full-length (FL) PHFs (blue, solid) and SFs (green, solid). (E) Cross-sections of the PHF and SF cryo-EM structures (with fitted atomic models) (all the panels in this figure are adapted from Fitzpatrick et al., 2017).

As the most common neurodegenerative illness, Alzheimer's disease is characterised by abnormal aggregations of neurofibrillary lesions and neuritic plaques in a patient's cerebral cortex. The lesions consist of paired helical and straight tau filaments (PHFs and SFs). Helical reconstruction in RELION has revealed the first high-resolution structures of such filaments. Fitzpatrick et al., 2017 purified samples of cerebral cortex from a patient diagnosed with the disease and they processed the data sets on their own. Negative staining results showed that the PHFs are $70 \text{ \AA} - 150 \text{ \AA}$ wide with crossover distances of $650 \text{ \AA} - 800 \text{ \AA}$ and the SFs are $\sim 100 \text{ \AA}$ wide uniformly with crossovers of $700 \text{ \AA} - 900 \text{ \AA}$. The cores of both PHFs and SFs consist of two proto-filaments, each of which is a helical stack of C-shaped subunits, whereas the N- and C-terminal regions of tau proteins extend outward to form

disordered fuzzy coats. The diffraction patterns of 2D class averages (figure 5.4A) show little information between the equator and the only layer line at ~ 4.7 Å (figure 5.4B). The absence of signals at medium-range resolutions posed challenges to resolve the structural details along the helical axis of such filaments. The authors set greater T values (≥ 8) to force RELION to make use of higher spatial frequencies during 3D alignment. The methods of incorporating helical prior information (section 2.2.6) were crucial for the determination of the PHF and SF structures. The authors also reported that the helical data processing benefits from additional settings such as the polarity correction, fixed tilt (θ) priors and the local averaging of alignment parameters of neighbouring segments. The high-resolution 3D reconstructions demonstrate that the C-shaped subunits are successive rungs of β -strands related by a helical rise of ~ 4.7 Å and a twist of $\sim -1^\circ$. The two proto-filaments in PHFs has an approximate 2_1 screw symmetry while such feature does not exist in SFs (figure 5.4E). The study is the first to elucidate the ordered core of proto-filament pairs of PHFs and SFs in a patient's brain of Alzheimer's disease. The structural information of this core, necessary for the seeding and propagation of tau filaments, is useful for the study of molecular mechanism and inhibitor design of tau related pathologies (Fitzpatrick et al., 2017). Helical reconstruction in RELION provides a powerful tool for cryo-EM structural determination of amyloid protein structures related to neurodegenerative diseases.

Apart from the four structures discussed above, helical reconstruction in RELION has also been used in the following research projects of structural determination of helices: Izoré et al., 2016; Short et al., 2016; Daum et al., 2017; Unwin, 2017.

6 FUTURE PROSPECTS

In conclusion, my PhD project has developed a general workflow of single-particle based helical processing in RELION 2.0. The thesis describes the key computational implementations and results in detail, as well as optimal settings of user-defined parameters and pitfalls in helical reconstructions. The methodology has opened up new possibilities of determining some exciting and unusual helical structures, including microtubules with pseudo-helical symmetry, actin-like filaments and amyloid fibres at near-atomic resolution, with reduced user intervention. This open-source computational project provides a powerful and free tool for cryo-EM structural determination of helices. In addition, it does not require much expertise for the new researchers of general structural biology to enter the field of helical processing. In the future, with more advanced EM facilities, I expect that new helical structures could be easily resolved to near-atomic resolutions with the help of the helical processing workflow in RELION.

Structural determination of membrane proteins might benefit from cryo-EM helical processing in particular. Well-ordered 3D crystals of membrane proteins are generally hard to obtain for X-ray crystallography. NMR analysis requires a large amount of labelled, stable proteins in detergents, which might not be feasible as well for membrane proteins especially from eukaryotes (Vinothkumar, 2015). As for the single-particle cryo-EM study of membrane proteins, over-expression of such proteins in cells is more difficult compared to soluble proteins. Even if the proteins are well-expressed, there may still be obstacles in extracting them from membranes in stable and non-aggregated states into solutions (Rawson et al., 2016). However, some membrane proteins can be induced to form 2D crystals in the form of helical tubes. Such helical lattices generate information-rich diffraction patterns and their helical symmetry can be conveniently figured out through indexing. Moreover, their helical lattices are curled into tubes which provide a pseudo-membrane environment for the proteins. Thereby, the structures revealed by single-particle helical reconstruction might have more resemblance to their *in vivo* states (for example, the AChR tubes, Unwin, 2017). The induced helical lattices may consist of a mixture of slow changing helical twist and rise values or totally different symmetries. In such cases, careful 2D/3D classifications must be performed in order to achieve high-resolution reconstructions.

The sizes of helical assemblies can be much larger than the ordinary single molecules. For example, a typical AChR tube (Unwin, 2017) has a diameter of ~ 800 Å and a length of several thousand Angstroms. The long and wide hollow tubes can be flattened in the middle and located at different depths in the ice

layer along its spanning directions. From an image processing point of view, the curvature of the Ewald sphere limits the highest resolution to which images can be easily interpreted, especially for images of thick specimen. In high resolution cryo-EM, the near-focus images superimpose the fringes from Fourier components created by interference between the unscattered beam and the two beams scattered from the specimen in opposing directions from the beam axis (Russo and Henderson, 2018). Once recorded as an in-focus image at the detector, these two approximately Friedel-related Fourier components created by the diffracted beams overlap and become indistinguishable in both real and reciprocal space. The Ewald sphere describes the geometry of this diffraction (figure 6.1). The conventional CTF correction (section 1.2) applies to the Fourier transform of the whole image, but it is only accurate for a flat Ewald sphere. Russo and Henderson (2018) proposes the single-sideband CTF correction which is computationally only slightly less efficient and allows separate determination of the amplitudes and phases of the two pseudo-Friedel-related Fourier components, provided the images are recorded with a sufficiently high defocus. According to their estimations, the method is expected to improve the structural determination for particles $> 800 \text{ \AA}$ in diameter at resolutions beyond 4 \AA using 300 keV electrons, which will be particularly beneficial for wide helical tubes. It has the potential to extract higher resolution information from thicker specimen using lower energy electrons and figure out the absolute hand of the specimen without the need of performing tilt pair analysis. In addition, higher order lens aberrations need to be corrected physically with additional lenses in the microscope or computationally during data processing for cryo-EM structures resolved near or above 2 \AA resolutions.

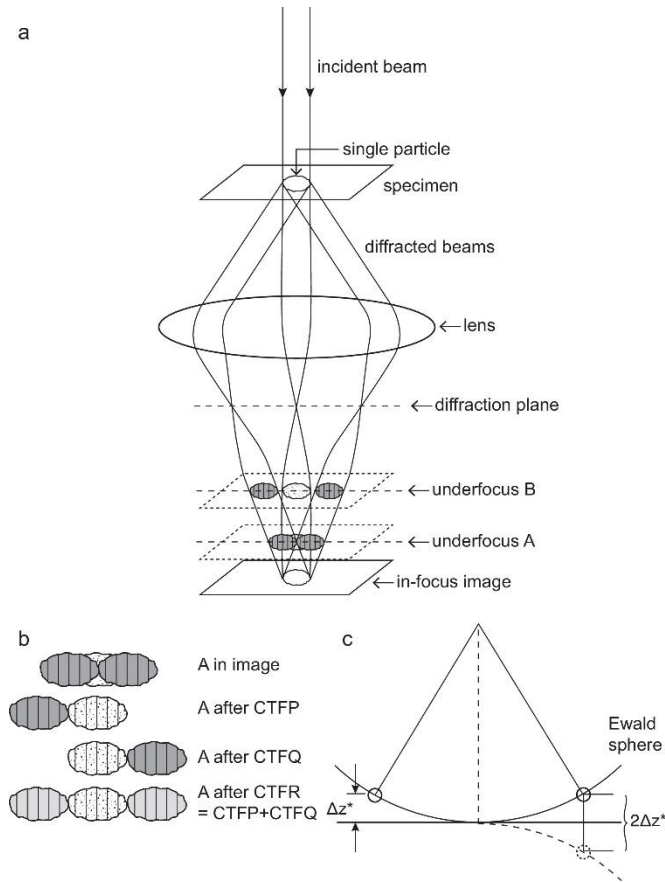


Figure 6.1. The Ewald sphere and Friedel-symmetry-related sidebands. (a) Diagram showing how underfocused images keep two approximately Friedel-related diffraction patterns completely separate so that they do not overlap in the image. Since they do not overlap physically in the image, their amplitude and phase can be recovered by an appropriate image processing procedure. At underfocus B, the two diffracted beams are fully resolved from the low-resolution shadow of the particle, whereas underfocus A is the minimum underfocus needed to avoid any overlap of the high-resolution information. (b) The result of applying the two CTFP and CTFQ procedures is shown in comparison with the result of applying conventional CTF correction. The conventional CTF correction, termed CTFR, is equivalent to the application of CTFP + CTFQ. (c) Diagram of the Ewald sphere construction, and how the deviation from the central section, Δz^* , is calculated. The two diffracted beams are shown intersecting the Ewald sphere on the right and left sides of the diagram. The intersection to the left is also shown as its Friedel-related partner, dashed, on the right. The $2\Delta z^*$ distance is the separation of the two Fourier components in reciprocal space (adapted from Russo and Henderson, 2018).

In the next few years, helical sub-tomogram reconstruction may provide more insights into cellular sub-structures combined with more advanced cryo-ET techniques (Koning et al., 2018). The achievable resolutions are being pushed higher for cryo-ET by ongoing developments of imaging and data collecting devices (Sharp et al., 2017), more accurate 3D-CTF models and improved tomographic reconstruction algorithms (Kunz and Frangakis, 2017; Turoňová et al., 2017). Cryo-ET has the potential to reveal any *in situ* cellular components in cells, and a variety of helical structures, such as filamentous and cytoskeletal assemblies, are attracting great interest in tomographic study. These assemblies are

often morphologically heterogeneous and such structural variations can be identified through helical sub-tomogram reconstructions across various tomograms (Milne et al., 2013). Their comparisons with cryo-EM helical reconstructions of *in vitro* filaments might also be beneficial to their structural and functional studies.

From the computational point of view, there is still room for improvement for single-particle helical processing. Currently, there is still not a single method which is reliable and universally applicable for initial model generations of all helices. Recent introduction of stochastic gradient descent into the RELION MAP-EM algorithm enables initial model generations of single-particle molecules, where initial 3D references can be generated using stringently low-pass filtered spherical densities as input and immediately fed into reconstructions jobs (Punjani et al., 2017). Unlike ordinary single particles, helical segments always span all the way along the helical axes and they are not at all similar to low-passed spheres. In addition, a strong low-pass filter risks wiping out almost all necessary information of the surface lattice, which makes correct convergence of such *de novo* initial model almost impossible. In addition, at the stage of initial model building, helical symmetry (and the number of asymmetrical units, H , within each segment) is not necessarily known or estimated, which makes the process even more difficult. The process may be further complicated by a mixture of slightly different helical symmetry parameters present in the data set. Section 2.4.3 in this thesis advises that there seems no one-size-fits-all solution to building initial references for helices and encourages researchers to explore various measures (including the traditional Fourier-Bessel analysis) which suit the properties of helical lattices, and helical symmetry should always be properly validated (using validation methods described in section 2.4.4). In the future, I am looking forward to new universal methods for building initial references of helices which do not rely on prior (symmetrical) information of the structure and can completely replace the Fourier-Bessel analysis, perhaps with the help of 2D class averages of helical segments.

RELION adopts Bayesian framework and many parameters of the statistical model are learned from the data, which leads to objective and high-quality results without the need for user expertise (Scheres, 2012). However, one free parameter, the regularisation parameter T , is currently not learnt from the input experimental images and can be set arbitrarily by the users to suit their specific needs. The regularisation parameter T determines the relative weight between the experimental data and the prior. A larger T value assigns more weight to the prior at every iteration in the reconstruction. Over-fitting effects can be remarkable if the T value is inappropriately large. Based on previous experiences, 2D and 3D classifications require values of 1-2 and 2-4 respectively to obtain better results for the reconstructions. In order to improve the single-particle analysis in general, it is possible that proper values of regularisation parameters T are estimated from the signal-to-noise ratios of both the input data set and references (during the iterative reconstruction) and the level of internal symmetry of the structure

(point-group, helical or local symmetry elements), given the dimensionality (2D or 3D) and the number of input particles and classes of the reconstruction. For highly symmetrical structures as helices, it is expected that larger values of T could speed up the convergence and still safely lead to reliable reconstruction results. Future plans of the development RELION might also involve auto-selection of 2D classes with the help of machine learning and/or pattern recognition techniques. For heterogeneous data sets of helices, it is desired to sort 2D class averages not only according to their image qualities, but also the helical symmetry (or diffraction patterns) and the filament widths. Currently as a common practice, researchers produce and inspect the class averages and diffraction patterns manually for cleaner subsets of data for 3D reconstructions. I expect reliable computation methods to speed up and automate the process. It could be beneficial to figure out different types of helices prior to 3D reconstructions, not only after rounds of careful 3D classifications.

Local resolution estimations of final post-processed maps always reveal that the best quality densities are located in the radially middle parts of the maps, while the outer and inner (if the curled helical lattices form hollow centres) parts are slightly blurred and yield lower local resolutions. The reason might be that the periphery of helices have direct contact with the solvent while the protein structures in the middle are surrounded by the surfaces therefore more stabilised. Particularly, the periphery (both inner and outer) of the helical wall provides positions for other proteins to bind to. So there are special interest in peripheral chain densities themselves and molecules which bind to them. Unfortunately, helical symmetry is assumed to be invariant for all voxels between the inner and outer diameters in the 3D volumes and any wobbling parts are smeared out due to the imposition of symmetry. Theoretically, signal subtraction might reveal some interesting details at the periphery, but hundreds of sub-masks around peripheral densities for every subunit are impractical to be generated, and there is no promise that this approach is going to work since the flexible parts might all take random orientations. Multi-body refinement (Nakane et al., 2018), which deals with the conformational changes and heterogeneities of the molecular structures, combined with the signal subtraction method may be feasible to explore the outer surfaces of helical structures.

Local symmetry was originally intended to tackle microtubules with seams in helical reconstructions, but it can also be applied to any structures with arbitrary numbers of identical regions. The performance of local symmetry still needs to be evaluated in single-particle reconstructions. In addition, the local symmetry operators can only be generated using regional masks on a density map at present. As an alternative, it might be necessary to enable calculations of operators based on fitted PDB models of the identical regions.

