TECHNISCHE UNIVERSITÄT MÜNCHEN
Max-Planck-Institut für Biochemie
Abteilung Molekulare Strukturbiologie

Entwicklung rechnergestützter Methoden für die Strukturanalyse von Makromolekülen durch die Kryoelektronentomographie

Thomas Hrabe

Vollständiger Abdruck der von der Fakultät für Chemie der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Michael Sattler
Prüfer der Dissertation:	1. Hon.-Prof. Dr. Wolfgang Baumeister
	2. Univ.-Prof. Dr. Sevil Weinkauf

Die Dissertation wurde am 2.5.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Chemie am 11.7.2012 angenommen.

Zusammenfassung

Die Kryoelektronentomographie (KET) ermöglicht die Abbildung und strukturelle Analyse makromolekularer Komplexe in ihrer nahezu natürlichen Umgebung. Allerdings erlaubt erst die Mittelung vieler identischer Proteinkomplexe zu einer Dichte den Einblick in strukturelle Details. Vor der Mittelung identischer Einzelpartikel müssen diese zuerst in Tomogrammen gefunden werden. Die tomographischen Einzelpartikel (Subtomogramme) müssen als nächstes in die identische Orientierung ausgerichtet werden, damit im Mittel eine Auflösung von $(15 - 30)\text{Å}^{-1}$ erreicht werden kann. Zudem können Subtomogramme durch Klassifikation anhand struktureller Unterschiede in sub-Populationen aufgeteilt werden. Kontinuierliche Verbesserungen von Instrumenten und Datenverabeitung führen zu qualitativ höherwertigen Resultaten. Deshalb wird ein wohl geordneter, rechnergestützter Arbeitsablauf für die erfolgreiche und schnelle KET Datenverarbeitung essenziell.

In dieser Arbeit wurde die neue *open source* Software-Plattform PyTom entwickelt, mit dem Ziel, die Arbeitsabläufe der Datenverarbeitung zu standardisieren und ebenso neue Methoden für die Verarbeitung von Tomogammen bereitzustellen. Dem Benutzer stehen parallelisierte Prozeduren für die schnelle Datenverarbeitung zur Verfügung, welche durch standardisierte Schnittstellen den raschen Fortschritt zum nächsten Prozess ermöglichen. Implementierte Prozeduren können leicht angepasst und durch neue Methoden erweitert werden. Die Ausrichtung von Subtomogrammen wurde durch die Verwendung adaptiver Parameter gegenüber der klassischen Variante verbessert und ebenso vereinfacht. Die in dieser Arbeit implementierte stochastische Klassifikationsmethode von Subtomogrammen, basierend auf *Simulated Annealing*, ist im Vergleich zu deterministischen Klassifikationsmethoden genauer. Testergebnisse von Ausrichtung und Klassifikation von Subtomogrammen wurden auf simulierten Daten, wie auch auf experimentellen Datensätzen (GroEL_{14}/$\text{GroEL}_{14}\text{GroES}_7$) bestimmt. Um das erfolgreiche Zusammenspiel der einzelnen Methoden zu demonstrieren, wurden Tomogramme eines *S. cerevisae*-Lysats prozessiert. Außerdem untermauern die Ergebnisse einer Studie von, an die Membran des Endoplasmatischen Retikulums gebundenen, Ribosomen die Zuverlässigkeit von PyTom, mit dessen Hilfe der Translokationsvorgang in das ER bei 30^{-1}Å^{-1} untersucht werden konnte.

Abstract

Cryo-electron tomography (CET) is a three-dimensional imaging technique for structural studies of macromolecules under close-to-native conditions. Averaging of subtomograms, each containing a copy of a macromolecule of interest, provides substantially higher resolution insights $(15 - 30)\text{Å}^{-1}$ into macromolecules than CET alone. In-depth analysis of macromolecule populations depicted in tomograms requires identification of subtomograms corresponding to putative particles, averaging of subtomograms to enhance their signal, and classification to capture the structural variations among them. With further advances in hard- and software looming, subtomogram analysis will play an increasingly important role in structural biology and streamlined software protocols will become key requirements towards rapid data-processing.

In this work, the open-source platform PyTom is introduced that unifies standard tomogram processing steps into a python toolbox. PyTom enables parallelized processing of large numbers of tomograms, but also provides a convenient, sustainable environment for algorithmic development. For subtomogram averaging, an adaptive adjustment of scoring and sampling was implemented that clearly improves the resolution of averages compared to static strategies. Furthermore, a novel stochastic classification method based on Simulated Annealing yields significantly more accurate classification results than two deterministic approaches. Faithful alignment and classification results determined by the PyTom toolbox are obtained by processing simulated and experimental subtomograms of ribosomes and GroEL_{14}/$\text{GroEL}_{14}\text{GroES}_7$, respectively. Robustness of all implemented procedures is demonstrated by processing the whole workflow on an experimental dataset constituted of *S. cerevisae* lysate tomograms. Furthermore, current studies of Ribosomes associated to membranes of the rough, endoplasmatic reticulum utilize PyTom routines and give insight how complexes involved into translocation are arranged at beyond 30^{-1}Å^{-1} resolution.

Teile dieser Arbeit wurden veröffentlicht

Hrabe, T. und Förster, F. (2011). Structure Determination by Single Particle Tomography. *Encyclopedia of Life Sciences*, DOI: 10.1002/9780470015902.a0023175

Hrabe, T., Chen, Y., Pfeffer, S., Cuellar, L. K., Mangold, A.-V. und Förster, F. (2012). PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *Journal of Structural Biology*, DOI:10.1016/j.jsb.2011.12.003

Chen, Y., Hrabe, T., Pfeffer, S., Pauly, O., Mateus, D., Navab, N. und Förster, F. (2012). Detection and Identification of Macromolecular Complexes in Cryo-Electron Tomograms using Support Vector Machines. *IEEE International Symposium on Biomedical Imaging*

Pfeffer, S., Brandt, F., Hrabe, T., Eibauer, M., Lang, S., Zimmermann, R. und Förster, F. (2012). Structure and 3D arrangement of ER membrane associated ribosomes, im Druck

Inhaltsverzeichnis

1 Einleitung

Die strukturelle Analyse von intakten Proteinkomplexen innerhalb von Zellen gibt Aufschluss über die Funktion einzelner Makromoleküle und über die Interaktionsketten von Proteinkomplexen untereinander. Derartige Informationen sind essentiell, um zu bestimmen, wie makromolekulare Maschinen innerhalb von Zellen funktionieren. Mikroskope ermöglichen den hierfür vergrößernden Einblick in die Organisation von Zellen. Mit Hilfe der Lichtmikroskopie ist es z.B. möglich, die Morphologie von Zellen zu untersuchen. Allerdings ist die Auflösung durch die Wellenlänge des Lichtes auf wenige $100nm$ begrenzt. Bekannte Signale entsprechend markierter Moleküle können aber z.B. mittels Fluoreszenzemission viel genauer lokalisiert werden ($\leq 100nm$). Mit Hilfe der Lichtmikroskopie ist es folglich nicht möglich, Informationen über die quarternäre Struktur makromolekularer Komplexe zu erhalten, und Interaktion dieser innerhalb der Zellen zu analysieren.

Die Struktur von Makromolekülen wird in der Regel durch die Röntgenkristallographie bestimmt. Die erreichte Auflösung liegt hier bei wenigen Angström, so dass die Daten auf atomarer Ebene interpretiert werden können. Ist das Makromolekül nicht kristalisierbar, so ist die Kernspinresonanzspektroskopie eine alternative Hochdurchsatzmethode. Durch sie kann die Struktur von Proteinen bis zu einer Größe von $50kDa$ bestimmt werden. Durch die beiden Methoden können allerdings nur Proteine untersucht werden, die vorher aufgereinigt wurden.

Im Vergleich zu Photonen vermögen Elektronen aufgrund der kürzeren Wellenlänge die strukturellen Details von Makromolekülen abzubilden. Die Struktur größerer ($\geq 300kDa$) Proteinkomplexe (z.B. das 26S Proteasom) kann mittels der Kryoelektronenmikroskopie gelöst werden. Das schnelle Einfrieren (Vitrifizieren) des biologischen Materials erhält die Struktur während der Aufnahme. Proben werden im Transmissionselektronenmikroskop durchstrahlt und so zweidimensionale Projektionen aufgereinigter Komplexe aufgenommen. Wurden genügend verschieden orientierte Projektionen identischer Makromoleküle aufgenommen, so kann mit Hilfe der Einzelpartikelanalyse *Single Particle Analysis* (SPA) die dreidimensionale Struktur *in silico* rekonstruiert werden [Frank, 2002]. Die hierbei erreichbare Auflösung liegt im sub-Nanometer Bereich [Armache et al., 2010, Sinkovits und Baker, 2011]. Folglich können α-Helices oder β-Faltblätter des Komplexes visualisiert werden. In die durch die SPA bestimmten Dichteverteilungen können allerdings atomare Strukturen von bekannten Domä-

nen eingepasst werden, um so pseudo-atomare Modelle des Makromoleküls zu erhalten [Rossmann et al., 2005].

Die Interaktion makromolekularer Komplexe innerhalb von Zellen kann nur in ihrer natürlichen Umgebung analysiert werden. Für die Analyse von Proteinen mit den im letzten Abschnitt erwähnten Methoden müssen die Proben methodenspezifisch aufgereinigt werden, so dass die Makromoleküle nicht in ihrer natürlichen Umgebung abgebildet werden können. Somit scheiden die Röntgenkristallographie und die Kernspinresonanzspektroskopie für diese Aufgabe aus. Projektionen durch dünne Zellen können mit dem Kryoelektronenmikroskop aufgenommen werden. Mit diesen Daten ist es aber unmöglich, eine dreidimensionale Rekonstruktion der makromolekularen Struktur mittels der SPA zu ermitteln. Um dieses Hindernis zu Umgehen, werden mehrere Projektionen durch die Probe aufgenommen. Hierbei wird die Probe in unterschiedliche Positionen gekippt, so dass eine Kippserie von Projektionen entsteht. Dieser Aufnahmevorgang ähnelt der Computertomographie (z.B. der medizinischen Computertomographie) und basiert ebenfalls auf der Idee, aus zweidimensionalen Projektionen dreidimensionale Volumen *in silico* zu berechnen.

Tomographische Aufnahmen mit dem Kryoelektronenmikroskop (Kryoelektronentomographie (KET)) ermöglichen die dreidimensionale Visualisierung von Teilen intakter Zellen und folglich die Analyse interagierender Makromoleküle. Die hierbei erzielten Auflösungen sind typischerweise im Bereich von $5^{-1} - 10^{-1} nm^{-1}$ [Lucić et al., 2005], was für die örtliche Bestimmung von interagierenden Makromolekülen ausreichend sein kann, jedoch im wesentlichen nicht ausreicht, um neue Einblicke in die Quarternärstrukturen von aktiven Proteinkomplexen zu bekommen. Die Gesamtdosis von Elektronen darf bei biologischen Proben eine kritische Grenze (max. $100 e/\text{Å}^2$) nicht überschreiten, da sonst das Präparat beschädigt wird. Wegen der limitierten Gesamtdosis werden Strukturen der Makromoleküle von einem hohen Rauschen überdeckt. Um die Auflösung auf unter $3^{-1} nm^{-1}$ zu verbessern, bedarf es der Mittelung vieler identischer Markromoleküle in gleicher Orientierung [Hrabe und Förster, 2011]. Die Verbesserung der Aufnahmequalität und die Entwicklung robuster, computergestützter Analysemethoden hat den rasanten Fortschritt auf in diesem Feld in dem letzten Jahrzehnt erst ermöglicht.

Die computergestützte Analyse von KET-Tomogrammen. Der Arbeitsablauf, nachdem die Kippserie aufgenommen wurde, ist im Groben: (*i*) die Rekonstruktion des Tomogramms (*ii*) das Auffinden potentieller Makromoleküle durch *Template Matching* (*iii*) das Ausrichten aller gefundenen Makromoleküle in eine identische Orientierung, um ein gemeinsames Mittel zu bestimmen (*iv*) die Klassifikation heterogener Moleküle (Abb. 1.1). Das Auffinden von Makromolekülen sowie die Klassifikation können interaktiv durchgeführt werden, erfordern aber ein intensive Interaktion mit den Daten und ist Resultate

sind nicht notwendigerweise reproduzierbar. Je nach Anwendung müssen nicht alle einzelnen Schritte rechnergestützt ausgeführt werden, sondern können interaktiv durchgeführt werden. Eine interaktive Prozessierung der Tomogramme ist vor allem für das Auffinden von Makromolekülen sowie die Klassifikation von Partikeln möglich, allerdings nicht empfehlenswert, da aufgrund des hohen Rauschens Ergebnisse subjektiv sein können und deshalb nicht notwendigerweise reproduzierbar sind.

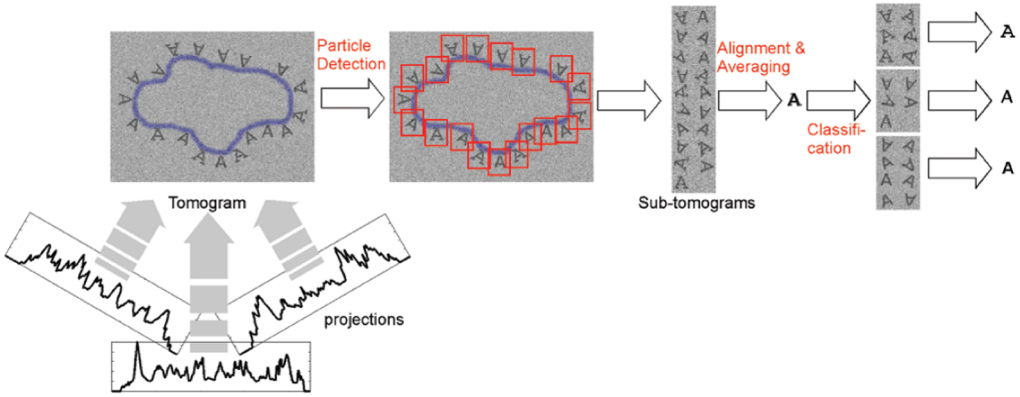

Abbildung 1.1: Die typischen Arbeitschritte in der Analyse von KET-Tomogrammen: die Tomogramme werden zuerst aus Projektionen rekonstruiert. Mittels *Template Matching* werden potentielle Makromoleküle (Subtomogramme) lokalisiert und rekonstruiert. Alle Partikel werden in einem weiteren Schritt in die gleiche Orientierung und Position ausgerichtet, um durch Mittelung das SNR zu verbessern. In letzter Instanz können die ausgerichteten Partikel klassifiziert werden, um strukturelle Unterschiede sichtbar zu machen.

Für diese Aufgaben wurden bereits Softwarepakete zur Verfügung gestellt, mit denen die Daten souverän verarbeitet werden können. Allerdings sind diese nicht notwendigerweise zueinander kompatibel, was für Benutzer und Entwickler ein unnötiges Hindernis darstellt. In vielen Forschungsgruppen haben sich Abläufe mittlerweile standardisiert. Diese sind aber von Labor zu Labor unterschiedlich. In der Abteilung für Molekulare Strukturbiologie das Max-Planck-Institutes für Biochemie werden für die Rekonstruktion die entwickelten Prozeduren aus der TOM (TOM)-Toolbox [Nickell et al., 2005] benutzt, die Lokalization läuft in MOLMATCH [Foerster et al., 2010], das Ausrichten und die Klassifikation in AV3 (AV3) [Foerster et al., 2005]. Ein großer Nachteil dieser hauptsächlich in Matlab implementierten Sammlung jedoch ist die schlechte Erweiterbarkeit und Wartung, da keine Versionsverwaltung wie Subversion (http://subversion.tigris.org) oder GiT (http://git-scm.com) benutzt wurde. Des Weiteren basieren TOM und AV3 auf nicht kostenlos zugänglicher Software. Trotzdem ist TOM und AV3 weit verbreitet und wird durchaus von Laboren für die Aufnahme und Verarbeitung von Tomogrammen sowie als Grundlage für die Entwicklung neuer Methoden benutzt [Xu et al., 2011,

Yu und Frangakis, 2011, Castaño Díez et al., 2012]. Natürlich wurden auch alternative Software-Lösungen in anderen Laboren entwickelt und verbreitet (XMIPP - [Sorzano et al., 2004], IMOD - [Kremer et al., 1996], BSOFT - [Heymann et al., 2008]). Allerdings bieten diese Pakete (bis auf TOM) nicht die Möglichkeit, den ganzen Arbeitsablauf in einer Umgebung durchzuführen, da Teilkomponenten nicht vorhanden sind. Basierend auf einer beliebigen Plattform werden heutzutage neue Methoden der Datenverarbeitung (z.B. [Heumann et al., 2011, Yu und Frangakis, 2011, Castaño Díez et al., 2012]) für die KET implementiert und veröffentlicht, ohne dass sie in einem Zusammenhang zum gesamten Arbeitsablauf präsentiert werden. Der Benutzer ist somit auf eine Kollaboration mit dem jeweiligen Labor angewiesen und kann nicht direkt von der veröffentlichten Methode Gebrauch machen.

Um mit etablierten Methoden wie der Röntgenkristallographie Schritt halten zu können, ist es deshalb essentiell die Datenverarbeitung möglichst kohärent zu strukturieren. In der Einzelpartikelanalyse sind aus diesem Grund verschiedene SPA-Softwaresammlungen zu einer zusammengefasst (Appion [Lander et al., 2009]). Basierend auf einer Datenbank sind alle Teilprozesse logisch verkettet, so dass ein Hochdurchsatzprogramm entsteht, mit dem mehr Daten schneller analysiert werden können [Moeller et al., 2012]. Der Benutzer hat zwar die Auswahl zwischen mehreren Methoden für eine Teilaufgabe, diese sind aber in ein geordnetes Protokoll integriert und ermöglichen so die einfache Benutzung.

Zielsetzung dieser Arbeit. Ziel dieser Arbeit ist es eine Software-Plattform für die KET zu entwickeln, in der die wichtigsten Schritte (Abb. 1.1) der Verarbeitung von Tomogrammen vereinheitlicht sind. Zum einen soll die Effizienz existierender Methoden verbessert zum anderen sollen neue Methoden für die Mittelung, das Ausrichten und die Klassifikation von Partikeln in die Plattform integriert werden. Numerisch robustere Interpolationsmethoden sollen die bisher in TOM und AV3 verwendete lineare Interpolation ablösen und somit eine akkuratere Bildverarbeitung auf der numerischen Ebene ermöglichen. Erweiterungen und neue Optimierungsmethoden sollen das Ausrichten sowie die Klassifikation von einzelnen Partikeln (Subtomogrammen) im Vergleich zu anderen Softwarepaketen ebenfalls genauer machen. Die bis dato implementierten Ausrichtungs- und Klassifikationsmethoden basierten alle auf einem deterministischen Optimierungsschema. Ein Verbesserungspunkt ist deshalb die Integration von stochastischen Methoden in den Optimierungsprozess, wie es zum Beispiel für die Vorhersage von Proteinstrukturen benutzt wird [Simons et al., 1997]. Außerdem soll die Plattform kostenlos zugänglich sein (*open source*). Benutzer der Plattform sollen einen möglichst einfachen Zugang zu diesen Methoden bekommen und schnell ihre Daten analysieren können. Die Verwendung einer einfachen Programmiersprache soll es dem Laien ermöglichen, individuelle Änderungen

an Methoden vorzunehmen. Trotzdem sollen durch Versionsverwaltung und Modultests Defekte erkannt werden, damit die Funktionalität der Software bewahrt bleibt. Darüberhinaus sind Kompatibilität und Transparenz weitere Anforderungen an die Plattform, um den Benutzer die Möglichkeit zu geben, externe Programme möglichst einfach in den Ablauf zu integrieren.

Biologische Applikationen. Parallel zur Entwicklung der Software-Plattform wurden zwei Projekte gestartet, die sich mit der Translation und Translokation von Proteinen am rauen Endoplasmatisches Retikulum (ER) beschäftigten [Mangold, 2010, Pfeffer, 2010](Abb. 1.2). Korrekt gefaltete Proteinketten im ER werden, abhängig von ihrer Aufgabe, in Vesikeln zum Golgi-Apparat und von dort aus zu ihrem Bestimmungsort transportiert [Johnson und van Waes, 1999].

Für die strukturelle Analyse der Translation und Translokation von Proteinen in das raue ER sind Mikrosomen gute Modellsysteme. Für die Mikrosomen spricht, dass sie *in vitro* translations- und translokationskompetente Fragmente des ER sind. Das Lumen und die membrangebundenen Ribosomen bleiben während der Aufreinigung intakt. KET basierende Untersuchungen an diesen ER-Vesikeln können demnach auf dünneren Proben der Mikrosomen anstatt auf dicken Proben ganzer Zellen ablaufen. Das verbessert die Qualität der Tomogramme signifikant.

Die Grundlage für die computergestützte Analyse des Translations- und Translokationsprozesses am ER waren Komponenten der in dieser Arbeit entwickelten Software-Plattform. Implementierte Methoden wurden basierend auf diesen experimentellen Daten getestet und waren wiederum das nötige Werkzeug für die weiteren Analyseschritte.

Gliederung dieser Arbeit

Kapitel 2. In diesem Kapitel werden zunächst die Grundlagen der Bildaufnahme im Elektronenmikroskop kurz erörtert. Des Weiteren werden die Rekonstruktion von Tomogrammen sowie die für deren digitale Verarbeitung benötigten Methoden eingeführt. Im Fokus stehen Methoden für die Ausrichtung und Klassifikation von Subtomogrammen.

Kapitel 3. Die Präparation der jeweiligen Tomogramme wird beschrieben. Außerdem werden Verbesserungen und neue Methoden für die Ausrichtung und Klassifikation von Subtomogrammen vorgestellt.

Kapitel 4. In diesem Kapitel wird die Implementierung der entwickelten Plattform PyTom erläutert. Es werden verbesserte numerische Methoden dargestellt und die Laufzeit-

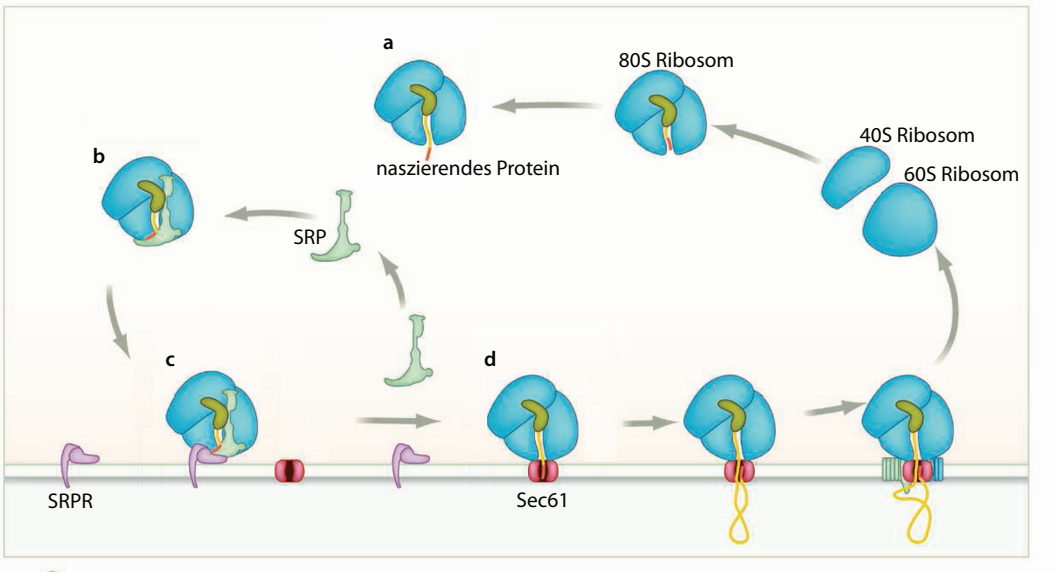

Abbildung 1.2: Eine schematische Darstellung der Proteintranslokation in das Lumen des rauen ERs, entnommen aus [Kampmann und Blobel, 2009]: (a) Das eukaryotische Ribosom mit einem naszierenden Protein. (b) Das *Signal Recognition Particle* (SRP) erkennt die Signalsequenz am N-Terminus des naszierenden Proteins (Rot) und bindet an diese. (c) Der membrangebundene *Signal Recognition Particle Receptor* (SRPR) erkennt das SRP. (d) Das translatierende Ribosom dockt an den Translokationskanal (Sec61) an und das Polypeptid wird in das Lumen translokalisiert.

komplexität der implementierten Algorithmen theoretisch analysiert und anhand gemessener Laufzeiten verifiziert.

Kapitel 5. Die implementierten Algorithmen werden einzeln an simulierten und experimentellen Daten getestet. Des Weiteren wird das Zusammenspiel der in PyTom implementierten Algorithmen an zwei experimentellen Datensätzen präsentiert.

Kapitel 6. Die bestimmten Ergebnisse werden diskutiert und geben einen Ausblick auf die Verwendung von PyTom und mögliche Erweiterungen.

2 Grundlagen

2.1 Kryoelektronentomographie

Der Aufbau eines Transmissionelektronenmikroskops ist mit dem eines klassischen Lichtmikroskops vergleichbar, allerdings werden andere Komponenten verwendet, um die Vergrösserung der Probe zu erreichen. Anstelle von Licht (Photonen) werden Elektronen, die mit der Probe interagieren, als Informationsträger benutzt. Anders als im Rasterelektronenmikroskop durchstrahlen die Elektronen die Probe, werden von elektromagnetischen Linsen ausgerichtet und in Photonen umgewandelt, welche schließlich durch eine CCD Kamera aufgenommen werden. Dieses Kapitel soll eine Übersicht über die Vorgänge vor, während und nach der Bildentstehung schaffen.

2.1.1 Probenpräparation für die Transmissionselektronenmikroskopie

Damit sich biologische Proben während der tomographischen Kippung innerhalb des Transmissionelektronenmikroskops nicht verschieben oder anderweitig verformen, müssen diese vorher in einen festen Zustand gebracht werden. Eine wässrige Probe würde ohne diese Fixierung auf dem Probenhalter unmittelbar in das Hochvakuum des Elektronenmikroskops evaporieren. Eine der gängigsten Methoden ist es, die Probe zu vitrifizieren [Dubochet et al., 1988], indem das biologische Material auf einem löchrigen *lacey* Kohlefilm in flüssigem Ethan (ca. $-190°C$) schockgefroren wird. Vor dem Vitrifizieren wird überschüssige Flüssigkeit mittels eines Filterpapiers entzogen, um eine Probendicke von weniger als $1\mu m$ zu erreichen.Während der Abkühlung erstarrt das Wasser der Probe in einen gläsernen Zustand, da Wasser bei extrem schneller Abkühlung ($10^4 K/s$) auf unter $140°K$ keine Eiskristalle formt. Eiskristalle könnten Druck auf die Probe ausüben, und somit strukturelle Änderungen innerhalb des Präparats hervorrufen.

Dosisbeschränkung für biologische Proben. Eingefrorene, biologische Proben sind extrem sensibel im Bezug auf die Elektronendosis, der sie im Elektronenmikroskop ausgesetzt werden. Da in der KET Projektionen des selben Probenauschnitts wiederholt aufgenommen werden, wird die tolerierbare Gesamtdosis auf alle Projektionen aufgeteilt. Wird die tolerierbare Dosis überschritten, so können entstehende Radikale die Probe beschädigen und unbrauchbar machen. Die Gesamtdosis ist demnach abhängig von der Zielset-

zung, unter der die Probe untersucht werden soll. Möchte man eine möglichst detaillierte makromolekulare Struktur z.B. durch Kryoelektronenmikroskopie (KEM) Einzelpartikelanalyse erreichen, so sollte die Dosis nicht höher sein als $10e/Å^2$ [Henderson, 1995]. Ist man allerdings an der Bestimmung von Interaktionen makromolekularer Komplexe durch die KET interessiert, so darf die Gesamtdosis $100e/Å^2$ nicht überschreiten [Foerster und Hegerl, 2006]. Des Weiteren beeinflusst die Dosis das SNR in den Projektionen und somit auch die Qualität der rekonstruierten Tomogramme.

Kolloides Gold als Markerpunkte für die Rekonstruktion. Kontraststarke Markerpunkte werden für die korrekte Ausrichtung der Projektionen vor der Rekonstruktion benötigt, damit Verschiebungen und Rotationen, welche während der Aufnahme entstehen, vor der Rekonstruktion ausgeglichen werden können [Mastronarde, 2006]. Vor der Vitrifikation werden deshalb kolloide Goldpartikel mit einem Durchmesser von ca. $10nm$ dem Präparat beigemischt [Zsigmondy und Thiessen, 1925].

2.1.2 Bildentstehung im Elektronenmikroskop

Beschleunigung der Elektronen durch die *Field Emission Gun*. Als Elektronenquelle in der KEM und folglich auch KET haben sich die Feldemissions-Kathoden (FEG) als Standard durchgesetzt, da sie gegenüber Glühkathoden bessere Strahleigenschaften haben. Die durch FEGs generierten Elektronenstrahlen sind räumlich und temporär kohärenter und können demnach höhere Auflösungen produzieren. Nach dem Austritt aus der Kathode wird der Primärstrahl durch ein Anodensystem beschleunigt.

Elastische- und inelastische Streuung. Durch die Wechselwirkung des Elektronenstrahls mit dem Präparat nimmt der Elektronenstrahl die strukturelle Information der Probe auf. Die Elektronen werden an den Atomen der Probe gestreut und die weitere Wechselwirkung der abgelenkten Elektronen mit dem Primärstrahl entspricht der aufgenommenen Information. Man unterscheidet zwischen zwei Streumechanismen, (i) der elastischen und (ii) der inelastischen Streuung.

(i) Kein Energieverlust tritt bei der elastischen Streuung von Elektronen auf, die Bahn des jeweiligen Elektrons wird durch das elektrische Feld der Atomhülle verändert und äußert sich durch eine geringfügige Richtungsänderung. Folglich legen die abgelenkten Elektronen einen längeren Weg bis zur Objektivlinse als der Primärstrahl zurück. Betrachtet man das Elektron als eine Welle, so impliziert die veränderte Trajektorie eine Phasenverschiebung des Elektrons.

(ii) Bei der inelastischen Streuung verliert das Elektron bei der Wechselwirkung mit Atomen der Probe Energie. Vitrifizierte, biologische Präparate nehmen hierbei Schaden,

da Atome ionisiert werden, die Probe sich aufheizt und somit strukturelle Änderungen entstehen können.

Biologische Proben werden in der Regel als schwache Streuobjekte (*weak phase objects*) bezeichnet, da sie aus Elementen mit niedrigen Ordnungszahlen (z.B. H,O,N und C) bestehen. Aus diesem Grund werden Elektronen von dünnen (ca. 50 - 100nm), vitrifizierten Proben hauptsächlich elastisch gestreut (ca. 90%), ungefähr 10% der Streuvorgänge sind inelastische Wechselwirkungen [Orlova und Saibil, 2011].

Phasenkontrast. Phasenkontrast entsteht, wenn elastisch gestreute Elektronen mit dem Primärstrahl interferieren. Da der elastische Streuprozess die Interaktion der Elektronen mit der Probe dominiert, muss während der Bildaufnahme der Phasenkontrast gemessen werden. Dieser ist allerdings kaum messbar, da die Phasenverschiebung durch biologische Proben minimal ist. Um den Phasenkontrast messbar zu machen, werden die Imperfektionen (Aberrationen) der verwendeten elektromagnetischen Linsen ausgenutzt, so dass die Phasenverschiebung der elastisch gestreuten Elektronen verstärkt wird. Die verstärkte Interferenz ist wiederum über die Amplitude messbar.

Kontrasttransfer im Elektronenmikroskop. Der Kontrasttransfer im Elektronenmikroskop wird durch die Kontrasttransferfunktion (KTF) beschrieben und moduliert die Amplitude und das Vorzeichen des fouriertransformierten Elektronenstrahls an der Ortsfrequenz u. Charakteristische Parameter der KTF sind der Defokus Δz, die sphärische Aberrationskonstante der Objektivlinse C_S und die Wellenlänge λ der Elektronen.

$$KTF(u) = sin(\frac{\pi}{2}(C_S\lambda^3u^4 - 2\Delta z\lambda u^2)) \qquad . \tag{2.1}$$

Frequenzbereiche mit positiven Funktionswerten der KTF weisen einen negativen Phasenkontrast auf (Dichte erscheint hell), Frequenzbereiche mit negativen Funktionswerten der KTF weisen entsprechend positiven Phasenkontrast auf (Dichte erscheint dunkel). An den Nullstellen der KTF wird keine Information übertragen. Ein wichtiger Parameter der KTF ist der Defokus Δz. Werden Projektionen bei betragsmäßig hohen Defokuswerten ($\Delta z \geq -8\mu m$) aufgenommen, so ist der Kontrast der tiefen Frequenzen besonders stark. Jedoch limitiert diese Einstellung die theoretisch erreichbare Auflösung, da die erste Nullstelle der KTF früh erreicht wird. Bei niedrigen Defokuswerten ($\Delta z < -8\mu m$) verbessert sich der Übertrag der hohen Frequenzen, da der Abstand zur ersten Nullstelle größer wird. Hingegen wird der höhere Übertrag mit einem niedrigeren Kontrast erkauft, so dass die Projektionen für den Betrachter verrauschter erscheinen.

