Title
CryoEM-Enabled Approaches to Structure Determination of Endogenous Protein Complexes Implicated in the Pathogenesis of the Malaria Parasite Plasmodium falciparum

Author
Ho, Chi-Min

Publication Date
2018

UNIVERSITY OF CALIFORNIA

Los Angeles

CryoEM-Enabled Approaches to Structure Determination of Endogenous Protein Complexes Implicated in the Pathogenesis of the Malaria Parasite *Plasmodium falciparum*

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Molecular Biology

by

Chi-Min Ho

2019

ABSTRACT OF THE DISSERTATION

CryoEM-Enabled Approaches to Structure Determination of Endogenous Protein Complexes Implicated in the Pathogenesis of the Malaria Parasite *Plasmodium falciparum*

by

Chi-Min Ho

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2019

Professor Z. Hong Zhou, Chair

The complexity and breadth of the host-cell remodeling machinery in the malaria parasite *P. falciparum* make it a rich and exciting system for the study of host-pathogen interfaces, particularly as many of the molecular mechanisms underlying this parasite's ability to hijack human red blood cells remain unclear. Furthermore, the *P. falciparum* proteome has proven recalcitrant to structural and biochemical characterization using recombinant methods, making it an intriguing model system for the development of new methods that leverage recent advances in cryoEM to enable structural studies of previously intractable systems at near-atomic resolution. The work presented here makes significant contributions in both these regards.

First, we use a targeted, CRISPR-enabled "top down" approach to determine near-atomic resolution structures of the unique malaria parasite translocon PTEX, which we purified directly from *P. falciparum* parasites in multiple functional states, yielding the first near-atomic resolution cryoEM structures of a protein isolated directly from an endogenous source using an epitope tag inserted into the endogenous locus with CRISPR-Cas9 gene editing.

We then developed a "bottom up" endogenous structural proteomics method whereby protein complexes enriched directly from the cellular milieu are identified by imaging and

structure determination using cryoEM and mass spectrometry. As a proof of principle, we successfully used this approach to obtain near-atomic resolution structures of multiple protein complexes from the *P. falciparum* proteome, which has previously proven recalcitrant to expression in recombinant systems, precluding structure determination by X-ray crystallography or NMR.

The body of work described here addresses a known need for methods that overcome the limitations of structural biology approaches that depend on recombinant systems, opening the door for high resolution structure determination of a vast number of previously intractable biological systems.

The dissertation of Chi-Min Ho is approved.

Patricia J. Johnson

Daniel E. Goldberg

Feng Guo

Jeffrey Floyd Miller

Todd O. Yeates

Hong Zhou, Committee Chair

University of California, Los Angeles

2019

Table of Contents

List of Figures

List of Tables

Acknowledgements

My deepest thanks to my thesis advisor and mentor, Dr. Z Hong Zhou, for his unwavering support and guidance throughout the past four years. He has surpassed my expectations at every turn, allowing me the freedom to explore, make mistakes, and grow as a scientist and as an individual in the pursuit of an ambitious, high risk project. He works endlessly to provide the environment and the resources needed to accomplish cutting edge research in cryoEM, and I firmly believe that the work I have accomplished under his mentorship would simply not have been possible anywhere else. Not only does he make scientific contributions at an elite level, but he manages to do so while remaining a generous, giving scientist who believes in second chances and taking risks, and who does everything in his power to provide his people with the support and resources they need to succeed. He is the kind of scientist I aspire to be, and I have been truly fortunate to have him as my mentor.

Thank you to Dr. Pascal Egea for his support. Thank you to past and present members of the Zhou lab, for their support, encouragement, and camaraderie during my time at UCLA. Thank you to Dr. Patricia Johnson and my fellow Microbial Pathogenesis Training Grant Trainees, for your support and guidance. Thank you to my committee members, for their continued time, support, and guidance.

I would like to thank to my collaborators, Dr. Daniel Goldberg and Dr. Josh Beck, for being a delight to work with, and for welcoming me into the world of parasitology with open arms.