7 REFERENCES

- Akabas, M. H. (2011). Ion channels: Manipulating the munchies in mice. *Nature Chemical Biology*, 7(11), 759-760.
- Alushin, G. M., Ramey, V. H., Pasqualato, S., Ball, D. A., Grigorieff, N., Musacchio, A., & Nogales, E. (2010). The Ndc80 kinetochore complex forms oligomeric arrays along microtubules. *Nature*, 467(7317), 805.
- Alvarez, F. J. D., He, S., Perilla, J. R., Jang, S., Prevelige, P. E., Engelman, A., ... & Zhang, P. (2017). Cryoem Structure of the Full Length Dynamin-Like MxB at 4.6 Å Resolution. *Biophysical Journal*, 112(3), 334a.
- Andrews, K. W. (1981). A table of maxima and minima of the Bessel functions $J_n(Z)$ for $n=0$ to $n=30$. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 37(5), 765-766.
- Atherton, J., Farabella, I., Yu, I. M., Rosenfeld, S. S., Houdusse, A., Topf, M., & Moores, C. A. (2014). Conserved mechanisms of microtubule-stimulated ADP release, ATP binding, and force generation in transport kinesins. *Elife*, 3, e03680.
- Bai, X. C., Fernandez, I. S., McMullan, G., & Scheres, S. H. (2013). Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife*, 2, e00461.
- Bai, X. C., McMullan, G., & Scheres, S. H. (2015). How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, 40(1), 49-57.
- Bai, X. C., Rajendra, E., Yang, G., Shi, Y., & Scheres, S. H. (2015). Sampling the conformational space of the catalytic subunit of human γ -secretase. *Elife*, 4, e11182.
- Bai, X. C., Yan, C., Yang, G., Lu, P., Ma, D., Sun, L., ... & Shi, Y. (2015). An atomic structure of human γ -secretase. *Nature*, 525(7568), 212.
- Behrmann, E., Tao, G., Stokes, D. L., Egelman, E. H., Raunser, S., & Penczek, P. A. (2012). Real-space processing of helical filaments in SPARX. *Journal of Structural Biology*, 177(2), 302-313.
- Beroukhi, R., & Unwin, N. (1997). Distortion correction of tubular crystals: improvements in the acetylcholine receptor structure. *Ultramicroscopy*, 70(1-2), 57-81.

- Bharat, T. A., Murshudov, G. N., Sachse, C., & Löwe, J. (2015). Structures of actin-like ParM filaments show architecture of plasmid-segregating spindles. *Nature*, 523(7558), 106.
- Bharat, T. A., Russo, C. J., Löwe, J., Passmore, L. A., & Scheres, S. H. (2015). Advances in single-particle electron cryomicroscopy structure determination applied to sub-tomogram averaging. *Structure*, 23(9), 1743-1753.
- Bichsel, H. (1988). Straggling in thin silicon detectors. *Reviews of Modern Physics*, 60(3), 663.
- Bluemke, D. A., Carragher, B., & Josephs, R. (1988). The reconstruction of helical particles with variable pitch. *Ultramicroscopy*, 26(3), 255-270.
- Brilot, A. F., Chen, J. Z., Cheng, A., Pan, J., Harrison, S. C., Potter, C. S., ... & Grigorieff, N. (2012). Beam-induced motion of vitrified specimen on holey carbon film. *Journal of Structural Biology*, 177(3), 630-637.
- Chen, S., McMullan, G., Faruqi, A. R., Murshudov, G. N., Short, J. M., Scheres, S. H., & Henderson, R. (2013). High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy*, 135, 24-35.
- Cheng, A., Henderson, R., Mastronarde, D., Ludtke, S. J., Schoenmakers, R. H., Short, J., ... & Winn, M. (2015). MRC2014: Extensions to the MRC format header for electron cryo-microscopy and tomography. *Journal of Structural Biology*, 192(2), 146-150.
- Cochran, W., Crick, F. H., & Vand, V. (1952). The structure of synthetic polypeptides. I. The transform of atoms on a helix. *Acta Crystallographica*, 5(5), 581-586.
- Crowther, R. A., Henderson, R., & Smith, J. M. (1996). MRC image processing programs. *Journal of Structural Biology*, 116(1), 9-16.
- Daum, B., Vonck, J., Bellack, A., Chaudhury, P., Reichelt, R., Albers, S. V., ... & Kühlbrandt, W. (2017). Structure and in situ organisation of the *Pyrococcus furiosus* archaellum machinery. *Elife*, 6.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 1-38.
- Deng, X., Fink, G., Bharat, T. A., He, S., Kureisaite-Ciziene, D., & Löwe, J. (2017). Four-stranded mini microtubules formed by *Prostheco bacter* BtubAB show dynamic instability. *Proceedings of the National Academy of Sciences*, 201705062.
- DeRosier, D. J., & Klug, A. (1968). Reconstruction of three dimensional structures from electron micrographs. *Nature*, 217(5124), 130-134.

- DeRosier, D. J., & Moore, P. B. (1970). Reconstruction of three-dimensional images from electron micrographs of structures with helical symmetry. *Journal of Molecular Biology*, 52(2), 355-369.
- Desfosses, A., Ciuffa, R., Gutsche, I., & Sachse, C. (2014). SPRING—an image processing package for single-particle based helical reconstruction from electron cryomicrographs. *Journal of Structural Biology*, 185(1), 15-26.
- Diaz, R., William, J. R., & Stokes, D. L. (2010). Chapter Five-Fourier–Bessel Reconstruction of Helical Assemblies. *Methods in Enzymology*, 482, 131-165.
- Dobbins, J. T., Ergun, D. L., Rutz, L., Hinshaw, D. A., Blume, H., & Clark, D. C. (1995). DQE (f) of four generations of computed radiography acquisition devices. *Medical Physics*, 22(10), 1581-1593.
- Dominguez, R., & Holmes, K. C. (2011). Actin structure and function.
- Dubochet, J., Adrian, M., Chang, J. J., Homo, J. C., Lepault, J., McDowell, A. W., & Schultz, P. (1988). Cryo-electron microscopy of vitrified specimens. *Quarterly Reviews of Biophysics*, 21(2), 129-228.
- Egelman, E. H. (2000). A robust algorithm for the reconstruction of helical filaments using single-particle methods. *Ultramicroscopy*, 85(4), 225-234.
- Egelman, E. H. (2007). The iterative helical real space reconstruction method: surmounting the problems posed by real polymers. *Journal of Structural Biology*, 157(1), 83-94.
- Egelman, E. H. (2010). Chapter Six-Reconstruction of Helical Filaments and Tubes. *Methods in Enzymology*, 482, 167-183.
- Egelman, E. H. (2014). Ambiguities in helical reconstruction. *Elife*, 3, e04969.
- Egelman, E. H., Wu, S. S., Amrein, M., Portner, A., & Murti, G. (1989). The Sendai virus nucleocapsid exists in at least four different helical states. *Journal of Virology*, 63(5), 2233-2243.
- Elmlund, H., Elmlund, D., & Bengio, S. (2013). PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. *Structure*, 21(8), 1299-1306.
- Erickson, H. P., & Klug, A. (1970). The Fourier transform of an electron micrograph: effects of defocussing and aberrations, and implications for the use of underfocus contrast enhancement. *Berichte der Bunsengesellschaft für physikalische Chemie*, 74(11), 1129-1137.
- Fernandez-Leiro, R., & Scheres, S. H. (2016). Unravelling biological macromolecules with cryo-electron microscopy. *Nature*, 537(7620), 339-346.
- Fernandez-Leiro, R., & Scheres, S. H. (2017). A pipeline approach to single-particle processing in RELION. *Acta Crystallographica Section D: Structural Biology*, 73(6).

- Finch, J. T. (1972). The hand of the helix of tobacco mosaic virus. *Journal of Molecular Biology*, 66(2), 291-294.
- Fitzpatrick, A. W., Falcon, B., He, S., Murzin, A. G., Murshudov, G., Garringer, H. J., ... & Scheres, S. H. (2017). Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature*, 547(7662), 185-190.
- Fourniol, F. J., Sindelar, C. V., Amigues, B., Clare, D. K., Thomas, G., Perderiset, M., ... & Moores, C. A. (2010). Template-free 13-protofilament microtubule-MAP assembly visualized at 8 Å resolution. *The Journal of Cell Biology*, 191(3), 463-470.
- Frank, J. (2006). Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. Oxford University Press.
- Frank, J. (Ed.). (2013). *Electron tomography: three-dimensional imaging with the transmission electron microscope*. Springer Science & Business Media.
- Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., & Leith, A. (1996). SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *Journal of Structural Biology*, 116(1), 190-199.
- Fromm, S. A., Bharat, T. A., Jakobi, A. J., Hagen, W. J., & Sachse, C. (2015). Seeing tobacco mosaic virus through direct electron detectors. *Journal of Structural Biology*, 189(2), 87-97.
- Ge, P., Scholl, D., Leiman, P. G., Yu, X., Miller, J. F., & Zhou, Z. H. (2015). Atomic structures of a bactericidal contractile nanotube in its pre-and postcontraction states. *Nature Structural & Molecular Biology*, 22(5), 377.
- Ghosal, D., Chang, Y. W., Jeong, K. C., Vogel, J. P., & Jensen, G. J. (2017). In situ structure of the Legionella Dot/Icm type IV secretion system by electron cryotomography. *EMBO reports*, 18(5), 726-732.
- Grange, M., Vasishtan, D., & Grünewald, K. (2017). Cellular electron cryo tomography and in situ sub-volume averaging reveal the context of microtubule-based processes. *Journal of Structural Biology*, 197(2), 181-190.
- Grant, T., & Grigorieff, N. (2015). Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *Elife*, 4, e06980.
- Hagen, W. J., Wan, W., & Briggs, J. A. (2017). Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. *Journal of Structural Biology*, 197(2), 191-198.
- Hall, S. R. (1991). The STAR file: a new format for electronic data transfer and archiving. *Journal of Chemical Information and Computer Sciences*, 31(2), 326-333.

- He, S., & Scheres, S. H. (2017). Helical reconstruction in RELION. *Journal of Structural Biology*.
- Hegner, M., Smith, S. B., & Bustamante, C. (1999). Polymerization and mechanical properties of single RecA–DNA filaments. *Proceedings of the National Academy of Sciences*, 96(18), 10109-10114.
- Henderson, R. (2013). Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences*, 110(45), 18037-18041.
- Henderson, R., Chen, S., Chen, J. Z., Grigorieff, N., Passmore, L. A., Ciccarelli, L., ... & Rosenthal, P. B. (2011). Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy. *Journal of Molecular Biology*, 413(5), 1028-1046.
- Henderson, R., Sali, A., Baker, M. L., Carragher, B., Devkota, B., Downing, K. H., ... & Jiang, W. (2012). Outcome of the first electron microscopy validation task force meeting. *Structure*, 20(2), 205-214.
- Heymann, J. B., Chagoyen, M., & Belnap, D. M. (2005). Common conventions for interchange and archiving of three-dimensional electron microscopy information in structural biology. *Journal of Structural Biology*, 151(2), 196-207.
- Heymann, J. B., & Belnap, D. M. (2007). Bsoft: image processing and molecular modeling for electron microscopy. *Journal of Structural Biology*, 157(1), 3-18.
- Hohn, M., Tang, G., Goodyear, G., Baldwin, P. R., Huang, Z., Penczek, P. A., ... & Ludtke, S. J. (2007). SPARX, a new environment for Cryo-EM image processing. *Journal of Structural Biology*, 157(1), 47-55.
- Holmes, K. C. (2009). Structural biology: actin in a twist. *Nature*, 457(7228), 389-390.
- Ichikawa, M., Liu, D., Kastiris, P. L., Basu, K., Hsu, T. C., Yang, S., & Bui, K. H. (2017). Subnanometre-resolution structure of the doublet microtubule reveals new classes of microtubule-associated proteins. *Nature Communications*, 8.
- Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J., & Patwardhan, A. (2016). EMPIAR: a public archive for raw electron microscopy image data. *Nature Methods*, 13(5), 387-389.
- Kimanius, D., Forsberg, B. O., Scheres, S. H., & Lindahl, E. (2016). Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *Elife*, 5, e18722.
- Kleywegt, G. J. (1996). Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallographica Section D: Biological Crystallography*, 52(4), 842-857.
- Klug, A., Crick, F. H. C., & Wyckoff, H. W. (1958). Diffraction by helical structures. *Acta Crystallographica*, 11(3), 199-213.

- Koning, R. I., Koster, A. J., & Sharp, T. H. (2018). Advances in cryo-electron tomography for biology and medicine. *Annals of Anatomy-Anatomischer Anzeiger*, 217, 82-96.
- Knuth, D. E. (1997). *The Art of Computer Programming. Volume 1: Fundamental Algorithms. Volume 2: Seminumerical Algorithms*. Bull. Amer. Math. Soc.
- Kremer, J. R., Mastronarde, D. N., & McIntosh, J. R. (1996). Computer visualization of three-dimensional image data using IMOD. *Journal of Structural Biology*, 116(1), 71-76.
- Kudryashev, M., Wang, R. Y. R., Brackmann, M., Scherer, S., Maier, T., Baker, D., ... & Basler, M. (2015). Structure of the type VI secretion system contractile sheath. *Cell*, 160(5), 952-962.
- Kunz, M., & Frangakis, A. S. (2017). Three-dimensional CTF correction improves the resolution of electron tomograms. *Journal of Structural Biology*, 197(2), 114-122.
- Li, X., Mooney, P., Zheng, S., Booth, C. R., Braunfeld, M. B., Gubbens, S., ... & Cheng, Y. (2013). Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods*, 10(6), 584-590.
- Linck, R. W., & Amos, L. A. (1974). The hands of helical lattices in flagellar doublet microtubules. *Journal of Cell Science*, 14(3), 551-559.
- Löwe, J., He, S., Scheres, S. H., & Savva, C. G. (2016). X-ray and cryo-EM structures of monomeric and filamentous actin-like protein MamK reveal changes associated with polymerization. *Proceedings of the National Academy of Sciences*, 201612034.
- Lu, A., Magupalli, V. G., Ruan, J., Yin, Q., Atianand, M. K., Vos, M. R., ... & Egelman, E. H. (2014). Unified polymerization mechanism for the assembly of ASC-dependent inflammasomes. *Cell*, 156(6), 1193-1206.
- Lyumkis, D., Brilot, A. F., Theobald, D. L., & Grigorieff, N. (2013). Likelihood-based classification of cryo-EM images using FREALIGN. *Journal of Structural Biology*, 183(3), 377-388.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press. pp. 549-551.
- Mastronarde, D. N. (2003). SerialEM: a program for automated tilt series acquisition on Tecnai microscopes using prediction of specimen position. *Microscopy and Microanalysis*, 9(S02), 1182-1183.
- May, A., & Huehns, E. R. (1976). The mechanism and prevention of sickling. *British Medical Bulletin*, 32(3), 223-233.
- McMullan, G., Clark, A. T., Turchetta, R., & Faruqi, A. R. (2009). Enhanced imaging in low dose electron microscopy using electron counting. *Ultramicroscopy*, 109(12), 1411-1416.