Die *Point Spread*-Funktion. Die *Point Spread*-Funktion (PSF) beschreibt im Folgenden den Effekt der KTF und einer einhüllenden Funktion (ENV) auf das entstehende Bild im Realraum. Die ENV beschreibt den dämpfenden Einfluss von Strahl-Inkohärenz und Linsenaberrationen auf hohe Frequenzen. Sei Pr eine Projektion mit einem abgebildeten Dirac-Stoß an der Position $1, 1$.

$$PSF(Pr) = \mathcal{F}^{-1}(\mathcal{F}(Pr) \cdot KTF \cdot ENV) \qquad . \tag{2.2}$$

Durch die KTF-Oszillationen und Nullstellen, ebenso wie die durch die ENV beschriebene Dämpfung der hohen Frequenzen, verschwimmt die scharfe Kante des Dirac-Stoßes. Analog hierzu verschwimmen Details in Projektionen biologischer Proben.

Aufnahme der Projektion. Letztendlich entsteht das digitale Bild, indem der Elektronenstrahl durch einen Szintilator vor der CCD-Kamera in Photonen übersetzt wird, welche in Glasfasern auf den CCD-Chip geleitet und durch Halbleiter ausgelesen werden [De Ruijter, 1995].

2.2 Aufnahme und Rekonstruktion von Tomogrammen

Das Projektions-Schnitt-Theorem ist die Grundlage für die tomographische Rekonstruktion von Volumen aus zweidimensionalen Projektionen wie sie im Elektronenmikroskop aufgenommen werden. Aus diesem Grund wird in diesem Kapitel das Projektions-Schnitt-Theorem sowie die gewichtete Rückprojektion (*Weighted Backprojection*) (WB) motiviert, da die in dieser Arbeit verwendeten Tomogramme ausschließlich mit der WB rekonstruiert wurden.

Die limitierte Kippung des Probenhalters. Im Gegensatz zur vollständigen Winkelabtastung in der medizinischen Computertomographie, ist die Winkelabtastung in der KET typischerweise auf das Intervall zwischen $[-70°, 70°]$ beschränkt. Grund für diese Limitierung ist der Probenhalter, bei Kippwinkeln von mehr als $70°$ in beide Richtungen wird die biologische Probe durch den Probenhalter verdeckt. Durch diese Einschränkung fehlen Projektionen für die vollständige Rekonstruktion jedes Tomogramms und führen zu Elongationen des rekonstruierten Objekts entlang der Strahlachse und folglich zu einer anisotropen Auflösung. Der fehlende Frequenzbereich selbst ist keilförmig und wird in der Literatur als *Missing Wedge* (W) bezeichnet (Abb. 2.1).

Die PSF (Formel 2.2) kann durch den *Missing Wedge* zu

$$PSF(V) = \mathcal{F}^{-1}(\mathcal{F}(V) \cdot W \cdot KTF \cdot ENV) \tag{2.3}$$

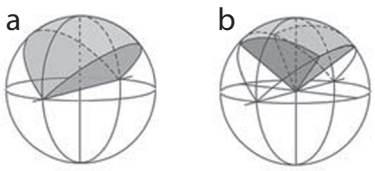

Abbildung 2.1: (a) Die Form des fehlenden Bereiches aufgrund der limitierten Winkellabtastung zu maximal $[-70°, 70°]$ entspricht einem Keil. (b) Der fehlende Bereich kann auf eine Pyramide reduziert werden, sofern die nötige Drehmechanik im Probenhalter vorhanden ist. (Abbildung aus [Lucić et al., 2005])

erweitert werden.
Um den fehlenden Bereich im Fourieraum zu minimieren, kann nach einer 90° Drehung der Probe im Probenhalter eine Doppelkippserie aufgenommen werden, so dass folglich der fehlende Keil auf eine fehlende Pyramide reduziert wird [Nickell et al., 2003]. Diese Strategie ist durch die Gesamtdosis limitiert und setzt außerdem die benötigte Drehmechanik im Probenhalter voraus.

Das Projektions-Schnitt-Theorem. Die Grundlage aller Rekonstruktionsmethoden ist das Projektions-Schnitt-Theorem, welches besagt, dass die Fouriertransformierte einer $n-1$ dimensionalen Projektion dem zentralen Schnitt durch die Fouriertransformierte des originalen, n dimensionalen Objekts entspricht [Radon, 1917]. Kippt man das Objekt um den Kippwinkel γ, so kann das abgebildete Objekt bei ausreichender Abtastung aus den Projektionen vollständig rekonstruiert werden.
Für den diskreten, zweidimensionalen Fall eines Objektes $o(x, y)$ ist die eindimensionale Projektion $p(x)$ entlang der y-Achse definiert als

$$p_0(x) = \sum_{y=0}^{N_Y} o(x, y) \qquad . \tag{2.4}$$

Werden Projektionen von o unter variablen Kippwinkeln γ aufgenommen, so muss 2.4 durch die entsprechende Kippung erweitert werden.

$$p_\gamma(x) = \sum_{y=0}^{N_Y} o(\mathcal{M}_\gamma \begin{pmatrix} x \\ y \end{pmatrix}) \tag{2.5}$$

$\mathcal{M}_\gamma$ ist Rotationsmatrix der Kippung um den Winkel γ, durch welche die x- und y-Koordinaten in die gekippte Position transformiert werden. Im Regelfall sind die neuen Koordinaten nicht mehr diskret und der Wert an dieser kontinuierlichen Stelle wird

durch Interpolation (Kap. 4.1.2) bestimmt. $P_\gamma(u)$ sei die Diskret-Fouriertransformierte von Projektion p_γ, u ist die entsprechende Ortsfrequenz

$$P_\gamma(u) = \sum_{x=0}^{N_X-1} p_\gamma(x) e^{-2\pi i \frac{xu}{N_X}} \quad . \tag{2.6}$$

Die Diskret-Fouriertransformierte O vom abgebildeten Objekt o ist definiert als

$$O(u,v) = \sum_{x=0}^{N_X-1} \sum_{y=0}^{N_Y-1} o(x,y) e^{-2\pi i \left(\frac{xu}{N_X} + \frac{yv}{N_Y}\right)} \quad . \tag{2.7}$$

Betrachtet man den zentralen Schnitt $O(u,0)$ für $\gamma = 0$ durch O, so ist

$$O(u,0) = \sum_{x=0}^{N_X-1} \left(\sum_{y=0}^{N_Y-1} o(x,y)dy \right) e^{-2\pi i \frac{xu}{N_X}} = \sum_{x=0}^{N_X-1} p_0(x) \; e^{-2\pi i \frac{xu}{N_X}} = P_0(u) \quad . \tag{2.8}$$

2.8 gilt ebenfalls auch für alle gekippten Projektionen, da Koordinatendrehungen der Form $\mathcal{M}_\gamma \begin{pmatrix} x \\ y \end{pmatrix}$ direkt in den Fourierraum übertragbar sind.

Gewichtete Rückprojektion Die WB ist eine oft benutzte Rekonstruktionsmethode, die Bestandteil vieler KET-Softwarepakete (z.B. Spider, TOM, XMIPP) ist und dadurch in vielen Projekten zum Einsatz kommt. Die Berechnung der WB findet sowohl im Real-, wie auch im Fourierraum statt. Aus diesem Grund kann man die WB in zwei Hauptschritte aufteilen: (i) die Gewichtung im Fourierraum und (ii) die Rückprojektion im Realraum.

Gewichtung im Fourierraum. Eine Gewichtung der einzelnen Projektionen ist notwendig, da sonst im rekonstruierten Tomogramm die tiefen Frequenzen überrepräsentiert sind. Ursache hierfür ist die Dicke der Probe. Im tieffrequenten Bereich überlappen sich die Frequenzen der Projektionen während der Rekonstruktion. Abhängig von der Dicke existiert eine Frequenz (Crowther Kriterium [Crowther et al., 1970]), ab der sich tiefe Frequenzen nicht mehr überlappen. Diese Frequenz bestimmt auch die maximale Auflösung. Mittels einem Rampenfilter R werden die sich überlagernden Frequenzen im Fourierraum normiert, damit das ganze Frequenzspektrum gleich gewichtet ist (analytische Gewichtung). Frequenzen außerhalb der möglichen Abtastung werden zu Null gesetzt.

$$R(x,y) = \begin{cases} \frac{r}{ny_{0.5}} & ; \; \frac{r}{ny_{0.5}} \leq 1 \\ 0 & ; \; \frac{r}{ny_{0.5}} > 1 \end{cases} \qquad Pr_{j,w} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(Pr_j)}{R}\right) \tag{2.9}$$

r ist der Abstand zum Ursprung entlang der x-Achse ($r = \sqrt{(x - center_X)^2}$). Die in 2.9 beschriebene Gewichtungsfunktion ist nur auf Projektionen aus einer Kippserie um die y-Achse anwendbar, da die Gewichtung senkrecht zur Kippachse berechnet werden muss [Radermacher, 2006].

Rückprojektion im Realraum. Die einzelnen Voxelwerte im rekonstruierten Objekt werden bei der Rückprojektion berechnet, indem Pixelwerte an den entsprechenden Positionen in den gewichteten Projektionen summiert werden. Hierbei werden die Koordinaten im rekonstruierten Objekt analog zu der Projektion so transformiert, dass die Pixelwerte in den Projektionen ausgelesen werden können.

$$o(x, y) = \sum_{\gamma} p_\gamma(\mathcal{M}_\gamma \begin{pmatrix} x \\ y \end{pmatrix}) \tag{2.10}$$

2.3 Mittelung und Ausrichtung von Subtomogrammen

Aufgrund des extrem hohen Rauschens in Subtomogrammen können strukturelle Details der abgebildeten Makromoleküle nur durch ein gemeinsames Mittel n identischer Makromoleküle zum Vorschein kommen. Diese Methode wurde, analog zur Einzelpartikel-Analyse in KEM, als erstes von [Knauer et al., 1983] und [Oettl et al., 1983] für die Analyse ribosomaler Untereinheiten mit Hilfe von KET angewandt. Seitdem ist dieses Verfahren ein elementarer Bestandteil der Analyse von Proteinkomplexen mittels KET.

2.3.1 Mittelung von Subtomogrammen

Unter der Annahme, dass jedes Subtomogramm P_i aus Signal S und additivem Rauschen N_i besteht, gilt

$$P_i = S + N_i \qquad . \tag{2.11}$$

Um ein möglichst hoch aufgelöstes Mittel (im folgendem *Average* genannt) der Subtomogramme zu erhalten, muss jedes Subtomogramm P_i individuell rotiert und translatiert werden, damit alle Dichten identisch ausgerichtet (im folgendem *aligniert* genannt) sind. Nur so entsteht im *Average* der transformierten Subtomogramme eine Dichte A, mit einem maximierten SNR, in dem Details in Erscheinung treten. Sei ρ_i die gesuchte Rotation und ν_i die gesuchte Translation von P_i. Beide Variablen werden hier vereinfacht durch eine Variable $\theta_i = (\rho_i, \nu_i)$ repräsentiert. Der resultierende Average der ausgerichteten Partikel ist

$$\tilde{A} = \sum_{i}^{n} \mathcal{T}(P_i, \theta_i) \qquad . \tag{2.12}$$

wobei $\mathcal{T}$ dem Transformationsoperator entspricht. Formel 2.12 ist allerdings noch eine extrem vereinfachte Darstellung des *Averages*. Bedingt durch den *Missing Wedge* fehlt in 2.12 die Gewichtung der spektralen Werte. Der spektrale Abtastungsbereich W jedes Subtomogrammes entspricht dem des rekonstruierten Tomogramms, in dem es gefunden wurde. Folglich haben alle Subtomogramme aus dem selben Tomogramm die identische Gewichtung.

$$A = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\tilde{A})}{\sum_{i}^{n}\mathcal{T}(W_i,\theta_i)}\right) \tag{2.13}$$

Formel 2.13 erlaubt eine exakte Gewichtung jeder Ortsfrequenz. $\mathcal{F}$ ist hier die diskrete Fouriertransformation. Nach der Mittelung aller Subtmogramme steigt somit das SNR, da das konstante Signal S im Gegensatz zum variablen Rauschen N_i verstärkt wird [Hrabe und Förster, 2011].

$$SNR(A) = \frac{\langle S_i^2\rangle}{\langle N_i^2\rangle} = \frac{n^2\cdot S^2}{n\cdot N^2} = n\cdot SNR_i \tag{2.14}$$

Die Bestimmung von A erscheint trivial, ist es aber nur unter der Voraussetzung, dass θ für jedes Partikel bekannt ist. Typischerweise ist diese Information *a priori* nicht oder nur zum Teil, bekannt und wird durch einen komplexen Prozess angenähert. Diese Optimierung besteht in der Regel aus zwei Komponenten: (i) eine Funktion, um die Transformation θ eines Partikels zu bewerten und (ii) eine effiziente Methode, um alle möglichen Transformationen aller Subtomogramme abzutasten.

2.3.2 *Scoring* - Bewertungsmethode für das Alignment

Die Bewertungsfunktion, die die Ähnlichkeit von zwei diskreten Funktionen (Subtomogrammen) numerisch abbildet, wird im folgenden *Score* ($\mathcal{S}$) genannt. Die wohl einfachste Methode die Ähnlichkeit zweier Subtomogramme zu messen besteht darin, deren euklidische Distanz d zu bestimmen. Hierfür summiert man die Differenz jeweils entsprechender Voxel der beiden Subtomogramme

$$d_{AP_i} = \sqrt{\sum_{x,y,z}(A(x,y,z) - P_i(x,y,z))^2} \quad . \tag{2.15}$$

Wird 2.15 ausmultipliziert, so ergibt sich nach der Binomialformel

$$d^2_{AP_i} = \sum_{x,y,z}(A^2(x,y,z) - 2A(x,y,z)P_i(x,y,z) + P^2(x,y,z)) \quad . \tag{2.16}$$

Kreuz-Korrelation. Als einziger variabler Term in 2.16 bestimmt $2A(x,y,z)P_i(x,y,z)$ die Ähnlichkeit der beiden Dichten. Bei der Anwendung dieser Formel ergibt sich außerdem das Problem, dass beide Dichten nicht notwendigerweise in der selben Position abgebildet werden. Deswegen wird 2.16 um die drei Translationen $\nu = (\nu_x, \nu_y, \nu_z)$ erweitert und der Korrelationsterm isoliert

$$xc_\nu(A, P_i) = \sum_{x,y,z} A(x-\nu_x, y-\nu_y, z-\nu_z)P_i(x,y,z) \tag{2.17}$$

Dieser wird auch als Kreuz-Korrelation (XC) bezeichnet. In Formel 2.17 werden die diskreten Funktionen A und P_i gefaltet, was durch das Faltungstheorem auch im Fourierraum effizient $xc(A,P_i) = \mathcal{F}^{-1}(\mathcal{F}(P_i)\cdot\mathcal{F}(A)^*)$ berechnet werden kann. * symbolisiert die Konjugation komplexer Zahlen. Die Kreuzkorrelation ist im Prinzip ein Filter mit einem beliebigem Muster [Kumar et al., 2006].

Normierte Kreuz-Korrelation. Formel 2.17 bildet zwar die Ähnlichkeit von A und P_i numerisch ab, allerdings nicht auf ein vordefiniertes Intervall, sondern auf alle Zahlen in $\mathbb{R}$. Um sicherzustellen, dass xc_{AP_i} auf ein vordefiniertes Intervall abbildet, teilt man die XC durch die Standartabweichungen der beiden Dichten A, P_i, die vor der Berechnung der XC mittelwertfrei $(A-\bar{A})$ gemacht werden müssen.

$$nxc_\nu(A,P_i) = \frac{xc_\nu(A-\bar{A}, P_i-\bar{P_i})}{\sigma_A\sigma_{P_i}} \quad . \tag{2.18}$$

Für die so normierte Kreuz-Korrelation (NXC) gilt $nxc_\nu(A,P_i) \to [-1;1]$, wobei 1 die Identität (Autokorrelation), 0 die Unähnlichkeit und -1 die inverse Identität indiziert.

Lokal Normierte Kreuz-Korrelation. Mit Hilfe einer Maske (M) kann man die NXC auf einen beliebigen Teilbereich in der Dichte beschränken. Dies ist vor allem dann von Vorteil, wenn man bestimmte Teilkomponenten zweier Dichten korrelieren will. Außerdem verbessert dieser Ansatz das Auffinden von Makromolekülen in großen Tomogrammen, da der Korrelationsbereich auf die Größe des gesuchten Makromoleküls beschränkt wird. [Roseman, 2003] schlug eine sehr effiziente Berechnung der lokalen Normierung im Fourierraum vor, auf deren Basis die lokal normierte Kreuz-Korrelation (LNXC) hier auch implementiert wurde. Darüberhinaus wurde in [Frangakis et al., 2002] die LNXC an das in der KET vorherrschende *Missing Wedge*-Problem W_i angepasst.

$$lnxc_v(A,P_i) = \frac{xc_\nu(((A-\bar{A})\otimes PSF(W_i))\cdot M, (P_i-\bar{P_i})\cdot M)}{\sqrt{\sum_{x,y,z}((A-\bar{A})\otimes\mathcal{F}^{-1}(W_i)\cdot M)^2}\sqrt{\sum_{x,y,z}((P_i-\bar{P_i})\cdot M)^2}} \quad . \tag{2.19}$$

Die *Constrained Correlation* Die bereits beschriebenen Methoden sind für das Auffinden (*Localization*) von Makromolekülen in Tomogrammen und für deren Ausrichtung (im folgendem *Alignment* genannt) geeignet. Sollen jedoch bereits alignierte Subtomogramme paarweise verglichen werden, so sind die fehlenden Bereiche W_i und W_j im Fourierraum nicht notwendigerweise gleich ausgerichtet. Die Berechnung des Korrelationskoeffizienten kann deshalb nur im Bereich der Fourierkoeffizienten erfolgen, die in beiden Subtomogrammen auch beobachtet wurden (*Constrained Correlation* (CC) [Foerster et al., 2008]). Deshalb entspricht der, für die Berechnung der Korrelation zulässige, Bereich dem Produkt aus W_i und W_j. Um sicherzustellen, dass das Signal nur in dem sich überschneidenden Fourierbereich korreliert wird, werden beide Subtomogramme P_i und P_j mit jeweils W_i und W_j des anderen Subtomogrammes gefiltert.

$$cc(P_i, P_j) = nxc(\mathcal{T}(P_i \otimes PSF(W_j), \theta_i), \mathcal{T}(P_j \otimes PSF(W_i)), \theta_j)) \tag{2.20}$$

Fourier-Ring-Korrelation. Ein Spezialfall unter allen in der KET gängigen Korrelationsmethoden ist die Fourier-Ring-Korrelation (FRK) [Saxton und Baumeister, 1982], die üblicherweise nicht innerhalb eines Optimierungsprozesses zum Messen von Ähnlichkeiten genutzt wird, sondern um die erreichte Auflösung r_A eines Averages A zu bestimmen. Es handelt sich hierbei um eine Methode, welche die Konsistenz der Daten in Form ihrer Auflösung misst.

Um die Auflösung zu bestimmen, werden aus dem Datensatz die Mittel A_g und A_u gebildet (Formel 2.13). A_g wird aus allen P_i mit geradem Index bestimmt, A_u entspricht dem Mittel aus allen P_i mit ungeradem Index. Korreliert werden A_u und A_g in Bändern, deren Breite b frei wählbar ist. Es wird ein Korrelationskoeffizient für jedes Band errechnet:

$$fsc(A_g, A_u, b) = \frac{\sum\limits_{j \in b} \mathcal{F}(A_g)_j \cdot \mathcal{F}(A_u)_j^*}{\sqrt{\sum\limits_{j \in b} |\mathcal{F}(A_g)_j|^2 \cdot \sum\limits_{j \in b} |\mathcal{F}(A_u)_j|^2}} \quad . \tag{2.21}$$

So entsteht ein Vektor von Korrelationskoeffizienten, der nach dem Bandindex aufsteigend sortiert ist. Die Auflösung r_A wird in dem Band bestimmt, in dem der Koeffizient zum ersten Mal unter ein vordefiniertes Konsistenzkriterium r_{Cutoff} fällt:

$$r_{r_{Cutoff}}(A) = \arg_b(fsc(A_g, A_u, b) \leq r_{Cutoff}) \quad . \tag{2.22}$$

Der Verlauf der FRK-Kurve zeigt auf, bis zu welcher Auflösung der Datensatz das vorgegebene Konsistenzkriterium einhält (Abb. 2.2). Außerdem kann man dem Verlauf der FRK-Kurve entnehmen, ob Signal oder Rauschen die Optimierung der Datenparameter maßgeblich beeinflusst haben [Penczek, 2010].

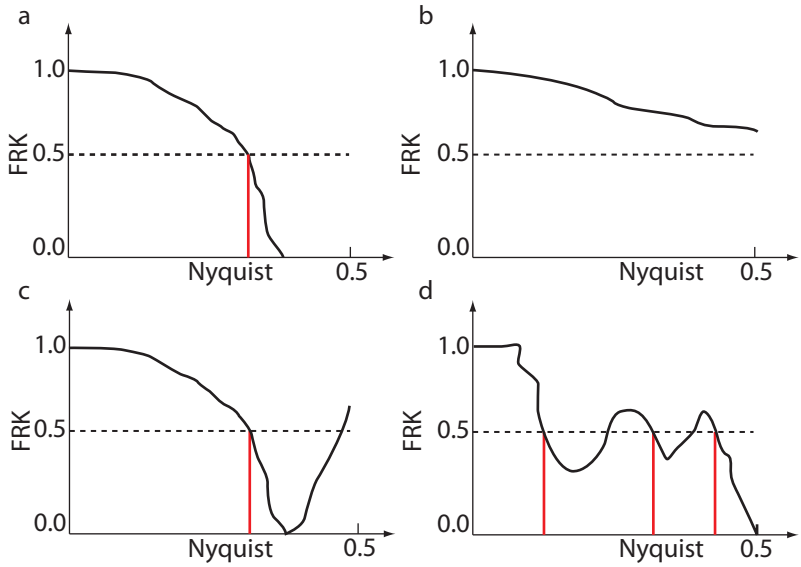

Abbildung 2.2: Vier mögliche FRK-Kurven ($r_{Cutoff} = 0.5$): (a) Eine FRK-Kurve durch die die Auflösung des Mittels eindeutig an der roten Linie bestimmt werden kann. Hier fällt ein Koeffizient zum ersten Mal unter das FRK-Kriterium r_{Cutoff}. (b) Kein FRK-Koeffizient fällt unter r_{Cutoff}, die Auflösung ist nicht bestimmbar. Dieser Verlauf indiziert ein vom Rauschen beeinflusstes Alignment (hochfrequentes Rauschen korreliert gut), oder dass die verwendete Pixelgröße zu groß war (*Undersampling*). (c) Der Anstieg im hochfrequentem Bereich zeigt an, dass während des Alignments falsche Parameter verwendet wurden (zu hohe Winkelinkremente, scharfe Maskierung), oder das der Prozess fehlerhaft ist und deshalb auf z.B. Rotationsartefakte aligniert wurde. (d) Daten mit verschiedenen Aufnahmeparametern wie Pixelgröße oder Defokus wurden vermischt und gemittelt. Die Auflösung kann nicht eindeutig bestimmt werden. (Abbildung adaptiert aus [Penczek, 2010].)

2.3.3 *Sampling* - Abtastung des Allignierungsraumes

Die Größe des Transformationsraumes Θ wächst mit der Feinheit der Suche nach der optimalen Rotation und Translation $\theta_{opt} \in \Theta$. Für eine effiziente Abtastung (im folgenden *Sampling* genannt) des Transformationsraumes werden drei Eulerwinkel[1] $\rho = (\rho_{z1}, \rho_x, \rho_{z2})$ und drei Translationen $\nu = (\nu_x, \nu_y, \nu_z)$ unabhängig auf diskreten Gittern abgetastet. Wählt man einen Winkelabstand von 10° für jeden der drei Eulerwinkel und beschränkt den Translationsraum auf 10 Pixel entlang jeder Dimension, so ergeben sich $36 \times 18 \times 36 \times 10 \times 10 \times 10 = 23.328.000$ mögliche Transformationen für jedes Subtomogramm, die es abzusuchen gilt. Zusätzlich wächst der Suchraum exponentiell mit der

[1] Die Spezifikation der drei Eulerwinkel entspricht dem ZXZ-Referenzsystem wie es in [Hegerl, 1996] definiert wurde.

Anzahl n der Partikel, da $\theta_{opt,i}$ für jedes Subtomogramm unabhängig ist.

Da die Scoringfunktion im Fourierraum berechnet wird, reduziert sich die vollständige Abtastung aller Translationen auf eine einzelne Faltung im Fourierraum [Bracewell, 2000]. Die Anzahl an Transformationen im obigen Beispiel wird um den Faktor 1000 auf die Anzahl der Winkel reduziert. Folglich müssen alle Winkel explizit abgesucht werden.

Diskretes Winkelgitter in SO(3). Das diskrete Gitter für das Sampling der Winkel wird im Eulerraum selbst bestimmt, indem jeder Eulerwinkel um ein vordefiniertes Winkelinkrement ($\Delta\alpha$) erhöht wird [Foerster et al., 2005, Winkler et al., 2009]. Die Iterationen der einzelnen Winkel sind unabhängig, werden aber geschachtelt. Diese Intervalle werden für die globale Suche nach ρ benutzt. Man kann diese Intervalle allerdings auch eingrenzen, um die Umgebung einer Rotation verstärkt abzusuchen (lokale Suche). Beim Sampling mit dieser Methode ergeben sich allerdings mehrere Probleme, die sich nachteilig auf das Sampling-Verhalten auswirken.

(i) Beschreibt ρ_x eine polare Position, indem es den Wert 0 annimmt, dann addieren sich ρ_{z1} und ρ_{z2} zu einer Rotation. Somit ist die Darstellung von Rotationen mittels Eulerwinkel eine surjektive Funktion, da unterschiedliche Belegungen der Eulerwinkel eine identische Rotation darstellen können.

(ii) Bestimmt man das diskrete Gitter für die Rotationswinkel nach der oben genannten Methode, so kann man eine Überabtastung an den Polen beobachten [Kuffner, 2004]. Folglich entsteht keine homogene Verteilung der abgetasteten Winkel und die Suche nach ρ_{max} wird auf die polaren Regionen fokussiert, wohingegen Winkel außerhalb der polaren Regionen unterrepräsentiert werden.

Diskretes Winkelgitter in SO(4). Eine weitere Alternative zum Sampling im Eulerraum ist durch Einheitsquarternionen auf einer Hyperkugel möglich [Kuffner, 2004]. Basierend auf Algorithmen aus der Computergrafik präsentiert [Stölken et al., 2010] eine Methode für ein homogenes Sampling von ρ. Punkte auf der Hyperkugel im $SO(4)$ Raum werden mittels eines simulierten Kraftfeldes solange verschoben, bis für alle der Abstand zu seinen nächsten Nachbarn $\Delta\alpha$ entspricht. Auf diese Weise erhält man eine Liste von Eulerwinkeln, die im Gegensatz zum Sampling in $SO(3)$ homogen verteilt sind und somit eine effiziente, globale Suche ermöglichen. Quantitativ entnimmt man aus Abbildung 2.3, dass für diese Methode bei gleichem Winkelinkrement weniger Winkel abgesucht werden als in $SO(3)$.

Lokal fokussiertes Winkelgitter. Eine effiziente Vorgehensweise für die lokale Suche, in der Regionen um einen festen Winkel abgesucht werden, wurde in [Foerster et al., 2005] vorgestellt. Hierfür wird zunächst ρ_{z1} vollständig abgesucht, dann werden Breiten- (ρ_x)

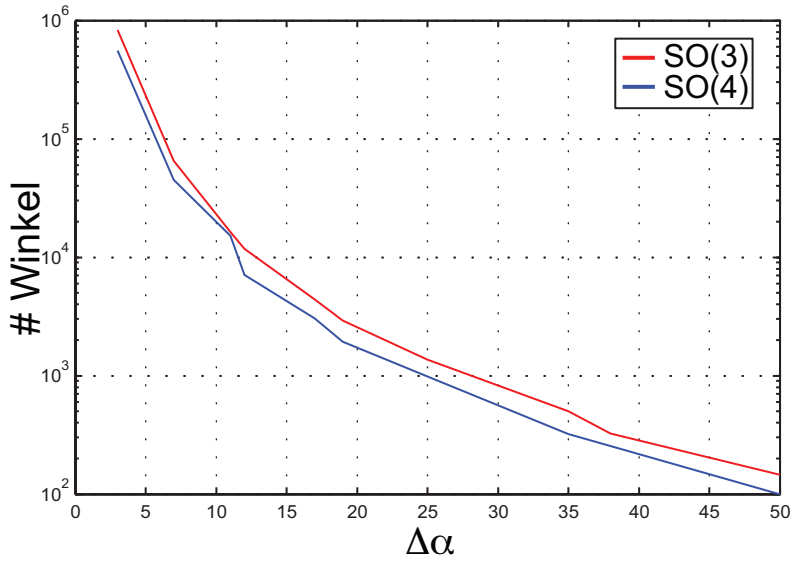

Abbildung 2.3: Die Anzahl der abgesuchten Winkel bestimmt durch: (Rot) das diskrete Winkelgitter im Eulerraum ($SO(3)$), (Blau) das diskrete Gitter im Quarternionenraum ($SO(4)$).

und Längengrad (ρ_{z2}) um $\Delta\alpha$ gekippt und es wird ein Ring (durch ρ_x und ρ_{z2} bestimmt) um den originalen Rotationsvektor abgesucht. In dieser neuen Position wird ρ_{z1} wieder vollständig abgesucht, dann werden ρ_x und ρ_{z2} modifiziert, bis alle Punkte im Abstand von $\Delta\alpha$ auf dem Kreis abgesucht wurden (Abb. 2.4).

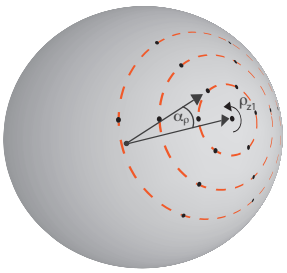

Abbildung 2.4: *Sampling* durch das lokal fokussierte Winkelgitter für die schrittweise Verfeinerung von Rotationswinkeln.

2.3.4 Optimierung durch *Expectation Maximization*

Der *Expectation Maximization*-Algorithmus (EM) wird verwendet, wenn für eine Abbildung $\mathcal{T}: P \mapsto A$ Beobachtungen aus P in einen Zustand A die unbekannten Abbildungsparameter $\theta \in \Theta$ bestimmt werden sollen [Forsyth und Ponce, 2003]. Angenommen

die für $\mathcal{T}$ benötigten Parameter θ sind bekannt, so beträgt die Wahrscheinlichkeit [2]

$$L(P, A) = log \prod_i l(P_i, A|\theta_i) = \sum_i log(l(P_i, A|\theta_i)) \quad . \tag{2.23}$$

dass P_i durch θ_i auf A abgebildet werden kann. Überträgt man das EM-Grundproblem auf das Alignment von Subtomogrammen, so entsprechen die Beobachtungen hier den Subtomogrammen P_i, die unbekannten Parameter (*hidden variables*) zu jedem P_i sind die Transformationsparameter θ_i und der Zustand A ist der Average aller P_i. Ähnlich wurde der EM-Algorithmus für die iterative Rekonstruktion von Makromolekülen in der KEM formuliert [Sorzano et al., 2006].

Im EM-Algorithmus werden Score und Sampling letztendlich kombiniert. Da die Wahrscheinlichkeiten l allerdings nicht bekannt sind, werden diese durch den Score gemessen und approximiert.

$$\mathcal{S}(P, A) = \frac{\sum_i^N \mathcal{S}(P_i, A, \theta_i)}{N} \tag{2.24}$$

Die Menge an Transformationsparametern Θ wird durch eine beliebige Sampling-Strategie (Kap. 2.3.3) bestimmt. Da man an der optimalen Belegung von θ_i und somit an der Maximierung von 2.24 interessiert ist, wird θ_i durch

$$\theta_i = argmax_{\theta \in \Theta}(\mathcal{S}(P_i, A, \theta)) \tag{2.25}$$

optimiert (für Korrelationsfunktionen maximiert). Wurde die optimale Belegung $\theta_{i,j}$ in einer Iteration j bestimmt, so kann A_{j+1} aktualisiert werden.

$$A_{j+1} = \mathcal{F}^{-1} \left(\frac{\mathcal{F}(\sum_i^n \mathcal{T}(P_i, \theta_{i,j}))}{\sum_i^n \mathcal{T}(W_i, \theta_{i,j})} \right) \tag{2.26}$$

Die Optimierung aller θ ist demnach ein Zweischrittverfahren, in dem in jeder Iteration j (i) die „Erwartung“ (*Expectation*) A_j für alle θ_{j-1} bestimmt wird und (ii) $\mathcal{S}$ für die „versteckten Variablen“ $\theta_j \in \Theta$ maximiert (*Maximization*) wird [Forsyth und Ponce, 2003]. Die beiden Schritte werden wiederholt, bis die Parameter θ in einen festen Zustand konvergiert sind (Abb. 2.5). Alle bisher publizierten Alignment-Methoden von Subtomogrammen basieren auf der *Expectation Maximization* (EM) Optimierung [Walz et al., 1997, Foerster et al., 2005, Bartesaghi et al., 2008, Scheres et al., 2009, Winkler et al., 2009, Stölken et al., 2010].