I have been very fortunate to have had a long history of strong, supportive, and giving mentors throughout my career, who have advocated for me as a scientist and as a person. Thank you to my dear friend and mentor, Dr. Min Li, for always believing in me and advocating fiercely on my behalf, and without whom I would not be here today. Thank you to my wonderful mentors and friends from Novartis, Dr. Min Li, Dr. Janet Sim, and Dr. Isabel Zaror, who enthusiastically advocated for me, and to my entire family at Novartis, who I miss every day.

I am deeply grateful to Dr. Robert Stroud for his unending support and mentorship over the past thirteen years, and for introducing me to the joys of membrane proteins and structural biology. Special thanks to Suzan Betheil, for all of her help, support, and friendship over the past thirteen years.

Thank you to the UCLA MBIDP staff and administration, including Dr. Luisa Iruela-Arispe, Jennifer Miller, Helen Houldsworth, Ashley Terhorst, and Stephanie Cuellar, as well as

Brian Phan and Weiling Chen, for their constant support in administrative matters pertaining to the UCLA MBIDP, the Department of Biological Chemistry and the Microbial Pathogenesis Training Grant.

Thank you to my friends for their understanding and support. A special thank you to my friend Josh Laniado, for being my person throughout this journey, and for countless intense, ridiculous, and inspiring conversations, scientific and otherwise. I would also like to thank my friend and roommate Lynn Barstow for being my anchor to the real world and making sure I remembered to eat. Thank you to one of my oldest friends, Bobak Pezeshki, for his unflagging belief in me.

My sincerest thanks to my family – to my parents, I-Hwa and Dr. Kai-Ming Ho, for their unconditional support, patience, and understanding, and to my sister, Shing Shing Ho, for always being my voice of reason. Thank you to my best friend, partner in crime, and human RELION-Wiki resource Dr. Anthony W. P. Fitzpatrick, who convinced me to give cryoEM a try, and has been with me every step of the way ever since. And finally, thank you to my dog, Mighty, who has been my constant companion and a loyal source of emotional support and comic relief throughout this journey.

Chapter 2 is adapted from Ho CM, Beck JR, Lai M, Cui Y, Goldberg DE, Egea PF, Zhou ZH. Malaria Parasite Translocon Structure and Mechanism of Effector Export. *Nature* **561**, 70-75 (2018); doi:10.1038/s41586-018-0469-4. Nature Research does not require authors of original (primary) research papers to assign the copyright of their published contributions. Authors grant Nature Research an exclusive licence to publish, in return for which they can reuse their papers in their future printed work without first requiring permission from the publisher of the journal. This research was supported in part by grants from National Institutes of Health (R21AI125983 to P.F.E., R01GM071940/AI094386/DE025567 to Z.H.Z. and K99/R00 HL133453 to J.R.B.). P.F.E. is the Alexander and Renee Kolin Endowed Chair in Molecular Biology and Biophysics. CM.H. acknowledges funding from the Ruth L. Kirschstein National Research Service Award (AI007323). We thank the UCLA Proteome Research Center for assistance in mass spectrometry and acknowledge the use of instruments in the Electron Imaging Center for Nanomachines supported by UCLA and grants from NIH (S10RR23057, S10OD018111 and U24GM116792) and NSF (DBI-1338135 and DMR-1548924). We thank Anthony W. P. Fitzpatrick for input on cryoEM aspects of the project and Judy Su for helping with Figure 1a. **Author Contributions:**

CMH, PFE and ZHZ initiated the project; JRB generated parasite lines, harvested parasites, performed complementation experiments and helped write the paper; CMH purified the sample, screened purified samples by negative stain, prepared cryoEM grids, acquired and processed the cryoEM data, interpreted the structures and wrote the paper; ML built and refined the atomic models and helped interpret the structures; YC helped with sample freezing; ZHZ supervised the cryoEM aspects of the project, interpreted the structures and wrote the paper; PFE supervised biochemical aspects of the project and helped interpret the structures; DEG supervised parasitology aspects of the project. DEG, ML, and PFE helped edit the paper.