- Metlagel, Z., Kikkawa, Y. S., & Kikkawa, M. (2007). Ruby-Helix: an implementation of helical image processing based on object-oriented scripting language. *Journal of Structural Biology*, 157(1), 95-105.
- Midgley, P. A., & Weyland, M. (2003). 3D electron microscopy in the physical sciences: the development of Z-contrast and EFTEM tomography. *Ultramicroscopy*, 96(3), 413-431.
- Nakane, T., Kimanius, D., Lindahl, E., & Scheres, S. H. (2018). Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife*, 7, e36861.
- Nigg, E. A. (2001). Mitotic kinases as regulators of cell division and its checkpoints. *Nature Reviews. Molecular Cell Biology*, 2(1), 21.
- Nogales, E. (2000). Structural insights into microtubule function. *Annual Review of Biochemistry*, 69(1), 277-302.
- Nogales, E., & Scheres, S. H. (2015). Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Molecular Cell*, 58(4), 677-689.
- Nogales, E., & Zhang, R. (2016). Visualizing microtubule structural transitions and interactions with associated proteins. *Current Opinion in Structural Biology*, 37, 90-96.
- Owen, C. H., Morgan, D. G., & DeRosier, D. J. (1996). Image analysis of helical objects: the Brandeis Helical Package. *Journal of Structural Biology*, 116(1), 167-175.
- Patwardhan, A. (2017). Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallographica Section D: Structural Biology*, 73(6).
- Penczek, P. A., & Asturias, F. J. (2014, October). Ab initio cryo-EM structure determination as a validation problem. In *Image Processing (ICIP), 2014 IEEE International Conference on* (pp. 2090-2094). IEEE.
- Penczek, P. A., Grassucci, R. A., & Frank, J. (1994). The ribosome at improved resolution: new techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. *Ultramicroscopy*, 53(3), 251-270.
- Penczek, P. A., Zhu, J., Schröder, R., & Frank, J. (1997). Three dimensional reconstruction with contrast transfer compensation from defocus series. *Scanning Microsc*, 11, 147-154.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13), 1605-1612.
- Punjani, A., Rubinstein, J. L., Fleet, D. J., & Brubaker, M. A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3), 290-296.

- Radermacher, M., Wagenknecht, T., Verschoor, A., & Frank, J. (1987). Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of *Escherichia coli*. *Journal of Microscopy*, 146(2), 113-136.
- Rohou, A., & Grigorieff, N. (2014). FREALIX: Model-based refinement of helical filament structures from electron micrographs. *Journal of Structural Biology*, 186(2), 234-244.
- Rohou, A., & Grigorieff, N. (2015). CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology*, 192(2), 216-221.
- Rosenthal, P. B., & Henderson, R. (2003). Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology*, 333(4), 721-745.
- Russo, C. J., & Henderson, R. (2018). Ewald sphere correction using a single side-band image processing algorithm. *Ultramicroscopy*, 187, 26-33.
- Sachse, C., Chen, J. Z., Coureux, P. D., Stroupe, M. E., Fändrich, M., & Grigorieff, N. (2007). High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus. *Journal of Molecular Biology*, 371(3), 812-835.
- (Submitted) Salih, O., He, S., Planamente, S., Stach, L., MacDonald, J. T., Manoli, E., Scheres, S. H., Filloux, A., & Freemont, P. S. (2017). Atomic structure of Type VI contractile sheath from *Pseudomonas textbfæuruginosa*. *EMBO Reports*.
- Scheres, S. H. (2012). A Bayesian view on cryo-EM structure determination. *Journal of Molecular Biology*, 415(2), 406-418.
- Scheres, S. H. (2012). RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3), 519-530.
- Scheres, S. H. (2014). Beam-induced motion correction for sub-megadalton cryo-EM particles. *Elife*, 3, e03665.
- Scheres, S. H. (2015). Semi-automated selection of cryo-EM particles in RELION-1.3. *Journal of Structural Biology*, 189(2), 114-122.
- Scheres, S. H. (2016). Chapter Six-Processing of Structurally Heterogeneous Cryo-EM Data in RELION. *Methods in Enzymology*, 579, 125-157.
- Scheres, S. H., & Chen, S. (2012). Prevention of overfitting in cryo-EM structure determination. *Nature Methods*, 9(9), 853-854.

- Schur, F. K., Obr, M., Hagen, W. J., Wan, W., Jakobi, A. J., Kirkpatrick, J. M., ... & Briggs, J. A. (2016). An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation. *Science*, 353(6298), 506-508.
- Shang, Z., Zhou, K., Xu, C., Csencsits, R., Cochran, J. C., & Sindelar, C. V. (2014). High-resolution structures of kinesin on microtubules provide a basis for nucleotide-gated force-generation. *Elife*, 3, e04686.
- Sharp, T. H., Faas, F. G., Koster, A. J., & Gros, P. (2017). Imaging complement by phase-plate cryo-electron tomography from initiation to pore formation. *Journal of Structural Biology*, 197(2), 155-162.
- Short, J. M., Liu, Y., Chen, S., Soni, N., Madhusudhan, M. S., Shivji, M. K., & Venkitaraman, A. R. (2016). High-resolution structure of the presynaptic RAD51 filament on single-stranded DNA by electron cryo-microscopy. *Nucleic Acids Research*, 44(19), 9017-9030.
- Silflow, C. D., & Lefebvre, P. A. (2001). Assembly and motility of eukaryotic cilia and flagella. Lessons from *Chlamydomonas reinhardtii*. *Plant Physiology*, 127(4), 1500-1507.
- Sindelar, C. V., & Downing, K. H. (2007). The beginning of kinesin's force-generating cycle visualized at 9-Å resolution. *Journal of Cell Biology*, 177(3), 377-385.
- Smith, J. M. (1999). Ximdisp—a visualization tool to aid structure determination from electron microscope images. *Journal of Structural Biology*, 125(2-3), 223-228.
- Sorzano, C. O. S., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J. R., Scheres, S. H., Carazo, J. M., & Pascual-Montano, A. (2004). XMIPP: a new generation of an open-source image processing package for electron microscopy. *Journal of Structural Biology*, 148(2), 194-204.
- Sosa, H., Hoenger, A., & Milligan, R. A. (1997). Three different approaches for calculating the three-dimensional structure of microtubules decorated with kinesin motor domains. *Journal of Structural Biology*, 118(2), 149-158.
- Stephens, R. E., & Edds, K. T. (1976). Microtubules: structure, chemistry, and function. *Physiological Reviews*, 56(4), 709-777.
- Steven, A. C., & Aebi, U. (2003). The next ice age: cryo-electron tomography of intact cells. *Trends in Cell Biology*, 13(3), 107-110.
- Stewart, M. (1988). Computer image processing of electron micrographs of biological structures with helical symmetry. *Microscopy Research and Technique*, 9(4), 325-358.
- Sunde, M., Serpell, L. C., Bartlam, M., Fraser, P. E., Pepys, M. B., & Blake, C. C. (1997). Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *Journal of Molecular Biology*, 273(3), 729-739.

- Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., & Ludtke, S. J. (2007). EMAN2: an extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1), 38-46.
- Turoňová, B., Schur, F. K., Wan, W., & Briggs, J. A. (2017). Efficient 3D-CTF correction for cryo-electron tomography using NovaCTF improves subtomogram averaging resolution to 3.4 Å. *Journal of Structural Biology*, 199(3), 187-195.
- Unwin, N. (2017). Segregation of lipids near acetylcholine-receptor channels imaged by cryo-EM. *IUCrJ*, 4(4).
- Vale, R. D. (1987). Intracellular transport using microtubule-based motors. *Annual Review of Cell Biology*, 3(1), 347-378.
- Van Heel, M. (1987). Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy*, 21(2), 111-123.
- van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R., & Schatz, M. (1996). A new generation of the IMAGIC image processing system. *Journal of Structural Biology*, 116(1), 17-24.
- von der Ecken, J., Müller, M., Lehman, W., Manstein, D. J., Penczek, P. A., & Raunser, S. (2015). Structure of the F-actin-tropomyosin complex. *Nature*, 519(7541), 114.
- Vulović, M., Voortman, L. M., van Vliet, L. J., & Rieger, B. (2014). When to use the projection assumption and the weak-phase object approximation in phase contrast cryo-EM. *Ultramicroscopy*, 136, 61-66.
- Whittaker, M., Carragher, B. O., & Milligan, R. A. (1995). PHOELIX: a package for semi-automated helical reconstruction. *Ultramicroscopy*, 58(3-4), 245-259.
- Williams, D. B., & Carter, C. B. (1996). The transmission electron microscope. In *Transmission electron microscopy* (pp. 3-17). Springer, Boston, MA.
- Williams, D. B., & Carter, C. B. (2009). *Transmission electron microscopy: a textbook for materials science*.
- Xu, H., He, X., Zheng, H., Huang, L.J., Hou, F., Yu, Z., de la Cruz, M.J., Borkowski, B., Zhang, X., Chen, Z.J., Jiang, Q.-X., 2014. Structural basis for the prion-like MAVS filaments in antiviral innate immunity. *Elife*, 3, e01489.
- Xu, H., He, X., Zheng, H., Huang, L. J., Hou, F., Yu, Z., ... & Jiang, Q. X. (2015). Correction: Structural basis for the prion-like MAVS filaments in antiviral innate immunity. *Elife*, 4, e07546.
- Zernike, F. (1955). How I discovered phase contrast. *Science*, 121(3141), 345-349.

- Zhang, K. (2016). Gctf: Real-time CTF determination and correction. *Journal of Structural Biology*, 193(1), 1-12.
- Zhang, R., Alushin, G. M., Brown, A., & Nogales, E. (2015). Mechanistic origin of microtubule dynamic instability and its modulation by EB proteins. *Cell*, 162(4), 849-859.
- Zheng, S. Q., Palovcak, E., Armache, J. P., Verba, K. A., Cheng, Y., & Agard, D. A. (2017). MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature Methods*, 14(4).
- Zheng, W., Wang, F., Taylor, N. M., Guerrero-Ferreira, R. C., Leiman, P. G., & Egelman, E. H. (2017). Refined Cryo-EM Structure of the T4 Tail Tube: Exploring the Lowest Dose Limit. *Structure*, 25(9), 1436-1441.

8 APPENDIX

Full text of:

He, S., & Scheres, S. H. (2017). Helical reconstruction in RELION. *Journal of Structural Biology*.



Helical reconstruction in RELION



Shaoda He, Sjors H.W. Scheres*

MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, CB2 0QH Cambridge, UK

ARTICLE INFO

Article history:

Received 19 December 2016

Received in revised form 24 January 2017

Accepted 9 February 2017

Available online 11 February 2017

Keywords:

Helical symmetry

Single-particle analysis

Cryo-electron microscopy

Image processing

Regularised likelihood optimisation

ABSTRACT

We describe a new implementation for the reconstruction of helical assemblies in the empirical Bayesian framework of RELION. Our approach calculates optimal linear filters for the 3D reconstruction by embedding helical symmetry operators in Fourier-space, and deals with deviations from perfect helical symmetry through Gaussian-shaped priors on the orientations of individual segments. By incorporating our approach into the standard pipeline for single-particle analysis in RELION, our implementation aims to be easily accessible for non-experienced users. Although our implementation does not solve the problem that grossly incorrect structures can be obtained when the wrong helical symmetry is imposed, we show for four different test cases that it is capable of reconstructing structures to near-atomic resolution.

© 2017 MRC Laboratory of Molecular Biology. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The first biological structure to be determined by three-dimensional electron microscopy (3D-EM), the extended tail of the T4 bacteriophage, had helical symmetry (De Rosier and Klug, 1968; DeRosier and Moore, 1970). An important advantage for structure determination of helical objects over asymmetrical particles lies in the observation that a single projection image of a helical specimen will typically contain all necessary information to perform a 3D reconstruction. Moreover, for helical reconstructions the number of parameters to be determined is, in principle, strongly reduced compared to single-particle analysis. That is because in single-particle analysis one needs to determine the relative orientations for every individual particle projection image, while for helical structures many copies of the repeating, asymmetrical unit have fixed relative orientations. Therefore, once the parameters describing helical symmetry and the orientation of the image of an entire filament with respect to that symmetry have been determined, one can reduce the experimental noise efficiently by averaging over a large number of asymmetrical units.

The determination of helical symmetry and the subsequent calculation of a 3D reconstruction were initially performed in Fourier

space. The mathematical description of the Fourier transform of an object with helical symmetry was first proposed by Crick, Cochran and Vand (Cochran et al., 1952), and Klug generalised the theory afterwards (Klug et al., 1958). The initial step is to inspect 2D Fourier diffraction patterns of, sometimes averaged, and preferably long and straight, helical filaments. The helical lattice can be considered as a 2D surface lattice that is curled into a cylinder (also see Fig. 1). The curling effect causes a convolution of the Fourier transform of the 2D surface lattice with a cylindrical harmonic called the Bessel function. In the 3D Fourier transform of a helical object, discrete horizontal lines, called layer lines, arise from the periodicity along the helical axis. Through the central slice theorem, the radially oscillating ring-like amplitudes on each plane perpendicular to the z-axis in 3D Fourier space give rise to symmetrical maxima across the meridian on each layer line in the Fourier transforms of 2D projections. The position of the maxima along these lines, combined with the real-space width of the helical object, can be used to infer the 2D lattice parameters in a process called Fourier-Bessel indexing. After the Fourier transform of a helix has been indexed, a 3D reconstruction can be obtained through Fourier inversion (DeRosier and Moore, 1970). For more detailed reviews on Fourier-Bessel analysis, the reader is referred to (Diaz et al., 2010; Stewart, 1988).

Computer programs for Fourier-Bessel analysis were first developed in the MRC image processing package (Crowther et al., 1996), and have since been adapted in many other packages, such as the Brandeis Helical Package (Owen et al., 1996), Phoelix (Whittaker et al., 1995) and for objects with a break in helical symmetry

Abbreviations: CARD, caspase activation and recruitment domain; EM, electron microscopy; FSC, Fourier shell correlation; IHRSR, iterative helical real-space reconstruction; MAVS, mitochondrial antiviral signaling protein; TMV, tobacco mosaic virus.

* Corresponding author.

E-mail address: scheres@mrc-lmb.cam.ac.uk (S.H.W. Scheres).