[2] Normalerweilse werden Wahrscheinlichkeiten durch p,P symbolisiert. Da P in dieser Arbeit bereits für Partikel oder Subtomogramm steht, wurden l und L (Likelihood) benutzt.

Der gewählte Score $\mathcal{S}$ (Kap. 2.3.2) und die gewählte Samplingstrategie Θ (Kap. 2.3.3) generieren die Energielandschaft, in der θ optimiert wird. Von der Wahl dieser beiden Methoden hängt ab, ob das globale Optimum der Energielandschaft tatsächlich den besten Average des Datensatzes abbildet. Des Weiteren ist aufgrund des extrem geringen SNR, dem *Missing Wedge* und der KTF die von $\mathcal{S}$ und Θ aufgespannte Energielandschaft nicht konvex, sondern durchsetzt von lokalen Optima. Der EM-Algorithmus konvergiert deshalb nicht notwendigerweise in das globale, sondern viel wahrscheinlicher in ein lokales Optimum. Der Startpunkt des Alignment-Prozesses, die initiale Referenz A_0, beeinflusst ebenfalls das Konvergenzverhalten des Alignments. Es ist daher schwierig eine valide Referenz A_0 zu bestimmen, von der aus das Alignment in das globale Optimum konvergiert, falls es von $\mathcal{S}$ und Θ richtig abgebildet wurde. Ist das nicht der Fall, so kann der Alignment-Prozess in ein lokales Optimum konvergieren, welches trotz optimaler Scores und guter FRK nicht das beste Average des Datensatzes wiedergibt.

2.4 Klassifikation von Subtomogrammen

Für die Bestimmung hoher Auflösungen spielt die Klassifikation von Subtomogrammen im Analyseprozess von Makromolekülen durch die KET eine entscheidende Rolle, da ein ganzer Datensatz bestehend aus mehreren tausend Partikeln üblicherweise heterogen ist. Homogene Datensätze sind in der Anwendung extrem unwahrscheinlich. Die Hauptanforderung an einen Klassifikationsalgorithmus in der KET ist die Zuverlässigkeit erstens gegenüber dem niedrigem SNR und zweitens gegenüber dem *Missing Wedge*.

2.4.1 Klassifikation durch *Constrained Principal Component Analysis*

Methoden für die Klassifikation von Subtomogrammen haben historisch gesehen ihren Ursprung in der KEM und deshalb in der Multivariaten Statistischen Analyse (MSA) [van Heel und Frank, 1981]. Rauschen in Subtomogrammen wird hier mittels Hauptkomponentenanalyse und anschliessender Projektion gedämpft, so dass die Partikel nach ihren charakteristischen Merkmalen klassifiziert werden können. Für die sukzessive Klassifikation wurden ursprünglich *K-Means* und die hierarchische Klassifikation benutzt. Die erste Studie [Walz et al., 1997], die sich mit der Klassifikation von Subtomogrammen beschäftigte, hat genau diese Methode angewandt, um alignierte Subtomogramme von Thermosomen zu klassifizieren. Wie in der MSA üblich wurde hier die Klassifikation im Pixelraum durchgeführt, so dass der *Missing Wedge*-Effekt der einzelnen Subtomogramme nicht berücksichtigt werden konnte. Das Hauptproblem bei der Klassifikation von alignierten Subtomogrammen ohne *Missing Wedge*-Korrektur ist, dass Subtomogramme nach der Orientierung durch den *Missing Wedge*-Effekt entstehenden Elongationen klas-

sifiziert werden und nicht nach ihren strukturellen Unterschieden.

Die ersten Klassifikationsansätze, die den *Missing Wedge* in Subtomogrammen berücksichtigten, sind [Bartesaghi et al., 2008] und [Foerster et al., 2008]. Beide benutzen die CC (Kap. 2.3.2) um die Berechnung des Scores auf die sich überschneidenden Regionen in den Subtomogrammen zu reduzieren. Während in [Bartesaghi et al., 2008] eine Klassifikationsmethode wie in Kapitel 2.4.2 beschrieben gewählt wurde, orientiert sich die *Constrained Principal Component Analysis* (CPCA) in [Foerster et al., 2008] an der MSA Methode. Der Hauptunterschied zu [Walz et al., 1997] ist der gewählte Raum, in dem die Eigenwerte bestimmt werden. Ist es für die MSA der Pixelraum, so ist es für den in [Foerster, 2005, Foerster et al., 2008] beschriebenen Ansatz der Partikelraum. Der Vorteil des Partikelraums ist, dass hier durch die CC der *Missing Wedge* in die Bestimmung der Eigenvektoren berücksichtigt werden kann. Es wird zuerst eine Ähnlichkeitsmatrix (K) aller Subtomogramme erstellt, die als Grundlage für die Eigenwertanalyse dient.

$$K_{ij} = cc(P_i, P_j) \tag{2.27}$$

K wird als nächstes durch die Eigenwertzerlegung (*Principal Component Analysis*) (PCA) in die respektiven Eigenwerte λ_{Ki} und Eigenvektoren μ_{Ki} zerlegt, so dass gilt

$$K\mu_{Ki} = \lambda_{Ki} \cdot \mu_{Ki} \qquad . \tag{2.28}$$

Der erste Eigenvektor μ_1 ist der Mittelwert aller Subtomogramme, μ_2 die größte Varianz u.s.w.. Die Klassifikation findet folglich im Eigenraum der Subtomogramme statt. Um die Subtomogramme aus dem karthesischen Raum in den Eigenraum zu transformieren, wird jedes Subtomogramm mit der Auswahl der m längsten Eigenvektoren multipliziert. Beschreiben die ersten m Eigenvektoren die charakteristischen Merkmale des Datensatzes, dann gilt:

$$\sum_{i}^{m} \lambda_{Ki} \leq s \qquad . \tag{2.29}$$

Für den Wert s, zu dem sich die Eigenwerte λ_{Ki} addieren, gilt $s \leq 1$. Die anschliessende Klassifikation wurde in [Foerster et al., 2008], wie für die MSA typisch, durch *K-Means* oder hierarchische Klassifikation durchgeführt.

2.4.2 Klassifikation durch *Multiple Correlation Optimization*

Eine Alternative zur CPCA ist die direkte Klassifikation von Subtomogrammen im Pixelraum mittels *Multiple Correlation Optimization* (MCO-EM). Im Gegensatz zur CPCA kommt dieser Ansatz ohne die Korrelationsmatrix aus und ähnelt im Prinzip der *K-Means* Klassifikation. MCO-EM ist ein weit verbreiteter Ansatz und wurde bisher in verschiede-

nen Varianten für die Klassifikation benutzt [Bartesaghi et al., 2008, Scheres et al., 2009, Winkler et al., 2009, Stölken et al., 2010]. Ein konzeptzioneller Vorteil von MCO-EM gegenüber CPCA ist, dass man das Alignment zwischen die Klassifikationsschritte integrieren kann, so dass eine *Mutli Reference Alignment* (MRA)-Prozedur entsteht (Abb. 2.5).

MCO-EM ohne Alignment. Analog zu dem in Kapitel 2.3.4 vorgestellten Subtomogramm-Alignment basiert MCO-EM auch auf *Expectation Maximization*. In dem hier dargestellten, einfachen Fall ohne integrierte Alignmentschritte beschränken sich die *hidden variables* auf das Klassenattribut κ eines jeden Subtomogrammes. Ausgehend von einer initialen Zuweisung[3] aller bereits alignierten Subtomogramme zu K Klassen werden Klassenmittel CA_k aus allen zugehörigen Subtomogrammen $P_i \in C_k$ gebildet (*Expectation*). Die Subtomogramme P_i werden einer der K Klassen zugewiesen, für die der optimalste Score bestimmt wurde (*Maximization*)

$$\kappa_{P_i} = argmax_k^K \mathcal{S}(P_i, CA_k) \qquad . \tag{2.30}$$

Wie beim Alignment wird MCO-EM wiederholt bis θ konvergiert oder eine maximale Anzahl von Iterationen überschritten wurde.

MCO-EM mit Alignment (*Mutli Reference Alignment*). Für alle in Kapitel 2.4.2 aufgelisteten Klassifikationsalgorithmen ist MCO-EM ein Spezialfall, da in allen Algorithmen das Alignment integriert ist. Sie unterscheiden sich allerdings in den benutzten *Scores* und in der Art, wie das *Alignmentsampling* implementiert ist. Beim MRA werden sieben Parameter optimiert: die sechs Alignment-Parameter θ und das Klassenattribut κ.

$$\kappa_{P_i}, \theta_{P_i} = argmax_\kappa^K (argmax_\theta^\Theta \mathcal{S}(P_i, \mathcal{T}(CA_\kappa, \theta))) \qquad . \tag{2.31}$$

MRA basierend auf Formel 2.31 ist aufwendig zu berechnen, da jedes Subtomogramm P_i mit jedem Klassenaverage CA_k aligniert wird. Abbildung 2.5 verdeutlicht, wie Alignment und Klassifikation durch MRA kombiniert werden können und dass die beiden unabhängigen Prozesse (MCO-EM und Alignment) Sonderfälle von MRA sind.

Das Alignment und die Klassifikation mit Hilfe von *Multi Reference Alignment*, ebenso wie die beiden Sonderfälle, basieren auf der Hypothese, dass die versteckten Parameter $\theta \in \Theta$, $\kappa \in K$ für jedes Subtomogramm *a priori* gleich wahrscheinlich und von allen anderen Subtomogrammen unabhängig sind. Die in MRA integrierte Optimierung des Scores schließt eine nähere Betrachtung der Verteilung der versteckten Parameter aus,

[3] Falls keine initiale Klassenzuweisung existiert, so ist eine initiale Zuweisung anhand der Gleichverteilung empfehlenswert [Bishop, 2006].

da nur der optimalste Score und die dazugehörigen Parameter gespeichert werden. Dieser Ansatz kann durch die *Maximum Likelihood* Optimierung erweitert werden, indem man für die versteckten Parameter θ_i und κ_i komplexere Wahrscheinlichkeitsverteilungen wie die Gauß-Verteilung postuliert. In [Scheres et al., 2009, Stölken et al., 2010] wird z.B. eine Gleichverteilung für die Klassenzugehörigkeit, wie auch für die Rotationsparameter vorausgesetzt. Die Translationsparameter werden allerdings als Gauß-verteilt angenommen. Für diesen Ansatz ist der Score keine Korrelation sondern $\mathcal{X}^2$. Des Weiteren müssen in jeder Iteration alle Scores für das Aufstellen der Verteilung zwischengespeichert werden, was die Komplexität des Algorithmus erweitert. Während den Iterationen werden θ und κ, sowie die, den Wahrscheinlichkeitsverteilungen zugrundeliegenden Parameter, optimiert. Nach mehreren Iterationen konvergieren die Verteilungsfunktionen in scharfe Deltafunktionen und somit in ein Optimum.

2.5 Erstellung von Referenzen für die Lokalisierung und Alignment

Referenzen werden in der Verarbeitung von Subtomogrammen in der Regel gebraucht, um einen Startpunkt des Alignments festzulegen.

Erstellung von Referenzen aus atomaren Modellen. Ist man daran interessiert die Interaktion bekannter Proteinkomplexe *in vivo* oder *in vitro* abzubilden, oder zu ermitteln, wie sie an Membranen gebunden sind, so kann es ausreichend sein, eine Referenz aus einem *Protein Databank* (PDB)-Eintrag zu generieren. Um eine Dichte aus einer PDB-Datei zu erstellen, wird der in der PDB bestimmte Raum zuerst auf eine vorgegebene Größe diskretisiert. Hierbei entspricht jede diskrete Einheit (Voxel) einer beliebigen Voxelbreite in Å. Als nächstes werden die Ladungszahlen Z aller Atome des Proteins summiert, die sich in einem Voxel befinden. Das Resultat ist eine Dichte, die als initiale Referenz für die Lokalisierung von Makromolekülen oder für das Alignment von Subtomogrammen benutzt werden kann.

Model Bias. Ist jedoch die Aufklärung einer bisher unbekannten, makromolekularer Struktur das Ziel, so muss man die Strukturbestimmung ausgehend von einem bekannten Komplex kritisch betrachten. Rauschen kann die strukturbestimmenden Methoden signifikant beeinflussen, so dass von der initialen Referenz unterschiedliche Subtomogramme trotzdem in ein Resultat aligniert werden, das der ursprünglichen Referenz entspricht (*Model Bias*).

2.6 Lokalisierung von Subtomogrammen

Die Lokalisierung von Makromolekülen in Tomogrammen wurde in dieser Arbeit nur marginal bearbeitet und wird hier deshalb nach dem Alignment und der Klassifikation beschrieben. Im gewöhnlichem KET-Arbeitsablauf wird es jedoch vor dem Alignment durchgeführt (Abb. 1.1). Eine detaillierte Beschreibung der *Template Matching* (TM)-Implementierung ist in [Chen et al., 2012] zu finden und entstand in Kollaboration im Rahmen dieser Arbeit.

Der für das spätere Alignment und die Klassifikation benötigte Datensatz wird durch das automatisierte TM bestimmt [Frangakis und Rath, 2006]. Dieser Vorgang wurde ebenfalls aus der KEM übernommen und an das *Missing Wedge*-Problem in der KET angepasst [Frangakis et al., 2002]. In diesem Fall wird ein verhältnismässig kleines Partikel P ($32 \times 32 \times 32$ Voxel) in einem großen Tomogramm T ($512 \times 512 \times 128$ Voxel) gesucht.

Wie auch im Alignment und der Klassifikation wird beim TM ebenfalls in Score $\mathcal{S}$ optimiert. Dieser ist in der Regel LNXC (Formel 2.19), um die lokale Invarianz durch die Fouriertransformation auszunutzen. Der spektrale Bereich von P wird nicht nur durch W beschränkt, sondern auch durch die KTF, damit das Rauschen in den nicht abgetasteten Frequenzen nicht in den Score einfließt.

$$S_{TM} = argmax_{\rho,x,y,z}\mathcal{S}(\mathcal{T}(P \otimes PSF(KTF, W), \rho), T) \qquad . \tag{2.32}$$

In der hier vorgestellten Implementierung werden alle Rotationen ρ aus in $SO(4)$ bestimmten Winkellisten abgetastet (Kap. 2.3.3), was im Gegensatz zu früheren Implementierungen (z.B. MOLMATCH [Foerster et al., 2010]) effizienter ist (Abb. 2.3).

Das Resultat der Lokalisierung S_{TM} ist ein Volumen mit identischer Größe zu T und enthält die bestimmten Maxima an den Positionen x, y, z. Der entscheidende Schritt für das weitere Vorgehen ist die Bestimmung der Anzahl von Subtomogrammen. Es werden die N höchsten Scores in S_{TM} bestimmt und Subtomogramme um diese Positionen rekonstruiert. In der Literatur findet man bislang drei Methoden mit denen N bestimmt werden kann:

Handedness check. TM wird für das Partikel P sowie für eine, an einer Koordinatenebene gespiegelten Kopie P_{mirr} durchgeführt [Ortiz et al., 2006, Foerster et al., 2010]. Es werden die Werte aller Scores beider Läufe gegen ihren Index aufgetragen und N am Schnittpunkt beider Kurven bestimmt. Der Kerngedanke dieses Verfahrens ist die Hypothese, dass ab dem Schnittpunkt der Score nicht mehr signifikant durch strukturelle Details, sondern durch Rauschen bestimmt werde. Dieser Methode muss vorausgesetzt werden, dass P nicht symmetrisch ist.

Bestimmung einzelner Untereinheiten. Kann man das gesuchte Partikel P in Untereinheiten $P = \sum_i^M P_{sub,i}$ zerlegen, so ist es möglich, die einzelnen Untereinheiten unabhängig voneinander zu lokalisieren [Foerster et al., 2010]. Werden alle Untereinheiten in der richtigen Orientierung und in der richtigen, relativen Nähe zueinander gefunden, so kann man diese Position zur Liste der Subtomogrammen hinzufügen. Diese Methode funktioniert allerdings nur, wenn die Untereinheiten ausreichend groß sind, um in T gefunden zu werden. Außerdem müssen jeweils enge Masken M_i benutzt werden, damit die zusammenhängenden Untereinheiten auch unmittelbar neben anderen Untereinheiten gefunden werden.

Statistische Abschätzungen. Generiert man mit den Werten aller Scores ein Histogramm, so kann man, falls in dem Histogramm ein lokales Maximum S_{lmax} auffindbar ist, eine Gauß-Kurve G in die entstandene Verteilung approximieren [Foerster, 2005, Ortiz et al., 2006]. N kann mit Hilfe der approximierten Kurve statistisch bestimmt und auch mehreren Konfidenzintervallen zugeordnet werden:

1. $N = |\{P_i | S_{P_i} \geq S_{lmax}\}|$
2. $N = |\{P_i | S_{P_i} \geq S_{lmax} - \sigma_G\}|$
3. $N = |\{P_i | S_{P_i} \geq S_{lmax} - 2\sigma_G\}|$

Darüberhinaus kann mit Hilfe der ganzen Gauß-Kurve G eine *Receiver Operating Characteristics* (ROC) erstellt werden [Forsyth und Ponce, 2003, Langlois und Frank, 2011], mit deren Hilfe man das Verhälnis von echten (*true positives*) und falschen (*false positives*) Makromolekülen N abschätzen kann.

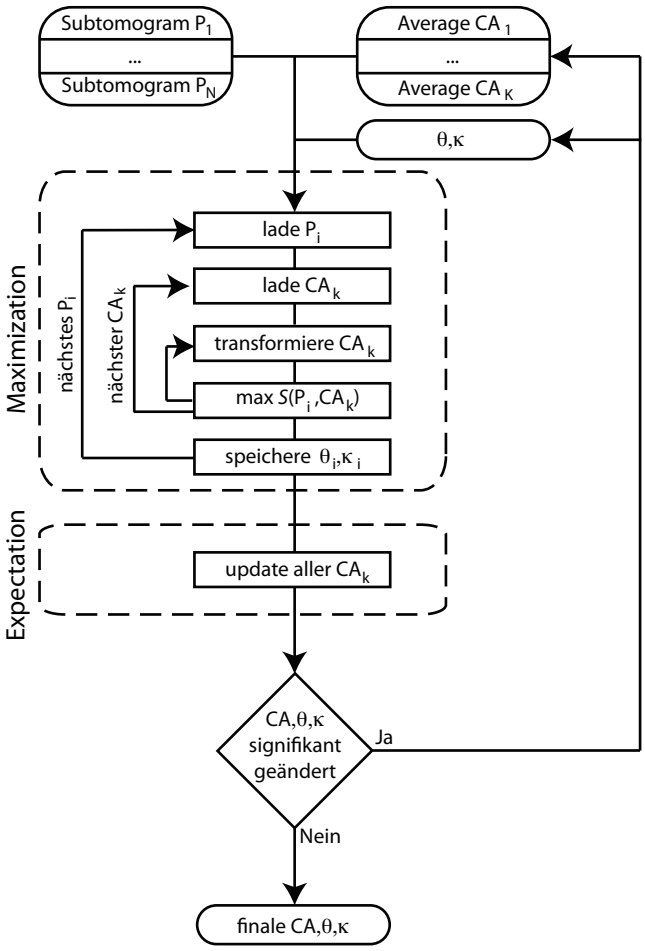

Abbildung 2.5: Flußdiagram der *Multi Reference Alignment*-Prozedur von N Subtomogrammen mit K Klassen. Im *Maximization*-Schritt wird für jedes Klassenmittel CA_k die optimale Transformation θ_i, und die Klassenzugehörigkeit κ_i für alle Subtomogramme P_i bestimmt. Nach dieser Optimierung werden die Klassenmittel CA im *Expectation*-Schritt aktualisiert. Haben sich alle CA, θ und κ nicht signifikant geändert oder wurde eine maximale Anzahl an Iterationen überschritten, wird der Prozess beendet. Anderenfalls wird *Maximization* und *Expectation* wiederholt. Für das auf EM basierte Alignment ohne Klassifikation ($K = 1$) wird nur auf ein Average CA aligniert, so dass die *nächster* CA_k-Schleife entfällt. Bei der MCO-EM Klassifikation ohne Alignment entfällt die innere Transformationsschleife.

3 Material und Methoden

3.1 Entstehung der verwendeten Daten

3.1.1 Simulationen des *S. cerevisae* 80S-Ribosoms in verschiedenen Stadien

Ein synthetischer Datensatz wurde generiert, um die in dieser Arbeit entwickelten Klassifikationsalgorithmen umfassend zu testen. Als Modellpartikel wurde das 80S-Ribosom aus *S. cerevisae* verwendet. Es wurde vier Stadien des 80S-Ribosomes generiert, um die Signalsequenz während der Proteintranslation aus dem Zytosol in das Lumen des rauen ERs zu simulieren [Alberts et al., 2010] (Abb. 1.2).

Das eukaryotische 80S-Ribosom. Um die Dichte des *S. cerevisae* 80S-Ribosomes zu erhalten, wurden atomare Modelle des selbigen aus der PDB-Datenbank verwendet. Die PDB Einträge 3IZS (Proteindichten) und 3IZF (rRNS-Dichten) für die 60S-Untereinheit und 3IZB (Proteindichten) und 3IZE (rRNS-Dichte) für die 40S-Untereinheit ergeben das vollständige 80S-Ribosom [Armache et al., 2010]. Das Elektronenpotential aller vier Modelle wurde in einem $100 \times 100 \times 100$ Voxel großes Volumen mit einer Voxelgröße von 4.7Å erstellt (Abb. 3.1).

80S-Ribosom mit dem *Signal Recognition Particle*. Eine weitere Population von 80S-Ribosomen mit dem *Signal Recognition Particle*(SRP) (PDB 1RY1) wurde generiert. Bei der Proteintranslokation aus dem Zytosol in das ER-Lumen bindet das SRP an das Ribosom, falls das translatierte Polipeptid mit der passenden Signalsequenz am N-Terminus gekennzeichnet ist (Abb. 3.1).

80S-Ribosom mit dem SRP-Rezeptor. Der *Signal Recognition Particle Receptor*(SRPR) wurde ebenfalls an das 80S-Ribosom (PDB 2GO5) appliziert. Das SRPR ist in der ER-Membran eingebettet und erkennt an Ribosomen gebundene SRP-Moleküle. Durch diesen Mechanismus werden Ribosomen an einen Translokationskanal (Sec61) in der Membran gebunden (Abb. 3.1).

80S-Ribosom mit dem Sec61-Partikel. Als letzte Klasse wurde zum generierten 80S-Ribosom das Sec61-Partikel assembliert, wie es während der Translokation von Polypeptiden von [Becker et al., 2009] beobachtet wurde (PDB: 2WWB) (Abb. 3.1).

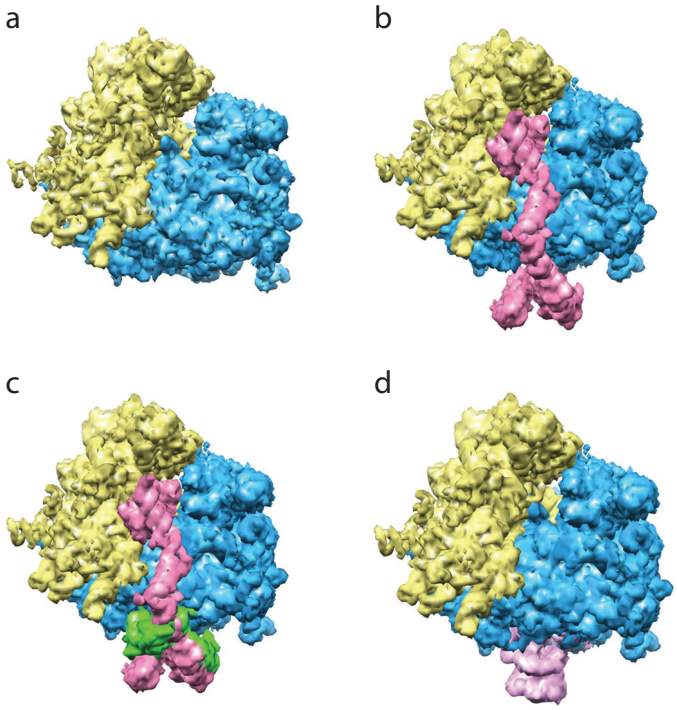

Abbildung 3.1: Als Testdatensatz für die Klassifikationsmethoden wurden vier Konformationen des 80S-Ribosomes generiert, um die Signalsequenz während der Proteintranslation aus dem Zytosol in das Lumen des ER zu simulieren. (a) Das freie 80S-Ribosom, die große Untereinheit in Blau, die kleine Untereinheit in Gelb. (b) Das 80S-Ribosom mit dem *Signal Recognition Particle* (Rot). (c) Der *Signal Recognition Particle Receptor* (Grün) am 80S-Ribosom. (d) Das Sec61-Partikel (Magenta) am 80S-Ribosom.

Simulation der Tomographie. Alle synthetisch generierten 80S-Modelle wurden zufällig rotiert ρ_{rand} und zufällig translatiert ν_{rand}, um möglichst echte Simulationen zu produzieren.

$$\rho_{rand} \in SO(3) \qquad \nu_{rand} \in [\mathbb{R}_{[-5;5]}, \mathbb{R}_{[-5;5]}, \mathbb{R}_{[-5;5]}]$$

Wie in [Foerster et al., 2008] beschrieben, wurden Projektionen Pr aus den entstandenen Modellen generiert. Das simulierte Kippinterval betrug $[-60°, 60°]$ und der Kippwinkel zwischen den einzelnen Projektionen war 5°. Zu jeder Projektion Pr_p wurde Gaußverteiltes Rauschen $\mathcal{G}(\mu_{Pr_p}, 0.5)$ addiert. Im Fourierraum wurde eine simulierte KTF appliziert,

um den Signalübertrag im Elektronenmikroskop möglichst wirklichkeitsnah nachzubilden. Die KTF bestimmenden Parameter betrugen:

- Defokus $\lambda = -5\mu m$
- Pixelgröße $v = 4.7$ Å
- Beschleunigungsspannung $av = 300kV$
- sphärische Aberration $C_S = 2$

Nach der Faltung mit der KTF wurde erneut GaußVerteiltes Rauschen $\mathcal{G}(\mu_{Pr_p}, 0.5)$ zu den Projektionen addiert und das Resultat wurde an der ersten Nullstelle der KTF tiefpassgefiltert.

$$Pr_{p,final} = F(\mathcal{G}(\mu_{Pr_p}, 0.5) + \mathcal{F}^{-1}(KTF \times \mathcal{F}(\mathcal{G}(\mu_{Pr_p}, 0.5) + Pr_p))) \tag{3.1}$$

Abschließend wurden aus den finalen Projektionen die simulierten Subtomogramme mittels gewichteter Rückprojektion (Kap. 2.2) rekonstruiert.

3.1.2 GroEL$_{14}$ und GroEL$_{14}$/GroES$_7$ als Quasi-Standard-Testdatensatz

In den letzten Jahren ist der GroEL$_{14}$, GroEL$_{14}$/GroES$_7$ zum Quasi-Standard-Testdatensatz von KET-Klassifikationsmethoden geworden [Foerster et al., 2008, Scheres et al., 2009, Yu und Frangakis, 2011, Heumann et al., 2011].

Das GroEL$_{14}$/GroES$_7$ Chaperon. Das *Escherichia coli* GroEL$_{14}$/GroES$_7$-Chaperon unterstützt ungefaltete Polypeptide bei der Faltung in den energetisch günstigsten Zustand. Die GroEL-Untereinheit (57kDa) selbst besteht aus drei Domänen: der equatorialen Domäne, der zwischen Domäne (*intermediate*) und der apikalen Domäne [Sigler et al., 1998]. Des Weiteren formen jeweils sieben GroEL-Monomere den *Cis*- und den *Trans*-Ring, aus welchen sich der GroEL$_{14}$-Komplex (798kDa) zusammensetzt. Beide Ringe sind mit dem Rücken zueinander angeordnet und durch den C-Terminus der apikalen Domäne voneinander getrennt, so dass die Faltung von Polypeptiden an beiden Seiten assistiert werden kann. Die isolierte Umgebung innerhalb des GroEL$_{14}$-Komplexes ermöglicht es somit einer ungefallteten Peptidkette sich in ihre optimale Struktur zu falten. Sieben GroES-Untereinheiten (10kDa) formen das heptamere GroES$_7$-Co-Chaperon, das ungefaltete Polypeptidketten an ihren hydrophoben Komponenten bindet und in den *Cis*-Ring befördert. Während der *Cis*- assistierten Faltung verschliesst das GroES$_7$ den *Cis*-Ring. Durch die Bindung von GroES$_7$ an den *Cis*-Ring ändert sich die Konformation der apikalen Domäne der GroEL-Monomere, so dass der *Cis*-Ring gestreckt wird und sich das Volumen innerhalb des Ringes annähernd verdoppelt [Chaudhuri et al., 2009]. Die atomare

Struktur von $GroEL_{14}$ und $GroEL_{14}/GroES_7$ selbst wurde bereits in [Braig et al., 1994] kristallographisch gelöst (Abb. 3.2).

Bedeutung von $GroEL_{14}/GroES_7$ für die KET. $GroEL_{14}$ (Länge entlang der Längsachse 142Å, Molekulargewicht 798kDa) und $GroEL_{14}/GroES_7$-Chaperone (Länge entlang der Längsachse 184Å, Molekulargewicht 868kDa) sind im Vergleich zu *S. cerevisae* 80S-Ribosomen (Durchmesser ca. 250Å, Molekulargewicht ca. 3.6MDa) ca. fünf mal kleiner, was die Qualität der Subtomogramme nachteilig beinflußt, und deshalb einen typischen KET-Anwendungsfall darstellt. Die Analyse der unterschiedlichen Konformationen von $GroEL_{14}$ und $GroEL_{14}/GroES_7$ soll die Genauigkeit von Alignment- und Klassifikationsmethoden messen.

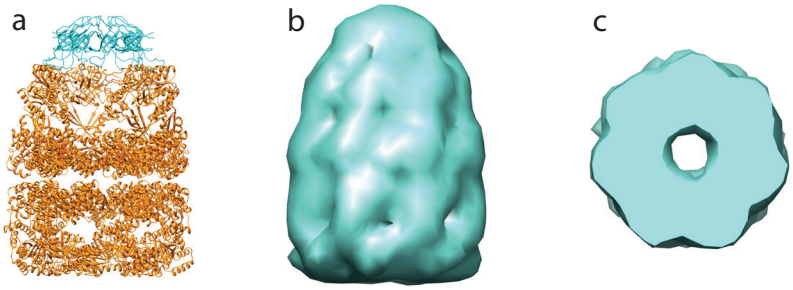

Abbildung 3.2: (a) Die atomare Struktur von $GroEL_{14}/GroES_7$ (PDB-ID: 1GRL [Braig et al., 1994]). (b) Elektronenpotential generiert aus der 1GRL Struktur, interpoliert auf eine Pixelgröße von 12Å. (c) Betrachtet man einen Schnitt senkrecht zur Längsachse der Dichte, so wird die C7-Symmetrie von $GroEL_{14}$ deutlich sichtbar.

Der Testdatensatz besteht aus 786 Subtomogrammen, die manuell in den rekonstruierten Tomogrammen gefunden wurden [Foerster et al., 2008]. Hierbei stammen 214 Subtomogramme aus Tomogrammen einer Probe mit ausschliesslich $GroEL_{14}$-Komplexen. Die restlichen 572 Subtomogramme stammen aus Tomogrammen einer Probe, in der $GroEL_{14}$- und $GroES_7$-Komplexe in einem Verhältniss von 1 : 10 vermischt wurden [Foerster et al., 2008].

3.1.3 Tomogramme eines *S. cerevisae*-Lysates

Es wurde ein Lysat aus *S. cerevisae* erstellt, in dem translationskompetente 80S-Ribosomen sowie deren einzelne Untereinheiten (40S und 60S) zu finden waren (Abb. 3.3).

Aufreinigung und Vorbereitung der Probe für KET. Das *S. cerevisae*-Lysat wurde nach dem in [Beckmann et al., 2001] beschriebenen Protokoll von Claudia Szalma erstellt. Die *S. cerevisae* Zellen wurden zuerst bei $30°C$ inkubiert und bis zu einer OD_{600} von $2-3$ kultiviert. Mit Zymolase (MP-BioMedicals) wurde die Glukanzellwand der Zellen zerstört. Die Sphäroblasten wurden zentrifugiert ($20min$, $4°$,$3.300 \times g$) und in $1ml$ Lysis-Puffer ($20mM$ Hepes, $100mM$ KOAc, $2mM$Mg(OAc)$_2$, $1mM$ DTT, $0.5mM$ PMSF) pro $5g$ Zellen resuspendiert. Des Weiteren wurden die Sphäroblasten mit bis zu 30 Stößen mit einem Glass-Douncer aufgeschlossen. Das Lysat wurde zentrifugiert ($15min$, $28.714 \times g$), um Zellorganellen und Zytosol zu trennen. Der Überstand wurde isoliert und wiederum $30min$ bei $140.531 \times g$ zentrifugiert. Abschliessend wurden das Lysat in flüssigem Stickstoff vitrifiziert und bei $-70°$ gelagert.