Chapter 3 is adapted from a manuscript submitted to Nature Methods on January 28, 2019: Ho CM, Li X, Lai M, Terwilliger TC, Beck JR, Wohlschlegel J, Goldberg DE, Fitzpatrick AWP, Zhou ZH. Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu. Manuscript submitted to Nature Methods, January 2019. This work was supported in part by grants from National Institutes of Health (R21AI125983 to P.F.E., R01GM071940/AI094386/DE025567 to Z.H.Z. and K99/R00 HL133453 to J.R.B.). CM.H. acknowledges funding from the Ruth L. Kirschstein National Research Service Award (AI007323). XL acknowledges funding from the China Scholarship Council (CSC). We thank the UCLA Proteome Research Center for assistance in mass spectrometry and acknowledge the use of instruments in the Electron Imaging Center for Nanomachines supported by UCLA and grants from NIH (S10RR23057, S10OD018111 and U24GM116792) and NSF (DBI-1338135 and DMR-1548924). **Author Contributions:** CMH, AWPF and ZHZ initiated the project; JRB cultured and harvested parasite material; CMH purified the sample from parasite pellets, screened purified samples by negative stain, optimized sample freezing conditions for cryoEM, acquired and processed the cryoEM data, interpreted the structures, designed the endogenous structural proteomics workflow, helped design *cryoID*, and wrote the paper; ML built and refined the atomic models and helped interpret the structures; CMH, XL and ML designed the

cryoID workflow. XL developed and benchmarked *cryoID* and helped write the paper. XL and CMH worked with TCT to write and optimize PHENIX.*sequence_from_map*. ZHZ supervised the cryoEM aspects of the project, interpreted the structures and wrote the paper; DEG supervised parasitology aspects of the project. AWPF, ML, TCT, JRB, and DEG helped edit the paper; all authors approved the paper.

VITA

Chi-Min Ho

Education

- **University of California, Los Angeles:** PhD in Molecular Biology, dissertation defended Dec 14, 2018, degree expected Mar 2019.
- **University of California, Berkeley:** BA in Molecular and Cell Biology, Biochemistry and Molecular Biology Emphasis, Aug 2004

Career History

- **University of California, Los Angeles, Thesis Advisor: Hong Zhou**, PhD candidate, Molecular Biology Institute (Oct 2014 – Dec 2018)
  CryoEM-Enabled Approaches to Structure Determination of Endogenous Protein Complexes Implicated in the Pathogenesis of the Malaria Parasite *Plasmodium falciparum*

- **Novartis Institutes for BioMedical Research, Protein Sciences Group, PI: Dr. Isabel Zaror,** Scientific Associate II, Scientist I (Oct 2011 – Feb 2014, Mar 2014 – Aug 2014)
  Infectious Diseases Drug Discovery.

- **University of California, San Francisco, Robert Stroud Group,** Staff Research Associate II (Jan 2006 - May 2011)
  Membrane Protein Purification and Characterization, X-ray Crystallography.

- **University of California, Berkeley, Jamie H.D. Cate Group,** Undergraduate Research Assistant (Sep 2002 - May 2004)
  X-ray Crystallography of spliceosomal core proteins.

- **University of California, Berkeley** (Sep 2002 to Dec 2004)
  General Biology Tutor, Study Group Leader – University of California, Berkeley, Student Learning Center.

- **Iowa State University**
  - **Gloria Culver Group:** (Jun 2002 – Aug 2002) Undergraduate Research Assistant – *E. coli* Ribosomal Assembly, Ribosomal Protein Purification
  - **Eve Wurtele Group:** (Jun 2001 – Aug 2001) Undergraduate Research Assistant – Metabolic Networking.
  - **Elizabeth and Steven Lonergan Group:** (Jun 2000 – Aug 2000) Undergraduate Research Assistant – Meat Science.