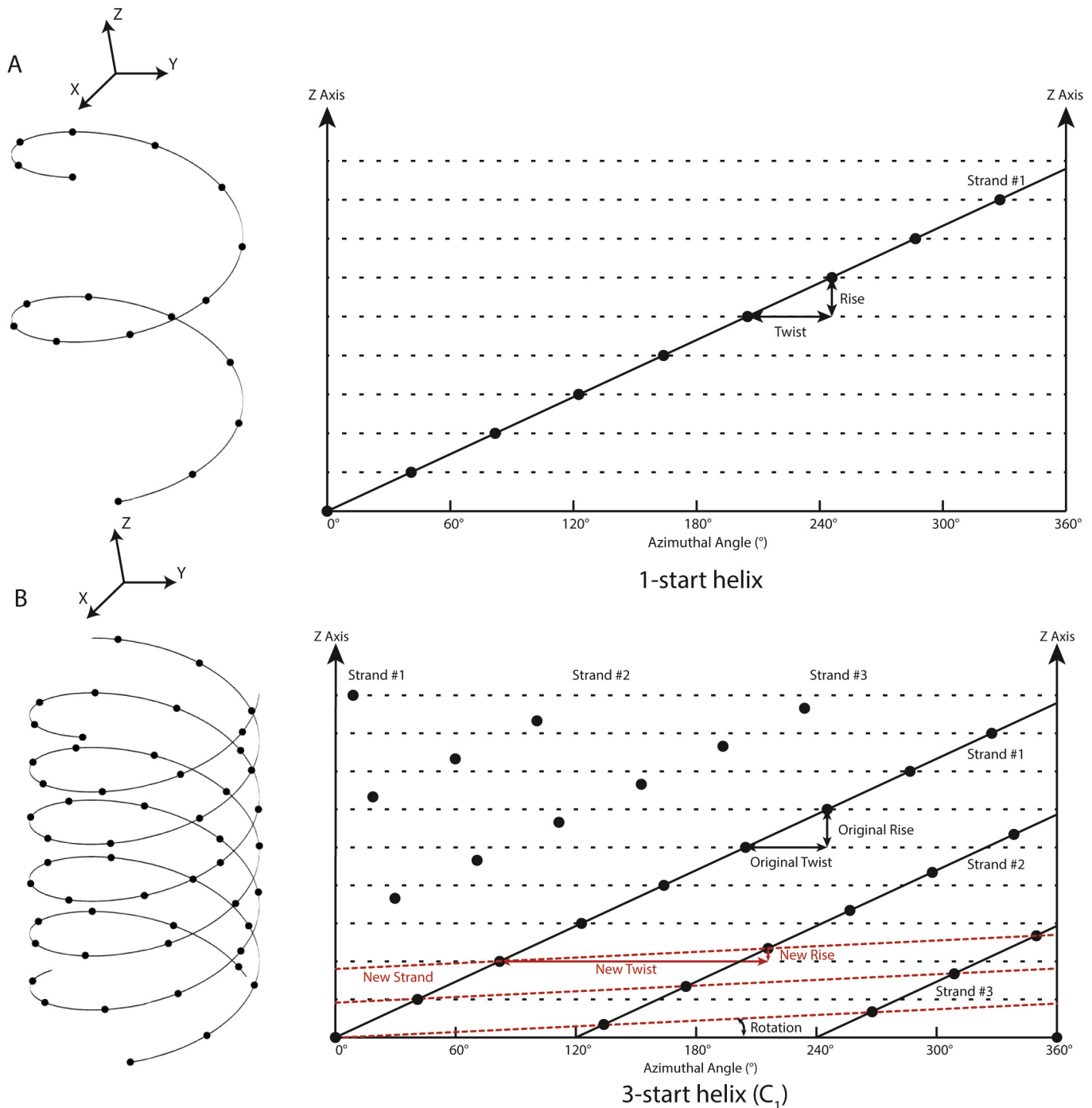


Fig. 1. Definition of helical symmetry for n -start helices. A. Example of a 1-start helix with no point group symmetry (C_1). The 3D helical structure on the left can be 'unfolded' into a 2D plot of the z -height against the azimuthal angle ϕ on the right. The dots represent the centres of the subunits in the 2D surface lattice. B) Example of a 3-start helix (with C_1 point group symmetry), where instead of using the original twist and rise of the individual three strands (in black), RELION expresses the helical symmetry using a third of the original rise (in red).

known as a seam in Ruby-helix (Metlagel et al., 2007). Fourier-Bessel analysis of helices is often considered laborious and difficult, as small inaccuracies in the indexing may lead to incorrect reconstructions. Indexing is particularly prone to errors when the filaments are not lying flat, the lattice is somehow distorted by bending or flattening of the filaments, or by short-range disorder in the helical symmetry. Therefore, the Fourier-Bessel approach only performs well when the specimen adopts a close to perfect helical structure. In practice, many helical assemblies are far from perfect due to molecular flexibility and distortions induced by the sample preparation process. To reduce the effects of long-range

distortions on the Fourier-Bessel reconstruction of AChR filaments, Beroukhi and Unwin introduced the idea of dividing helical filaments into short segments, which are independently aligned against a reference structure (Beroukhi and Unwin, 1997).

Analogous to single-particle analysis, one can also align separate helical segments in real-space against projections of a 3D reference model. Early applications of real-space helical reconstruction were performed on haemoglobin fibrils (Bluemke et al., 1988) and on microtubules (Sosa et al., 1997). The introduction of the iterative helical real-space reconstruction (IHRSR) algorithm by Egelman opened up new horizons for application of this

approach to a wide range of specimens (Egelman, 2000). The IHRSR algorithm is based on iterative projection matching, where computer-generated projections of a 3D reference in many different orientations are compared with small segments that are cut out from the micrographs in overlapping square boxes along the helical filaments. After the optimal orientations of all segments have been determined in this way, the aligned segments are back-projected into an (asymmetric) 3D reconstruction. The helical twist and rise are then determined by least-square fit within a user-defined search range, and the best symmetry is imposed to generate a symmetric 3D reference for the next iteration. For favourable cases, this method was shown to converge to the correct solution when starting from a solid cylinder (Egelman, 2007), thereby reducing human intervention and enabling helical reconstruction without expertise in Fourier-Bessel analysis. The original IHRSR approach was implemented in the SPIDER package (Frank et al., 1996). More recent variations on the single-particle like approach to helical reconstruction were made in FREALIGN (Alushin et al., 2010; Sachse et al., 2007), SPARX (Behrmann et al., 2012), FREALIX (Rohou and Grigorieff, 2014) and SPRING (Desfosses et al., 2014).

In this paper, we describe an implementation of single-particle-like analysis for helical assemblies in the RELION program (Scheres, 2012a). RELION is based on an empirical Bayesian approach in Fourier-space, where a marginalised likelihood function is augmented with a Gaussian prior on the reconstruction (Scheres, 2012b). Optimisation of the corresponding regularised likelihood target is performed using the expectation maximisation algorithm (Dempster et al., 1977). Because parameters of a statistical model for both the signal and the noise are estimated from the data themselves, optimal filters for both the reconstruction and the alignment are determined automatically. Thereby the need for user expertise in devising *ad hoc* filters is strongly reduced. Although helical reconstructions using RELION have been described previously (Clemens et al., 2015; Short et al., 2016; Xu et al., 2015), these approaches relied on external software to impose helical symmetry, so that RELION could not incorporate the additional averaging into its calculation of optimal filters. Moreover, previous versions of RELION were unaware of the signal extending to the edges of the boxed images, and did not exploit statistical priors on the orientations of individual segments to describe deviations from perfect helical symmetry.

Below, in the Methods section, we first describe the modifications to RELION that were made to implement helical reconstruction. Then, in the Results section, we describe how this implementation behaves for four different helical test specimens. We start the Results section with a description of the structure determination procedures for these four test cases, and we describe more detailed analyses of various features of the RELION implementation in the second half of the Results section. Finally, in the Discussion section we analyse the benefits and the remaining pitfalls of our implementation. The approach described in this paper has been implemented in RELION version 2.0, and was accelerated for execution on graphics cards along with the code for standard single-particle analysis of asymmetric particles (Kimanius et al., 2016). RELION is distributed as open-source and can be downloaded for free by both academic and non-academic users from <http://www2.mrc-lmb.cam.ac.uk/relion>.

2. Methods

2.1. Definition of helical symmetry

Helical symmetry in RELION is described by two parameters (Fig. 1A): helical *twist* (between -180 and +180 degrees, with

positive values corresponding to right-handed helices along the z-axis) and helical *rise* (>0 Å, along the z-axis). Therefore, for any point in real space, the voxel values of a perfectly helical object have the following relationship under cylindrical coordinates with radial distance r , angular coordinate ϕ , and height z (also see Fig. 1A):

$$\rho(r, \phi, z) = \rho(r, \phi + h * \text{twist}, z + h * \text{rise}) \text{ for any integer value } h.$$

Apart from helical symmetry, additional point group symmetry may also be present. That is, the sample may have an additional n -fold rotational symmetry axis along the helical axis (C_n -symmetry with $n \geq 2$), and/or it may have a 2-fold rotational axis perpendicular to the symmetry axis (D_n -symmetry with $n \geq 1$, where $D1$ stands for a single 2-fold axis perpendicular to the helical axis but no rotational symmetry along the helical axis). Helices with D_n symmetry lack polarity, whereas helices with C_n symmetry typically consist of n separate strands of molecules that are packed against each other and that all start at the same level along the z-axis.

In some helical structures the subunit shape and packing may make a particular n -start helical family especially prominent, even though the structure does not have C_n point group symmetry. In such a case the n separate helical strands do not all start at the same level along the z-axis. To maintain an intact helical lattice, the difference in translation between the n strands have to be multiples of the rise/n . Thereby, it becomes possible to describe these so-called n -start helical families by a 1-start helix with the following twist and rise (Fig. 1B):

$$\text{rise} = \text{rise}_{n\text{-start}}/n$$

$$\text{twist} = (\text{twist}_{n\text{-start}} + 360^\circ)/n \text{ for } \text{twist}_{n\text{-start}} > 0 \text{ or}$$

$$\text{twist} = (\text{twist}_{n\text{-start}} - 360^\circ)/n \text{ for } \text{twist}_{n\text{-start}} < 0$$

As is done in alternative programs, RELION uses the 1-start description of such structures. Although mathematically correct, the 1-start helical lattice may not have an intuitive biological interpretation. The handedness of the 1-start lattice is not necessarily the same as for the n -start lattice, and neighbouring molecules do not necessarily contact each other on the 1-start helical lattice.

2.2. Fourier-space symmetrisation and the inter-box distance

In order to estimate the signal-to-noise ratio in the experimental images, which is an intricate part of the empirical Bayesian approach, RELION needs to know over how many asymmetric units a reconstruction is calculated. For asymmetrical particles, each experimental projection contributes a single asymmetric unit to the reconstruction. For helical segments, the amount of overlap between adjacent boxes defines how many new asymmetric units are contained in each new box. Therefore, upon the extraction of helical segments from the micrographs, RELION will ask the user for an inter-box distance that is an integer H times the helical rise (also see below), so that each new box will contain H new asymmetric units. At early stages of processing, when an exact rise may not yet be known, a best guess may be used instead.

For values of H larger than one, during the alignment of each segment in the expectation step one could in principle set the 2D Fourier slice H times back into the 3D Fourier transform, each time applying a rotation of $h * \text{twist}$ and phase shifts that would correspond to a translation of $h * \text{rise}$ along Z , with $h = 0, \dots, H - 1$ (Fig. 2). Such a Fourier-space symmetrisation scheme has been implemented for example in FREALIGN (Alushin et al., 2010; Sachse et al., 2007). RELION implements the computationally cheaper, mathematically equivalent option to apply these rotations and phase shifts to the accumulated sum of all slices in 3D Fourier

space at the end of every iteration, immediately prior to performing the reconstruction. A similar scheme was also implemented in FREALIX (Rohou and Grigorieff, 2014).

Since the information content of a given set of micrographs is invariant of the inter-box distance, different values of the latter should in principle lead to identical signal-to-noise estimates and 3D reconstructions. Larger inter-box distances, or larger values of H , will result in fewer segments and are thus computationally cheaper. However, in RELION the useful inter-box distance is limited in practice by artefacts caused by our symmetrisation operations in Fourier-space. For values of H larger than 1, translations through phase shifts in the 3D Fourier transform will cause densities at the top of the helix in real-space to “wrap around” to the bottom and *vice versa*. Because the densities at the top and the bottom will typically not coincide, this will lead to additional blurring at the helix ends. Consequently, only the central z-slices in the real-space reconstruction will be intact after the symmetrisation in Fourier space, which places a strict maximum on the inter-box distance of 50% of the box size.

2.3. Real-space symmetrisation and local optimisation of twist and rise

Apart from the artefacts caused by our symmetrisation in Fourier space, other factors will further blur the reconstructed density: helical segments that are not perfectly straight, have varying twist, rise, shrinkage or expansion along the helical axis and inaccuracies in the alignment. Bent helices can be described better by using smaller inter-box distances, which will lead to more segments, and thus more orientational parameters to describe the deviations from a straight helix. But because bent helices can only be approximated with a straight helix in a sufficiently small central section, the top and bottom parts of the reconstructed helix will be more blurred than the central part. The same happens with alignment errors, where the density that is further away from the centre of rotation will be more blurred. As a result, the central part of the reconstructed helix will be better defined than the top and bottom ends. Therefore, it is beneficial to calculate a symmetric structure using only the central z-sections in real-space, and to extend the symmetric structure over the length of the entire box.

As was originally introduced for the IHRSR approach (Egelman, 2000), this real-space symmetrisation can also be accompanied with a local optimisation of the helical twist and rise. This may be necessary, as estimates from traditional Fourier-Bessel indexing

or other methods are often not accurate enough for high-resolution reconstruction. Therefore, RELION optionally performs a two-dimensional grid search of the helical twist and rise to find the smallest variance between the averaged sections of the map. Initial search range and step size parameters are provided by the user, and the 2D grid search is iterated with increasingly fine steps around the best twist and rise from the previous iteration. Because the refinement of twist and rise is a local optimisation process, and the energy landscape may contain many local minima, the optimisation of helical parameters should not be considered as a substitute for Fourier-Bessel indexing, asymmetric refinements from featureless cylinders, or experimental approaches like tomography.

2.4. 2D and 3D masks

Masks are used during refinement on both the 2D experimental images and on the 3D reconstruction. Masking the 2D experimental images helps to reduce the influence of noise or the neighbouring particles on the alignment in the expectation step, and the same mask is also used to determine the mean and standard deviation in the background area for the normalisation of boxed segments. Masking the 3D reconstruction reduces the noise in the solvent region, and thus leads to a better reference for the next iteration. For the analysis of helical structures, a circular 2D mask from standard single-particle analysis is multiplied with an additional rectangular mask across the entire particle image with a width according to the *outer_tube_diameter* parameter (Fig. 3A). This mask is no longer rotationally invariant, but is rotated according to the most likely in-plane rotation angle for each experimental image. Prior to any refinement, this in-plane rotation is estimated from the manual or template-based particle picking procedures (also see below).

Likewise, a spherical solvent mask on the 3D reconstruction from standard single-particle analysis is multiplied with a hollow cylinder mask with the same *outer_tube_diameter* parameter describing the width of the 2D mask. An optional *inner_tube_diameter* parameter also allows to mask away densities inside hollow helices (Fig. 3B).

One could argue that because helical segments extend all the way across the image, one would not need the circular or spherical components of these masks. However, as all reference projections are taken as 2D slices from the 3D Fourier transform of the reference map, real-space artefacts may again arise from non-empty

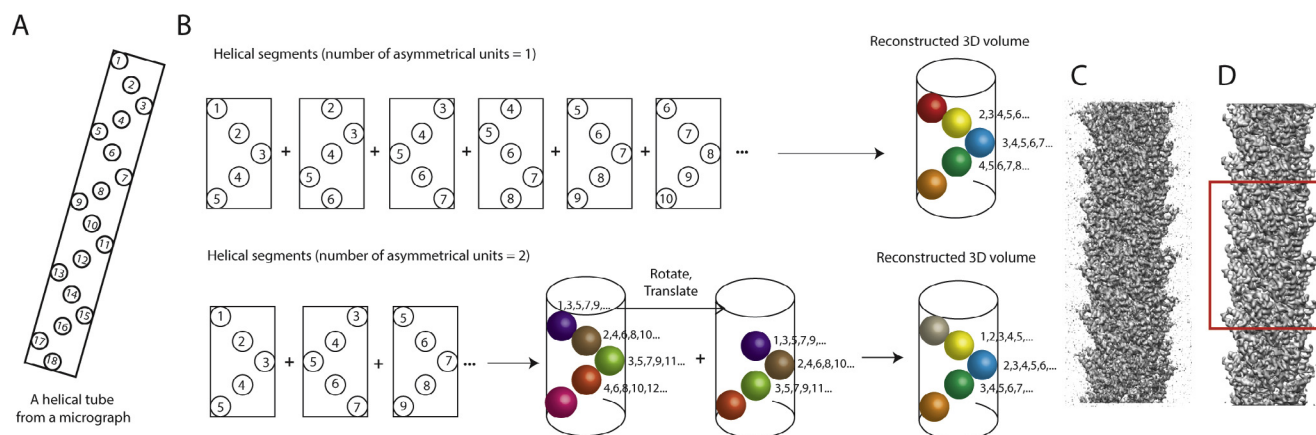


Fig. 2. A. A schematic representation of a helical filament with 18 asymmetrical units. B. Extracted segments with 1 (top) or 2 (bottom) new asymmetrical units per box lead to similar reconstructed volumes after symmetrisation in Fourier-space. Fourier-space symmetrisation is not necessary if the inter-box distance is one helical rise. Otherwise, the reconstructed volume from the helical segments is rotated and translated according to the helical symmetry, and summed up in Fourier space. C. The reconstruction of MAVS from manually picked filaments and an inter-box distance corresponding to 6 new asymmetrical units per segment. Without Fourier-space symmetrisation, one does not average over all asymmetrical units. D. The MAVS reconstruction after Fourier-space symmetrisation. The central section along the helical axis (within the red box) is averaged over all available asymmetrical units. At the top and the bottom of the helix, blurring is caused by Fourier artefacts (also see main text).