Aufnahmeparameter der Tomographie und Rekonstruktion. Die Aufnahme erfolgte in einem *FEI Tecnai Polara*-Transmissionselektronenmikroskop. Die Beschleunigungsspannung des Elektronenstrahls betrug 300kV, die Bildaufnahme erfolgte durch eine 2048×2048 Pixel CCD-Kamera (Gatan, Pleasanton, USA), ausgestattet mit einem Gatan GIF 2002-Energiefilter. Der Defokus lag bei $-8\mu m$, die in der Bildebene entstandene Vergrößerung betrug 27.000 und resultierte in einer abgebildeten Voxelbreite von 4.7Å. Für alle sechs Kippserien betrug das Kippintervall $[-60°; 60°]$ und der Kippwinkel zwischen den einzelnen Projektionen war $3°$.

Manuell selektierte Positionen der abgebildeten, kollodialen Goldpartikel dienten als Grundlage für das Alignment der Einzelprojektionen [Mastronarde, 2006], aus denen die einzelnen Tomogramme mittels gewichteter Rückprojektion rekonstruiert wurden (Kap. 2.2). Die Rekonstruktion wurde in TOM durchgeführt.

3.1.4 Tomogramme von ER-Mikrosomen aus *S. cerevisae*

Probenvorbereitung für die KET. Die Probe wurde nach der in [Ramezani-rad et al., 1985] beschriebenen Methode von Ann-Victoria Mangold erstellt [Mangold, 2010]. Der *S. cerevisae*-Stamm YWO-343 wurde bei $30°C$ und $170rpm$ $12h$ lang kultiviert. Bei einer OD_{600} von 1 wurden die Zellen fünf Minuten lang bei $4.000 \times g$ pelletiert. Das entstandene Pellet wurde in einem Homogenisierungspuffer gelöst und für ca. $30min$ bei $30°C$ inkubiert. Der Puffer bestand aus $20mM$ Hepes, $100mM$ KaOAc, $2mM$ MgOAc, $1M$ Sorbit und 20% Glycerin. Die entstandenen Sphäroblasten wurden nochmals bei $4.000 \times g$ fünf Minuten lang pelletiert (1,9mM Sorbitol) und anschliessend im Homogenisierungspuffer gelöst. In einer *French press* ($3min$, $100.000kPa$, $4°C$, Avestin-Canada) wurden die Zellen aufgeschlossen. Die Zellorganellen wurden mittels sukzessiver Zentrifugationsschritte ($3.000 \times g$ und $5min$,

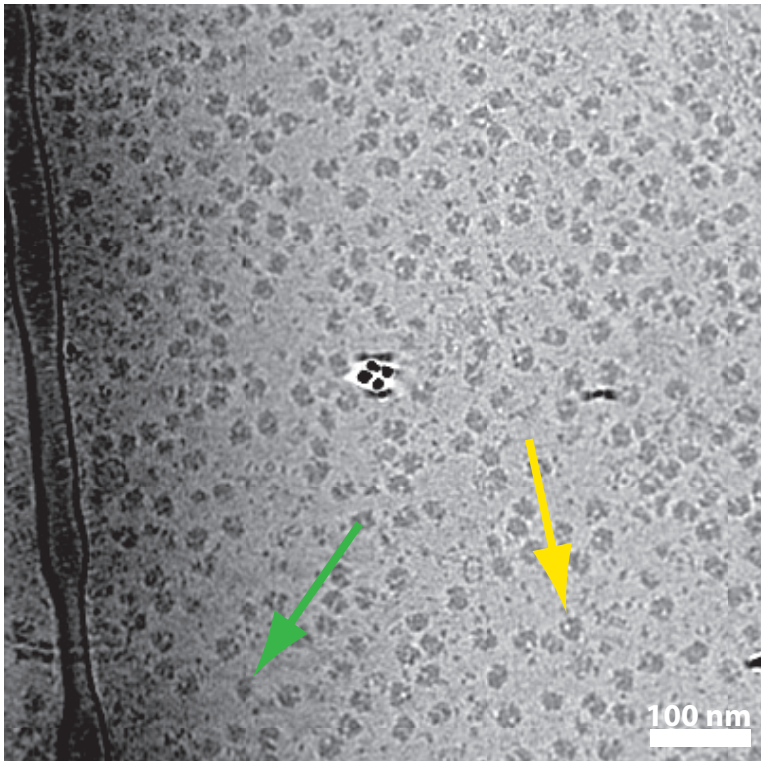

Abbildung 3.3: Schnitt durch eines der sechs *S. cerevisae*-Lysat-Tomogramme. Der gelbe Pfeil markiert die Position eines 80S-Ribosoms, der grüne Pfeil die Position einer 60S-Untereinheit.

16.000 $\times$ g und $10min$, 90.000 $\times$ g und $60min$) separiert, so dass die Fraktion des ER mittels Ultrazentrifugation weiter aufgereinigt werden konnte. Hierfür wurde ein vierstufiger Sacharosegradient erstellt: (i) 1ml 2M Saccharose im HKM Puffer ($20mM$ HEPES, $100mM$ KaOAc, $5mM$ MgOAc, $2mM$ DTT), (ii) $1,5ml$ $1,75M$ Saccharose im HKM Puffer, (iii) $1,5ml$ $1,33M$ Saccharose im HKM, und (iv) $1ml$ $0,2M$ Saccharose und HKM. Das glatte ER sedimentierte am Saccharoseübergang von $0,2M$ zu $1,33M$, das raue ER zwischen $1,33M$ und $1,75M$. Um die Saccharose Konzentration für die KET auf weniger als 5% zu reduzieren, wurde die Interphase bei $1,33M$ und $1,75M$ (Saccharose Konzentration bei $250\mu l$ 49%) bei 100.000 $\times$ g zwei Stunden zentrifugiert und resuspendiert [Mangold, 2010]. Die so verdünnte Probe wurde bei $-180°C$ auf einem Lacey-Karbon-Grid vitrifiziert und im Elektronenmikroskop untersucht.

Aufnahmeparameter der Tomographie und Rekonstruktion. Der Aufnahme- und Rekonstruktionsvorgang war identisch zu dem in Kapitel 3.1.3 beschriebenen Vorgang.

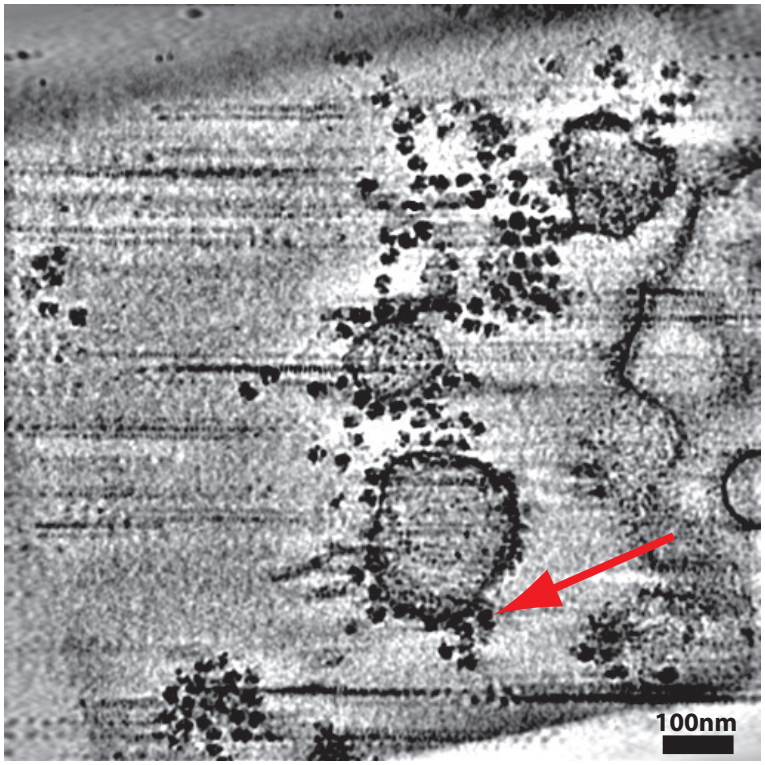

Abbildung 3.4: Schnitt durch das *S. cerevisae*-Mikrosomen-Tomogramm. Der rote Pfeil zeigt auf die Position eines an die Mikrosomenmembran gebundenen 80S-Ribosoms. Man erkennt deutlich die beiden Untereinheiten bei vielen der abgebildeten Ribosomen.

3.1.5 Tomogramme von Ribosomen gebunden an *canine* ER

Probenvorbereitung für die KET. In Zusammenarbeit mit Sven Lang wurde die Probe erstellt und in [Pfeffer, 2010] verwendet. Der Hunde-Pankreas wurde gepresst, je $30ml$ in einen $60ml$ Motorpotter überführt und in drei Durchgängen ($300upm$) wurden die Zellwände aufgebrochen. Bei $2°C$ wurde anschliessend $10min$ zentrifugiert ($706 \times g$) und nicht aufgeschlossene Zellen und Zelltrümmer getrennt. Des Weiteren wurden die Zelltrümmer von großen Zellorganellen, wie den Zellkernen, getrennt ($10min$, $2°C$, $6.350 \times g$). Jeweils $17ml$ des Überstandes wurden auf einem $9ml$-Kissen ($52ml$ $2,5M$ Sacharose, $5ml$ $1M$ TEA, $1,25ml$ $4M$ Kalziumacetat, $0,6ml$ $1M$ $MgCl_2$, $0,5ml$ $0,2M$ EDTA)[1] gelegt und hindurch-zentrifugiert ($2,5h$, $2°C$, $149.008xg$). Die so pelletierten Mikrosomen wurden in $5ml$ K_A ($25mg$ Pepstatin A in $2,08ml$ DMSO, $25mg$ Leupeptin in $2,08ml$ DMSO, $25mg$ Antipain in $2,08ml$ DMSO, $25mg$ Chymostatin in $2,08ml$ DMSO, $174,2mg$ PMSF in $1ml$ Aceton , $3,08g$ DTT in $20ml$ Wasser) resuspendiert und in einem $15ml$-Potter

[1] Im Folgenden durch K_A abegekürzt

homogenisiert. Um glatte und raue Mikrosomen zu trennen, wurde wie in Kapitel 3.1.4 ein Saccharosegradient verwendet ($4ml$ $2,1M$ Saccharose, $4ml$ $1,75M$ Saccharose und $4ml$ $1,5M$ Saccharose). Das Pellet wurde mit $1,5M$ Saccharose vermischt und und zentrifugiert ($15-20h$, $2°C$, $149.008 \times g$). Mikrosomen des glatten ER sedimentierten am Saccharoseübergang zwischen $1,5M$ und $1,75M$, Mikrosomen vom rauen ER am Saccharoseübergang zwischen $1,75M$ und $2,1M$. Um die Saccharosekonzentration für die KET auf weniger als 5% zu reduzieren, wurden die rauen Mikrosomen $1:1$ mit K_A versetzt und auf einem Saccharosekissen ($2ml$ $1,3M$) pelletiert ($2h$, $2°$,$149.008 \times g$). Die Pellets wurden vitrifiziert und bei $-80°C$ gelagert.

Aufnahmeparameter der Tomographie und Rekonstruktion. Bis auf den Defokus ($-4\mu m$) waren die Parameter beim Aufnahme- und Rekonstruktionsvorgang ebenfalls identisch zu dem in Kapitel 3.1.3 beschriebenen Vorgang.

3.2 Generierung initialer Referenzen *de novo*

In der Praxis kann es einerseits vorkommen, dass für eine Strukturbestimmung von Makromolekülen kein PDB-Eintrag existiert, aus dem eine Referenz für die Lokalisation oder das Alignment generiert werden könnte (Kap. 2.5). Andereseits könnte man möglichst den strukturellen Einfluss einer hochaufgelösten Referenz vermeiden wollen, indem man keine aus PDB-Dateien generierte Referenzen benutzt. Aus diesen Gründen werden in diesem Kapitel zwei Methoden vorgestellt, durch die man das Alignment von Subtomogrammen referenzfrei (*de novo*) durchführen kann. Beide Methoden basieren auf der Annahme, dass die gesuchten Makromoleküle bereits lokalisiert und entsprechende Subtomogramme rekonstruiert wurden.

3.2.1 Initiale Referenz generiert aus Rotationsklassen

Rotationsklassen können für die Erstellung einer initialen Referenz benutzt werden, mit der ein Alignment gestartet werden kann. Die hier gewählte Strategie ist im Prinzip eine Vereinfachung der *Constrained Principal Component Analysis* (Kap. 2.4.1), da der *Missing Wedge* für alle Partikel gleich orientiert ist. Aus diesem Grund muss bei dieser Methode keine Korrelationsmatrix mit Hilfe der CC aufgestellt werden, um die Klassifikation im Eigenraum der Subtomogramme durchzuführen.

Alignment der Translationen. Für die spätere Klassifikation ist es ausschlaggebend, dass der Schwerpunkt aller Subtomogramme im Zentrum ist, damit unterschiedliche Translationen nicht als Charakteristik des Datensatzes in die Klassifikation einfließen. Um

ν_i zu approximieren, werden alle Subtomogramme translations-aligniert. Als Alignment-Referenz für diesen Schritt bietet sich das Average A_0 der nicht alignierten Subtomogramme an.

$$\begin{aligned} A_0 &= \sum_{i}^{N} \mathcal{T}(P_i, 0) \\ \nu_i &= argmax_\nu \mathcal{S}(A_0, P_i, \nu) \end{aligned} \tag{3.2}$$

Singulärwertzerlegung und Klassifikation. Ähnlich zur CPCA-Klassifikation werden im nächsten Schritt die Hauptkomponenten der Subtomogramme berechnet. Da für alle N Subtomogramme $\rho_i = 0$ gilt, muss keine Ähnlichkeitsmatrix wie in [Foerster et al., 2008] aufgestell werden, aus der die Hauptkomponenten errechnet werden. Alternativ können diese durch die Singulärwertzerlegung (*Singular Value Decomposition*) (SVD) bestimmt werden [Rade und Westergren, 2000]. Die für die SVD benötigte Datenmatrix M der Größe $(x \times y \times z,N)$ wird aus den linearisierten Subtomogrammen erzeugt, so dass jede Zeile einem linearisierten[2] Subtomogramm entspricht. Optional können die Subtomogramme vor der Linearisierung Tief- oder Bandpass-gefiltert werden, um Rauschen vor der Singulärwertzerlegung zu reduzieren.

$$M = QSP^T \tag{3.3}$$

S ist eine Matrix der Größe $(x \times y \times z,N)$ mit den Singulärwerten $\mu_{j,j}$ für die gilt: $\mu_{j,j} = \sqrt{\lambda_j}$. Diese Einträge entsprechen den Wurzeln der Eigenwerte von $M^T M$. Jede Spalte der (N, N) großen Matrix P entspricht den Eigenvektoren μ von $M^T M$.

Basierend auf den durch SVD bestimmten Eigenwerten und Eigenvektoren können alle Subtomogramme analog zur CPCA durch *K-Means* klassifiziert werden (Kap. 2.4.1). Da die Subtomogramme bereits in das gemeinsame Zentrum aligniert wurden, werden die in den Subtomogrammen abgebildeten Makromoleküle nach ihren Orientierungen klassifiziert, sofern der Datensatz homogen ist. Die Grundannahme hier ist, dass die abgebildeten Makromoleküle keine Vorzugsorientierung haben und deren Rotation somit gleichverteilt ist.

Growing Average. Ein wachsender Average (*Growing Average*) GA_j wird berechnet, indem der Average CA_k der Klasse k auf das vorhergehende GA_{j-1} aligniert wird (ähnlich zu Kap. 2.3.4 oder zu [Winkler et al., 2009]). Der initiale Average GA_1 ist die Rotationsklasse CA_1. In der nächsten Iteration j wird GA_{j-1} auf den Average CA_j der Klasse j registriert. GA_j wird aktualisiert, indem alle Subtomogramme mit $(\kappa = j, \theta_j)$ mit allen

[2] $k = x \times y \times z$, wobei x, y, z der Kantengröße eines Subtomogrammes entspricht

vorhergehenden Subtomogrammen in den Klassen $1, ..., j-1$ gemittelt werden.

$$GA_1 = CA_1 = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\sum_{P \in CA_1} \mathcal{T}(P,0))}{\sum_{P \in CA_1} \mathcal{T}(W_P,0)}\right)$$

$$\theta_j = argmax_\theta^\Theta \mathcal{S}(GA_{j-1}, CA_j, \theta) \tag{3.4}$$

$$GA_j = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\sum_{k}^{j} \sum_{P \in CA_k} \mathcal{T}(P,\theta_k))}{\sum_{k}^{j} \sum_{P \in CA_k} \mathcal{T}(W_P,\theta_k)}\right)$$

3.2.2 Alignment durch wiederholtes, globales *Sampling*

Initiale Referenzen können direkt aus den Subtomogrammen generiert werden, indem man die Rotationsparameter ρ_i jedes Subtomogrammes i randomisiert. Das resultierende Average ist eine nicht ideale Kugel. Diese Methode ist für globuläre Makromoleküle wie Ribosomen durchführbar und wurde z.B. in [Ortiz et al., 2010] präsentiert. Man kann die Randomisierung auf einen oder zwei Winkel beschränken oder gänzlich vernachlässigen, wenn *a priori* Information über die Orientierung der Makromoleküle bekannst ist. Durch Abschätzung der Oberflächengeometrie einer Membran kann der Suchraum für das Alignment von membrangebundenen Komplexen auf eine planare Rotation eingeschränkt werden. Allerdings haben aus den untransformierten Subtomogrammen generierte, initiale Referenzen jedoch den Nachteil, dass der *Missing Wedge* nicht aufgefüllt wird und dieser während dem Alignment einen starken Einfluss auf das Resultat haben kann.

Wiederholtes, globales *Sampling* Wurde eine initiale Referenz nach einer der präsentierten Strategien generiert, so müssen alle Subtomogramme zuerst global aligniert werden (Kap. 2.3.3), um eine grobe Abschätzung der richtigen Transformation θ_i zu bekommen. Wiederholt man das globale Sampling und aktualisiert A_j, so steigt die Wahrscheinlichkeit, dass θ_i für jedes Subtomogramm in die optimale Konfiguration registriert wird.

3.3 Adaptives Sampling des Alignierungsraumes

Bisher realisierte Alignmentprozeduren basierten auf gleichbleibenden (statischen) Parametern, die vor dem Ablauf durch den Benutzer spezifiziert wurden. Zum einen handelt es sich hier um den Tiefpassfilter $F(A, r)$ und zum anderen um das Winkelinkrement $\Delta\alpha$. Basierend auf Erfahrungswerten, wurden beide Parameter bislang nur subjektiv bestimmt und konnten höchstens bei einem Neustart der Prozedur aktualisiert werden. Um den Tiefpassfilter und das Winkelinkrement von der Eingabe des Benutzers unabhängig zu machen, wurden im Rahmen dieser Arbeit beide Parameter an die in jeder Iteration neu bestimmte Auflösung r_A gekoppelt und so adaptiv bestimmt.

3.3.1 Adaptiver Tiefpassfilter

Der Einfluss vom hochfrequenten Rauschen auf die Auflösung wird minimiert, wenn man den spektralen Bereich, in dem $\mathcal{S}$ berechnet wird, durch einen Tiefpassfilter $F(A, r)$ auf $[0, r_{r_{Cutoff}}(A_j)]$ beschränkt. Um allerdings die Zunahme von r_A durch höherfrequentes Signal zu gewährleisten, muss $F(A, r)$ durch ein zusätzliches Intervall erweitert werden $(1 + \Delta r)$.

$$F(A, r) = \begin{cases} 1, & b \in [0, r_{r_{Cutoff}}(A_j) \cdot (1 + \Delta r)] \\ 0 \end{cases} \tag{3.5}$$

F kann als zusätzlicher Faktor zur Berechnung eines beliebigen Scores im Fourierraum hinzugefügt werden.

3.3.2 Adaptiver Suchwinkel

Während der Optimierung von A kann das Winkel-Sampling angepasst werden [Foerster et al., 2005, Haller, 2008, Winkler et al., 2009]. Die benötigten Parameter müssen jedoch vor Prozessstart spezifiziert werden. Hierbei wird das Winkelgitter um die bereits bestimmte Rotation $\rho_{i,j-1}$ verfeinert. Für die nächste Iteration kann man das Winkelinkrement $\Delta\alpha_j$ halbieren oder anderweitig verfeinern. Es bietet sich deshalb an, nicht nur den Tiefpassfilter $F(A, r)$, sondern auch $\Delta\alpha_j$ an die bereits erreichte Auflösung r_{A_j} zu koppeln.

Adaptives Winkelsampling nach dem Crowther-Kriterium. Das Crowther-Kriterium besagt, dass die in einem rekonstruierten Tomogramm maximale Auflösung r durch die Dicke der Probe D und dem, während der Aufnahme im Elektronenmikroskop, benutzten

Kippwinkel $\Delta\phi$ abhängt [Crowther et al., 1970].

$$r = \frac{1}{D\Delta\phi} \tag{3.6}$$

Ersetzt man in 3.6 $\Delta\phi$ durch $\Delta\alpha_j$, r durch die in der Alignment-Iteration j erreichte Auflösung r_{A_j}, so kann man 3.6 zu

$$\Delta\alpha_j = f \cdot \frac{1}{D_P \cdot r_{A_j}} \tag{3.7}$$

umformen. Basierend auf dem Crowther-Kriterium wird $\Delta\alpha_j$ durch die bereits erreichte Auflösung auf eine sinnvolle Größe reguliert, da $\Delta\alpha$ einerseits für grob aufgelöste Dichten nicht unnötig klein gewählt wird, andereseits für Dichten mit hoher Auflösung entsprechend fein ist. D_P ist der Durchmesser des Makromoleküls, f ist ein Skalierungsfaktor, der dazu dient, ein feineres Winkelsampling zu erreichen ($f \leq 1$), damit höherfrequentes Signal aligniert werden kann.

3.4 Klassifikation von Subtomogrammen durch *Simulated Annealing*

Constrained Principal Component Analysis oder MCO-EM Verwandte sind deterministische Klassifikationsalgorithmen, die, abhängig von der Startverteilung, immer in das selbe Ergebnis konvergieren. Allgemein sind *Expectation Maximization* basierte Algorithmen zuverlässige Optimierungsmethoden, die schnell in eine Lösung konvergieren, welche aber nicht notwendigerweise das globale Optimum ist [Bishop, 2006]. Da alle in Kapitel 2.4.2 aufgeführten Algorithmen auf EM basieren, sind Klassifikationsergebnisse möglicherweise lokale Optima im abgesuchten Lösungsraum. Um zu verifizieren, dass das globale Optimum κ_{opt} gefunden wurde, kann man die Klassifikation aus einer anderen Konfiguration starten. Konvergieren diese erneuten Versuche in ein schlechteres Ergebnis, so steigt die Wahrscheinlichkeit, dass man das globale Optimum bereits gefunden hat. Eine hundertprozentige Sicherheit kann nur durch ein extensives Sampling aller Kombinationen gegeben werden, was allerdings aufgrund der rechnerischen Komplexität nicht möglich ist.

Im Rahmen dieser Arbeit wurde deshalb ein in der KET neuer Klassifikationsansatz entwickelt, in dem *Simulated Annealing* (SA) das Sampling des Lösungsraumes steuert, um aus lokalen Optima zu springen [Kirkpatrick et al., 1983, Černý, 1985]. SA ist ein weitverbreiteter Optimierungsalgorithmus [Aarts et al., 2005] und wurde mit gleicher Motivation bereits erfolgreich für die Bestimmung von Proteinfaltungen eingesetzt [Simons et al., 1997]. Der hier vorgestellte Algorithmus *Multiple Correlation Optimization*

(Annealing) (MCO-A) kapselt die deterministische MCO-EM-Klassifikation ein. Zuerst wird MCO-EM aus einer initialen Konfiguration gestartet. Nachdem dieser Prozess konvergiert ist, wird die Klassenzugehörigkeit jedes Partikels durch MCO-A randomisiert, so dass Klassen mit einem sub optimalen Score für die Klassifikation akzeptiert werden können. Die hierdurch entstandene Konfiguration ist der Start der nächsten MCO-EM-Klassifikation, auf die wieder ein weiterer *Annealing*-Schritt folgt. In den folgenden Iterationen von MCO-A sinkt die Sprungwahrscheinlichkeit eines Partikels in eine schlechtere Klasse.

Nach einer MCO-A Iteration j bestimmt die Sprungwahrscheinlichkeit L_{ij}, ob P_i zu einer schlechteren Klasse k' zugewiesen wird. Wurde P_i bereits Klasse k ($\kappa_{ij} = k$) zugewiesen, so ist die Sprungwahrscheinlichkeit von k nach k' zu springen nach dem Metropolis-Kriterium

$$L_{ij}(k,k') = \begin{cases} 1 & \Delta_i(k,k') \geq 0 \\ exp\left(\frac{-\Delta_i(k,k')}{T_j}\right) & \Delta_i(k,k') < 0 \end{cases} \quad . \tag{3.8}$$

$L_{ij}(k,k')$ hängt von zwei Größen ab:

(i) $\Delta_i(k,k')$ ist die Energiedifferenz der zwei Zustände $\kappa_{ij} = k$ und $\kappa_{ij} = k'$, und hier die Differenz der Scores

$$\Delta_i(k,k') = \mathcal{S}(P_i, CA_{kj}) - \mathcal{S}(P_i, CA_{k'j}) \quad . \tag{3.9}$$

Ist diese Differenz positiv, so wird P_i gleich der Klasse k' zugewiesen, da $L_{ij}(k,k') = 1$ gilt. Bei negativen Differenzen ist $L_{ij}(k,k') < 1$ und das sub optimale k' wird mit der Wahrscheinlichkeit $L_{ij}(k,k')$ akzeptiert.

(ii) T_j ist die Temperatur in Iteration j und wird über alle J Iterationen verringert. L_j ist also proportional zu T; wird T verringert, so sinkt auch die Sprungwahrscheinlichkeit L_{ij}.

$$T_j = s\left(1 - \frac{j}{J}\right) \cdot \sigma(\mathcal{S}_j) \tag{3.10}$$

$\sigma(\mathcal{S}_j)$ ist die Standardabweichung aller in Iteration j berechneten Scores, s ist ein freier Skalierungsparameter. Nach jeder *Annealing*-Iteration wird die beste, bisher gefundene Klassifikation κ_{opt} aktualisiert, so das gilt:

$$\sum_i^N \mathcal{S}(P_i, CA_{\kappa_{opt,i}}) \geq \sum_i^N \mathcal{S}(P_i, CA_{\kappa_{j,i}}) \tag{3.11}$$

Nach jedem *Annealing* Schritt wird die lokale Suche MCO-EM an der in Iteration j bestimmten Konfiguration κ_j gestartet und nach deren Konvergenz SA in Iteration $j+1$

wiederholt.

3.4.1 Konvergenzkriterium für die Klassifikation

Für die Klassifikation wurde zusätzlich ein Konvergenzkriterium eingeführt, um redundante Berechnungen der lokalen Optimierung zu vermeiden. Diese Eigenschaft ist vor allem für große Datensätze von großer Bedeutung. Das Konvergenzkriterium bezieht sich auf die MCO-EM Klassifikation, wird aber auch während der lokalen Suche in MCO-A eingesetzt. Verändert sich die Klassenzuweisung κ zwischen zwei MCO-EM Iterationen für einen gewissen Prozentsatz c der Subtomogramme nicht, so wird MCO-EM abgebrochen.

3.4.2 Klassenvereinigung mittels hierarchischer Klassifikation

Die Bestimmung der Klassenanzahl ist ein bekanntes Problem der Klassifikation und typischerweise auch in der KET *a priori* nicht bekannt. Aus diesem Grund wurde die Klassenanzahl in dieser Arbeit überabgeschätzt. Die MCO-EM-Methode sowie *K-Means*-Klassifikation haben implizit die Möglichkeit, die Klassenanzahl automatisch zu verringern, falls einer Klasse k kein einzelnes Subtomogramm zugewiesen wird. Formal gesehen ist das der Fall, wenn

$$\exists\, k\, \forall\, P_i \mid \mathcal{S}(P_i, CA_k) < \mathcal{S}(P_i, CA_j)\ j \neq k \tag{3.12}$$

In diesem Fall verringert sich K zu $K-1$ in der nächsten Iteration. Dieser Effekt ist hilfreich, falls $K_{opt} < K_{start}$. Die Methoden konvergieren aber nicht notwendigerweise in K_{opt}. Für diesen Fall wird die hierarchische Klassifikation benutzt, um mittels Klassenähnlichkeiten ein Dendrogramm aufzustellen, mit dessen Hilfe Klassen zusätzlich vermengt werden können. Das Dendrogramm wird aus einer Ähnlichkeitsmatrix K generiert, für die typischerweise die NXC oder LNXC als Ähnlichkeitsmaß benutzt werden sollte.

4 Implementierung von PyTom

PyTom wurde entwickelt, um eine benutzerfreundliche Plattform für die KET-Anwender bereitzustellen [Hrabe et al., 2012]. Außerdem sollten Entwickler eine Umgebung vorfinden, die nicht nur den Einstieg und schnellen Erfolg garantiert, sondern leicht erweiterbar und robust ist. Versionsverwaltung ist ein elementarer Bestandteil bei der Entwicklung von PyTom. Die Struktur dieses Kapitels orientiert sich an Abbildung 4.1 und beschreibt die in PyTom vorhandenen Softwareschichten.

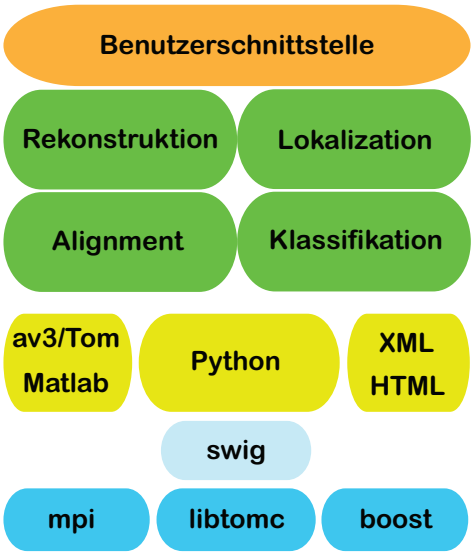

Abbildung 4.1: Schichtendarstellung von PyTom. Im Kern (blau) sind alle zeitkritischen Routinen für die Bildverarbeitung in C/C++ implementiert (libtomc). Außerdem werden extenene Bibliotheken (mpi, boost) eingebunden, um standardisierte Werkzeuge für die Entwicklung bereitzustellen. Prozesse wie *Scores* sind in Python implementiert (gelb). Swig ist die Schnittstelle zwischen libtomc und Python, wodurch kompilierte Funktionen direkt in Python benutzt werden können. Python selbst ermöglicht die Integration von XML und Matlab. Arbeitsabläufe für die Verarbeitung von Tomogrammen findet man in der grünen Schicht. Durch die Benutzerschnittstelle (orange) werden alle implementierten Algorithmen und Routinen angesprochen.

4.1 Numerische Methoden im Kern - *libtomc*

Die numerischen Methoden zur schnellen Datenverarbeitung sind in der Bibliothek *libtomc* implementiert. Das Grundgerüst von *libtomc* wurde in [Haller, 2008] implementiert und war auch die Basis für die in [Stölken et al., 2010] entwickelten Algorithmen. In PyTom wurde die *libtomc* mit den wachsenden Anforderungen an ein vielseitiges Softwarepaket sukzessive erweitert.

4.1.1 Ein- und Ausgabe

Ursprünglich konnten nur in EM gespeicherten KET-Daten [Hegerl, 1996] durch die *libtomc* gelesen werden. Diese Limitierung wurde aufgehoben, so dass nun in der aktuellen Version von PyTom EM, MRC und $CCP4$ gelesen und geschrieben werden können. Diese Funktionalität vereinfacht die Kompatibilität zu anderen Softwarepaketen.

4.1.2 Interpolationsmethoden für die Transformation

Um den bereits in Kapitel 2.3.1 eingeführten Transformationsoperator $\mathcal{T}$ zu implementieren, bedarf es einer Interpolationsmethode, welche die Voxelwerte eines transformierten Subtomogrammes P_t berechnet. In der aktuellen Version von PyTom stehen vier Interpolationsmethoden zur Verfügung, mit denen eine Transformation durchgeführt werden kann (*pytom.basic.transformation*).

- *Tri-Linear* (http://mathworld.wolfram.com/LagrangeInterpolatingPolynomial.html)
- *Tri-Cubic* (http://mathworld.wolfram.com/LagrangeInterpolatingPolynomial.html)
- *Tri-Cubic-Spline* (http://mathworld.wolfram.com/CubicSpline.html)
- *Tri-Fourier-Spline*

Die Wahl der richtigen Interpolationsmethode sollte nach der Laufzeit (Größe des Interpolations-Kernels) und der gewünschten Genauigkeit erfolgen, da die höhere Genauigkeit eine längere Laufzeit benötigt.