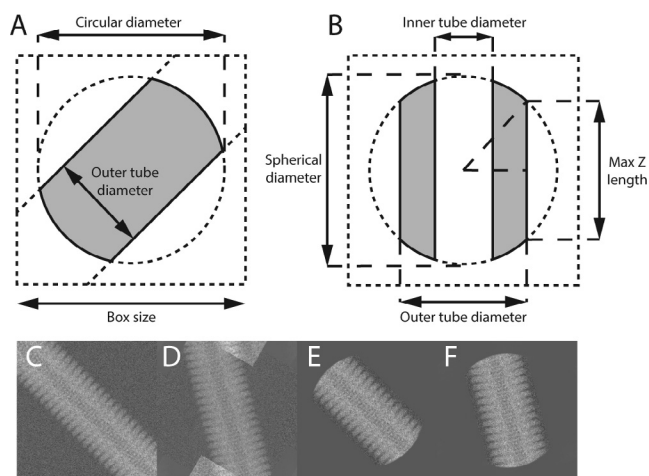


Fig. 3. Masks for 2D images (A) and 3D maps (B). Areas of a filament inside the mask are shown in grey, areas outside the mask in white. In panel A, the helical axis runs at a 45° angle with the horizontal direction, which represents an arbitrary in-plane rotation angle. The helical axis (i.e. the z-axis) runs parallel to the vertical direction in B. Panels C–F illustrate the need for the circular (or spherical) component of these masks. Rotation of an unmasked segment (C) leads to artefacts (top and bottom in D), due to wrapping effects caused by rotations in the Fourier domain. The application of a circular mask to the original image (E) allows rotation without artefacts (F).

corners of the reference upon rotation. To make comparisons between the reference projections and the experimental images consistent, the corners are removed from both the 3D map and the 2D images (Fig. 3C–F).

2.5. Incorporating prior information about the orientations

The statistical framework provides an elegant way to express prior information about the relative rotations and translations of each experimental image. For example, in both single-particle analysis and the processing of helical filaments, (manual or automated) particle picking will yield residual translational offsets around the approximate centre of the images. RELION expresses this information as a Gaussian prior probability function, and estimates its standard deviation from the data during the iterative refinement. For helical processing, this distribution is changed into an anisotropic function (Fig. 4A). Along the helical axis, the prior probability is a tophat distribution with non-zero values from $-rise/2$ to $+rise/2$ around the centre of the image. This distribution prevents segments from translating into the next asymmetric unit, which would lead to averaging multiple times over the same signal in overlapping segments. For helices with a very small rise, e.g. TMV in the examples below, the prior becomes very narrow. To still allow movements of individual segments along the helical direction, we impose a minimum of five sampling points for the translational searches along the Z-axis. In the orthogonal direction, a Gaussian distribution is used, for which the standard deviation is re-estimated at every iteration of the expectation maximisation algorithm in the same way as in the single-particle approach (Scheres, 2012a).

Unlike in single-particle analysis, there is also ample prior knowledge about the relative rotations of helical segments. Firstly, because the long filaments tend to lie approximately horizontal inside the ice layer, we use a Gaussian distribution that is centred at 90° on the tilt angle in RELION (Fig. 4B). Distributions with a wider standard deviation can be used to account for increasing degrees of rocking, but at present this parameter is fixed at a value provided by the user. Alternatively, the user can choose to allow the centre of the tilt prior for each particle to move to the most

likely angle from the previous iteration. In principle, if the helical twist and rise are known, there are also strong relationships for neighbouring segments between the rotations along the helical axis, i.e. the ‘rot’ angle in RELION. However, as the rot angle is strongly coupled with the translation along the helical axis, expressing this information into a prior is more complicated, and this is not done in the current implementation.

For the in-plane rotations, the ‘psi’ angle in RELION, we introduce a bimodal prior for helical segments that expresses the prior knowledge that the particle picking provides about the direction of the helix, but with an ambiguity in its polarity (Fig. 4C). This prior consists of two opposing Gaussian distributions that are clipped at 3 standard deviations. The ambiguity in polarity is absent for helices with D_n point group symmetry. For helices with C_n symmetry, this polarity is typically resolved during the refinement process. One can also express the prior knowledge that all segments from the same helix should have the same polarity. In our implementation, one starts the refinement with symmetrical priors on the in-plane rotations, where the two opposing Gaussian distributions have equal weights. As the refinement progresses, and particles accumulate a stronger probability for one of the two polarities, the difference in the weights on the two opposing Gaussians becomes correspondingly larger. In 3D auto-refinement, and sometimes also in 2D or 3D classification, initial iterations of exhaustive orientational searches are followed by iterations where only local orientational searches are performed. At this point, the local searches are only performed around the in-plane rotation with the largest weight.

For both the in-plane rotation and the tilt-angle, we also implemented an option to centre the prior on the optimal orientations from the previous iteration for all neighbouring segments within a user-defined distance. A Gaussian function (with a standard deviation of one third of a user-defined distance along the filament) is used to assign lower weights to segments that are further away. This local averaging option is implemented in the `relion_refine` program as `--helical_sigma_distance X`, where X is the user-defined distance. Since local averaging is performed within three standard deviations of the Gaussian function, X effectively defines the distance from the middle to the edge of the local averaging window. Local averaging reduces discontinuities in the orientational assignments along a single filament.

2.6. Particle picking and extraction

Apart from modifications to the refinement program that underlies the 2D/3D classification and 3D auto-refinement options of RELION, we also modified the programs for manual and template-based particle picking and particle extraction. The program for manual picking allows the user to identify the start and end points of helices by subsequent left-mouse clicks in the micrographs. For the template-based picking algorithm, the figure-of-merit calculation, which expresses the fit between the template and any local area in the micrograph, remains unchanged from the single-particle approach (Scheres, 2015). The adaptation of the algorithm lies in the detection of filaments and cross-overs between filaments. The latter requires additional parameters for the helix diameter, the inter-box distance (in multiples of the estimated helical rise), and the maximum curvature of the filaments (which is expressed as the fraction, between 0 and 1, of the curvature of the circular mask in 2D). The algorithm to detect filaments searches for peaks in the figure-of-merit map for the entire micrograph, and considers neighbouring peaks within a distance of half the template box size. Peaks that have too few neighbours, or neighbours that do not satisfy the maximum curvature requirement, are discarded. Then, the algorithm iteratively extends candidates for helical tracks in both directions, starting from the peak

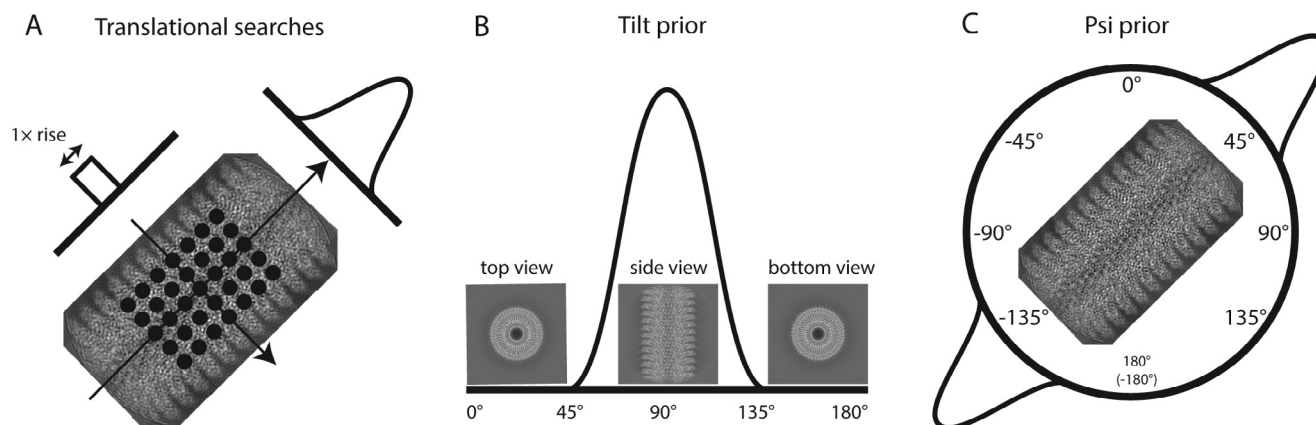


Fig. 4. Priors on the orientational parameters. A. A top-hat prior on the in-plane translations along the helical axis, and a Gaussian prior on the in-plane translations perpendicular to the helical axis. B. A mono-modal Gaussian prior on the tilt angle, which describes the out-of-plane rocking of filaments in the ice layer. C. A bi-modal Gaussian prior on the psi angle, which describes the in-plane rotation of the filament.

with the highest figure-of-merit, until no new segments can be at the ends. Once all helical tracks have been identified, the program calculates output segment coordinates according to the specified inter-box distance. At this stage, regions in the micrographs where two different helical filaments overlap are excluded.

Upon extraction of the segments from the micrographs, apart from the corresponding X-Y coordinates and the figure-of-merit, the particle extraction program also outputs for each segment its in-plane rotation along the helical track, a distinct number for the helix it belongs to, and its position within that helix. The latter information is then used to construct the relevant priors on the orientations of each segment as described above. This process does not require any additional user input when RELION is used for all steps from particle picking to refinement. However, if users provide extracted segments from outside the RELION workflow, then the additional metadata should be provided through an imported STAR file in RELION format. In that case, it may be more convenient to re-pick inside RELION. Alternatively, RELION can read coordinate files in EMAN(2) BOX-format (Tang et al., 2007), or as plain ASCII files with X and Y coordinates in the first two columns. Therefore, one can also conveniently provide the start and end coordinates of helices in such formats, and re-extract the individual segments inside RELION.

3. Results

3.1. Test data sets

To test the different aspects of our implementation we used four data sets (Table 1). Three of these data sets are available from the EMPIAR data base (Iudin et al., 2016): EMPIAR-10020 on tobacco mosaic virus (TMV) (Fromm et al., 2015); EMPIAR-10019 of the contractile sheath of the bacterial type-VI secretion system, which is composed of the proteins VipA/VipB (Kudryashev et al., 2015); and EMPIAR-10031 on prion-like aggregates of the mitochondrial antiviral signaling protein that are mediated by the N-terminal caspase activation and recruitment domain (MAVS/CARD) (Xu et al., 2015). The fourth data set was collected as part of an in-house research project on the bacterial actin homolog, MamK (Löwe et al., 2016). These test data sets were collected on different direct-electron detectors, while the filaments have different helical and point group symmetries and vary in their levels of order. TMV has been used as a test specimen for other helical reconstruction programs, e.g. (Desfosses et al., 2014; Rohou and Grigorieff, 2014), and consists of straight and regular helical assemblies,

although occasional end-to-end joints of helices do not obtain helical symmetry. VipA/VipB filaments are also relatively straight and regular, while MAVS/CARD and MamK filaments are more flexible.

3.2. Particle picking and 2D classification

Fig. 5 shows a representative micrograph for each of the four test samples. Because the end-to-end joints of TMV filaments are hard to distinguish using the template-based picking approach, we performed the picking of TMV filaments manually. The relatively straight VipA/VipB filaments were also picked manually. MAVS/CARD and MamK filaments are increasingly bent. Because the manual picking approach assumes straight filaments, template-based particle picking is more suitable for these cases. The corresponding templates were generated from initial 2D classification runs on segments that were obtained by manually picking of the straighter filaments in a subset of the micrographs.

After manual or template-based particle picking of the entire data sets, all extracted segments were subjected to a single round of reference-free 2D classification, which was used to separate sub-optimal segments from segments that contribute to high-resolution class averages (Fig. 5A–D, insets).

For the VipA/VipB and the MAVS/CARD filaments, we also compared how manual and template-based particle picking affected the resolution of the final reconstruction after 3D auto-refinement (Supplementary Table 1). For this comparison, we used the manually picked coordinates of MAVS/CARD that were available in the EMPIAR-10031 entry. For both test cases, auto-picking was again performed with 2D class averages from a small subset of manually picked segments. Template-based particle picking performed at least as well as manual particle picking. In particular for the more bent MAVS/CARD filaments, template-based particle picking selected many more particles, which gave rise to a higher resolution map after 3D auto-refinement.

3.3. Initial model generation

Because the expectation maximisation algorithm in RELION is a local optimiser, the initial 3D reference needs to be within the radius of convergence of the optimisation algorithm in order to yield a correct structure. This is also true for alternative IHRSR implementations. In particular, providing incorrect helical twist and rise has been associated with false minima that lead to grossly incorrect structures, even at apparently relatively high resolutions (reviewed by Desfosses et al., 2014; Egelman, 2014, 2010).

Table 1
Test data set characteristics.

	TMV	VipA/VipB	MAVS/CARD	MamK
Detector	K2	K2	Falcon-II	Falcon-III
Pixel size (Å)	1.126	1.0	1.05	1.34
EMPIAR entry	10,020	10,019	10,031	–
EMDB entry	2833	2699	6428	–
EMDB software	SPRING	IHRSR/SPIDER	RELION-1.2 (custom)	–
EMDB resolution (Å)	4.0	3.5	4.2	–
	RELION processing			
Box size (pix)	300	450	300	280
Inter-box distance (Å)	42.2	43.6	30.4	52.2
Nr. asym. units/box	30	2	6	1
Max. curvature	0.05	0.05	0.30	0.30
Nr. boxes (picked)	1761	11,114	227,176	118,911
Nr. boxes (final)	1761	9480	133,237	110,055
Point group symmetry	C ₁	C ₆	C ₁	C ₂
Helical rise (Å)	1.41	21.79	5.07	52.11
Helical twist (°)	22.03	29.41	−101.2	23.75
B-factor (Å ²)	−83	−91	−135	−119
Final resolution (Å)	3.9	3.3	4.0	3.9

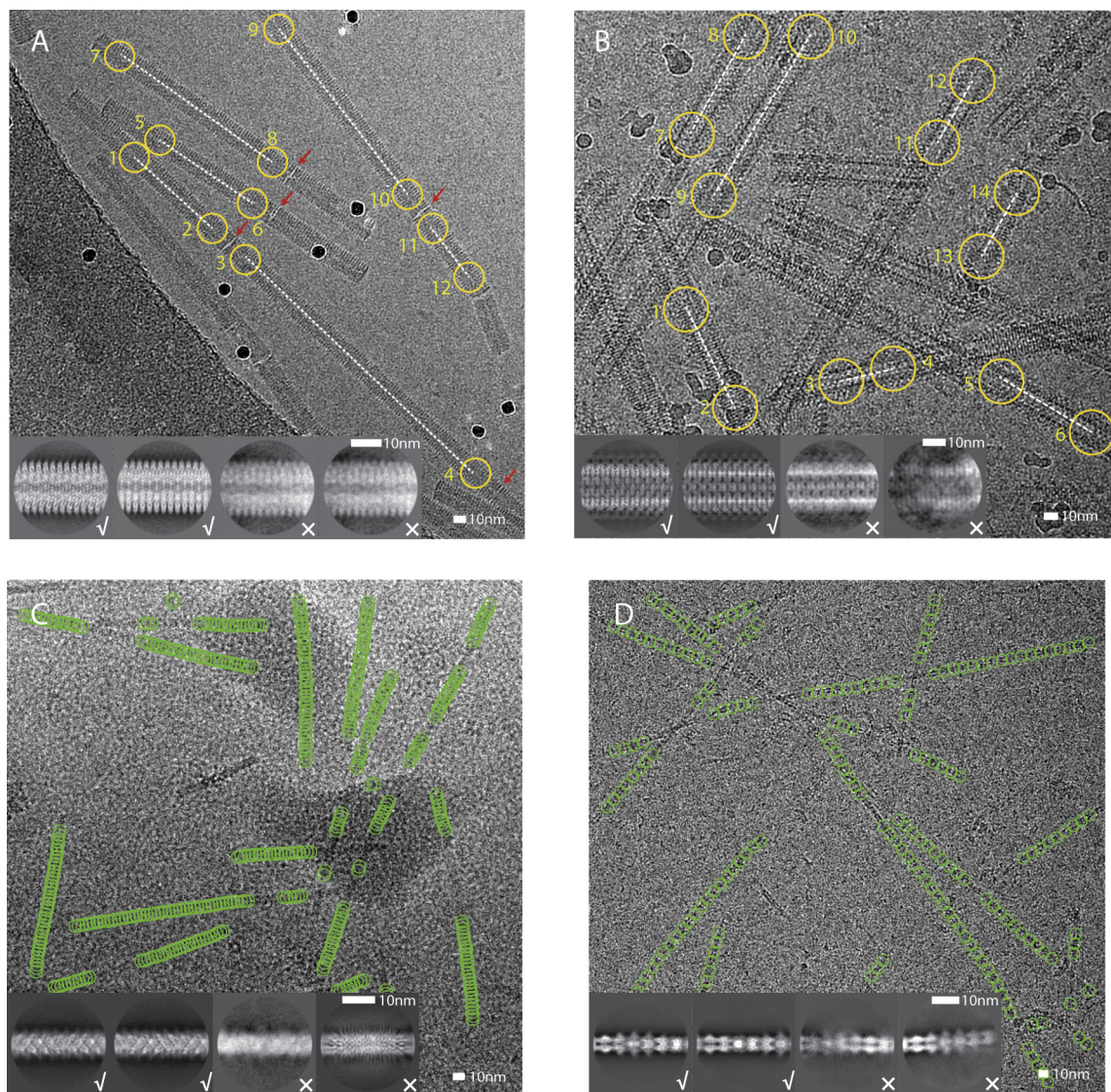


Fig. 5. Representative micrographs, particle picking and 2D classification results for the four test cases. A. Manually picked coordinates (yellow) for a TMV micrograph. Non-helical end-to-end joints are indicated with a red arrow. B. Manually picked coordinates (yellow) for a VipA/VipB micrograph. C. Auto-picked coordinates (green) for a MAVS/CARD micrograph. D. Auto-picked coordinates (green) for a MamK micrograph. For each test case, four 2D class averages are shown as insets. Good class averages are indicated with a tick, suboptimal ones with a cross.

Thereby, initial model generation and the corresponding estimation of helical parameters probably represent the most critical steps in successful helical reconstruction, and the user has multiple options here. Additional experimental data, e.g. from tomograms, sub-tomogram averages, or from homologous structures may be an efficient way to obtain initial 3D models. For example, in the original publication on the MamK structure (Löwe et al., 2016), we placed the atomic coordinates of a crystal structure of the MamK subunit in a helical lattice that was similar to actin, using the `--pdb_helix` option in the `relion_helix_toolbox` program. Nevertheless, for all four test data sets described here, we chose to calculate initial models *de novo* (Fig. 6).

For both TMV and VipA/VipB 3D refinements from a featureless cylinder did not work (also see MAVS/CARD and MamK below), and we obtained initial estimates for the helical twist and rise from indexing the diffraction pattern of a 2D class average image (Fig. 6A,B). In brief, we first calculate the helical repeat distance, which is the distance between two subunits which are (approximately) registered along the z-axis, from the reciprocal distance from the equator to the first layer-line (see blue dashed lines in Fig. 6A,B). Then, to determine how many subunits there are in a single turn, we mark the first maxima along each layer-line and find a 2D lattice that links half of these maxima, while the mirrored lattice across the meridian links the other half of the maxima. This mirror symmetry can be thought of as representing two overlapping lattices from the near and far sides of the cylindrical structure (shown with red and green boxes in Fig. 6A,B). For the two unit vectors that span this lattice, we then estimate the order of their Bessel functions from their distances to the meridian and the real-space diameter of the helix as observed in the 2D class averages. The Bessel orders and the geometry of the lattice in reciprocal space, are then used to calculate the helical pattern in real-space (Supplementary Fig. 1), from which the helical twist and rise can be derived.

For TMV, the unit vectors of the lattice on layer lines 1 and 3, have estimated Bessel orders +16 and −1, respectively. This leads to $16\frac{1}{3}$ subunits in a full turn in real-space, and hence an estimated

twist of 22.04° . The $16\frac{1}{3}$ subunits per turn results in the same structure being in register along the z-axis after 3 turns. Therefore, the estimated repeat distance of 71 Å is divided by $3 \times 16\frac{1}{3} = 49$ to yield an estimated rise of 1.45 Å. For VipA/VipB, the unit vectors of the lattice both lie on the first lattice line, and have estimated orders of +12. So initial estimates for twist and rise were 30° and 21.4 Å. Initial 3D models were then generated for both test cases by placing spheres in a helical pattern with the estimated twist and rise. The latter functionality has been implemented through the `--simulate_helix` option in the `relion_helix_toolbox` program, where the user also provides the diameter of each sphere (so that neighbouring spheres touch each other to form the helical lattice) and the diameter of the helix (as obtained from 2D class averages). Subsequent refinement of these multi-sphere models, while imposing helical symmetry and performing local optimisations of the helical twist and rise, then led to correctly converged structures at near-atomic resolution for both TMV and VipA/VipB. The final, refined values of twist and rise were 22.03° and 1.410 Å for TMV, and 29.41° and 21.79 Å for VipA/VipB.

For the MAVS/CARD and the MamK helices, we started 3D refinements from a featureless cylinder. The `relion_helix_toolbox` program implements functionality to create such models through the `--cylinder` option, where the user only needs to provide its diameter. For both test cases, refining a cylindrical model in C1, without imposing any helical symmetry (using the `--ignore_helical_symmetry` option in the `relion_refine` program), yielded a map in which helical symmetry was already apparent. The `--search` option in the `relion_helix_toolbox` program was then used to estimate the helical twist and rise from these asymmetrical reconstructions. This yielded estimates of -101.1° and 5.150 Å for MAVS/CARD and 18.29° and 55.24 Å for MamK. Using a symmetrised version of the asymmetrical map, while again imposing helical symmetry and performing local optimisations of the helical twist and rise, then also led to correct convergence. The final, refined values of twist and rise were -101.2° and 5.071 Å for MAVS/CARD, and 23.75° and 52.11 Å for MamK.

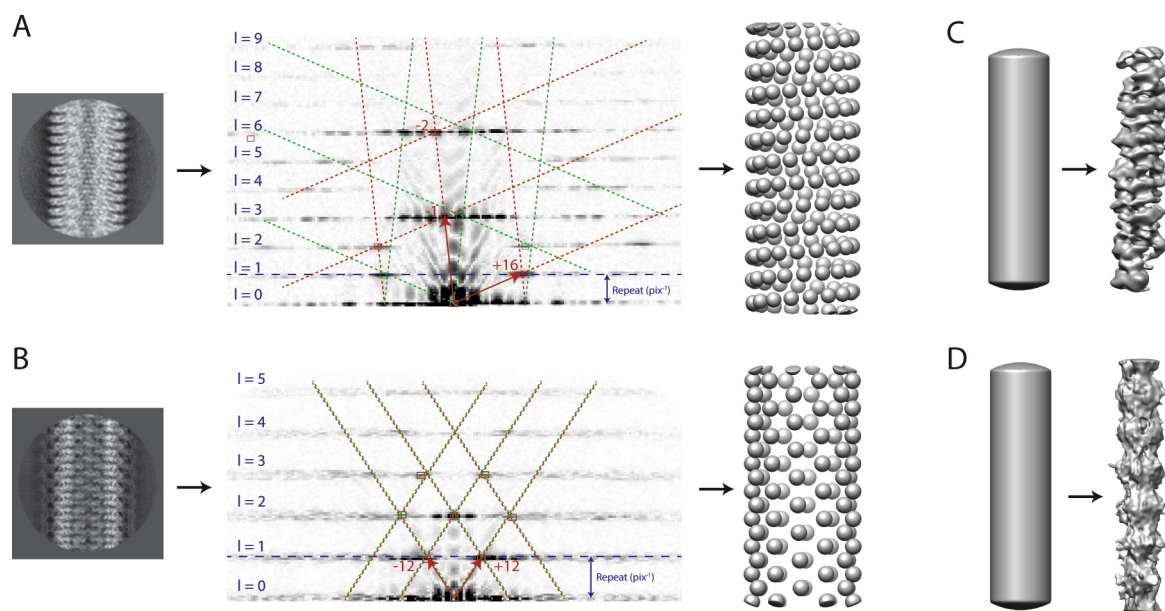


Fig. 6. Initial model generation for TMV (A), VipA/VipB (B), MAVS/CARD (C) and MamK (D). In A and B, on the left the amplitudes of Fourier-space diffraction pattern of a selected 2D class average image is shown, with the layer lines numbered and the vector pair that constructs the helical lattice indicated with red arrows. The helical lattice and its mirror counterpart are shown with red and green boxes, respectively. The corresponding initial spheres-only models are shown on the right. In C and D, the featureless cylinder (the width of which is estimated from the 2D class averages) that was used as an initial model is shown on the left, and the result from a 3D refinement without imposing any symmetry is shown on the right.

3.4. High-resolution structure determination

At this stage of the structure determination process, the workflow of high-resolution structure determination becomes more similar to the standard approach to single-particle analysis, see (Scheres, 2016) for a recent review. For each of the four test cases, we performed an initial 3D auto-refinement with the selected segments. For the TMV and MAVS/CARD data sets, we then extracted movie-frames for the selected particles and performed per-particle beam-induced motion correction and resolution dependent radiation-damage weighting for each movie frame, which is known as particle polishing in the RELION workflow (Scheres, 2014). Although movie frames were available for our MamK data set, and particle polishing was performed when we initially published this data set (Löwe et al., 2016), we chose not to re-process the movie frames of this large data set for this paper. The VipA/VipB EMPIAR entry does not contain movies, so no movie processing was performed.

After 3D auto-refinement of the polished particles, we attempted further selection of the best segments by 3D classification. In these classifications, we did not allow further adjustments of the orientations. However, for none of the four test cases 3D classification managed to identify a subset that yielded a higher resolution map than we had already calculated from all polished particles. Using RELION's post-processing job-type, we then applied automated B-factor sharpening (Rosenthal and Henderson, 2003) and corrected the FSC between the two independently refined maps of the 3D auto-refinement for the convolution effects of a solvent mask using phase randomisation (Chen et al., 2013). The final overall resolution estimate was based on the FSC = 0.143 threshold (Scheres and Chen, 2012). Variations in local resolution were estimated using phase-randomisation in combination with sliding a soft spherical mask across the entire volume (Fernandez-Leiro and Scheres, 2016a). After post-processing, we used the `relion_helix_toolbox` program with the `--impose` option to symmetrise the (asymmetric) postprocessed map according to the refined helical twist and rise. Fig. 7 shows the final maps, the solvent-corrected FSC curves and the local resolution estimates. Parameters specific to the helical reconstruction, estimated B-factors, overall resolution estimates and other refinement characteristics are summarised in Table 1. 2D classification or 3D auto-refinement for the smaller data sets of TMV and VipA/VipB took a few hours on 64–128 CPU cores. Similar calculations with the larger data sets of MAVS/CARD and MamK took up to an entire day using similar computer resources. A modern GPU (like the Nvidia GTX1080) processes the equivalent of approximately 100 CPU cores in these calculations.

3.5. Local optimisation of twist and rise

Although the local optimisation of helical twist and rise did converge to the correct solution in all four test cases, this is not necessarily the case in general. To provide insights into the robustness of the local optimisation of helical parameters, we performed multiple refinements with the MamK data set, starting from different initial estimates for the helical twist and rise (Supplementary Table 2). All these refinements were started from a cylindrical reference model that was low-pass filtered at 30 Å resolution. The auto-refinements converged within 15–25 iterations, according to the internal convergence criteria based on resolution and changes in orientations among the segments. Although the refined values of twist and rise are not part of the convergence criteria, we observed convergence onto a stable twist and rise in all refinements (e.g. see Supplementary Fig. 2). For convergence to the correct symmetry, values for the initial estimate of the twist need to lie within approximately 19–25°; values for the initial estimates

of the rise need to be within 42–62 Å. Interestingly, relatively large errors in the initial estimates can still lead to correctly refined values, but in some cases the resolution of the corresponding refined map is suboptimal. This behaviour is likely due to the switching from exhaustive orientational searches in the initial stages of the refinement to local searches later on. This means that, if in a given refinement the parameters for twist and/or rise change considerably, one should probably perform a second refinement starting from the refined values of the first run. Indications for incorrectly refined symmetries are a lower resolution of the final map, and convergence of twist and/or rise to values at the edge of the user-defined search range.

3.6. Pitfalls of imposing incorrect helical symmetry

To further illustrate the dangers of imposing an incorrect helical symmetry, we also re-refined the VipA/VipB data set using a twist of 64.9° and a rise of 3.63 Å. This symmetry corresponds to a 6-start helix with C1 symmetry, instead of the correct helix with C6 symmetry. After 3D auto-refinement and post-processing, RELION produces a map with a reported resolution of 4.5 Å (with refined values of twist = 65.07° and rise = 3.76 Å). Despite the relatively high nominal resolution, the map is devoid of protein-like features like alpha-helices or beta-strands that one would expect at this resolution (Fig. 8A–B, left). In addition, the FSC curve shows “spiky” artefacts (Fig. 8C). When the correct symmetry is imposed, the achieved resolution is 3.3 Å, the map does show the expected features (Fig. 8A–B, right), and the FSC curve is smooth.