Fourier-Spline-Interpolation. Die Fourier-Spline-Interpolationsmethode unterscheidet sich von den anderen Methoden dahingehend, dass sie im Fourierraum berechnet wird. Hierfür wird Real- und Imaginärteil unabhängig voneinander rotiert und mit Hilfe der *Cubic Spline*-Methode interpoliert. Eine Translation kann nur durch eine Phasenverschiebung der Fourierkoeffizienten vor oder nach der Rotation implementiert werden. Aufgrund der zusätzlichen Transformation in den Fourierraum, ist die Fourier-Spline Interpolation die langsamste Interpolationsmethode in PyTom, allerdings auch die genaueste (Abb.

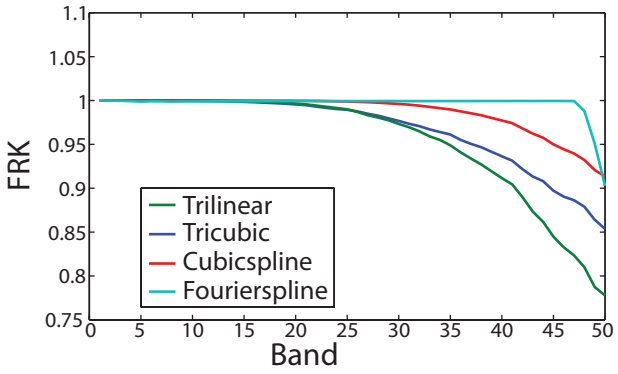

Abbildung 4.2: Konsistenzbestimmung der einzelnen Interpolationsmethoden. Es wurde ein $100 \times 100 \times 100$ großes Volumen $1\times$ rotiert und das Zwischenergebnis wieder in die originale Position zurücktransformiert. Die Konsistenz des jeweiligen Resultates wurde mit dem originalem Volumen durch Fourier-Ring-Korrelation bestimmt.

4.2). Um die Laufzeit jedoch zu optimieren, kann man die Rotation im Fourierraum aufgrund der hermiteschen Symmetrie nur in einer Hälfte des Volumens berechnen. Der jeweils entsprechnende Punkt bezüglich der Achsensymmetrie wird als konjugiert komplex des bereits bestimmten Wertes gesetzt.

4.1.3 Filter

Der Bandpass- und der *Missing Wedge*-Filter wurde erweitert. Beide Objekte sind notwendig, um die Berechungen im Fourierraum auf einen sinnvollen Bereich einzuschränken, wie in Kapitel 3.3 beschrieben. Der Bandpassfilter F (*pytom.basic.filter*) entspricht in seiner Spezifikation dem Bandpassfilter aus der TOM Toolbox und unterstützt ebenfalls einen Gaußschen Abfall an den Filterenden. Der *Missing Wedge* Filter (*pytom.basic.structures.WedgeInfo*) ist gegenüber der TOM Implementierung genauer geworden. Die PyTom Implementierung des *Missing Wedge* Filters unterstützt ebenfalls den Gaußschen Abfall an den Filterenden, wie auch eine asymmetrische Kippgeometrie, um den fehlenden Bereich im Fourierraum genauer zu maskieren.

4.2 Skripte in PyTom

Um die Lesbarkeit der in PyTom implementierten Algorithmen zu gewährleisten, wurden alle komplexen Abläufe in der Skriptsprache *Python* implementiert. Durch die Softwarebrücke *Swig* werden alle in *libtomc* implementierten Methoden in *Python* angesprochen. Diese Art von Architektur (Abb. 4.1) ist in wissenschaftlichen Softwarepaketen zuneh-

mend verbreitet [Sorzano et al., 2004, Tang et al., 2007], um eine möglichst hohe Transparenz und Benutzerfreundlichkeit für Laien zu garantieren.

4.2.1 Strukturierung von PyTom

Während der Entwicklung von PyTom lag das Hauptaugenmerk auf der Einfachheit und Lesbarkeit der generierten Skripte, damit fachfremde Wissenschaftler einen möglichst leichten Einstieg in das Verständnis von PyTom-Abläufen haben. Skripte und Algorithmen wurden deshalb nach Möglichkeit thematisch sortiert und in Python-Modulen gekapselt. Alle anderen in PyTom implementierten Methoden verarbeiten, die im Modul *py-*

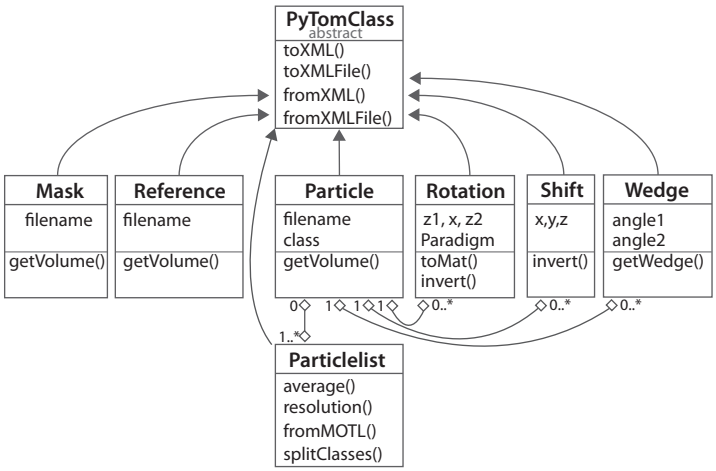

Abbildung 4.3: Das hier dargestellte UML-Diagramm beschränkt sich auf die wichtigsten PyTom-Klassen und ist somit ein kleiner Ausschnitt aus dem *pytom.basic.structures* Modul. Der Kern von PyTom ist die Klasse *PyTomClass*, von der fast alle Klassen in PyTom abgeleitet sind. Durch diesen Mechanismus erben Kinder der *PyTomClass* die Methoden während der Laufzeit in XML und aus XML übersetzt zu werden. Die *Particle*-Klasse, die zusätzlich die *Rotation*, *Shift* und den *Wedge* aggregiert, ist eine der am meisten benutzten Klassen in PyTom.

tom.basic.structures bereitgestellten, Klassen. In *pytom.reconstruction* findet man alle für die Rekonstruktion benötigten Algorithmen, in *pytom.basic.transformations* sind alle verfügbaren Transformationsroutinen gespeichert, u.s.w.. Durch die Module *pytom_volume*, *pytom_fourier* und *pytom_filter* können die in *libtomc* implementierten Routinen direkt angesprochen werden.

4.2.2 Datenspeicherung in PyTom

Nicht-nummerische PyTom-Daten werden in *eXtensible Markup Language* (XML) gespeichert, um möglichst hohe Transparenz und Kompatibilität zu anderen Softwarepaketen

anzubieten. Außerdem ermöglichen XML verwandte Sprachen wie *eXtensible Style Sheet Language* (XSLT) *http://www.w3.org/standards/xml/transformation* die Transfrormation von XML Daten in das HTML-Format für die Präsentation und *XML Path Language* (XPath) *http://www.w3.org/standards/techs/xpath* ermöglicht das Auslesen von XML wie für eine reguläre Datenbank. Ein weiterer Vorteil bei der Übersetzung von PyTom-Objekten in XML wird bei der Parallelisierung deutlich.

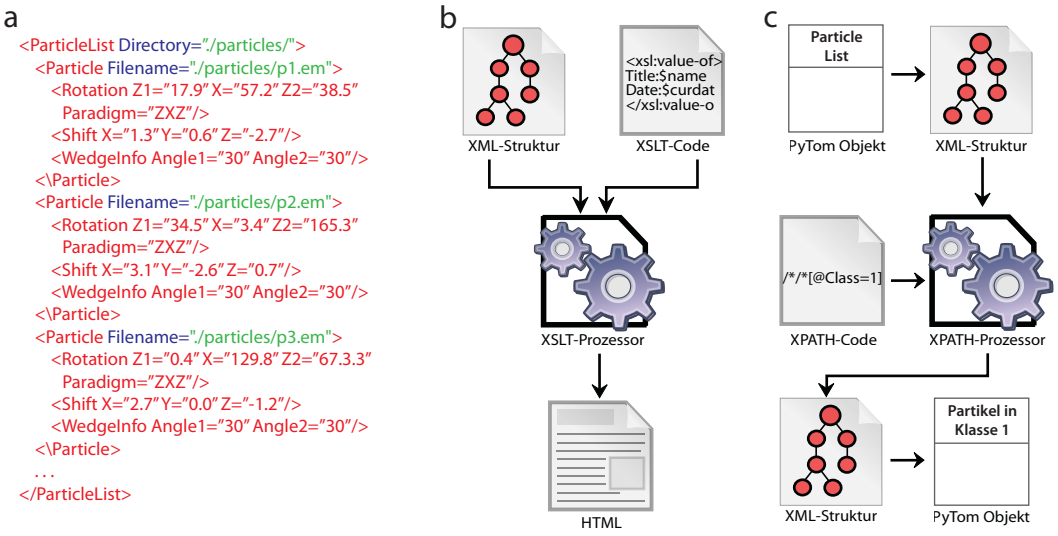

Abbildung 4.4: (a) Die Klasse *ParticleList* in XML übersetzt. Jeder Parameter von jedem Subtomogramm ist leicht lesbar und editierbar, sofern die XML-Struktur als Datei vorliegt. (b)Transformation von XML nach HTML mittels XSLT. (c) Ablauf einer XPath-Anfrage an eine Partikelliste, um alle Subtomogramme der Klasse 1 zu bekommen.

4.2.3 Parallelisierung

Implementierung. Das Modul *pytom.parallel* beinhaltet eine Klasse *ParallelWorker*, mit deren Hilfe sich die Verteilung von Berechnungen auf verschiedene Prozessoren einfach lösen läßt. Da für die parallele KET-Datenverarbeitung keine komplexen Kommunikationen zwischen einzelnen Rechenknoten benötigt werden, ist die Kommunikationstopologie sternförmig: ein zentraler Knoten (*Master*) verteilt Aufgaben an alle anderen (*Worker*). Um eine Berechnung zu parallelisieren, muss man lediglich eine Klasse für die Aufgabenbeschreibung erstellen, eine neue Klasse von *ParallelWorker* ableiten und abschliessend spezifizieren, wie die Aufgabe verarbeitet werden muss. Die Verteilung der Aufgaben ist bereits in *ParallelWorker* implementiert. Alle in PyTom parallelisierten Algorithmen sind nach dieser Topologie implementiert.

Serialisierung. Wie bereits erwähnt, ist XML für die Parallelisierung sehr hilfreich, da mit der Konvertierungsfähigkeit jedes PyTom-Objekts nach XML eine Möglichkeit bereitgestellt wird, um Aufgabenbeschreibungen einheitlich zu verschicken. Die Kommunikationssprache zwischen *Master* und *Worker* ist XML. Hierfür wird jedes zu sendende PyTom-Objekt in XML und dann in eine Zeichenkette übersetzt. Beim Empfang wird die Zeichenkette wieder in XML und in ein PyTom-Objekt konvertiert. Auf diese Weise wird sichergestellt, dass die gesendete Nachricht der empfangenen Nachricht entspricht.

4.2.4 Implementation des Winkel-Samplings

In Kapitel 2.3.3 wurden im Wesentlichen zwei Sampling-Strategien präsentiert: das globale und das lokale Sampling. Das globale Sampling wurde durch die Klasse *pytom.angles.fromFile.AngleListFromEM* realisiert, die gespeicherten Listen liegen jeweils im EM-Format vor. Das lokale Sampling wird dynamisch durch die Klasse *pytom.angles.angleList.LocalList* berechnet, allerdings muss man vorher die Anzahl der Ringe und eine feste Rotation spezifizieren. Objekte dieser Klassen können eigenständig für das Alignment von Subtomogrammen benutzt werden. Des Weiteren wurden zusätzliche Klassen implementiert, die von den beiden Sampling-Strategien abgeleitet sind. Die Klasse *pytom.angles.combined.GlobalLocalCombined* zum Beispiel ermöglicht den Wechsel zwischen beiden Sampling-Strategien in der jeweils nächsten Alignment-Iteration. Alternative Strategien können durch Ableitung von den bereits vorhandenen Klassen realisiert werden.

4.3 Algorithmen

4.3.1 Rekonstruktion von Tomogrammen

Alle in dieser Arbeit beschriebenen Rekonstruktionsalgorithmen sind durch das Modul *pytom.reconstruction* aufrufbar. Hier findet man auch zwei für die Rekonstruktion in PyTom wichtigen Klassen. In der Klasse *Projection* werden alle Informationen über eine Projektion gespeichert. Die *ProjectionList* speichert eine Sammlung von Projektionen und kann eine ganze Kippserie, wie auch mehrere tausend Projektionen eines einzelnen Partikels, repräsentieren.

Gewichtete Rückprojektion. Die gewichtete Rückprojektion wurde nach dem gleichen Prinzip wie in TOM implementiert, allerdings wurde der Programmcode hierfür neu aufbereitet und die Lesbarkeit verbessert. Nach dem in Kapitel 4.2.3 präsentierten Prinzip wurde die Rekonstruktion von einzelnen Subtomogrammen parallelisiert.

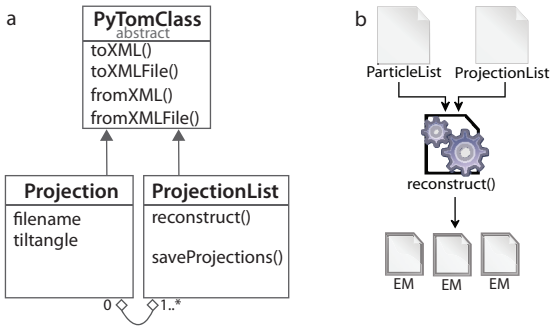

Abbildung 4.5: Wie alle Klassen in PyTom erben die Klassen *Projection* und *ProjectionList* die für die Serialisierung benötigten XML-Eigenschaften.

4.3.2 Lokalisierung von Subtomogrammen

Im Modul *pytom.localization* findet man alle für die Lokalisierung von Subtomogrammen (Kap. 2.6) benötigten Prozesse. Die Lokalisation ist nach dem in Kapitel 4.2.3 präsentierten Prinzip parallelisiert. Allerdings wurden hier drei verschiedene Lastverteilungen implementiert, deren theoretische und tatsächliche Laufzeit evaluiert wurden. Diese Strategien wirken sich jedoch nicht auf die Sensitivität der Lokalisierung aus, sondern nur auf die Laufzeit.

***Split Angles*-Verteilung.** Soll die Lokalisierung auf c *Worker* verteilt werden, so ist das Verteilen der abzusuchenden Winkelliste die einfachste Lastverteilungsstrategie. Hierfür wird die Winkelliste in c Teile geschnitten und auf die einzelnen *Worker* verteilt. Sind alle *Worker* fertig, so müssen die finalen Einzelergebnisse zu einem Endergebnis zusammengefügt werden. Für jedes Voxel im Tomogramm wird der Winkel mit dem höchstem Score gespeichert.

***Split Volume*-Verteilung.** Eine speichereffiziente Methode ist das Aufteilen des Tomogrammes selbst. Jeder *Worker* erhält so eigenständig einen Ausschnitt aus dem Tomogramm. Die Winkellisten werden bei dieser Methode nicht geteilt. Bei der *Split Volume*-Verteilung muss allerdings immer ein überlappender Teil o zusätzlich berechnet werden, damit beim finalen Vereinigen der Einzelergebnisse ein kontinuierliches Gesamtergebnis entsteht. Die Größe von o ist deshalb

$$\begin{aligned} o = (S_x - 1)v_x v_z r_x + \\ (S_y - 1)(v_x v_z r_y - (S_x - 1)r_x r_y v_z) + \\ (S_z - 1)(v_x v_y r_z - (S_x - 1)r_x r_z v_y - (S_y - 1)r_y r_z v_x) \end{aligned} \tag{4.1}$$

S_x, S_y, S_z repräsentiert die Anzahl der Subvolumen entlang jeder Dimension, v_x, v_y, v_z die Größe des Tomogrammes und r_x, r_y, r_z die Größe der benutzten Referenz. Der Speicherverbrauch geht mit den nun kleineren Tomogrammen zurück, allerdings wird o redundant Fouriertransformiert, was sich nachteilig auf die Gesamtlaufzeit auswirkt.

***Split Hybrid*-Verteilung.** Dieser Mechanismus vereinigt die beiden bisher vorgestellten Methoden. Die Winkelliste sowie das Tomogramm werden aufgeteilt. Das Tomogramm muss allerdings nicht unbedingt in c Subvolumina aufgeteilt werden. Hier werden zuerst alle Winkellisten für ein Subvolumen verteilt und die Teilergebnisse verglichen. Erst dann wird das nächste Subvolumen prozessiert. Der Speicherverbrauch sinkt durch das Aufteilen des Tomogrammes. Die Komplexität des Zusammenschreibens aller Einzelergebnisse nimmt allerdings zu.

Theoretische Laufzeitabschätzung. Die Laufzeit der drei Lastverteilungsstrategien kann für ein Tomogramm mit n Voxeln theoretisch abgeschätzt werden. Wird das Tomogramm in m Subvolumina aufgeteilt, und die Winkelliste mit a Winkeln in s Listen geteilt, so ist die Anzahl der benötigten Operationen:

$$N_{OP}(a, s, n, m, o) = \frac{a}{s} \cdot (\frac{n+o}{m} \cdot log_2 \frac{n+o}{m} + \frac{n+o}{m} + \frac{n}{m}) + \frac{n \cdot s}{m} \quad . \tag{4.2}$$

Bezüglich der Laufzeit ist die Komplexität der Fouriertransformation $\frac{n+o}{m} \cdot log_2 \frac{n+o}{m}$ die teuerste aller Operationen. Die Anzahl der für jedes Subvolumen benötigten Score-Operationen ($\frac{n+o}{m}$) und die Anzahl der benötigten Vergleiche ($\frac{n}{m}$) liegt unter der Komplexität der Fouriertransformation. Das finale Vereinigen der Einzelergebnisse kostet $\frac{n \cdot s}{m}$ Operationen.

Die Laufzeit kann für die *Split Angles*-Methode als $N_{OP}(a, c, n, 1, 0)$ abgeschätzt werden. Für die *Split Volume*-Strategie wird die Laufzeit durch $N_{OP}(a, 1, n, c, o)$ approximiert, wobei o nach Formel 4.1 bestimmt werden muss. Letztlich kann die Laufzeit für *Split Hybrid* als $N_{OP}(a, c, n, m, o) \cdot m$ angenähert werden.

Die theoretische Approximation der Laufzeit hat ergeben, dass die *Split Angles*-Strategie die schnellste Verteilungsstrategie für die Lokalisation sein sollte (Abb. 4.6).

Laufzeitmessung der Verteilungsarten. Die drei Lastverteilungsstrategien wurden unabhängig auf einem Computer-Cluster getestet, um deren tatsächliche Perfomanz zu messen. Die getestete Anzahl an Rechenknoten c lag zwischen 8 und 64 CPUs. Jeder einzelne Knoten hatte eine Taktfrequenz von $2.54 MHz$ und einen für sich verfügbaren Arbeitsspeicher von $2.75 GB$. Das identische Lokalisationsproblem wurde mehrfach mit einer variierenden Anzahl an Knoten prozessiert und die jeweilige Laufzeit gemessen (Abb. 4.6). Wie

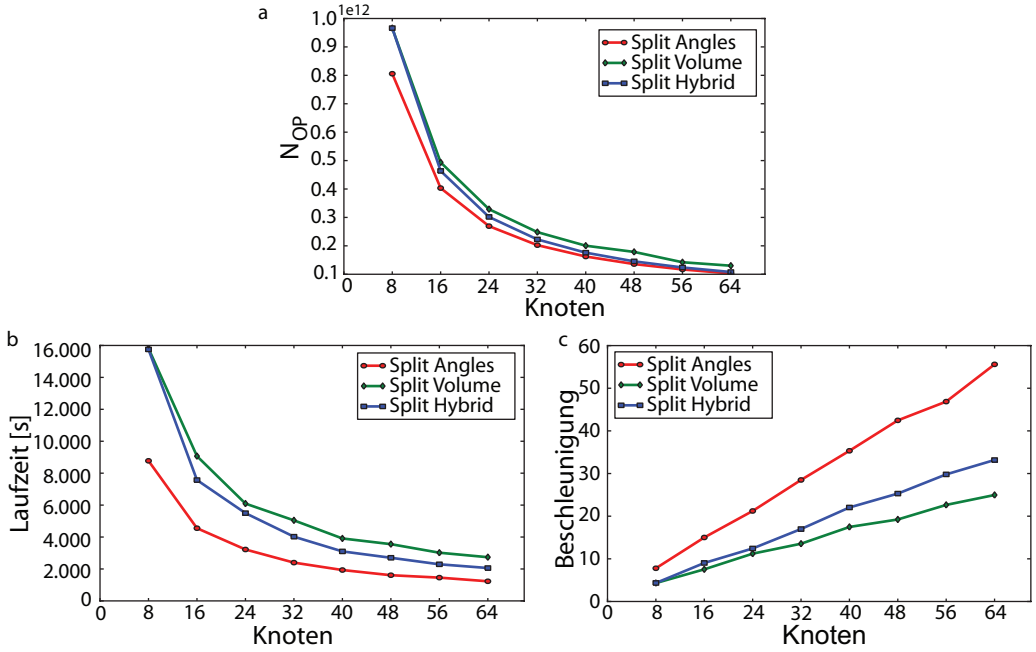

Abbildung 4.6: (a) Die theoretisch bestimmte Anzahl an Rechenoperationen für die drei Lastverteilungsstrategien: *Split Angles*, *Split Volume* und *Split Hybrid*. (b) Die tatsächlich gemessene Laufzeit auf einem Computercluster. (c) Die basierend auf der Laufzeit bestimmte Beschleunigung.

bereits theoretisch approximiert war die *Split Angles*-Strategie die schnellste Variante, da hier keine redundanten Berechnungen durch eine Überlappung auftraten. Gegenüber der *Split Volume*-Strategie ist *Split Hybrid* allerdings vorzuziehen, da diese schneller ist (Abb. 4.6) und weniger Speicher verbraucht. Zudem verzögert sich die Ausführung der beiden langsameren Strategien zusätzlich durch die erhöhte Anzahl an redundanten Leseoperationen. Die gemessene Beschleunigung [Tanenbaum, 2002] lag für *Split Angle* bei ca. $0.8 \cdot c$, wohingegen die Beschleunigungen für die anderen beiden Strategien ca. $0.5 \cdot c$ (*Split-Hybrid*) und ca. $0.3 \cdot c$ (*Split-Volume*) betrugen.

4.3.3 Alignment von Subtomogrammen

Das Alignment von Subtomogrammen (Kap. 2.3.4) ist im PyTom-Modul *pytom.alignment.ExMaxAlignment* implementiert. Das Alignment mit integrierter MCO-EM-Klassifikation (Kap. 2.4.2) steht in dem Modul *pytom.alignment.MultiRefAlignment* zur Verfügung. Das adaptive Sampling (Kap. 3.3) kann wahlweise aktiviert oder deaktiviert werden.

Beschleunigung durch parallelisiertes Alignment. Die Alignmentroutinen wurden ebenfalls parallelisiert, um ein möglichst schnelles Alignment zu garantieren. Die Alignmentprozedur besteht aus zwei Hauptprozessen, das Alignment selbst (Kap. 2.3.4) und die Mittelung der Subtomogramme (Kap. 2.3.1). Beide Prozesse wurden, wie in Kapitel 4.2.3 beschrieben, parallelisiert. Da für das Alignment und die Mittelung praktisch keine Kommunikation zwischen den Knoten stattfindet, sollte die Beschleunigung durch c CPUs proportional zu c sein. Ein Alignment mit 100 relativ kleinen Subtomogrammen ($32 \times 32 \times 32$ Voxel) über drei Alignmentiterationen wurde auf einem Rechenknoten mit maximal 7 CPUs getestet. Die Taktfrequenz jedes einzelnen CPUs betrug 2.8 GHz. Abbildung 4.7 bestätigt die Annahme, dass das Alignment zu fast einhundert Prozent parallelisierbar ist.

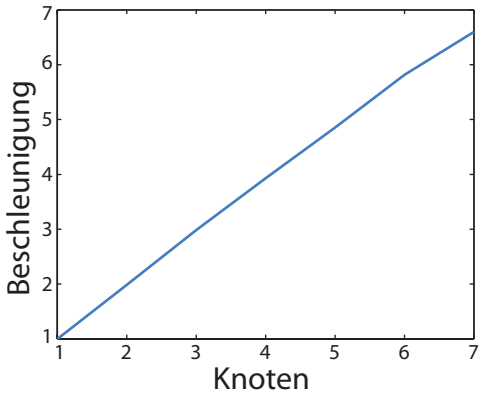

Abbildung 4.7: Die Beschleunigung des parallelisierten Alignments ist proportional zur Anzahl der benutzten Rechenknoten.

4.3.4 Klassifikation von Subtomogrammen

Die implementierten Klassifikationsalgorithmen können im Modul *pytom.cluster* gefunden werden. Für beide Algorithmen existieren eigenständige Klassen, wobei die MCO-A-Klassen von den MCO-EM-Klassen abgeleitet sind. CPCA ist ebenfalls in PyTom (*pytom.cluster*) enthalten, allerdings wurde in dieser Arbeit die Implementation aus AV3 in Matlab benutzt, um Implementationsunterschiede, die sich nachteilig auf die Genauigkeit auswirken könnten, auszuschliessen. Die gemessene Genauigkeit wird in Kapitel 5.3 geschildert.

Theroetische Laufzeitanalyse der Klassifikationsmethoden. Die MCO-Klassifikationsmethoden unterscheiden sich von der CPCA nicht durch die Klassifikationsergebnisse, sondern auch durch die Laufzeit. Im Gegensatz zur Lokalisation ist

die Bestimmung der Laufzeit der einzelnen Klassifikationsmethoden viel einfacher, da hier keine Subvolumina verteilt werden müssen. In der CPCA-Klassifikation ist der kostenintensivste Schritt die Bestimmung der Ähnlichkeitsmatrix aller N Subtomogramme. Da für jedes Subtomogramm ein Score mit jedem anderen Subtomogramm bestimmt werden muß , ist man gezwungen N_S Scores zu berechnen.

$$N_{S,CPCA} = \frac{N(N-1)}{2} \tag{4.3}$$

Für die MCO-EM-Klassifikation beschränkt sich $N_{S,MCO-EM}$ auf

$$N_{S,MCO-EM} = J_{MCO-EM} \cdot K \cdot N \qquad . \tag{4.4}$$

J_{MCO-EM} ist die Anzahl der Iterationen und K die Anzahl der Klassen. Für MCO-A wird dieser Term um J_{MCO-A} erweitert.

$$N_{S,MCO-EM} = J_{MCO-A} \cdot J_{MCO-EM} \cdot K \cdot N \tag{4.5}$$

Somit kann die Laufzeit mit Hilfe der Landau-Symbole für CPCA auf $O(CPCA) = N^2$ und $O(MCO-A) = N$ approximiert werden [Steger, 2002]. Da die Laufzeit proportional zu der Partikelzahl ist, die in der KET in der Regel über tausend liegt, sind die auf MCO basierenden Methoden der CPCA vorzuziehen. Der Effekt, dass Klassen währen MCO-EM nach Möglichkeit vereinigt werden, wirkt sich ebenfalls positiv auf die Gesamtlaufzeit aus, da K in den Folgeiterationen schrumpft.

4.4 Die Benutzerschnittstelle von PyTom

In der Benutzerschnittstelle verdeutlicht sich der Vorteil, XML zur Datenspeicherung zu verwenden. Wie in Kapitel 4.2.2 beschrieben, können XML Dateien mittels XSLT in HTML-Daten transformiert werden. Deshalb wurden die Benutzeroberfläche an den HTML5-Standard angepasst. Ein weiterer Vorteil der für die Benutzung von HTML als Benutzerschnittstelle spricht, ist die Plattformunabhängigkeit, da HTML auf unterschiedlichen Systemen identisch angezeigt werden kann. Jeder in PyTom implementierte Algorithmus hat eine eigenständige HTML-Seite, durch die eine Rekonstruktion, Lokalisierung, Alignment oder Klassifikation spezifiziert werden kann. Der Benutzer kann somit Aufgaben generieren oder bereits existierende XML-Dateien laden und modifizieren.

PyTom Client-Server-Struktur. Die HTML-Benutzerschnittstelle ist nach dem klassischem Client-Server-Prinzip implementiert. Die Sprache *Python* stellt Klassen (*http.server.BaseHTTPServer*) für eigenständige Webserver bereit, die auch in PyTom

Verwendung fanden. Da der Server so unmittelbar Zugriff auf die in PyTom implementierten Klassen und Prozesse hat, kann man Daten direkt im Browser anzeigen und mit PyTom-Funktionen interagieren (Abb. 4.8). Der Standard zur asynchronen Kommunika-

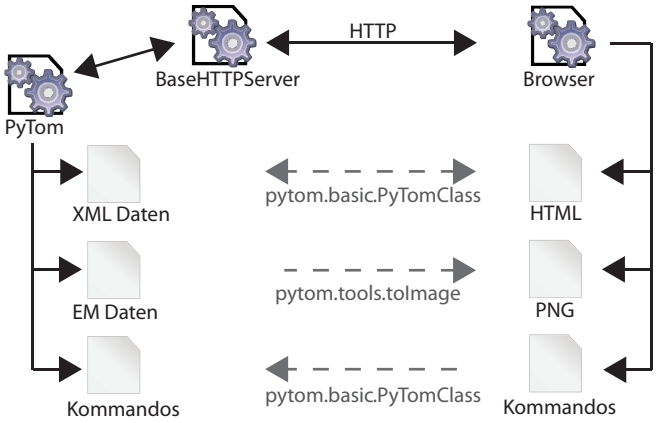

Abbildung 4.8: Schematischer Kommunikationsablauf zwischen Browser (rechts) und dem PyTom Webserver (links). Da der Server direkt Zugang zu den, in PyTom implementierten, Klassen und Prozessen hat, können KET spezifische Daten direkt im Browser angezeigt werden. Da die datenspezifischen Aufgaben für den Browser unsichtbar sind, verarbeitet dieser nur gängige Datenformate wie HTML-Webseiten und Bilder im PNG-Format.

tion zwischen einem Browser und eine HTTP-Server ist JavaScript mit AJAX (*Asynchronous JavaScript and XML*). Diese Technologie ermöglicht die Datenübertragung von Aufgabenbeschreibungen zwischen Browser und Server und wurde für die Benutzerinteraktion verwendet.

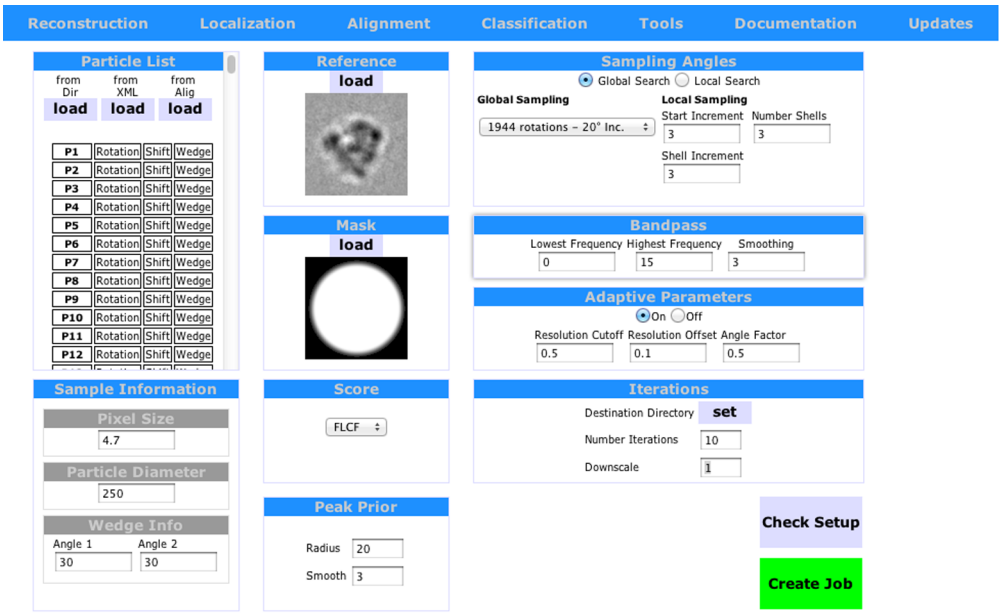

Abbildung 4.9: Eine typische PyTom Benutzeroberfläche. Hier die Alignment-Maske, durch die das Alignment von Subtomogrammen spezifiziert werden kann. Der Browser interagiert mit dem PyTom-Server und läd Daten sowie Bilder nach, wenn diese von dem Benutzer angefordert wurden.

5 Prozessierungsergebnisse der in PyTom implementierten Methoden

5.1 *De novo* Referenzen generiert aus *S. cerevisae*-Lysat-Tomogrammen

In diesem Kapitel wurden die zwei vorgestellten Methoden für die *de novo* Generierung von Referenzen aus Subtomogrammen der *S. cerevisae*-Lysat Tomogramme (Kap. 3.1.3) getestet: die Generierung von Referenzen anhand von Rotationsklassen sowie durch Alignment mit wiederholtem, globalem Winkel-Sampling.