3.7. The inter-box distance and mask diameter

The description of deviations from perfectly straight helices requires additional parameters in the form of more segments per unit-length, which may be obtained through shorter inter-box distances. In addition, smaller diameters of the circular and spherical masks in 2D and 3D reduce the dependencies between the alignments of the overlapping segments. To illustrate the effect of these two parameters, on the reconstruction results, we performed refinements with varying inter-box distances while keep the mask diameter fixed, and refinements with varying mask diameters while keeping the inter-box distance fixed for TMV and for MAVS/CARD (Supplementary Fig. 3). For straight helical tubes like TMV, using a relatively large inter-box distance is expected to save computation time without affecting the final resolution. Also, by using larger mask diameters, more mass inside the mask will contribute to the signal for alignment. For more flexible helices like MAVS/CARD, shorter inter-box distance and smaller mask diameters might be needed to describe the bends in the filaments more accurately. However, the results of our experiments suggest that larger inter-box distance still perform well, even for the more flexible MAVS/CARD filaments. In fact, in both cases inter-box distances in the range of 2–30% seem to yield structures of very similar quality. We do note that segments corresponding to the most highly curved MAVS/CARD filaments are rejected by our auto-picking algorithm. An obvious advantage of using larger inter-box distances is that the computational costs of the refinement scale linearly with the total number of extracted segments. With regards to the mask diameter, in most cases one should probably not use smaller values than the diameter of the helices, in order not to discard part of the signal in alignment. For TMV, reduces the mask diameter from 90% to 80% of the box size had no noticeable effect. For MAVS/CARD, using a mask diameter of 80% of the box size was marginally better than using 70% or 90%. However, the resolution of the reconstruction decreased markedly when using only 60%, presumably because too little signal remained within the mask to yield accurate alignments.

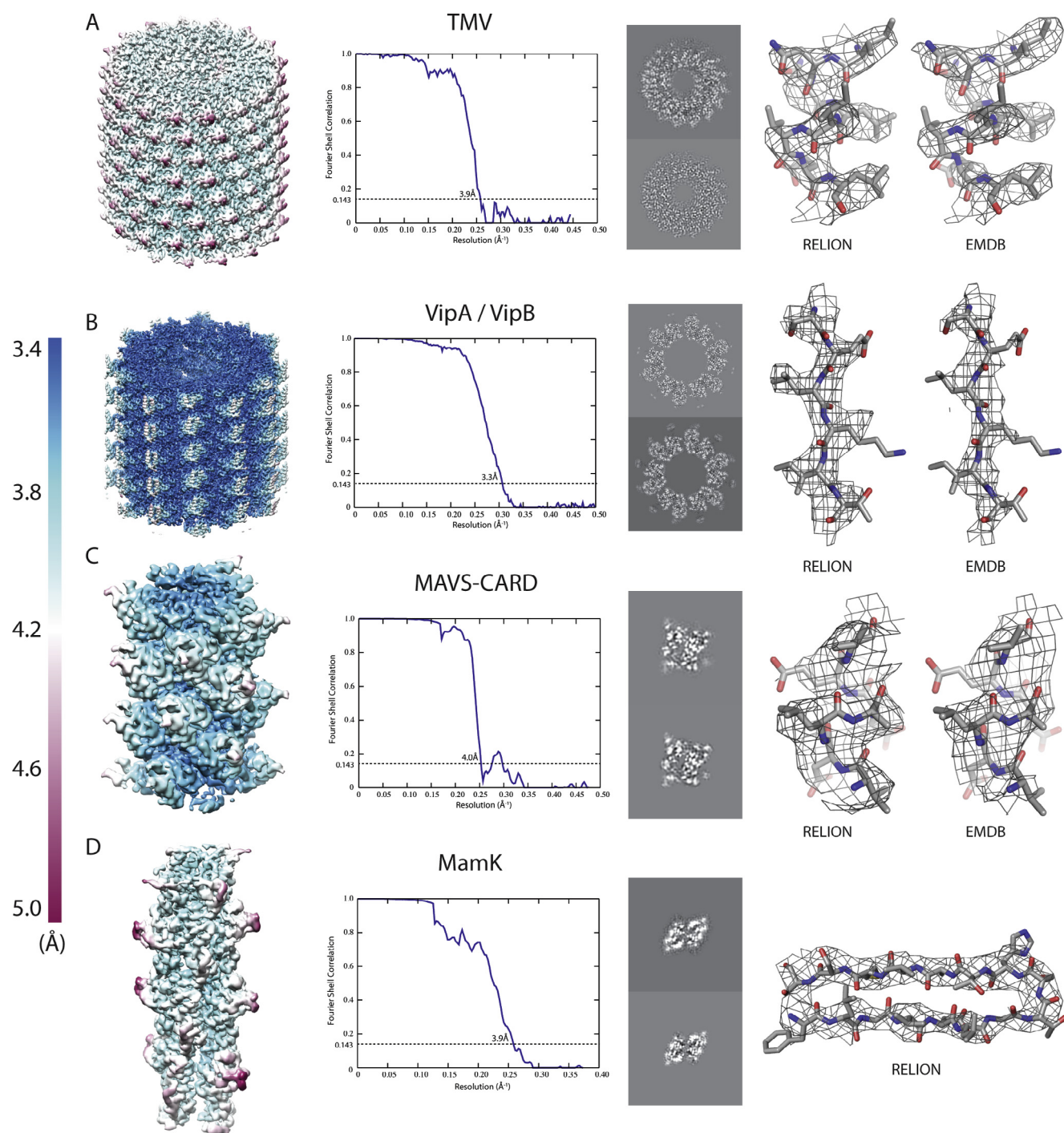


Fig. 7. Reconstructions for the four test cases: TMV (A), VipA/VipB (B), MAVS/CARD (C) and MamK (D). A view of the reconstructed helix coloured according to local-resolution estimates (from 3.5 Å in blue to 5 Å in maroon) is shown on the left. The FSC curve between the two independently refined half-sets is shown in the middle. A detail of the structure with the corresponding atomic model is shown on the right. For TMV, VipA/VipB and MAVS/CARD, the same detail from the EMDB map is also shown.

3.8. The half-sets for independent refinements

In 3D auto-refinement the data is divided into two half-sets, which are then refined independently in order to prevent an iterative built-up of noise due to overfitting (Scheres and Chen, 2012). For single-particle analysis, the half-sets are created randomly from the extracted particles. For helical segments, random generation of the half-sets may be problematic, as the signal in one segment will partially overlap with the signal in its neighbouring segment. This overlap could introduce dependencies between the half-set refinements, and thus potentially lead to overfitting.

There are two aspects of our helical implementation that prevent this type of overfitting. Firstly, the prior on the translations (Fig. 4A) only extends for a single rise in the direction of the helical axis. This will prevent the signal of any segment from translating on top of the identical, overlapping signal in the neighbouring segment. Secondly, to build in an additional safeguard against overfitting, our implementation of helical processing divides the data set into half-sets based on a random division of helical filaments. That is, at the particle extraction stage, all segments from a single filament will be assigned to the same half-set.

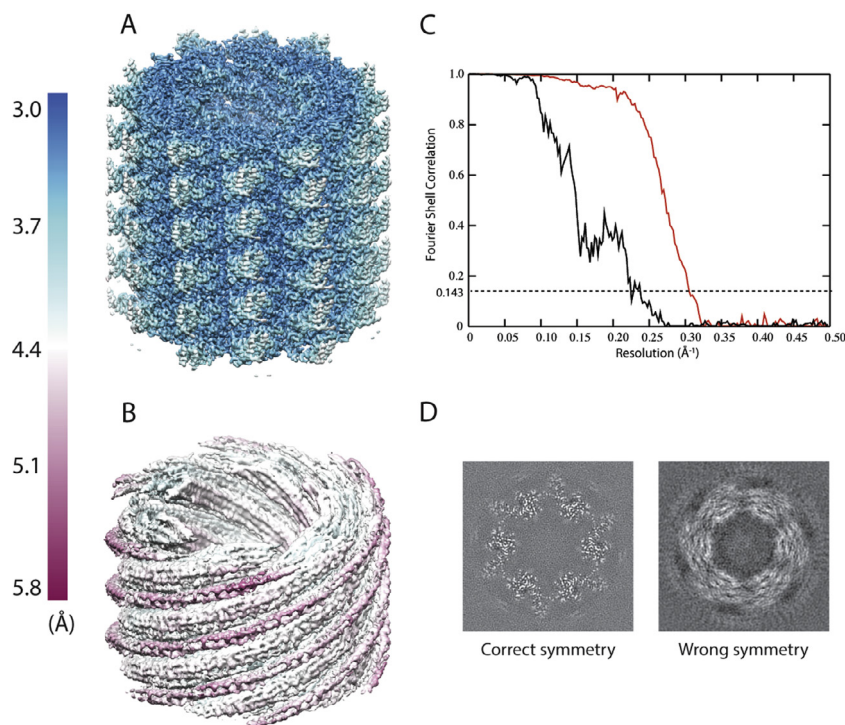


Fig. 8. Effects of imposing different helical symmetries for the VipA/VipB data set. A. Reconstruction with the correct helical twist and rise imposed. B. Reconstruction with the wrong helical symmetry imposed. The maps in A and B are coloured according to local-resolution estimates (from 4 Å in blue to 6 Å in maroon). C. FSC curves between two independently refined half-sets for the refinement with the correct helical symmetry (red) and the wrong helical symmetry (black). D. Central slices through the unsharpened map from the refinement with the correct helical symmetry (left) and the wrong helical symmetry (right).

To test both counter-overfitting measures, we performed 3D auto-refinements on the TMV data set (Supplementary Fig. 4). In the first refinement, we used our implementation as intended, by both using the translational prior in the helical direction and the division of the data set into halves on the filament level. The post-processed FSC curve indicates a final resolution of 4.1 Å, which is in agreement with the features visible in the map. (Note that this resolution estimate is somewhat lower than the one reported in Fig. 7, as non-polished particles were used in this test.) Importantly, beyond 3.8 Å, the FSC drops to values close to zero for the rest of the frequencies (note that the post-processed FSC in RELION does not report negative FSC values). Secondly, we performed a refinement where we switched off both safe guards, i.e. we used an isotropic translational prior and random half-sets at the segment-level. In this case, severe overfitting of the data occurs as expected, as is evident from the FSC curve, which never drops to zero. A third refinement, where the half-sets were made at the filament-level but isotropic translational priors were used, shows no signs of overfitting. This refinement yielded a lower resolution map than the first refinement, probably as a result of a suboptimal translations along the helical axis. Finally, in a fourth refinement we made the half-sets at the segment-level and only used the anisotropic translational prior. In this case, a small amount of overfitting is present, indicating that the anisotropic translational prior alone is not enough to rigorously prevent overfitting. Therefore, random division on helical segments alone is enough to prevent iterative overfitting, but because the translational priors will also prevent loss of data due to translations into neighbouring segments, the default in RELION was set to use both measures.

3.9. Angular and translational priors

In IHRSR, the concept of priors on the rotations or translations is expressed as local searches within a user-defined search range.

Within the statistical framework, this corresponds to imposing a prior with a top-hat shape (i.e. with a constant, non-zero value within the search range, and a zero value outside the search range). In general, the better the shape of the prior reflects the actual distribution of the corresponding orientational parameters, the better the program is expected to perform, and the difference in performance is expected to become larger with increasing levels of noise in the data. FREALIX (Rohou and Grigorieff, 2014) and FREALIGN (Alushin et al., 2010) implement Gaussian-like priors similar to the ones in RELION. Another advantage of the statistical framework, is that the optimal search range can be estimated from the data themselves, e.g. by optimising the variance in the translations perpendicular to the helical axis.

To test the influence of the shape of the prior on the refinement results, we compared standard refinements with Gaussian priors to refinements with top-hat priors. In the latter refinements, all top-hat priors had non-zero values from -3 to $+3$ standard deviations of the corresponding Gaussian priors in the standard runs. Supplementary Fig. 5 shows the resulting FSC curves after post-processing for these refinements on the MamK and VipA/VipB data sets. The difference between the Gaussian and the top-hat prior was not notable for MamK, but for VipA/VipB the Gaussian prior did yield a better structure.

4. Discussion

We introduce functionality for helical reconstruction in the RELION program, and show for four different data sets that it can calculate maps to near-atomic resolution. For the three test cases from the EMPIAR data base, we show that helical processing in RELION yields results that are comparable to those obtained in alternative programs. Our implementation is integrated into the standardised workflow engine of RELION version 2.0, which facilitates the design of image processing strategy for novice users, and

provides convenient tools for electronic bookkeeping and file management (Fernandez-Leiro and Scheres, 2016a). Because of the strong parallels with single particle analysis, users already familiar with RELION should find the transition to helical processing straightforward. The interaction with the user is similar to single-particle analysis of asymmetric particles, and parameters specific for helical reconstruction are mainly provided on dedicated tabs in the main GUI. Advantages that are inherent to the empirical Bayesian approach in RELION, such as the data-driven calculation of optimal filters for 3D reconstructions and alignment, are directly carried over from standard single-particle analysis to helical processing. In addition, the existing options for classification of images in 2D or in 3D, with or without masks, and with or without subtraction of partial signal (Bai et al., 2015; Scheres, 2016), are all available to be explored for helical processing of highly heterogeneous samples.

In the test cases presented in this paper, 2D classification proved useful in discarding suboptimal segments from the data set, but similar classification attempts in 3D were unsuccessful. These results contrast with common observations in single-particle analysis, where 3D classification is often key in selecting subsets of the data that yield the highest resolution maps (Fernandez-Leiro and Scheres, 2016b). On the one hand, these results may be attributed to the relatively high quality of the test data sets. On the other hand, there are intrinsic differences in classification of helices and individual asymmetric particles. The packing interactions inside the helical lattice, for example, may impose limitations on the structural heterogeneity of the individual protein subunits. And if structural heterogeneity does exist among the neighbouring subunits, for example because a floppy domain sticks out into the solvent, then classification of that heterogeneity may be difficult. The concept of a 3D helical class is particularly inadequate when the structural heterogeneity happens independently among individual segments along each filament. Only when different types of helical structures exist, i.e. when neighbouring segments in stretches of the filaments that are at least as long as the mask diameter all adopt the same conformation, is 3D classification expected to perform well. For such cases, we implemented the option to independently refine the helical twist and rise of each class. Although this option did not improve the maps for the test cases described in this paper, we envision that it will be useful in cases where these types of structural heterogeneity do exist.

We explicitly stress that our implementation does not provide a generally applicable solution to the problem of local minima in the optimisation of helical parameters. Similar to alternative implementations of the IHRSR approach, RELION is based on an optimisation algorithm that will converge to the nearest local minimum, and the energy landscape of helical reconstruction, in particular with respect to the twist and rise parameters, is known to be highly complex (Desfosses et al., 2014; Egelman, 2010, 2014). Thereby, the generation of a reliable initial 3D model with reasonable estimates for the helical twist and rise is likely to be the most difficult hurdle in helical reconstruction. For our test cases we employed two different methods. For MAVS/CARD and MamK, we successfully refined a featureless cylinder without imposing any point group or helical symmetry, and then estimated the helical twist and rise from the asymmetric reconstruction. This method is more likely to work for relatively simple helices, i.e. helices with few asymmetric subunits per box and few interactions between the subunits. An advantage of this method is that it has minimal user input. A drawback is that the initial few iterations of such refinements are relatively slow, as the probabilities for many orientations are significant, and these orientations thus all need to be considered in the reconstruction. For more complicated helices, asymmetric refinement of a featureless cylinder often does

not work. For example, asymmetric refinement of a featureless cylinder never yielded a structure with discernable subunits for TMV or VipA/VipB. Fortunately, more complicated helices often yield more information-rich Fourier transforms, especially when these can be averaged together into high-resolution 2D class averages. Therefore, for complicated helices Fourier-Bessel indexing is a viable way to determine the initial estimate for helical twist and rise. For both TMV and VipA/VipB, placing spheres into the corresponding helical lattice then yielded an initial model within the radius of convergence of the refinement algorithm. Although computationally cheaper than refinement of a featureless cylinder, a drawback of this method is its dependence on user expertise in indexing diffraction patterns. Still, Fourier-Bessel indexing typically yields a better understanding of the structure, and may therefore help to avoid errors.