5.1.1 *De novo* Referenzen durch Rotationsklassen

Die Bestimmung der initialen Referenz durch Rotationsklassen (Kap. 3.2.1) wurde mit 80S-Ribosomen aus den *S. cerevisae*-Lysat Tomogrammen getestet. Durch die Verwendung eines 80S-Ribosoms als Alignment-Referenz konnten die Transformationsparameter der Subtomogramme nahezu perfekt bestimmt werden. Die so bekannten Rotationsparameter wurden als Grundlage für die statistische Analyse ($\overline{\rho_{opt}}$) der Rotationsklassifikation benutzt. Als Gütemaß wurde der Rotationsabstand $\overline{\rho_k}$ aller Subtomogramme innerhalb einer Klasse k verwendet.

Analyse der Rotationklassen. Wie bereits in Kapitel 3.2.1 erwähnt, hat man an zwei Stellen die Möglichkeit, durch Adaption von Parametern, Einfluss auf die Klassifikation zu nehmen. Einerseits kann man den Tief- oder Bandpassfilter vor der Klassifikation der Subtomogramme variieren, andererseits kann die Klassenzahl unterschiedlich gewählt werden. Unter der Annahme, dass die durchschnittliche Klassengröße von mindestens dreißig Subtomogrammen nicht unterschritten werden sollte, da sonst das SNR in der Klasse für das spätere *Growing Average*-Alignment zu niedrig ist, wurde die Klassenanzahl auf $K = 60$ festgelegt. Nach jedem Versuch wurde jeweils der durchschnittliche Winkelabstand $\overline{\Delta\rho}$ aller, einer Klasse zugewiesenen, Subtomogramme bestimmt. [1] Bei einer zufälligen Verteilung aller Subtomogramme zu 60 Klassen war $\overline{\Delta\rho} = 127°$. Um die

[1] Die Distanz zweier Rotationen $\Delta\rho$ wurde im Quarternionenraum berechnet [Kuffner, 2004].

bestmögliche Klassenbelegung zu bestimmen, wurde eine Distanzmatrix aller ρ_{opt} erstellt und mittels *K-Means* zu 60 Klassen klassifiziert. Hierbei betrug $\overline{\Delta\rho}_{opt} = 33°$. Der maximale Winkelabstand innerhalb einer Klasse betrug $\Delta\rho_{max} = 77°$. In Tabelle 5.1 ist $\overline{\Delta\rho}$

Tabelle 5.1: Durschschnittliche Rotationsabstände nach *K-Means* mit $K = 60$. Die obere Zeile entspricht dem verwendeten Tiefpassfilter (in $Å^{-1}$). Die Spalte gibt den verwendeten Hochpassfilter (in $Å^{-1}$) aller Subtomogamme vor der Linearisierung an. 0 bedeutet: kein Hochpassfilter wurde verwendet.

	67^{-1}	58^{-1}	52^{-1}	47^{-1}	42^{-1}
0	105°	104°	104°	106°	104°
470^{-1}	108°	108°	107°	107°	107°
235^{-1}	117°	117°	118°	116°	117°

nach der Klassifizierung durch die Singulärwertzerlegung und *K-Means* zu 60 Klassen aufgetragen. Die bestimmten Werte lagen deutlich über $\overline{\Delta\rho}_{opt}$ und auch über $\Delta\rho_{max}$. In keiner der resultierenden Klassen wurde eine mit $\overline{\Delta\rho}$ von weniger als 60° bestimmt. Im Anschluss wurde die Klassifikation wiederholt, diesmal allerdings mit einer variierenden Klassenanzahl (Tab. 5.2). Der Bandpassfilter während der Linearisierung der Subtomogramme wurde auf das Intervall $[0; 58^{-1}]Å^{-1}$ eingestellt.

Tabelle 5.2: Durschschnittliche Rotationsabstände nach *K-Means*-Klassifikation bei variierender Klassenanzahl K.

K	40	50	60	70	80	90	100
	107°	106°	104°	105°	103°	102°	102°

Die in Tabelle 5.2 aufgetragenen Werte lagen ebenfalls deutlich über $\overline{\Delta\rho}_{opt}$ sowie über $\Delta\rho_{max}$, so dass in beiden Experimenten keine zufriedenstellende Verteilung bezüglich der Rotation der Subtomogramme bestimmt werden konnte.

Vom *Growing Average* zu einer Startreferenz. Weiterhin wurde das *Growing Average* für die Bestimmung einer initialen Alignment-Referenz durchgeführt. Der *Growing Average*-Algorithmus wurde für die optimalen Rotationsklassen bei $K = 60$ sowie für die bestimmten Rotationsklassen bei $K = 60$, Bandpass-Filter bei $[0; 58^{-1}]Å^{-1}$ erstellt. In jeder Iteration wurden die Rotationen aus einer globalen Winkelliste mit $\Delta\alpha = 38°$[2] abgesucht.

[2]Ebenso wurden Winkel mit $\Delta\alpha = 25°$ und $\Delta\alpha = 19°$ abgesucht, allerdings mit ähnlichem Ergebnis

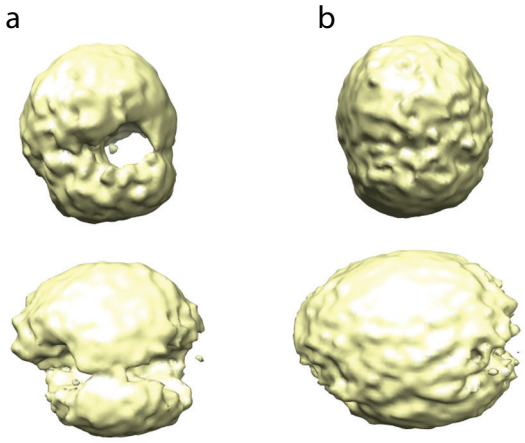

Abbildung 5.1: Die nach Rotationsklassifikation und *Growing Average* erstellten, initialen Referenzen. (a) Basierend auf einer optimalen Belegung der Rotationsklassen ähnelte das finale *Growing Average*-Ergebnis einem Ribosom. (b) Das finale *Growing Average* Ergebnis basierend auf Rotationsklassen, welche durch Singulärwertzerlegung und *K-Means*-Klassifikation bestimmt wurden.

Das finale *Growing Average*-Ergebnis basierend auf einer optimalen Belegung der Rotationsklassen ähnelte einem Ribosom (Abb. 5.1). Zumindest konnte man den mRNS Tunnel zwischen der großen und der kleinen Untereinheit ausmachen. Im Gegensatz hierzu konnten im *Growing Average*-Ergebnis basierend auf den durch Singulärwertzerlegung und *K-Means*-Klassifikation bestimmten Rotationsklassen keine markanten, strukturellen Eigenschaften des Ribosoms erkannt werden.

5.1.2 *De novo* Referenzen durch wiederholtes, globales Winkel-Sampling

Eine Referenz wurde ebenfalls durch Alignment mit wiederholtem, globalem Winkel-Sampling (Kap. 3.2.2) generiert. Wie auch im Rotationsklassen-Experiment wurde der selbe Datensatz, bestehend aus reinen Ribosomen, benutzt. Die Rotationsparameter ρ_i wurden vorab für jedes Subtomogramm randomisiert (Gleichverteilung) und eine initiale Referenz wurde aus der zufälligen Verteilung generiert. Die Winkel-Sampling-Strategie (*pytom.angles.combined.GlobalLocalCombined*) alternierte nach jeder Iteration zwischen globalem und lokalem Sampling. Man konnte nach fünf Alignment-Iterationen eine deutliche Ähnlichkeit des Averages zum 80S-Ribosom ausmachen (Abb. 5.2), nach zehn Iterationen wurde die Auflösung des Averages auf ca. 45^{-1}Å^{-1} ermittelt.

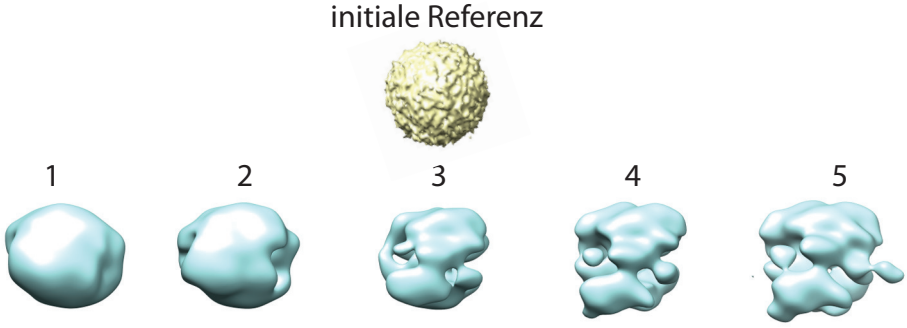

Abbildung 5.2: Nach fünf Alignment-Iterationen mit wiederholtem, globalemn Winkel-Sampling des Datensatzes ohne kolloide Goldpartikel hatte der entstandene Average eine deutliche Ähnlichkeit zum Ribosom.

5.2 Alignment mit adaptivem Sampling

Um den Einfluss des adaptiven Samplings (Kap. 3.3) auf das Alignment von Subtomogrammen (Kap. 2.3.4) zu testen, wurden Subtomogramme von an *S. Scerevisae*-Mikrosomen gebundenen 80S-Ribosomen (Kap. 3.1.4) prozessiert.

Lokalisierung der ribosomalen Einheiten. Mit der Struktur des *Homo sapiens* Ribosoms (EMDB-ID: 1093) wurden Ribosome im Mikrosomen-Tomogramm lokalisiert. Es wurden insgesamt 800 Subtomogramme mittels der gewichteten Rückprojektion (Kap. 2.2) rekonstruiert und ein Average A_{pre} bestimmt. Die Rotationsparameter für den vorläufigen Average waren die durch die Lokalisation bestimmten Winkel.

Alignment mit adaptivem Sampling. Für den Test der Alignment-Strategie mit aktiviertem, adaptivem Sampling wurden mit der Startreferenz A_{pre} zehn mal iteriert. Dabei wurde zuerst eine globale Winkelabtastung durchgeführt und θ anschliessend mittels lokalem Winkelsampling optimiert. Die für die adaptive Abtastung notwendigen Parameter waren $r_{cutoff} = 0.5$, $\Delta r = 0.1$, $f = 0.25$. Des Weiteren wurden diese Werte als Standardeinstellung in das Alignment durch PyTom festgelegt. Das Alignment der 800 Subtomogramme konvergierte in eine Dichte mit einer Auflösung von 40^{-1}Å^{-1}.

Alignment mit statischem Sampling. Für den direkten Vergleich beider Methoden wurden die Subtomogramme mit deaktiviertem, adaptivem Sampling wiederholt prozessiert. Der statische Tiefpassfilter lag bei 18.8^{-1}Å^{-1}, das statische Winkelinkrement war $\Delta\alpha = 3°$ und entsprach somit dem besten, von der adaptiven Strategie bestimmten Parameter. Bis auf diese Unterschiede waren alle anderen Parameter identisch zum adaptiven Test. Nach zehn Alignment-Iterationen lag die Auflösung bei 42^{-1}Å^{-1}.

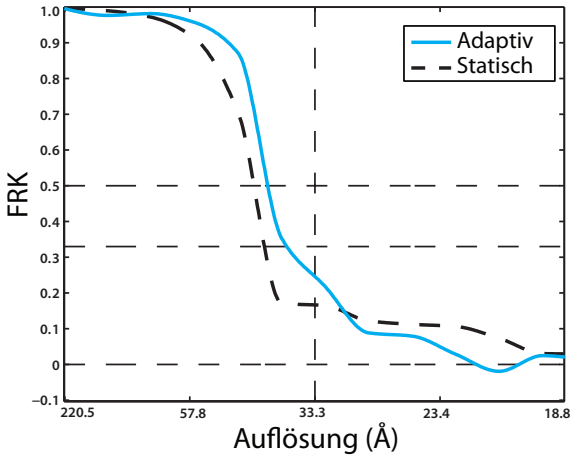

Abbildung 5.3: Die FRK-Kurven bestimmt für Averages nach dem Alignment mit adaptiven (Blau) und statischen Parametern (Schwarz).

5.3 Klassifikationsergebnisse von CPCA, MCO-EM und MCO-A

Die neu implementierten Klassifikationsmethoden MCO-EM und MCO-A wurden auf simulierten Subtomogrammen (Kap. 3.1.1) getestet, damit sie mit gängigen Methoden wie der CPCA-Klassifikation verglichen werden konnten. Hierfür wurden jeweils 100 Subtomogramme aus den, in Kapitel 3.1.1 generierten, vier ribosomalen Partikelklassen zu einem Datensatz von 400 Subtomogrammen vermengt und jeweils durch die drei Methoden klassifiziert. Es wurden außerdem sechs verschiedene SNR Situationen untersucht, um die Toleranz der Klassifikationsmethoden gegenüber abnehmendem SNR zu analysieren.

5.3.1 Prozessierungsparameter

Während der Klassifikation wurde jedes Subtomogramm auf die Auflösung von ca. 23^{-1}Å^{-1} gefiltert, um die Klassifikation möglichst realistisch durchzuführen. Für jeden der sechs unterschiedlichen SNR Datensätze wurden die Klassifikationsmethoden 100 Mal aus verschiedenen Initialkonfigurationen gestartet. Die Initialkonfigurationen sahen vor, dass alle Subtomogramme zufällig auf zehn Klassen verteilt wurden. Für MCO-EM und MCO-A wurde das Konvergenzkriterium $c = 0\%$ gewählt, so dass der Algorithmus abbrach, wenn sich κ nach einer Iteration für alle Subtomogramme nicht geändert hatte (Kap. 3.4.1). Die eigenständige MCO-EM-Klassifikation wurde zehn Mal iteriert. Die MCO-A-Klassifikation wurde ebenfalls zehn Mal iteriert, allerdings wurde die Anzahl an lokalen Optimierungsschritten (MCO-EM) auf fünf Iterationen beschränkt. Für die CPCA wurde die *K-Means*-Klassifikation auf zehn Iterationen beschränkt, um ähnliche

Ausgangsbedingungen für alle drei Methoden zu schaffen.

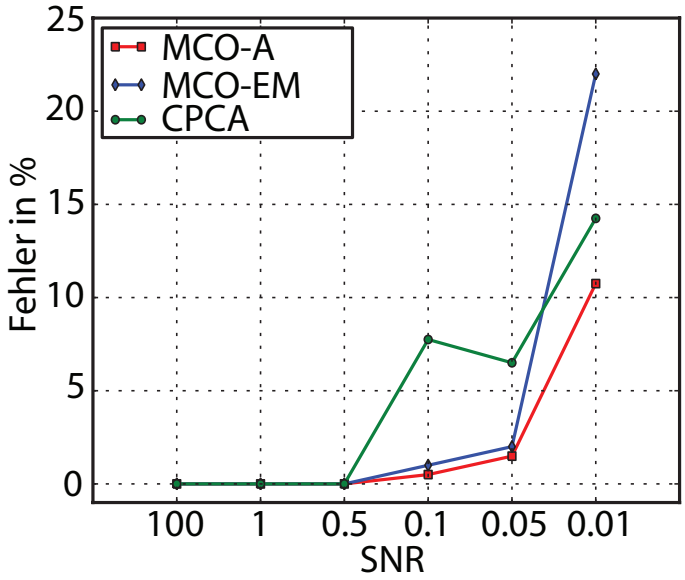

Abbildung 5.4: Aufgetragen ist hier der Klassifikationsfehler der besten Klassifikation aus 100 Wiederholungen. Es wurden Simulationen verwendet, um das Verhalten der drei Klassifikationsalgorithmen bei unterschiedlichen SNRs zu testen.

5.3.2 Ergebnisse der Klassifikationsmethoden

Für hohe SNR bis 0.5 konvergierte die Klassenanzahl zu genau vier Klassen, so dass die optimale Klassenzuweisung automatisch von allen drei Methoden bestimmt werden konnte (Abb. 5.4). Für die verbleibenden SNR-Datensätze ($0.1 - 0.01$) stieg der Klassifikationsfehler an. Folglich konnte weder die optimale Klassenzuweisung noch die Klassenanzahl bestimmt werden. In diesem Fall wurden die resultierenden Klassen mittels hierarchischer Klassifikation auf vier Klassen zusammengeführt (Kap. 3.4.2).

Aus den gemessenen Klassifikationsfehlern (Abb. 5.4) konnte geschlossen werden, dass die Kombination aus MCO-A und hierarchischer Klassifikation von den drei Methoden den niedrigsten Klassifikationsfehler erreicht hat (10.8% bei SNR 0.01). Die Fehlerrate bei SNR 0.01 betrug für CPCA 14.5%, für MCO-EM 22.6%.

5.4 Alignment und Klassifikation von $GroEL_{14}$ und $GroEL_{14}/ES_7$

Der bereits in Kapitel 3.1.2 erwähnte $GroEL_{14}$-, $GroEL_{14}/ES_7$-Datensatz zum Testen und Vergleich von KET-Alignment- und Klassifikationsmethoden wurde ebenfalls prozessiert. Alignment und Klassifikation wurden sequenziell (Kap. 5.4.1) oder kombiniert im *Multi Reference Alignment* (Kap. 5.4.2) durchgeführt.

5.4.1 Alignment und Klassifikation in sequenziellen Schritten

Alignment. Die Startreferenz für das Alignment wurde aus den 786 unalignierten Subtomogrammem generiert. Des Weiteren wurden alle Subtomogramme zuerst global ausgerichtet ($\Delta\alpha \sim 19°$) und in neun lokalen Sampling-Schritten verfeinert. Während des Alignments wurde $\Delta\alpha$ nach der in Kapitel 3.3.2 beschriebenen Strategie adaptiv bestimmt. Beim lokalen Sampling wurden insgesamt sechs Ringe um die bereits bestimmte Rotation abgesucht. Die Kombination aus globalem und lokalem Sampling wurde drei Mal wiederholt, um zu gewährleisten, dass möglichst alle $GroEL_{14}/ES_7$-Dichten in einer identischen Orientierung zum Liegen kommen. Nach den insgesamt 30 Alignment-Runden konnte die charakteristische C7-Symmetrie des $GroEL_{14}/ES_7$-Komplexes gemessen werden (Abb. 5.5).

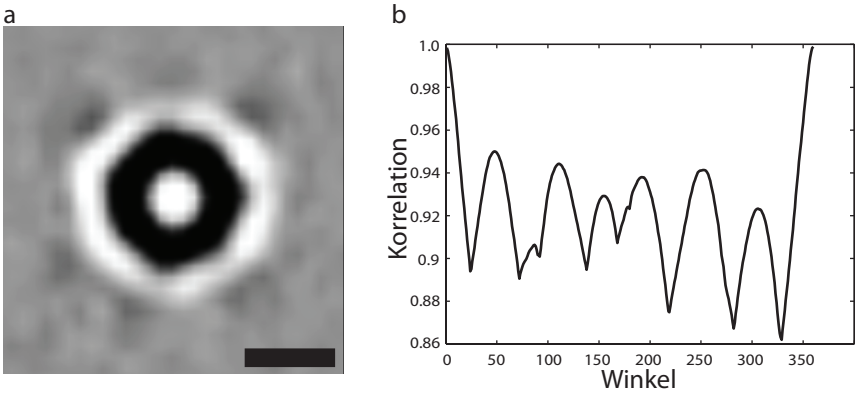

Abbildung 5.5: (a) Schnitt (xy Ebene) durch das alignierte $GroEL_{14}/GroES_{14}$-Average. Die C7-Symmetrie entlang der Z-Achse ist deutlich sichtbar. (b) Mittels Kreuzkorrelation wurde die 7-fache Achsensymmetrie methodisch nachgewiesen.

Klassifikation. Während der anschliessenden MCO-A-Klassifikation wurde die C7-Symmetrie auf die entstehenden Klassenmittel appliziert, um das SNR zu verbessern. MCO-A wurde mit zehn Initialklassen gestartet. Die Klassenanzahl konvergierte nach 30

Iterationen zu drei Klassen[3]. Es wurde eine reine $GroEL_{14}$- sowie eine reine $GroEL_{14}/ES_7$-Klasse gefunden, und ebenfalls auch eine dritte Dichte, die wesentlich kleiner war als das $GroEL_{14}/ES_7$ (Abb. 5.6). Die ermittelte Auflösung lag für die $GroEL_{14}$- sowie für die $GroEL_{14}/GroES_7$-Klassen bei ca. 42^{-1}Å^{-1} nach dem 0.5-Kriterium. Des Weiteren wurde die Klassenreinheit der drei Klassen analog zu [Yu und Frangakis, 2011] ermittelt. Die ermittelten Werte stimmten mit Klassifizierungsergebnissen in anderen Arbeiten überein (Tab. 5.3). Die durch in andere Methoden bestimmten prozentualen Verteilungen wurden entsprechend aus den Publikationen übernommen[4].

Tabelle 5.3: Reinheit der bestimmten $GroEL_{14}$- und $GroEL_{14}/ES_7$-Klassen im Vergleich mit anderen KET Klassifikationsmethoden.

Methode	**GroEL**	**GroEL/ES**	**GroES Fragment**
MCO-A	97%	98%	100% (GroEL/ES)
Maximum Likelihood	96%	86%	100% (GroEL/ES)
KerDenSom3D	95%	98%	96% (GroEL/ES)
WMD-PCA	94%	99%	unbekannt

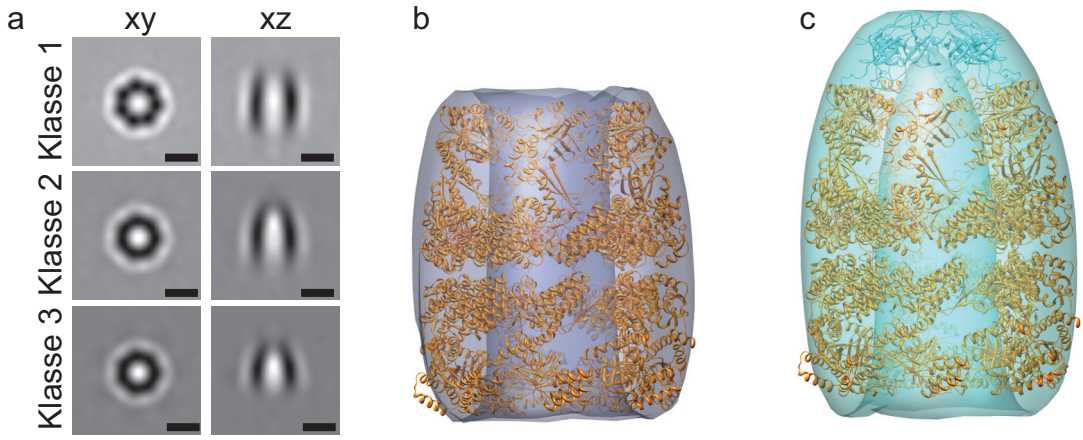

Abbildung 5.6: Der $GroEL_{14}$- und $GroEL_{14}/ES_7$-Datensatz nach der MCO-A-Klassifikation. (a) Schnitte (xy Ebene, xz Ebene) durch die drei resultierenden Klassenmittel. (b) Die atomare Struktur von $GroEL_{14}$ registriert in die Dichte von Klasse 1. (c) Die atomare Struktur von $GroEL_{14}/ES_7$ konnte ebenfalls in der Dichte von Klasse 2 registriert werden. Die ermittelte Auflösung lag für die $GroEL_{14}$- sowie für die $GroEL_{14}/GroES_7$-Klassen bei ca. 42^{-1}Å^{-1} nach dem 0.5-Kriterium.

[3] Zum Vergleich, in [Scheres et al., 2009] wurden 25 Iterationen durchgeführt.
[4] *Maximum Likelihood* [Scheres et al., 2009], *KerDenSOM3D* [Yu und Frangakis, 2011], *WMD-PCA* [Heumann et al., 2011]

5.4.2 Kombiniertes Alignment und Klassifikation durch *Multi Reference Alignment*

Alternativ zum sequentiellen Vorgehen im vorhergehenden Kapitel, wurde der $GroEL_{14}$-, $GroEL_{14}/ES_7$-Datensatz durch das, in Kapitel 2.4.2 vorgestellte, *Multi Reference Alignment* prozessiert, in dem Alignment und Klassifikation direkt integriert sind (Abb. 2.5). MRA wurde ebenfalls aus zehn Klassen gestartet. Im Gegensatz zum sequenziellen Alignment und Klassifikation wurde das globale und lokale Winkelsampling abwechselnd durchgeführt (*pytom.angles.combined.GlobalLocalCombined*). Dieser Zyklus wurde insgesamt fünf Mal wiederholt. Der MRA-Algorithmus konvergierte, wie im sequenziellem Versuch, ebenfalls in drei unterschiedliche Klassen (Abb. 5.7), allerdings war die resultierende Auflösung schlechter als im sequenziellen Versuch. Die ermittelte Auflösung lag für die $GroEL_{14}$- sowie für die $GroEL_{14}/GroES_7$-Klassen bei ca. $54^{-1}Å^{-1}$ nach dem 0.5-Kriterium.

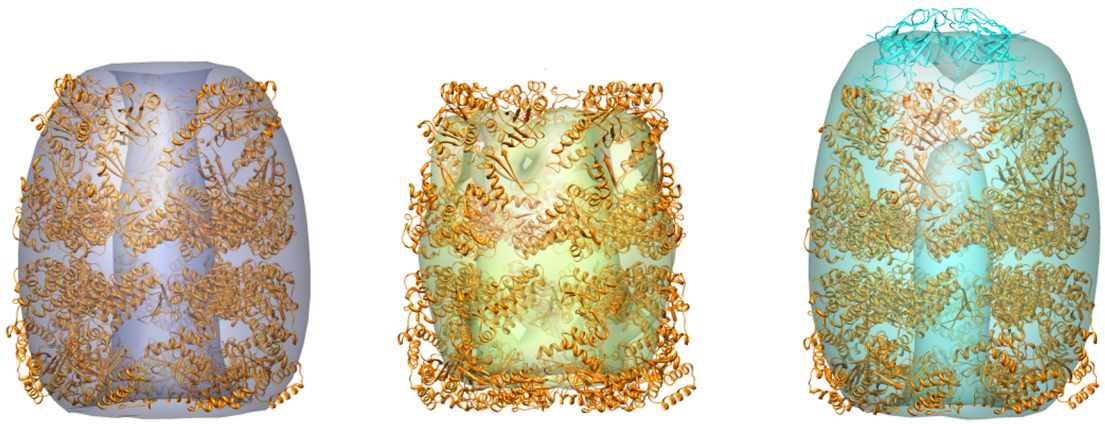

Abbildung 5.7: Alignment- und Klassifikationsergebnis von *Multi Reference Alignment* auf dem $GroEL_{14}$- und $GroEL_{14}/ES_7$-Datensatz. Nach abwechselndem globalem und lokalem Sampling wurden ebenfalls drei unterschiedliche Klassen bestimmt. Die ermittelte Auflösung lag für die $GroEL_{14}$- sowie für die $GroEL_{14}/GroES_7$-Klassen bei ca. $54^{-1}Å^{-1}$ nach dem 0.5-Kriterium.

5.5 Analyse von *S. cerevisae* 80S-Ribosomen mit PyTom

Der in Abbildung 1.1 gezeigte Arbeitsablauf wurde in PyTom getestet, um 80S-Ribosomen (Kap. 3.1.3) in sechs Tomogrammen zu finden und deren Struktur zu bestimmen. Lediglich die Rekonstruktion der Tomogramme wurde in TOM durchgeführt, da diese bis zu diesem Zeitpunkt noch nicht in PyTom integriert war. Tests der in PyTom implementierten Rekonstruktionsmethoden verifizieren allerdings, dass die rekonstruierten Tomo-

gramme identisch sind.

5.5.1 Lokalisierung von Ribosomen mit der 60S-Untereinheit

Subtomogramme von Ribosomen mussten zuerst in den sechs Lysat-Tomogrammen bestimmt werden (Kap. 2.6). Um sicherzugehen, dass die Ergebnisse dieses Versuchs nicht durch *Model Bias* entstehen, wurde nicht nach dem ganzen *S. scerevisiae* 80S-Ribosom gesucht; die Referenz der Suche war nur die 60S-Untereinheit. Die Referenz wurde außerdem auf die Pixelgröße von 18.8^{-1}Å^{-1} interpoliert, mit der für die Tomogramme charakteristischen KTF gefaltet und an der ersten Nullstelle der KTF tiefpassgefiltert. Während des *Matchings* betrug die Winkeldistanz 12.85°, so dass insgesamt 7.112 Winkel abgetastet wurden (Kap. 2.3.3).

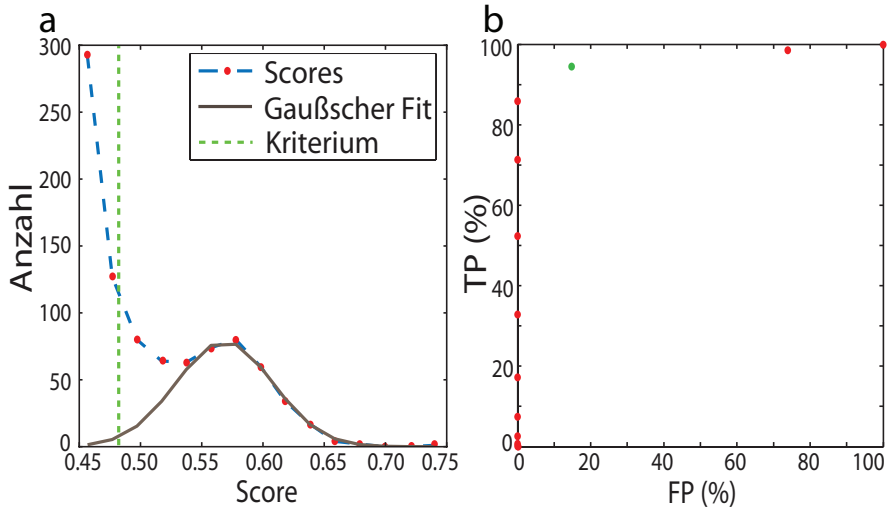

Abbildung 5.8: (a) Das Score-Histogramm (Blau) mit der approximiertern Gauß-Kurve (Braun). Die senkrechte Linie (Grün) entspricht dem durch die *Receiver operating curve* (b) bestimmten Kriterium $S = 0.48$, bis zu dem Subtomogramme in einem Tomogramm akzeptiert wurden.

Mit Hilfe der approximierten ROC-Kurve wurde die Anzahl, der in den sechs Tomogrammen erwarteten, Subtomogramme auf insgesammt $N = 2700$ bestimmt. Die erwartete Rate an *True Positives* betrug 95% ; die zu erwartende Rate an *False Positives* betrug 15% (Abb. 5.8). Der Score, an dem dieser Wert bestimmt wurde, betrug $S = 0.48$. Mittels der Gaußschen-Approximation (Kap. 2.6) wurde die zu erwartende Anzahl von 2084 echten Subtomogrammen im extrahierten Datensatz bestimmt.

5.5.2 Alignment aller ribosomalen Subtomogramme

Basierend auf den groben Winkelparametern nach der Lokalisierung wurde ein vorläufiges Average A_{pre} aller Subtomogramme bestimmt, welches wiederum als Startreferenz für das Alignment verwendet wurde. Trotz der groben Alignment-Parameter θ_{pre} konnte in A_{pre} bereits eine markante Dichte an der Stelle der 40S-Untereinheit ausgemacht werden. Das deutete schon vorab darauf hin, dass die Mehrheit aller Subtomogramme 80S-Ribosome sein müssen (Abb. 5.9).

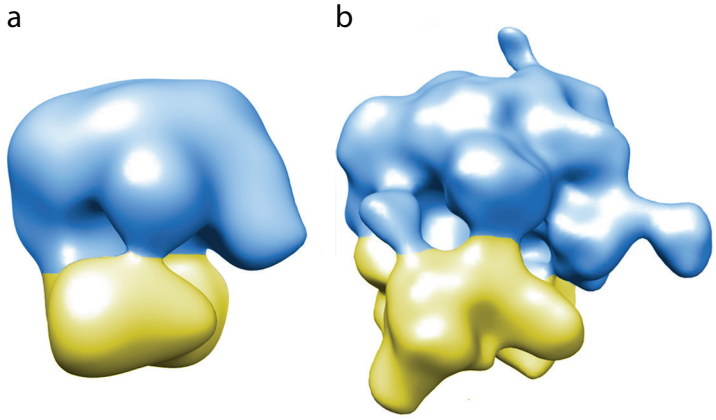

Abbildung 5.9: (a) Average der 2700 Subtomogramme vor dem Alignment. (b) Nach zehn Alignierungsschritten wurde für das entstandene Average eine Auflösung von ca. 34^{-1}Å^{-1} bestimmt.