Our refinement of the VipA/VipB data set as a 6-start helix in C1 instead of the correct helical symmetry in C6 illustrates the pitfalls of imposing incorrect helical symmetry. Despite the incorrect symmetry, 3D auto-refinement yielded two half-maps that correlated with each other up to a resolution of 4.5 Å. Instead of directly measuring resolution, the FSC merely measures the consistency between two maps. By imposing the same, incorrect symmetry in both half-set refinements, the resulting maps become incorrect in a consistent manner, which results in a near-atomic resolution estimation of a completely incorrect map. At closer inspection, the reconstructed map did not reveal any protein-like features, and also the FSC curve itself showed artefacts. However, the indications that something is wrong may not necessarily be so clear. Actually, the MAVS/CARD test case was the subject of controversy itself, as an initial structure by the same authors was proven to be incorrect (Egelman, 2014; Xu et al., 2014, 2015). In our hands, starting MAVS/CARD refinements without imposing any symmetry from a featureless cylinder still led to the correct structure (Fig. 6). Moreover, when starting from the incorrect helical parameters described in the original MAVS/CARD paper, the helical twist and rise consistently converged to the edge of the search ranges (not shown), which is a strong indication of incorrect estimates. Nevertheless, in the general case, validation of helical symmetry will remain difficult until one can reconstruct maps to near-atomic resolution, protein features like separated beta-strands and amino acid side chains can be readily distinguished, and a stereochemically sound atomic model can be refined inside the map. For cases where such maps cannot be obtained, doubts about the correct helical symmetry will remain relevant.

Apart from the initial model and the initial estimate for the helical twist and rise, important user-defined parameters are the inter-box distance and the different options to impose priors on the orientations. Smaller inter-box distances and small mask diameters allow modelling more flexible filaments, but smaller inter-box distances result in more segments and thus come at considerable computational cost, while smaller mask diameters may lead to too little signal for accurate alignment. Our results with TMV and MAVS/CARD indicate that relatively larger inter-box distances (i.e. up to 30%) may indeed be used without noticeable loss in resolution. When in doubt, one could start processing with a relatively large inter-box distance, and then re-run refinements near the end with a smaller inter-box distance to make sure the best possible resolution is obtained. Our results with VipA/VipB indicate that Gaussian-shaped priors may add useful information to the refinement when compared to top-hat priors, which are typically employed in alternative programs. This difference however was not noticeable in the case of MamK. Still, because these priors come at virtually no extra computational cost and Gaussian distributions are more likely to describe the orientational distributions in experimental data, the Gaussian priors in RELION are switched on by default.

Within the field of single-particle analysis, considerable concern has arisen about the accurate estimation of resolution in 3D reconstructions (Henderson et al., 2012). One approach to prevent the iterative build-up of high-resolution noise in the reconstruction is the independent refinement of two halves of the data (Scheres and Chen, 2012). In standard single-particle analysis, the two half sets are often generated randomly or based on odd/even particle numbers. When extracting overlapping segments from helical filaments, the overlap in signal between neighbouring segments could introduce spurious dependencies between the two half-sets. Our results show that a random division of the data set based on filaments rather than segments, combined with a tophat prior on the translations along the helical axis, effectively yield maps with reliable resolution estimates.

In conclusion, we present an implementation of the IHRSR approach to helical processing in the RELION program. The object-oriented character of RELION's code facilitates the re-use of many of its functions. This makes it relatively straightforward to extend the helical processing approach presented here to 3D data, such as sub-tomograms. This extension is something we will pursue for a future version. Another possible extension is the introduction of functionality to deal with breaks in the helical symmetry, such as the seam in microtubules (Metlagel et al., 2007). The code available in the current 2.0 version implements all features described in this paper, and has already been useful in calculating several new structures in our own lab. We hope this paper will help to make it similarly useful to others.

Acknowledgements

We thank Nigel Unwin, Tony Crowther, Jude Short and Jan Löwe for stimulating discussions, and Jake Grimmett and Toby Darling for assistance with high-performance computing. This work was funded by the UK Medical Research Council (MC_UP_A025_1013).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jsb.2017.02.003>.

References

- Alushin, G.M., Ramey, V.H., Pasqualato, S., Ball, D.A., Grigorieff, N., Musacchio, A., Nogales, E., 2010. The Ndc80 kinetochore complex forms oligomeric arrays along microtubules. *Nature* 467, 805–810. <http://dx.doi.org/10.1038/nature09423>.
- Bai, X., Rajendra, E., Yang, G., Shi, Y., Scheres, S.H., 2015. Sampling the conformational space of the catalytic subunit of human γ -secretase. *eLife* 4. <http://dx.doi.org/10.7554/eLife.11182>.
- Behrmann, E., Tao, G., Stokes, D.L., Egelman, E.H., Raunser, S., Penczek, P.A., 2012. Real-space processing of helical filaments in SPARX. *J. Struct. Biol.* 177, 302–313. <http://dx.doi.org/10.1016/j.jsb.2011.12.020>.
- Beroukhim, R., Unwin, N., 1997. Distortion correction of tubular crystals: improvements in the acetylcholine receptor structure. *Ultramicroscopy* 70, 57–81. [http://dx.doi.org/10.1016/S0304-3991\(97\)00070-3](http://dx.doi.org/10.1016/S0304-3991(97)00070-3).
- Bluemke, D.A., Carragher, B., Josephs, R., 1988. The reconstruction of helical particles with variable pitch. *Ultramicroscopy* 26, 255–270. [http://dx.doi.org/10.1016/0304-3991\(88\)90226-4](http://dx.doi.org/10.1016/0304-3991(88)90226-4).
- Chen, S., McMullan, G., Faruqi, A.R., Murshudov, G.N., Short, J.M., Scheres, S.H.W., Henderson, R., 2013. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* 135, 24–35. <http://dx.doi.org/10.1016/j.ultramicro.2013.06.004>.
- Clemens, D.L., Ge, P., Lee, B.-Y., Horwitz, M.A., Zhou, Z.H., 2015. Atomic structure of T6SS reveals interlaced array essential to function. *Cell* 160, 940–951. <http://dx.doi.org/10.1016/j.cell.2015.02.005>.
- Cochran, W., Crick, F.H., Vand, V., 1952. The structure of synthetic polypeptides. I. The transform of atoms on a helix. *Acta Crystallogr.* 5, 581–586. <http://dx.doi.org/10.1107/S0365110X52001635>.
- Crowther, R.A., Henderson, R., Smith, J.M., 1996. MRC image processing programs. *J. Struct. Biol.* 116, 9–16. <http://dx.doi.org/10.1006/jsbi.1996.0003>.
- De Rosier, D.J., Klug, A., 1968. Reconstruction of three dimensional structures from electron micrographs. *Nature* 217, 130–134. <http://dx.doi.org/10.1038/217130a0>.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum-likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–38.
- DeRosier, D.J., Moore, P.B., 1970. Reconstruction of three-dimensional images from electron micrographs of structures with helical symmetry. *J. Mol. Biol.* 52, 355–369.
- Desfosses, A., Ciuffa, R., Gutsche, I., Sachse, C., 2014. SPRING – an image processing package for single-particle based helical reconstruction from electron cryomicrographs. *J. Struct. Biol.* 185, 15–26. <http://dx.doi.org/10.1016/j.jsb.2013.11.003>.
- Diaz, R., Rice, W.J., Stokes, D.L., 2010. Fourier-bessel reconstruction of helical assemblies. In: *Methods in Enzymology*. Elsevier, pp. 131–165.
- Egelman, E.H., 2000. A robust algorithm for the reconstruction of helical filaments using single-particle methods. *Ultramicroscopy* 85, 225–234.
- Egelman, E.H., 2007. The iterative helical real space reconstruction method: Surmounting the problems posed by real polymers. *J. Struct. Biol.* 157, 83–94. <http://dx.doi.org/10.1016/j.jsb.2006.05.015>.
- Egelman, E.H., 2010. Reconstruction of helical filaments and tubes. *Methods Enzymol.* 482, 167–183. [http://dx.doi.org/10.1016/S0076-6879\(10\)82006-3](http://dx.doi.org/10.1016/S0076-6879(10)82006-3).
- Egelman, E.H., 2014. Ambiguities in helical reconstruction. *eLife* 3, e04969. <http://dx.doi.org/10.7554/eLife.04969>.
- Fernandez-Leiro, R., Scheres, S., 2016. A pipeline approach to single-particle processing in RELION (No. bioRxiv:078352v1).
- Fernandez-Leiro, R., Scheres, S.H.W., 2016b. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* 537, 339–346. <http://dx.doi.org/10.1038/nature19948>.
- Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., Leith, A., 1996. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* 116, 190–199. <http://dx.doi.org/10.1006/jsbi.1996.0030>.
- Fromm, S.A., Bharat, T.A.M., Jakobi, A.J., Hagen, W.J.H., Sachse, C., 2015. Seeing tobacco mosaic virus through direct electron detectors. *J. Struct. Biol.* 189, 87–97. <http://dx.doi.org/10.1016/j.jsb.2014.12.002>.
- Henderson, R., Sali, A., Baker, M.L., Carragher, B., Devkota, B., Downing, K.H., Egelman, E.H., Feng, Z., Frank, J., Grigorieff, N., Jiang, W., Ludtke, S.J., Medalia, O., Penczek, P.A., Rosenthal, P.B., Rossmann, M.G., Schmid, M.F., Schröder, G.F., Steven, A.C., Stokes, D.L., Westbrook, J.D., Wriggers, W., Yang, H., Young, J., Berman, H.M., Chiu, W., Kleywegt, G.J., Lawson, C.L., 2012. Outcome of the first electron microscopy validation task force meeting. *Struct. Lond. Engl.* 1993 (20), 205–214. <http://dx.doi.org/10.1016/j.str.2011.12.014>.
- Iudin, A., Korir, P.K., Salavert-Torres, J., Kleywegt, G.J., Patwardhan, A., 2016. EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods* 13, 387–388. <http://dx.doi.org/10.1038/nmeth.3806>.
- Kimanius, D., Forsberg, B.O., Scheres, S.H., Lindahl, E., 2016. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* 5. <http://dx.doi.org/10.7554/eLife.18722>.
- Klug, A., Crick, F.H.C., Wyckoff, H.W., 1958. Diffraction by helical structures. *Acta Crystallogr.* 11, 199–213. <http://dx.doi.org/10.1107/S0365110X58000517>.
- Kudryashev, M., Wang, R.Y.-R., Brackmann, M., Scherer, S., Maier, T., Baker, D., DiMaio, F., Stahlberg, H., Egelman, E.H., Basler, M., 2015. Structure of the Type VI secretion system contractile sheath. *Cell* 160, 952–962. <http://dx.doi.org/10.1016/j.cell.2015.01.037>.
- Löwe, J., He, S., Scheres, S.H.W., Savva, C.G., 2016. X-ray and cryo-EM structures of monomeric and filamentous actin-like protein MamK reveal changes associated with polymerization. *Proc. Natl. Acad. Sci.* 113, 13396–13401. <http://dx.doi.org/10.1073/pnas.1612034113>.
- Metlagel, Z., Kikkawa, Y.S., Kikkawa, M., 2007. Ruby-Helix: an implementation of helical image processing based on object-oriented scripting language. *J. Struct. Biol.* 157, 95–105. <http://dx.doi.org/10.1016/j.jsb.2006.07.015>.
- Owen, C.H., Morgan, D.G., DeRosier, D.J., 1996. Image analysis of helical objects: the brandeis helical package. *J. Struct. Biol.* 116, 167–175. <http://dx.doi.org/10.1006/jsbi.1996.0027>.
- Rohou, A., Grigorieff, N., 2014. FREALIGN: model-based refinement of helical filament structures from electron micrographs. *J. Struct. Biol.* 186, 234–244. <http://dx.doi.org/10.1016/j.jsb.2014.03.012>.
- Rosenthal, P.B., Henderson, R., 2003. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* 333, 721–745.
- Sachse, C., Chen, J.Z., Coureux, P.-D., Stroupe, M.E., Fändrich, M., Grigorieff, N., 2007. High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus. *J. Mol. Biol.* 371, 812–835. <http://dx.doi.org/10.1016/j.jmb.2007.05.088>.
- Scheres, S.H.W., 2012a. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* 180, 519–530. <http://dx.doi.org/10.1016/j.jsb.2012.09.006>.
- Scheres, S.H.W., 2012b. A Bayesian view on cryo-EM structure determination. *J. Mol. Biol.* 415, 406–418. <http://dx.doi.org/10.1016/j.jmb.2011.11.010>.
- Scheres, S.H.W., 2014. Beam-induced motion correction for sub-megadalton cryo-EM particles. *eLife* 3, e03665.
- Scheres, S.H.W., 2015. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* 189, 114–122. <http://dx.doi.org/10.1016/j.jsb.2014.11.010>.
- Scheres, S.H.W., 2016. Processing of structurally heterogeneous Cryo-EM data in RELION. *Methods Enzymol.* 579, 125–157. <http://dx.doi.org/10.1016/bs.mie.2016.04.012>.

- Scheres, S.H.W., Chen, S., 2012. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* 9, 853–854. <http://dx.doi.org/10.1038/nmeth.2115>.
- Short, J.M., Liu, Y., Chen, S., Soni, N., Madhusudhan, M.S., Shivji, M.K.K., Venkitaraman, A.R., 2016. High-resolution structure of the presynaptic RAD51 filament on single-stranded DNA by electron cryo-microscopy. *Nucleic Acids Res.* 44, 9017–9030. <http://dx.doi.org/10.1093/nar/gkw783>.
- Sosa, H., Hoenger, A., Milligan, R.A., 1997. Three different approaches for calculating the three-dimensional structure of microtubules decorated with kinesin motor domains. *J. Struct. Biol.* 118, 149–158. <http://dx.doi.org/10.1006/jsbi.1997.3851>.
- Stewart, M., 1988. Computer image processing of electron micrographs of biological structures with helical symmetry. *J. Electron Microsc. Tech.* 9, 325–358. <http://dx.doi.org/10.1002/jemt.1060090404>.
- Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., Ludtke, S.J., 2007. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* 157, 38–46. <http://dx.doi.org/10.1016/j.jsb.2006.05.009>.
- Whittaker, M., Carragher, B.O., Milligan, R.A., 1995. PHOELIX: a package for semi-automated helical reconstruction. *Ultramicroscopy* 58, 245–259.
- Xu, H., He, X., Zheng, H., Huang, L.J., Hou, F., Yu, Z., de la Cruz, M.J., Borkowski, B., Zhang, X., Chen, Z.J., Jiang, Q.-X., 2014. Structural basis for the prion-like MAVS filaments in antiviral innate immunity. *eLife* 3. <http://dx.doi.org/10.7554/eLife.01489>.
- Xu, H., He, X., Zheng, H., Huang, L.J., Hou, F., Yu, Z., de laCruz, M.J., Borkowski, B., Zhang, X., Chen, Z.J., Jiang, Q.-X., 2015. Correction: structural basis for the prion-like MAVS filaments in antiviral innate immunity. *eLife* 4, e07546. <http://dx.doi.org/10.7554/eLife.07546>.