Die Verfeinerung der Alignment-Parameter wurde über zehn Iterationen durchgeführt. Hierbei war das adaptive Alignment aktiviert (Kap. 3.3), die jeweiligen Skalierungswerte betrugen $\Delta r = 0.1$ und $f = 0.25$. Nach zehn Iterationen konvergierte das Alignment in einen Zustand, in dem sich die durchschnittliche Differenz der Rotationen $\Delta\rho$ und Translationen $\Delta\nu$ nur noch marginal geändert haben:

$$\Delta\rho_{9,10} = 0.03^\circ \qquad \Delta\nu_{9,10} = 0.01 voxel \quad .$$

***De novo* Alignment der Subtomogramme.** Um die Zuverlässigkeit der *de novo* Referenzgenerierung weiter zu testen, wurde in einem weiteren Versuch der Datensatz auf eine durch wiederholtes, globales Sampling generierte Referenz aligniert. Analog zu Kapitel 5.1.2 alternierte die Winkel-Sampling-Strategie (*pytom.angles.combined.GlobalLocalCombined*) nach jeder Iteration zwischen globalem und lokalem Sampling. Nach zehn Iterationen konnte man die Struktur des 80S-Ribosoms deutlich ausmachen (Abb. 5.10). Da der vollständige Datensatz noch kolloide Goldpartikel enthielt, wurden diese in der Nähe der zentralen Potruberanz der 60S-Untereinheit

platziert, was man an der kugelförmigen Ausbuchtung im Average erkennt. Nach zehn Iterationen wurde die Auflösung des Averages auf ca. 43^{-1}Å^{-1} ermittelt. Basierend auf den

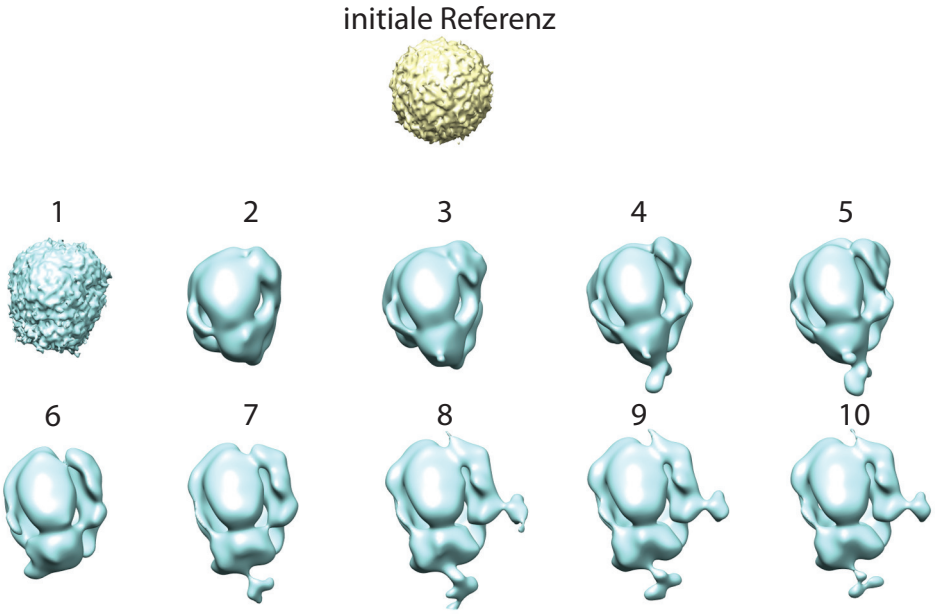

Abbildung 5.10: *De novo* Alignment durch wiederholtes, globales Sampling. Ausgehend von der initialen Referenz wurde das Alignment aller 2700 Subtomogramme (Kap. 3.1.3) über zehn Iterationen bestimmt. Nach jeder Iteration konnte man die Struktur des 80S-Ribosoms immer deutlicher ausmachen.

bestimmten Alignment-Parametern wurde die MCO-A-Klassifikation durchgeführt. Die Klassifikation wurde analog zu dem im folgenden Kapitel beschriebenen Prozess durchgeführt. Die Klassifikationsergebnisse unterschieden sich nur minimal.

5.5.3 Klassifikation aller alignierten Subtomogramme

Die Subtomogramme wurden nach dem Alignment mittels MCO-A (Kap. 3.4) klassifiziert, um den Datensatz in Ribosom, kolloides Gold, Karbonkanten und Schmutz aufzutrennen. Zu Beginn der Klassifikation wurden die Subtomogramme zufällig auf 25 Klassen verteilt (Gleichverteilung). Die lokale Optimierung durch MCO-EM (Kap. 2.4.2) wurde auf maximal fünf Iterationen beschränkt, die Anzahl der SA-Runden betrug insgesamt 30. Außerdem wurden alle Subtomogramme während der Klassifikation auf die, im Alignment bestimmte, Auflösung von ca. 34^{-1}Å^{-1} gefiltert. Während der Klassifikationsrunden wurden mehrere Klassen durch den Algorithmus automatisch vermengt, so dass die Subtomogramme nach der Klassifikation nur noch auf 22 Klassen verteilt wurden.

In Abbildung 5.11 kann man deutlich zwischen 80S-Ribosom (Klassen 1,7,8,10,12,15), 60S-Untereinheit (Klasse 9), kolloides Gold (Klassen 5,6,11,13,14), Karbonkante (Klas-

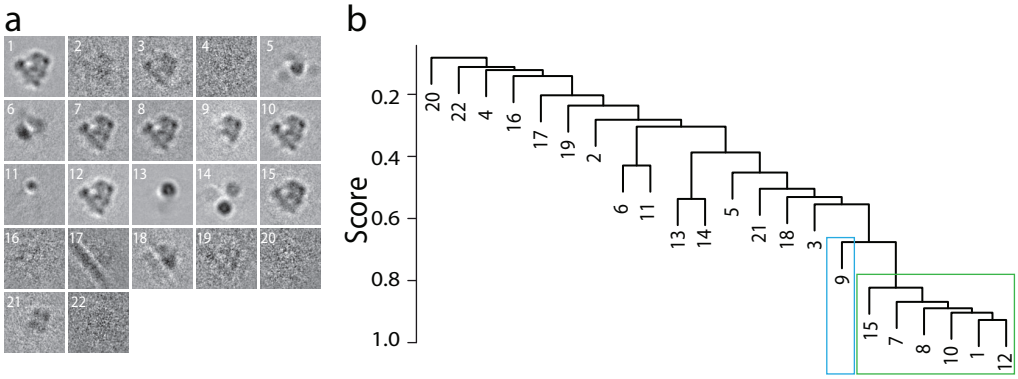

Abbildung 5.11: (a) Zentrale Schnitte durch die finalen, 22 Klassenmittel nach 30 MCO-A Iterationen. Man erkennt das 80S-Ribosom sowie kolloides Gold, Karbonkanten oder Rauschen deutlich. (b) Basierend auf einer Ähnlichkeitsmatrix der Klassenmittel und Hierarchischer Klassifikation wurden im entstandenem Dendrogramm die 80S-Klassen in den gleichen Teilbaum verteilt (grüne Box). Die 60S-Klasse (blau) konnte als direkter Nachbar identifiziert werden.

se 17) und Rauschen (Klassen 4,16,20,22) unterscheiden. Um die Klassen methodisch zusammenzuführen, wurde eine Ähnlichkeitsmatrix (NXC) aller Klassenmittel erstellt (Kap. 3.4.2). Diese Matrix war Grundlage für die hierarchische Klassifikation. Es stellte sich heraus, dass die Klassenmittel mit der größten Ähnlichkeit die 80S-Averages waren. Basierend auf dieser Ähnlichkeitsverteilung wurden alle Subtomogramme dieser Klassen zu einem Average zusammengefügt.

5.5.4 Validierung des Alignments und der Klassifikation

Validierung durch Auflösungsbestimmung. Wie in der KET üblich, wurde die Auflösung des 60S- und 80S-Klassenmittels mittels der FRK und der Halbsatzmethode bestimmt. Aufgrund der geringen Klassengröße der 60S-Klasse (57 Subtomogramme) wurde hier die Auflösung von ca. 52^{-1}Å^{-1} erreicht. Die finale 80S-Klasse bestand aus 1806 Subtomogrammen, so dass hier die Auflösung auf ca. 32^{-1}Å^{-1} bestimmt werden konnte. Als Kontrolle wurden aus den 2700 Subtomogrammen 1806 zufällig ausgewählt und die resultierende Auflösung auf ca. 39^{-1}Å^{-1} bestimmt (Abb. 5.12).

In den aufgetragenen FRK-Kurven erkennt man die Zunahme des alignierten Signals deutlich, da für die klassifizierten 80S-Ribosome das Signal zwischen den KTF-Nullstellen höher korreliert, als für den gesamten Datensatz. Außerdem kommen die KTF-Nullstellen selbst schärfer zum Vorschein.

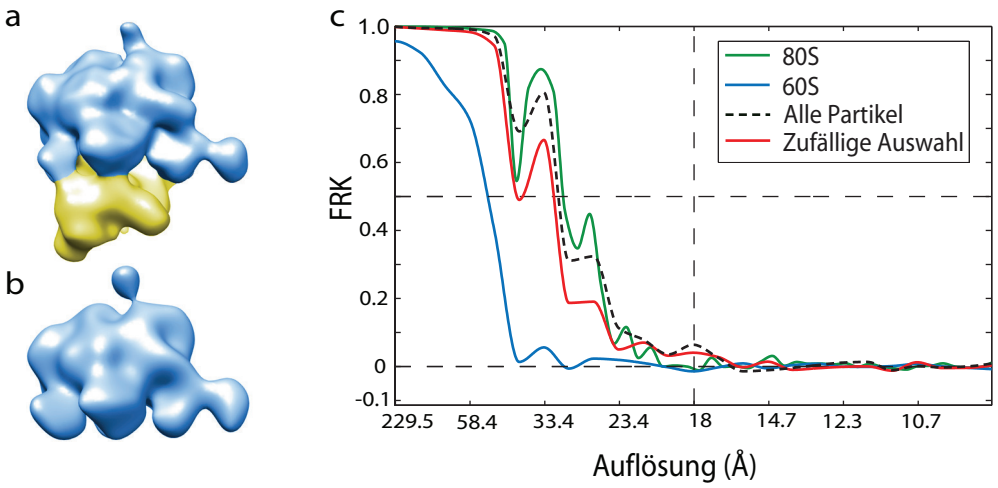

Abbildung 5.12: (a) Der Average aus 1806 als 80S-Ribosom klassifizierten, Subtomogrammen, gefiltert auf 32^{-1}Å^{-1}. (b) Der für die 60S-Klasse bestimmte Average bestand aus 56 Subtomogrammen. (c) Die entsprechenden FRK-Kurven für alle 2700 Subtomogramme (Schwarz), die klassifizierten 80S-Ribosome (Grün), die 60S-Ribosome (Blau). Als Kontrolle wurden aus den 2700 Subtomogrammen 1806 zufällig ausgewählt (Rot).

Photometrische Validierung. Die Klassifikation der Subtomogramme wurde mit der photometrisch bestimmten Konzentration der ribosomalen Einheiten im Lysat verglichen, um die Klassifikationsergebnisse nochmals zu validieren. Hierfür wurde das Lysat auf einem linearen Zuckergradienten ($15\% - 45\%$) und einem Puffer ($5mM$ $MgCK_2$, $140mM$ KCL, $10mM$ Hepes $7.4pH$, $1mM$ DTT, PI) suspendiert und zentrifugiert ($2h$, $38.000rpm$, SW41 Rotor). Die optische Dichte der sedimentierten Lösung wurde bei einer Wellenlänge von $254nm$ bestimmt. [5] Es entstand ein Profil, aus dem man direkt den prozentualen Anteil von 40S- (2.35%), 60S- (2.91%) und 80S-Ribosomen (94.74%) im Lysat ablesen konnte (Abb. 5.13). Die so bestimmten Werte waren in Übereinstimmung mit der, durch die MCO-A Klassifikation bestimmten, Verteilung (60S (3.1%), 80S (96.9%)).

5.6 Analyse von an *canine* ER gebundenen Ribosomen mit PyTom

PyTom wurde benutzt, um die Tomogramme der an *canine* ER-Mikrosomen gebundenen Ribosomen zu analysieren [Pfeffer et al., 2012].

[5] Die Wellenlänge von $254nm$ entspricht dem Absorptionsmaximum von RNS.

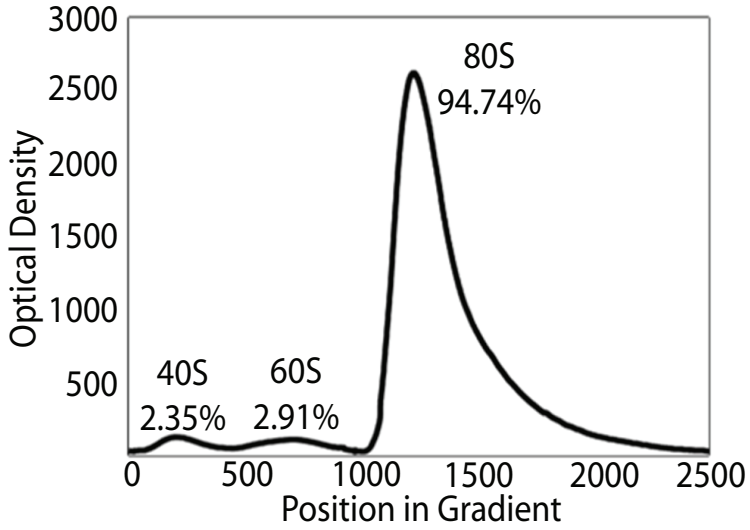

Abbildung 5.13: Nach Ultrazentrifugation des *S. cerevisae*-Lysats wurde ein UV-Profil erstellt, in dem die prozentuelle Verteilung der 40S-, 60S- und 80S-Ribosomen bestimmt werden konnte.

5.6.1 Lokalizierung und Alignment der Ribosomen

Für das Auffinden von potentiellen Ribosomen in den acht Tomogrammen wurde das humane 80S-Ribosom als Referenz benutzt. Wie in Kapitel 5.5 wurde die Referenz (80S-Ribosom) ebenfalls auf die entsprechende Voxelgröße von 18.8^{-1}Å^{-1} skaliert. Darüberhinaus wurde die Referenz mit einer approximierten KTF gefaltet und auf die Auflösung von von 40^{-1}Å^{-1} gefiltert. Während des *Matchings* betrug die Winkeldistanz 12.85° und es wurden insgesamt 7.112 Winkel abgetastet. Die Anzahl an potentiellen Ribosomen in den Tomogrammen wurde durch die *Handedness-check*-Methode (Kap. 2.6) abgeschätzt. Es wurden somit insgesamt ca. 8.000 Ribosome in den acht Tomogrammen lokalisiert und die entsprechenden Subtomogramme mit TOM rekonstruiert. Das humane 80S-Ribosom war ebenfalls die Startreferenz für das anschliessende Alignment der Subtomogramme.

5.6.2 Klassifikation der Ribosomen

Die alignierten Subtomogramme wurden durch CPCA (Kap. 2.4.1) mit AV3 klassifiziert. Eine globuläre Maske um die Ribosome und einen Teil der mikrosomalen Membran wurde benutzt, um membrangebundene Ribosome von kolloidem Gold, Membran wie auch von ungebundenen 60S Ribosomen zu trennen. Die verbliebenen 1.950 ER gebundenen Ribosomen wurden wiederholt klassifiziert (CPCA). Diesmal allerdings fokussierte eine andere Maske auf die lumenale Seite der Membran. 1.004 Subtomogramme membrangebundener Ribosome mit wohl definierten, lumenalen Dichten wurden wiederholt mit

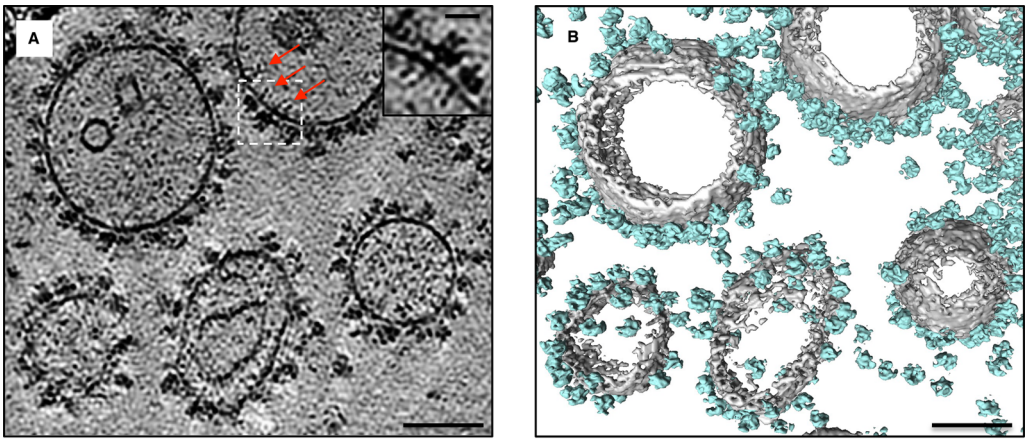

Abbildung 5.14: (a) Schnitt durch ein Tomogramm von *caninen* Mikrosomen des rauen ERs (Der Maßstab entsrpicht $100nm$). Oben rechts ist die eingerahmte Membran vergrössert gezeigt. (b) zeigt die manuell segmentierten Mikrosomen (Grau) mit 80S-Ribosomen (Blau). Die Position und Orientierung der Ribosomen wurde durch *Template Matching* in PyTom mit der humanen 80S-Ribosom Referenz bestimmt. Abbildungen aus [Pfeffer et al., 2012].

PyTom aligniert. Die Auflösung von ca. 31^{-1}Å^{-1} konnte hier nach dem 0.5 Kriterium bestimmt werden.

5.6.3 Interpretation der Dichte

Auf der zytosolischen Seite der ER-Membran konnte das 80S-Ribosom ausgemacht werden. Auf der lumenalen Seite der ER-Membran konnten zwei Dichten LD1 und LD2 (Abb. 5.15) ausgemacht werden. LD1 konnte leicht identifiziert werden, mit diesem Wissen konnte man auf die Identität der LD2 schliessen.

Die ribosomalen Expansions-Segmente ES27L und ES7L. Aus den Verbindungen zwischen Ribosom und Membran konnte man folgern, dass das ribosomale Expansions-Segment ES27L sowie ES7L an der Verbindung Ribosom-Membran beteiligt sein muss. Die Grundlage für diese Schlussfolgerung war das Einpassen des *caninen*-Ribosoms (EMDB 1480 - bestimmt mittels Einzelpartikelanalyse) in die Dichte des Ribosoms. Bei der Überlagerung beider Dichten konnte man anhand der Endposition der Referenzdichte schließen, dass es sich bei den stärksten Verbindungen zwischen Ribosom und Membran um ES27L sowie um ES7L handelt. Die Verbindung zwischen ES27L und Membran wurde in 80% der Subtomogramme in $15nm$ Entfernung vom ribosomalen Peptidausgang gefunden, bei 20% nur $9nm$ entfernt. Mit Hilfe der Dichte des *caninen*-Ribosoms (ebenfalls EMDB 1480) konnte, das an die 60S-Untereinheit gebundene, Expansions-Segment

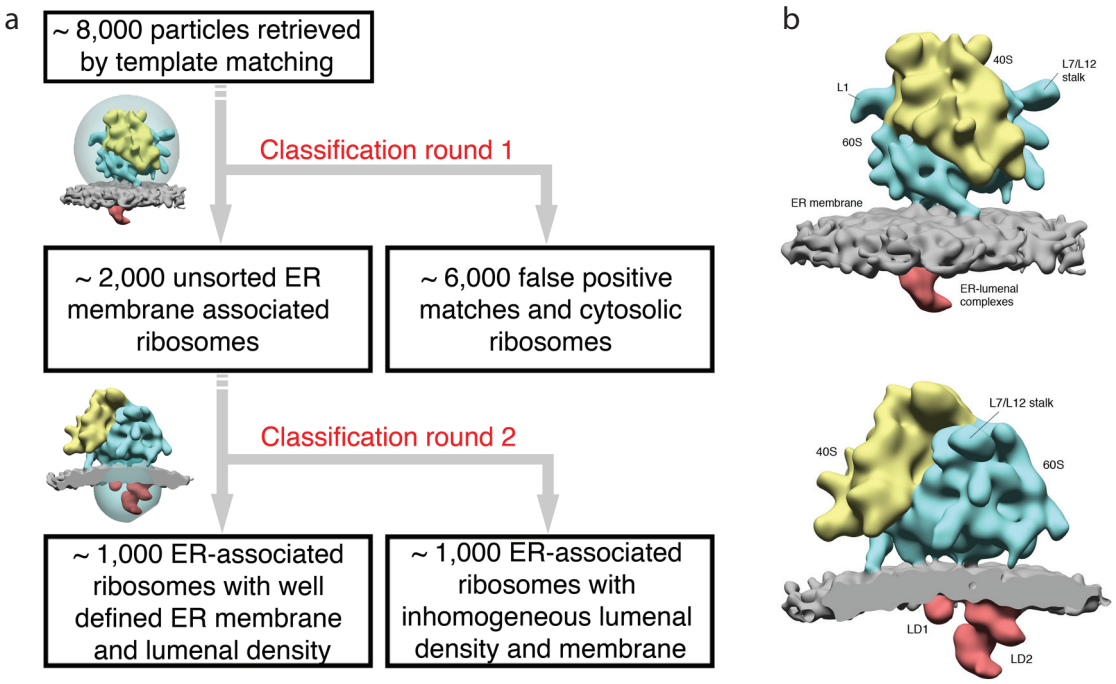

Abbildung 5.15: (a) Der Klassifikations und Alignment-Ablauf der Ribosom-Analyse. Ca. 8000 Subtomogramme wurden aligniert (PyTom). Klassifikation (AV3) mit einer globulären Maske hat ER-gebundene Ribosome von freien Ribosomen sowie von *false positives* getrennt. Im nächsten Klassifikationsschritt wurde mit einer fokussierten Maske klassifiziert, um gebundene Ribosomen mit vollständigen, lumenalen Dichten von unvollständigen zu trennen. (b) Die wiederholt alignierten (PyTom) Dichten der gebundenen 80S-Ribosomen gefiltert auf 31^{-1}Å^{-1}. Auf der lumenalen Seite der ER-Membran konnte man zwei Dichten (Rot - LD1 und LD2) ausmachen. Abbildungen aus [Pfeffer et al., 2012].

ES7L identifiziert werden. Die Verbindung zwischen ES7L und Membran wurde ohne Abweichungen in ca. 16nm Entfernung vom Peptidausgang lokalisiert.

LD1 ist der TRAP-Komplex. Das 80S-Ribosom, das Translokon (Sec61) und der Sec61-assoziierte Proteinkomplex TRAP (Elektronendichte aus EMDB 1528) wurden in das Average der 1.004 Subtomogramme eingepasst. Die Endposition der Referenzdichte des 80S-Ribosomes überdeckte die ribosomale Dichte im Average sehr gut. Ausserdem wurde das Sec61 als Translokationskanal korrekt in der Membran lokalisiert. Der TRAP-Komplex kam in der lumenalen Dichte LD1 zum Liegen, so dass diese als TRAP identifiziert werden konnte (Abb. 5.16).

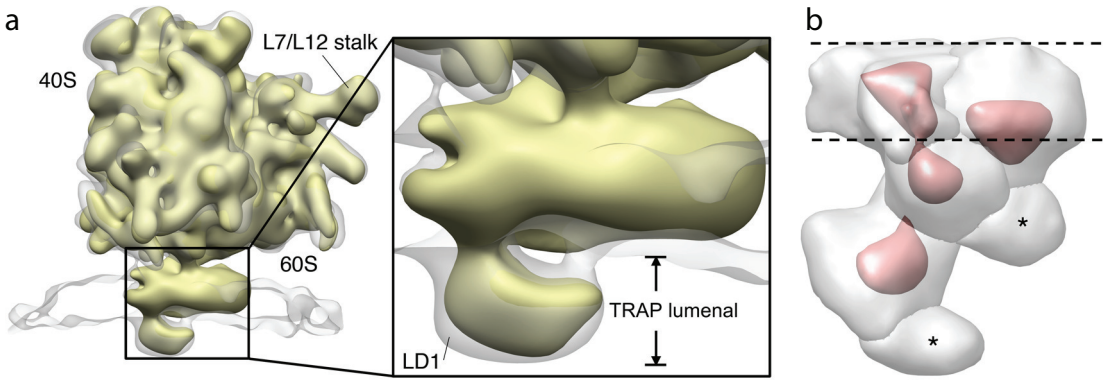

Abbildung 5.16: (a) Eine Dichte bestehend aus 80S-Ribosom, Sec61 und TRAP (EMDB 1528) wurde in die Dichte des Averages eingepasst. LD1 konnte mit der Endposition der Referenzstrukturen als TRAP-Komplex identifiziert werden. (b) LD2 unterteilt in sechs Dichten nach automatischer Dichtesegmentierung. Die mit * gekennzeichneten Dichten haben eine geringere Dichte als die rot eingefärbten Segmente und weisen somit weniger Masse auf. Abbildungen aus [Pfeffer et al., 2012].

LD2 ist höchstwahrscheinlich OST und SP? LD2 ragt ca. $9nm$ in das Lumen des Mikrosomes und ist deshalb aufgrund seiner Größe von Interesse. Die elongierte, lumenenale Dichte LD2 konnte jedoch nicht über bereits bekannte Dichten identifiziert werden. Nach der biochemischen Analyse ist aber bekannt, dass neben dem Sec61 und TRAP auch der Oligosaccharyltransferase-Komplex (OST) während der Translation an membrangebundene Ribosomen assoziiert ist [Potter und Nicchitta, 2002]. Da Sec61 und TRAP bereits gut in LD1 eingepasst werden konnten, kommt man zu der Schlussfolgerung, dass LD2 höchswahrscheinlich der OST-Komplex ist. Nichtsdestotrotz ist es möglich, dass LD2 auch andere Proteindichten enthalten könnte, da durch Analysen mit unterschiedlichen Visualisierungsschwellwerten variierende Dichteverteilungen in LD2 gefunden wurden (Abb. 5.16). Nach dem in [Chen et al., 2001] postulierten Modell verarbeitet das OST und die Signal Peptid Peptidase (SP) das naszierende Protein sequenziell. Der SP-Komplex schneidet die ER-Signalsequenz am naszierenden Protein ab, nachdem es translokalisiert wurde. Der SP-Komplex hat nur direkten Zugang zu dem durch OST glykosilierten Protein, wenn es sich in unmittelbarer Nähe zum OST befindet. Aus diesem Grund könnte das SP ebenfalls in LD2 enthalten sein, möglicherweise an den Stellen mit unterschiedlichen Dichteverteilungen.

6 Diskussion und Ausblick

Das wesentliche Ziel dieser Arbeit war es, eine KET-Software-Plattform zu entwickeln, in der die einzelnen Arbeitsschritte für die Prozessierung von Subtomogrammen integriert sind. Methoden für die strukturelle Analyse von Makromolekülen sollten des Weiteren algorithmisch verbessert werden. Die in dieser Arbeit erzielten Ergebnisse werden in den folgenden Abschnitten diskutiert.

Die Implementierung von PyTom. Die Prozessierungsschritte von KET-Daten wie Rekonstruktion, Lokalisation, Alignment und Klassifikation (Abb. 1.1) waren bis dato auf eigenständige Software-Pakete verteilt. Die üblichen Arbeitsschritte (in der Abteilung für molekulare Strukturbiologie, MPI) sind: Rekonstruktion von Tomogrammen mittels der TOM Toolbox [Nickell et al., 2005], Lokalisation von potentiellen Makromolekülen durch MOLMATCH [Foerster et al., 2010], Alignment der Subtomogramme durch AV3 [Foerster et al., 2005] und Klassifikation entweder durch AV3 [Foerster et al., 2005], tomcorr3d [Haller, 2008] oder MLTomo [Stölken et al., 2010]. Methoden aus anderen Plattformen wie z.B. EMAN2 [Tang et al., 2007], XMIPP [Sorzano et al., 2004], BSOFT [Heymann et al., 2008] oder IMOD [Kremer et al., 1996] sind in anderen Laboren entwickelte Alternativen. Obwohl sich die Methoden der einzelnen Plattformen nur in Details voneinander unterscheiden, wurden in keiner der hier bekannten Sammlungen alle Schritte zu einem kohärenten Ablauf vereinheitlicht. Das ist von essentiellem Nachteil für den Benutzer, da variierende Koordinatensysteme, Rotationskonventionen oder Einheitenspezifikationen den Fortschritt unnötig verzögern.

Die in dieser Arbeit entwickelte Softwaresammlung PyTom vereinheitlicht die genannten Prozessierungsschritte in eine benutzerfreundliche Umgebung [Hrabe et al., 2012]. Es wurde Wert auf Einfachheit und Transparenz gelegt, so dass Koventionen möglichst verständlich präsentiert werden, um Fehlerquellen wie Doppeldeutigkeiten auszuschliessen. Eine auf Webseiten basierende Schnittstelle bietet eine komfortable Möglichkeit, einzelne Arbeitsschritte ohne Programmierkenntnisse durchzuführen. Die Verwendung der verhältnismässig einfachen Programiersprache Python erlaubt dem Laien zudem einen schnellen Einstieg in die Prozessierung von Tomogrammen auf Programmebene. Die Einarbeitung in PyTom sollte deshalb Entwicklern nicht schwer fallen, da Klassen sowie Prozesse bezüglich ihrer Funktionalität in Modulen sortiert sind. Numerische Methoden wurden zumin-

dest im direktem Vergleich zu den, in der Abteilung entwickelten Methoden, qualitativ verbessert (Kap. 4.1).

Ergebnisse der *de novo*-Referenz-Generierung. Die Verwendung von Referenzen für die Alignierung von Subtomogrammen ist ein essenzieller Diskussionspunkt bei der Interpretation des finalen Averages. Um den *Model Bias* zu minimieren, sollten *de novo*, d.h. referenzfreie Aligment-Methoden bevorzugt verwendet werden [Scheres et al., 2009] . Damit ein *de novo*-Verfahren auch in PyTom vorhanden ist, wurden zwei Methoden (Kap. 3.2) implementiert, die eine datengetriebene Generierung von Referenzen ermöglichen. Im direkten Vergleich der zwei Methoden schnitt das wiederholte, globale Sampling grundsätzlich besser ab als die Kombination von Rotationsklassifikation und *Growing Average* (Kap. 5.1.1). Durch das globale Sampling konnte eine initiale Referenz bestimmt werden, in der bereits strukturelle Details des 80S-Ribosoms sichtbar wurden. Diese war somit eine geeignete, initiale Referenz für ein anschliessendes Fein-Alignment oder Klassifikation der Subtomogramme. Im Gegensatz zu diesem Ergebnis konnte keine zufriedenstellende Referenz durch die Rotationsklassifikation und das *Growing Average* erreicht werden, obwohl hier ausschliesslich Subtomogramme von 80S-Ribosomen prozessiert wurden. Bei einer genauen Analyse der Rotationsklassifikation konnte die optimale Klassenverteilung bei weitem nicht erreicht werden. Obwohl der durchschnittliche Rotationsabstand innerhalb der Klassen kleiner war als bei einer zufälligen Verteilung, so war er signifikant größer als bei der optimalen Klassenverteilung. Letztendlich hatte die, durch das *Growing Average* bestimmte, Referenz keine Ähnlichkeit zum 80S-Ribosom, sondern ähnelte einem aus zufällig rotierten Subtomogrammen erstellten Average. Deshalb führen die beobachteten Ergebnisse zu dem Schluss, dass die Kombination von Rotationsklassifikation und *Growing Average*, wie sie in Kapitel 3.2.1 präsentiert wurde nicht zuverlässig auf KET-Daten angewandt werden kann. Für die Bestimmung initialer Referenzen durch PyTom ist das wiederholte, globale Sampling vorzuziehen.

Subtomogramm-Alignment mit adaptiven Parametern. Das Alignment von Subtomogrammen wurde vereinfacht, um kritische Sampling-Parameter wie den Bandpassfilter oder das Winkelinkrement vom Benutzer unabhängig zu machen (Kap. 3.3). Hierfür wurden diese beiden Parameter an die jeweils erreichte Auflösung in einer Iteration gekoppelt und adaptiv für die nächste Iteration bestimmt. Durch das Koppeln der Parameter an die erreichte Auflösung wird eine objektive Strategie bereitgestellt, durch die der Einfluss einer subjektiven Parameterwahl auf das resultierende Alignment minimiert wird. Im direkten Vergleich von adaptiven und statischen Sampling-Parametern wird deutlich, dass das adaptive dem statischen Sampling vorzuziehen ist (Abb. 5.3). Erstens erkennt man die, durch die adaptive Strategie bestimmte, höhere Auflösung nach dem $FSC = 0.5$

Kriterium. Zweitens erkennt man dass hochfrequentes Rauschen weniger korreliert als in dem durch die statische Strategie bestimmten Average. Die sichtbare Dichte im finalen Average ist somit weniger von Rauschen beeinflusst und gibt die natürliche Struktur des Makromoleküls wieder. Darüberhinaus erlaubt diese vereinfachte Parameterwahl eine effizientere Abtastung der aufgespannten Energielandschaft (Kap. 2.3.4). Durch Rauschen oder durch kleine Winkelinkremente bedingte lokale Optima werden durch die adaptiven Parameter aus der Energielandschaft gefiltert und kommen somit nicht als potentielle Konvergenzpunkte in Frage. Außerdem ermöglicht die adaptive Parameterbestimmung das automatisierte Prozessieren von Subtomogrammen, da der Fortgang des Alignments so nur noch durch Aufnahmeparameter wie die Voxelgröße oder den Partikeldurchmesser beinflusst wird. Die adaptive Parameterbestimmung für das Alignment von Subtomogrammen erleichtert somit die Analyse, da dem Benutzer die Spezifikation des Bandpassfilters sowie des Winkelinkrements abgenommen wird.

Klassifikationsergebnisse von MCO-A auf Simulationen sowie auf experimentellen Daten. Mit dem MCO-A-Algorithmus (Kap. 3.4) wurden zum ersten Mal stochastische Komponenten in die Klassifikation von Subtomogrammen integriert. Während der Entwicklung wurde MCO-A auf simulierten Subtomogrammen getestet (Kap. 5.3), um die Lauffähigkeit sicherzustellen. Der Vergleich der drei in der Abteilung (molekulare Strukturbiologie, MPI) entwickelten Klassifikationsmethoden CPCA (Kap. 2.4.1 [Foerster et al., 2008]), MCO-EM (Kap. 2.4.2) und MCO-A fand ebenfalls auf simulierten Subtomogrammen statt. Hierfür wurden die Klassifikationsergebnisse unter verschiedenen SNR untersucht und die Vermutung bestätigt, dass stochastische Komponenten (wie *Simulated Annealing*) KET-Klassifikationsalgorithmen verbessern.

Des Weiteren hat die theoretische Laufzeitanalyse ergeben, dass bei großen Datensätzen MCO-A mit weniger Rechenoperationen auskommt. Die Komplexität von CPCA wächst quadratisch mit der Anzahl der Subtomogramme während für MCO-EM und MCO-A die Komplexität linear zur Anzahl der Subtomogramme wächst.

An dieser Stelle muss allerdings erwähnt werden, dass Tests auf simulierten Daten nicht optimal sind, da es extrem schwierig bis unmöglich ist, vitrifizierte Proben und die Bildentstehung im Elektronenmikroskop detailgetreu zu simulieren. Aufgrund der Tatsache, dass die Simulation von Subtomogrammen nicht standardisiert ist, können Klassifikationsergebnisse, die auf simulierten Daten bestimmt wurden, nur einen Trend aufzeigen, den man im Fall von experimentellen Daten erwarten sollte. Simulierte Daten sollten deshalb vorrangig verwendet werden, um das Konvergenzverhalten von neuen Algorithmen bei variierendem SNR zu studieren. Vergleiche zu anderen Klassifikationsmethoden sind nur durch Tests auf standardisierten, experimentellen Daten möglich (Kap. 5.4.1).

Der $GroEL_{14}/GroES_7$-Datensatz wurde daher bereits in anderen Arbeiten verwen-

det und stellt somit eine Möglichkeit dar, die Genauigkeit der unterschiedlichen Alignment- und Klassifikationsmethoden zu vergleichen (*Maximum Likelihood Alignment* [Scheres et al., 2009], *KerDenSOM3D* [Yu und Frangakis, 2011] und *WMD-PCA* [Heumann et al., 2011]). Für den direkten Vergleich mit anderen Methoden wurde der Datensatz ebenfalls mit den, in PyTom implementierten, Alignment- und Klassifikationsmethoden prozessiert. Das Ergebnis vom adaptiven Alignment mit anschliessender Klassifikation durch MCO-A stimmte mit Ergebnissen aus anderen Arbeiten überein (Tab. 5.3). Bereits während des adaptiven Alignments kam die C7-Symmetrie des $GroEL_{14}$-Komplexes zum Vorschein. Durch die anschliessende MCO-A-Klassifikation konnten ebenfalls die, in allen vorhergehenden Arbeiten beobachteten, drei Klassen bestimmt werden: der $GroEL_{14}$-Komplex (42^{-1}Å^{-1}), der $GroEL_{14}/GroES_7$-Komplex (42^{-1}Å^{-1}) und eine kleinere Dichte (Abb. 5.6). Letztere wurde in [Heumann et al., 2011] als „unvollständiger oder fragmentierter Komplex“ bezeichnet und stammt, wie hier und in den anderen Arbeiten bestimmt, aus $GroEL_{14}/GroES_7$-Tomogrammen. Hierbei handelt es sich möglicherweise um einen $GroEL_7$ Ring mit einem gebundenen $GroES_7$ oder um ungebundene Heptamere.

Die mit Hilfe von PyTom auf dem $GroEL_{14}$, $GroEL_{14}/GroES_7$-Datensatz ermittelten Ergebnisse demonstrieren, dass die implementierten Alignment- und Klassifikationsmethoden dem aktuellen Standard und den Anforderungen in der KET entsprechen. Darüberhinaus wurde keine Referenz für das Alignment der $GroEL_{14}$-Subtomogramme benutzt, um zu zeigen, dass der Ablauf von Alignment und Klassifikation in PyTom ebenfalls *de novo* beziehungsweise referenzfrei erfolgen kann.

Die Prozessierung der *S. cerevisae*-Lysat-80S-Ribosomen. Der gesamte KET-Prozessierungsablauf, von der Lokalisation von Makromolekülen über das Alignment von Subtomogrammen, bis zu deren Klassifikation wurde in PyTom auf Tomogrammen eines *S. cerevisae*-Lysats durchgeführt. Während dieser Prozessierung kamen fast alle in dieser Arbeit implementierten Methoden zur Anwendung, wodurch ein gutes Zusammenspiel der einzelnen Methoden demonstriert werden konnte. Da sich die 60S-Untereinheit als die einzig benutzte Referenz klar vom finalen Average (80S-Ribosom) unterschied, war der *Model Bias* minimal. Des Weiteren wurde die Zuverlässigkeit der in PyTom implementierten Methoden durch die Auflösung des finalen Average von ca. 32^{-1}Å^{-1} sowie durch die gute Übereinstimmung der Klassengrößen mit der biochemischen Analyse des Lysats bekräftigt.

In einem weiteren Experiment wurde das Alignment der Subtomogramme wiederholt. Ausgegangen wurde diesmal von einer *de novo* generierten Referenz. Nach dem Alignment aller Subtomogramme und anschliessender MCO-A-Klassifikation konnte ebenfalls ein 80S-Ribosom aus den Subtomogrammen gemittelt werden. Da hier allerdings weniger

Subtomogramme als 80S-Ribosome klassifiziert wurden, lag die in diesem Versuch bestimmte Auflösung bei ca. 41^{-1}Å^{-1}, also etwa bei der ersten Nullstelle der KTF.

Die Klassifikation der *S. cerevisae*-Lysat-Subtomogramme in Kapitel 5.5 ermöglicht das direkte Testen anderer Algorithmen auf experimentellen Daten, da die bestimmten Alignmentparameter wie auch die Klassenzugehörigkeit der 2700 Partikel als korrekt angenommen werden können. Gegenüber dem $\text{GroEL}_{14}/\text{GroES}_7$-Datensatz haben die Subtomogramme aus dem Lysat keine Vorzugsausrichtung, was dem generellen Testen von Alignment-Strategien zu Gute kommt. Allerdings ist das *S. cerevisae* 80S-Ribosom (Kap. 3.1.2) ca. fünf mal größer als das GroEL_{14}-Molekül. Deswegen sollten trotzdem beide Datensätze für die Analyse von Klassifikationsalgorithmen benutzt werden. Da aber die Zwischenergebnisse aller Prozessierungsschritte vorliegen, können neue KET-Rekonstruktions-, Lokalizations-, Alignment- oder Klassifikationsalgorithmen auf den *S. cerevisae* Lysat-Daten getestet werden.

Die Analyse ER-Membran gebundener Ribosomen. Ein weiterer experimenteller Datensatz wurde mit Algorithmen aus PyTom prozessiert, um die Bindung von 80S-Ribosomen an die Membran des Endoplasmatischen Retikulums zu untersuchen (Kap. 3.1.5, Kap. 5.6). Die mit PyTom und AV3 bestimmte Dichte zeigt ein an die ER-Membran gebundenes 80S-Ribosom bei ca. 31^{-1}Å^{-1} auf der zytosolischen Seite der Membran. Die stärksten Verbindungen von Ribosom und Membran wurden als die Expansionssegmente ES27L und ES7L identifiziert. Darüberhinaus wurden auf der lumenalen Seite der Membran nach wiederholtem Alignment und Klassifikation Dichten aufgelöst, die nach detaillierter Analyse zum einen der Sec61- und TRAP-Komplex und zum anderen höchstwahrscheinlich der OST- und SP-Komplex waren.

Ausblick. Die entwickelte Software-Plattform PyTom ist in einem robusten Zustand und kann für die Verarbeitung von KET-Daten verwendet werden. Die Erweiterung von PyTom mit neuen KET-Verarbeitungsmethoden sollte aufgrund der intuitiven Strukturierung der Plattform dem Benutzer leicht fallen. Eine leicht durchfürbare Erweiterung wäre die Integration der MCO-A-Klassifikation in den *Mutli Reference Alignment*-Prozess, um den Determinismus im MRA aufzulösen. Ebenfalls könnte man erwägen, Annealing-Komponenten in das Alignment selbst zu integrieren. Ein aktuelles Forschungsprojekt ist die Beschleunigung des Alignments durch die Verwendung von Kugelflächenfunktionen und deren Erweiterung durch verbesserte *Scores* (Implementierung von Yuxiang Chen). Vorläufige Ergebnisse konnten die Beobachtungen aus [Xu et al., 2012] verifizieren, dass durch die Verwendung von Kugelflächenfunktionen das Alignment von Subtomogrammen signifikant beschleunigt werden kann. Diese Komponente ist zur Zeit in der Testphase, würde aber als Beschleunigung von *Mutli Reference Alignment* mit MCO-A-Iterationen

diese Kombination erst ermöglichen.

Einziger noch fehlender, aber essentieller KET-Verarbeitungsschritt in Pytom ist die KTF-Korrektur. Die Korrektur wurde bislang mit externen Methoden durchgeführt, sollte aber der Vollständigkeit und der Benutzbarkeit halber ebenfalls integriert werden.

In dieser Arbeit wurde demonstriert, dass PyTom einen Zustand erreicht hat, in dem es der interessierten Öffentlichkeit zur Verfügung gestellt werden kann. *Direct Detection Detectors* (DDD) [Milazzo et al., 2011] oder Phasenplatten [Danev et al., 2010] werden die Bildqualität in den aufgenommenen Projektionen und folglich auch die erreichte Auflösung durch KET-Analysemethoden in naher Zukunft weiter verbessern. Erste Ergebnisse mit einer DDD bestätigen diese Vermutung: eine zu den ER assoziierten Ribosomen (Kap. 3.1.5) identische Probe wurde tomographisch abgebildet und mittels PyTom prozessiert. Die Auflösung konnte hier auf $\leq 20^{-1}$Å $^{-1}$ bestimmt werden. Der einzige Unterschied zu dem in dieser Arbeit verwendeten Datensatz war die DDD-Kamera, alle anderen Prozessierungsparameter waren nahezu identisch.

Sollte das Interesse nach der KET basierenden Strukturanalyse deshalb wachsen, so wird die Nachfrage nach stabilen, computergestützten Hochdurchsatzmethoden wie in der Röntgenkristalographie oder der KEM ebenfalls größer. Mit PyTom steht somit eine Plattform bereit, die einfach erweiterbar ist, um den Anforderungen der nächsten Jahre gewachsen zu sein.

Literaturverzeichnis

[Aarts et al., 2005] Aarts, E., Korst, J., und Michiels, W. (2005). Simulated annealing. *Search Methodologies*.

[Alberts et al., 2010] Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., und Walter, P. (2010). *Essential Cell Biology*. 3.rd edition.

[Armache et al., 2010] Armache, J.-P., Jarasch, A., Anger, A. M., Villa, E., Becker, T., Bhushan, S., Jossinet, F., Habeck, M., Dindar, G., Franckenberg, S., Marquez, V., Mielke, T., Thomm, M., Berninghausen, O., Beatrix, B., Söding, J., Westhof, E., Wilson, D. N., und Beckmann, R. (2010). Cryo-EM structure and rRNA model of a translating eukaryotic 80S ribosome at 5.5-A resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 107(46):19748–53.

[Bartesaghi et al., 2008] Bartesaghi, A., Sprechmann, P., Liu, J., Randall, G., Sapiro, G., und Subramaniam, S. (2008). Classification and 3D averaging with missing wedge correction in biological electron tomography. *Journal of structural biology*, 162(3):436–450.

[Becker et al., 2009] Becker, T., Bhushan, S., Jarasch, A., Armache, J.-P., Funes, S., Jossinet, F., Gumbart, J., Mielke, T., Berninghausen, O., Schulten, K., Westhof, E., Gilmore, R., Mandon, E. C., und Beckmann, R. (2009). Structure of monomeric yeast and mammalian Sec61 complexes interacting with the translating ribosome. *Science (New York, N.Y.)*, 326(5958):1369–73.

[Beckmann et al., 2001] Beckmann, R., Spahn, C. M., Eswar, N., Helmers, J., Penczek, P. a., Sali, a., Frank, J., und Blobel, G. (2001). Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell*, 107(3):361–72.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Bracewell, 2000] Bracewell, R. N. (2000). *The Fourier transform and its applications*. McGraw-Hill series in electrical and computer engineering. McGraw Hill.

[Braig et al., 1994] Braig, K., Otwinowski, Z., Hegde, R., Boisvert, D. C., Joachimiak, A., Horwich, A. L., und Sigler, P. B. (1994). The crystal structure of the bacterial chaperonin GroEL at 2.8 A. *Nature*, 371(6498):578–86.

[Castaño Díez et al., 2012] Castaño Díez, D., Kudryashev, M., Arheit, M., und Stahlberg, H. (2012). Dynamo: A flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments. *Journal of structural biology*.

[Chaudhuri et al., 2009] Chaudhuri, T. K., Verma, V. K., und Maheshwari, A. (2009). GroEL assisted folding of large polypeptide substrates in Escherichia coli: Present scenario and assignments for the future. *Progress in biophysics and molecular biology*, 99(1):42–50.

[Chen et al., 2001] Chen, X., VanValkenburgh, C., Liang, H., Fang, H., und Green, N. (2001). Signal peptidase and oligosaccharyltransferase interact in a sequential and dependent manner within the endoplasmic reticulum. *The Journal of biological chemistry*, 276(4):2411–6.

[Chen et al., 2012] Chen, Y., Hrabe, T., Pfeffer, S., Pauly, O., Mateus, D., Navab, N., und Förster, F. (2012). Detection and Identification of Macromolecular Complexes in Cryo-Electron Tomograms using Support Vector Machines. In *IEEE International Symposium on Biomedical Imaging*, pages 1–4.

[Crowther et al., 1970] Crowther, R. A., DeRosier, D. J., und Klug, A. (1970). The Reconstruction of a Three-Dimensional Structure from Projections and its Application to Electron Microscopy. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 317(1530):319–340.

[Danev et al., 2010] Danev, R., Kanamaru, S., Marko, M., und Nagayama, K. (2010). Zernike phase contrast cryo-electron tomography. *Journal of structural biology*, 171(2):174–181.

[De Ruijter, 1995] De Ruijter, W. (1995). Imaging properties and applications of slow-scan charge-coupled device cameras suitable for electron microscopy. *Micron*, 26(3):247–275.

[Dubochet et al., 1988] Dubochet, J., Adrian, M., Chang, J.-J., Homo, J.-C., Lepault, J., McDowall, A. W., und Schultz, P. (1988). Cryo-electron microscopy of vitrified specimens. *Quarterly Reviews of Biophysics*, 21(02):129–228.

[Foerster, 2005] Foerster, F. (2005). *Quantitative Analyse von Makromolekuelen in Kryoelektronentomogrammen mittels Korrelationsmethoden*. PhD thesis.

[Foerster et al., 2010] Foerster, F., Han, B. G., und Beck, M. (2010). Visual proteomics. *Methods in enzymology*, 483:215–243.

[Foerster und Hegerl, 2006] Foerster, F. und Hegerl, R. (2006). Structure Determination In Situ by Averaging of Tomograms. pages 1–29.

[Foerster et al., 2005] Foerster, F., Medalia, O., Zauberman, N., Baumeister, W., und Fass, D. (2005). Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography. *Proceedings of the National Academy of Sciences of the United States of America*, 102(13):4729–4734.

[Foerster et al., 2008] Foerster, F., Pruggnaller, S., Seybert, A., und Frangakis, A. S. (2008). Classification of cryo-electron sub-tomograms using constrained correlation. *Journal of structural biology*, 161(3):276–286.

[Forsyth und Ponce, 2003] Forsyth, D. A. und Ponce, J. (2003). *Computer Vision*. Pearson, 1 edition.

[Frangakis et al., 2002] Frangakis, A. S., Böhm, J., Foerster, F., Nickell, S., Nicastro, D., Typke, D., Hegerl, R., und Baumeister, W. (2002). Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14153–14158.

[Frangakis und Rath, 2006] Frangakis, A. S. und Rath, B. K. (2006). Motif Search in Electron Tomography. In Frank, J., editor, *Electron Tomography*, pages 401–417. Springer, 2.nd edition.

[Frank, 2002] Frank, J. (2002). SINGLE-PARTICLE IMAGING OF MACROMOLECULES BY CRYO-ELECTRON MICROSCOPY. *Annu. Rev. Biophys. Biomol. Struct.*, (31):303–319.

[Haller, 2008] Haller, T. (2008). *Multi reference Alignment of 3D Data from Electron Tomography*. PhD thesis, Technische Universitaet Muenchen.

[Hegerl, 1996] Hegerl, R. (1996). The EM Program Package: A Platform for Image Processing in Biological Electron Microscopy. *Journal of Structural Biology*, 116(1):30–34.

[Henderson, 1995] Henderson, R. (1995). The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly Reviews of Biophysics*, 28(02):171–193.

[Heumann et al., 2011] Heumann, J. M., Hoenger, A., und Mastronarde, D. N. (2011). Clustering and variance maps for cryo-electron tomography using wedge-masked differences. *Journal of structural biology*, 175(3):288–299.

[Heymann et al., 2008] Heymann, J. B., Cardone, G., Winkler, D. C., und Steven, A. C. (2008). Computational resources for cryo-electron tomography in Bsoft. *Journal of structural biology*, 161(3):232–42.

[Hrabe et al., 2012] Hrabe, T., Chen, Y., Pfeffer, S., Cuellar, L. K., Mangold, A.-V., und Förster, F. (2012). PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *Journal of Structural Biology*.

[Hrabe und Förster, 2011] Hrabe, T. und Förster, F. (2011). Structure Determination by Single Particle Tomography. *Encyclopedia of Life Sciences*, pages 1–11.

[Johnson und van Waes, 1999] Johnson, A. E. und van Waes, M. A. (1999). A Dynamic Gateway at the ER Membrane. *Structure*, pages 799–842.

[Kampmann und Blobel, 2009] Kampmann, M. und Blobel, G. (2009). Biochemistry. Nascent proteins caught in the act. *Science (New York, N.Y.)*, 326(5958):1352–3.

[Kirkpatrick et al., 1983] Kirkpatrick, S., Gelatt, C. D., und Vecchi, M. P. (1983). Optimization by simulated annealing. *Science (New York, N.Y.)*, 220(4598):671–80.

[Knauer et al., 1983] Knauer, V., Hegerl, R., und Hoppe, W. (1983). Three-dimensional reconstruction and averaging of 30 S ribosomal subunits of Escherichia coli from electron micrographs. *Journal of molecular biology*, 163(3):409–430.

[Kremer et al., 1996] Kremer, J. R., Mastronarde, D. N., und McIntosh, J. R. (1996). Computer visualization of three-dimensional image data using IMOD. *Journal of structural biology*, 116(1):71–6.

[Kuffner, 2004] Kuffner, J. (2004). Effective sampling and distance metrics for 3D rigid body path planning. *IEEE International Conference on Robotics and Automation*, pages 3993 – 3998.

[Kumar et al., 2006] Kumar, B. V. K. V., Mahalanobis, A., und Juday, R. D. (2006). *Correlation Pattern Recognition*. Cambridge.

[Lander et al., 2009] Lander, G. C., Stagg, S. M., Voss, N. R., Cheng, A., Fellmann, D., Pulokas, J., Yoshioka, C., Irving, C., Mulder, A., Lau, P.-W., Lyumkis, D., Potter, C. S., und Carragher, B. (2009). Appion: An integrated, database-driven pipeline to facilitate EM image processing. *Journal of Structural Biology*, 166(1):95–102.

[Langlois und Frank, 2011] Langlois, R. und Frank, J. (2011). A clarification of the terms used in comparing semi-automated particle selection algorithms in Cryo-EM. *Journal of structural biology*, 175(3):348–52.

[Lucić et al., 2005] Lucić, V., Foerster, F., und Baumeister, W. (2005). Structural studies by electron tomography: from cells to molecules. *Annual review of biochemistry*, 74:833–865.

[Mangold, 2010] Mangold, A.-V. (2010). *Kryoelektronentomographische Untersuchungen an mikrosomalen Fraktionen aus Saccharomyces cerevisiae*. PhD thesis.

[Mastronarde, 2006] Mastronarde, D. N. (2006). Fiducial Marker and Hybrid Alignment Methods for Single and Double-axis Tomography. In Frank, J., editor, *Electron Tomography*, pages 163–185. Springer, 2.nd edition.

[Milazzo et al., 2011] Milazzo, A.-C., Cheng, A., Moeller, A., Lyumkis, D., Jacovetty, E., Polukas, J., Ellisman, M. H., Xuong, N.-H., Carragher, B., und Potter, C. S. (2011). Initial evaluation of a direct detection device detector for single particle cryo-electron microscopy. *Journal of structural biology*, 176(3):404–8.

[Moeller et al., 2012] Moeller, A., Zhao, C., Fried, M. G., Wilson-Kubalek, E. M., Carragher, B., und Whiteheart, S. W. (2012). Nucleotide-dependent conformational changes in the N-Ethylmaleimide Sensitive Factor (NSF) and their potential role in SNARE complex disassembly. *Journal of structural biology*, 177(2):335–43.

[Nickell et al., 2005] Nickell, S., Foerster, F., Linaroudis, A., Net, W. D., Beck, F., Hegerl, R., Baumeister, W., und Plitzko, J. M. (2005). TOM software toolbox: acquisition and analysis for electron tomography. *Journal of structural biology*, 149(3):227–234.

[Nickell et al., 2003] Nickell, S., Hegerl, R., Baumeister, W., und Rachel, R. (2003). Pyrodictium cannulae enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *Journal of Structural Biology*, 141(1):34–42.

[Oettl et al., 1983] Oettl, H., Hegerl, R., und Hoppe, W. (1983). Three-dimensional reconstruction and averaging of 50 S ribosomal subunits of from electron micrographs. *Journal of Molecular Biology*, 163(3):431–450.

[Orlova und Saibil, 2011] Orlova, E. V. und Saibil, H. R. (2011). Structural analysis of macromolecular assemblies by electron microscopy. *Chemical reviews*, 111(12):7710–48.

[Ortiz et al., 2010] Ortiz, J. O., Brandt, F., Matias, V. R., Sennels, L., Rappsilber, J., Scheres, S. H. W., Eibauer, M., Hartl, F. U., und Baumeister, W. (2010). Structure of hibernating ribosomes studied by cryoelectron tomography in vitro and in situ. *The Journal of cell biology*, 190(4):613–621.

[Ortiz et al., 2006] Ortiz, J. O., Foerster, F., Kürner, J., Linaroudis, A. A., und Baumeister, W. (2006). Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition. *Journal of structural biology*, 156(2):334–341.

[Penczek, 2010] Penczek, P. A. (2010). Resolution measures in molecular electron microscopy. *Methods in enzymology*, 482:73–100.

[Pfeffer, 2010] Pfeffer, S. (2010). *Kryoelektronentomographische Analyse membrangebundener eukaryotischer Polyribosomen*. PhD thesis, Max Planck Institute of Biochemistry.

[Pfeffer et al., 2012] Pfeffer, S., Brandt, F., Hrabe, T., Eibauer, M., Lang, S., Zimmermann, R., und Förster, F. (2012). Structure and 3D arrangement of ER membrane associated ribosomes. *submitted*.

[Potter und Nicchitta, 2002] Potter, M. D. und Nicchitta, C. V. (2002). Endoplasmic reticulum-bound ribosomes reside in stable association with the translocon following termination of protein synthesis. *The Journal of biological chemistry*, 277(26):23314–20.

[Rade und Westergren, 2000] Rade, L. und Westergren, B. (2000). *Springers Mathematische Formeln*. Springer, 3.nd edition.

[Radermacher, 2006] Radermacher, M. (2006). Weighted Back-projection Methods. In Frank, J., editor, *Electron Tomography*, chapter Weighted B, pages 245–275. Springer, 2.nd edition.

[Radon, 1917] Radon, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Berichte über die Verhandlungen der Sächsische Akademie der Wissenschaften*, 69:262–277.

[Ramezani-rad et al., 1985] Ramezani-rad, M., Käufer, N. F., Hasilik, A., und Lochmann, E.-R. (1985). In Vitro Studies With Rough Microsomes From Saccharomyces cerevisiae. 260:249–260.

[Roseman, 2003] Roseman, A. M. (2003). Particle finding in electron micrographs using a fast local correlation algorithm. *Ultramicroscopy*, 94(3-4):225–236.

[Rossmann et al., 2005] Rossmann, M. G., Morais, M. C., Leiman, P. G., und Zhang, W. (2005). Combining X-ray crystallography and electron microscopy. *Structure (London, England : 1993)*, 13(3):355–62.

[Saxton und Baumeister, 1982] Saxton, W. O. und Baumeister, W. (1982). The correlation averaging of a regularly arranged bacterial cell envelope protein. *Journal of microscopy*, 127(Pt 2):127–38.

[Scheres et al., 2009] Scheres, S. H. W., Melero, R., Valle, M., und Carazo, J. M. (2009). Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization. *Structure (London, England : 1993)*, 17(12):1563–1572.

[Sigler et al., 1998] Sigler, P. B., Xu, Z., Rye, H. S., Burston, S. G., Fenton, W. a., und Horwich, a. L. (1998). Structure and function in GroEL-mediated protein folding. *Annual review of biochemistry*, 67:581–608.

[Simons et al., 1997] Simons, K. T., Kooperberg, C., Huang, E., und Baker, D. (1997). Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. *Journal of Molecular Biology*, 268:209–225.

[Sinkovits und Baker, 2011] Sinkovits, R. S. und Baker, T. S. (2011). Structure Determination of Icosahedral Viruses Imaged by Cryo-electron Microscopy. *Structure*, (21):81–99.

[Sorzano et al., 2006] Sorzano, C. O. S., Marabini, R., Pascual-Montano, A., Scheres, S. H. W., und Carazo, J. M. (2006). Optimization problems in electron microscopy of single particles. *Annals Of Operations Research*, 148(1):133–165.

[Sorzano et al., 2004] Sorzano, C. O. S., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J. R., Scheres, S. H. W., Carazo, J. M., und 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.

[Steger, 2002] Steger, A. (2002). *Diskrete Strukturen 1.* Springer, 2 edition.

[Stölken et al., 2010] Stölken, M., Beck, F., Haller, T., Hegerl, R., Gutsche, I., Carazo, J. M., Baumeister, W., Scheres, S. H. W., und Nickell, S. (2010). Maximum likelihood based classification of electron tomographic data. *Journal of structural biology*, 173(1):77–85.

[Tanenbaum, 2002] Tanenbaum, A. S. (2002). *Moderne Betriebssysteme.* Pearson, 2 edition.

[Tang et al., 2007] Tang, G., Peng, L., Baldwin, P. R., Mann, D., Jiang, W., Rees, I., und Ludtke, S. (2007). EMAN2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46.

[van Heel und Frank, 1981] van Heel, M. und Frank, J. (1981). Use of multivariate statistics in analysing the images of biological macromolecules. *Ultramicroscopy*, 6(2):187–194.

[Černý, 1985] Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51.

[Verma et al., 2006] Verma, A., Schug, A., Lee, K. H., und Wenzel, W. (2006). Basin hopping simulations for all-atom protein folding. *The Journal of chemical physics*, 124(4):44515.

[Walz et al., 1997] Walz, J., Typke, D., Nitsch, M., Koster, A. J., Hegerl, R., und Baumeister, W. (1997). Electron tomography of single ice-embedded macromolecules: Three-dimensional alignment and classification. *Journal of structural biology*, 120(3):387–395.

[Williams und Carter, 1996] Williams, D. B. und Carter, B. (1996). *Transmission Electron Microscopy*. Plenum Publishing Coorporation, 1 edition.

[Winkler et al., 2009] Winkler, H., Zhu, P., Liu, J., Ye, F., Roux, K. H., und Taylor, K. A. (2009). Tomographic subvolume alignment and subvolume classification applied to myosin V and SIV envelope spikes. *Journal of structural biology*, 165(2):64–77.

[Xu et al., 2011] Xu, M., Beck, M., und Alber, F. (2011). Template-free detection of macromolecular complexes in cryo electron tomograms. *Bioinformatics (Oxford, England)*, 27(13):i69–76.

[Xu et al., 2012] Xu, M., Beck, M., und Alber, F. (2012). High-throughput subtomogram alignment and classification by Fourier space constrained fast volumetric matching. *Journal of Structural Biology*.

[Yu und Frangakis, 2011] Yu, Z. und Frangakis, A. S. (2011). Classification of electron sub-tomograms with neural networks and its application to template-matching. *Journal of structural biology*, 174(3):494–504.

[Zsigmondy und Thiessen, 1925] Zsigmondy, R. und Thiessen, P. A. (1925). Das kolloide Gold. Kolloidforschung in Einzeldarstellungen. *Archiv der Pharmazie*, (1).

Abkürzungsverzeichnis

$\Delta\alpha$ Winkelinkrement
K Ähnlichkeitsmatrix
M Maske
W *Missing Wedge*
AV3 AV3
CC *Constrained Correlation*
CPCA *Constrained Principal Component Analysis*
EM *Expectation Maximization*
ER Endoplasmatisches Retikulum
FRK Fourier-Ring-Korrelation
KEM Kryoelektronenmikroskopie
KET Kryoelektronentomographie
KTF Kontrasttransferfunktion
LNXC lokal normierte Kreuz-Korrelation
MSA Multivariaten Statistischen Analyse
MCO-A *Multiple Correlation Optimization (Annealing)*
MCO-EM *Multiple Correlation Optimization*
MRA *Mutli Reference Alignment*
NXC normierte Kreuz-Korrelation
PCA Eigenwertzerlegung (*Principal Component Analysis*)
PDB *Protein Databank*
ROC *Receiver Operating Characteristics*
SA *Simulated Annealing*
SNR Signal Rausch Verhältniss (*Signal to Noise Ratio*)
SVD Singulärwertzerlegung (*Singular Value Decomposition*)
TOM TOM
TM *Template Matching*
WB gewichtete Rückprojektion (*Weighted Backprojection*)
XC Kreuz-Korrelation
XML *eXtensible Markup Language*
XPath *XML Path Language*
XSLT *eXtensible Style Sheet Language*

Danksagung

Diese Arbeit entstand zwischen Oktober 2008 bis April 2012 in der Forschungsgruppe von Dr. Friedrich Förster in der Abteilung Molekulare Strukturbiologie des Max-Planck-Instituts für Biochemie in Martinsried. Ich möchte mich bei allen Kollegen für die Zusammenarbeit, Unterstützung und das angenehme Arbeitsklima bedanken.

Insbesondere danke ich meinem Mentor Dr. Friedrich Förster für die herausragende Betreuung während der ganzen Zeit. Seine Erfahrung und Ideen haben meine Arbeit entscheidend geprägt und ich hoffe, dass ich mir über die Zeit einiges habe abschauen können. Vielen Dank!

Herrn Prof. Baumeister möchte ich für die Möglichkeit danken, in seiner Abteilung promovieren können, ebenfalls für die Infrastruktur die ich benutzen konnte wie auch für die finanzielle Unterstützung über den ganzen Zeitraum.

Durch meinen Freund und späteren Kollegen Florian Beck habe ich den Kontakt zum MPI in 2006 erst bekommen. Im Prinzip war meine Assistenzstelle in der Gruppe von Stephan Nickell der Anfang der Promotion. Florian brilliert nicht nur auf den Brettern in den Bergen und am Meer, sondern auch am Keyboard. Vielen Dank für die Zusammenarbeit am MPI und Freundschaft seit Moliets 2002.

Yuxiang Chen möchte ich für die exzellente Zusammenarbeit danken, die er als Praktikant, Masterstudent und letztendlich als Doktorand am PyTom Projekt erbracht hat. Seine konstruktive Kritik und Entwicklungen hat das Projekt weit vorangebracht.

Stefan Pfeffer hat hauptsächlich die in dieser Arbeit verwendeten Daten produziert und als Benutzer wesentlich zur Verbesserung von PyTom beigetragen. Luis Kuhn Cuellar hat einen Großteil der Rekonstruktionsmethoden nach PyTom portiert.

Thomas Hoffman, Nicolas Schrod und Claudia Klingelhöfer haben sich die Zeit genommen diese Arbeit zu lesen und zu korrigieren.

Zum Schluss möchte ich meinen Eltern und meiner Frau Iris Hrabe-Ritzert für die Geduld während der letzten Jahre danken, ebenso für die Motivation und Unterstützung auf die ich mich immer verlassen kann.