Publications

Ho CM, Beck JR, Lai M, Cui Y, Goldberg DE, Egea PF, Zhou ZH. Malaria Parasite Translocon Structure and Mechanism of Effector Export. **Nature** 561, 70-75 (2018).

Ho CM, Li X, Lai M, Terwilliger TC, Beck JR, Wohlschlegel J, Goldberg DE, Fitzpatrick AWP, Zhou ZH. Endogenous Structural Proteomics of the Malaria Parasite *P. falciparum*. Manuscript under review.

Ho CM, Zhou ZH. Title To Be Determined. Invited review manuscript in preparation for Trends in Parasitology.

Gruswitz F, Chaudhary S, Ho JD, Schlessinger A, Pezeshki B, **Ho CM**, Sali A, Westhoff CM, Stroud RM. Structure of Human Rh based on structure of RhCG at 2.1Å. **Proc Natl Acad Sci U S A.** 2010 May 25;107(21):9638-43.

Li, M, Hays FA, Roe-Zurz Z, Vuong L, Kelly L, **Ho CM**, Robbins RM, Pieper U, O'Connell JD 3rd, Miercke LJ, Giacomini KM, Sali A, Stroud RM. Selecting optimum eukaryotic integral membrane proteins for structure determination by rapid expression and solubilization screening. **J Mol Biol**. 2009 Jan 23;385(3):820-30.

Miercke, Larry JW, Robbins, Rebecca A, **Ho, Mimi (Chi-Min)**, Sandstrom, Andrew, Bond, Rachel K, Stroud, Robert M. Monitoring and Optimizing Detergent Concentration for Membrane Protein Crystallization While Following Protein Homogeneity. **Biophysical Journal**. 2010 Jan: 98(3):50a-50a

---

Honors and Awards

Research Presentations

- Winner, Best Invited Short Talk, 2018 Biology of Host-Parasite Interactions Gordon Research Conference (Best of ~120 abstracts)
- 1st Prize, 2018 Frontiers and Careers in cryoEM Symposium Poster Competition (1st of 15)
- Winner, #1 Best Invited Full Length Talk, Molecular Parasitology Meeting 2018 (1st of 227 abstracts)
- ACMCIP Woods Hole Speaker Award: Travel Award and Invitation to Speak at American Society for Tropical Medicine and Hygiene Annual Meeting 2018

Funding

- NIH NRSA Institutional Pre-doctoral Training Grant (T32): Microbial Pathogenesis (2016-2019)
- Whitcome Pre-doctoral Fellowship in Molecular Biology (2016-2017)
- NSF GRSA Honorable Mention (2016)

Teaching

- UCLA Life Sciences Certificate of Distinction in Teaching (2015-2016)

Chapter 1

Introduction

1.1 *Plasmodium falciparum* and Malaria

Malaria is a deadly disease caused by *Plasmodium* parasites. Nearly half of the world's population, an estimated 3.8 billion people, are currently at risk of contracting *Plasmodium* parasites, with >200 million cases annually leading to almost 500 thousand deaths per year, predominantly in children five years and younger[1]. Of the five *Plasmodium* species known to cause malaria in humans, *Plasmodium falciparum* is the most lethal and accounts for the majority of severe and fatal cases.

1.1.1 *Plasmodium falciparum* Life Cycle and Biology

Plasmodium falciparum has a complex bimodal life cycle (**Fig. 1.1**) involving an invertebrate host, the *Anopheles* mosquito, and a vertebrate host, *Homo sapiens*. After undergoing sexual reproduction in the midgut of an *Anopheles* mosquito, sporozoite-stage parasites migrate to the salivary glands of the mosquito. The next time the mosquito bites a human, sporozoite-stage parasites are injected into the skin of the human. From there, the sporozoites quickly make their way through the bloodstream to the liver, where they invade and replicate inside hepatocytes, eventually maturing into a form known as the merozoite, which is capable of infecting human erythrocytes.

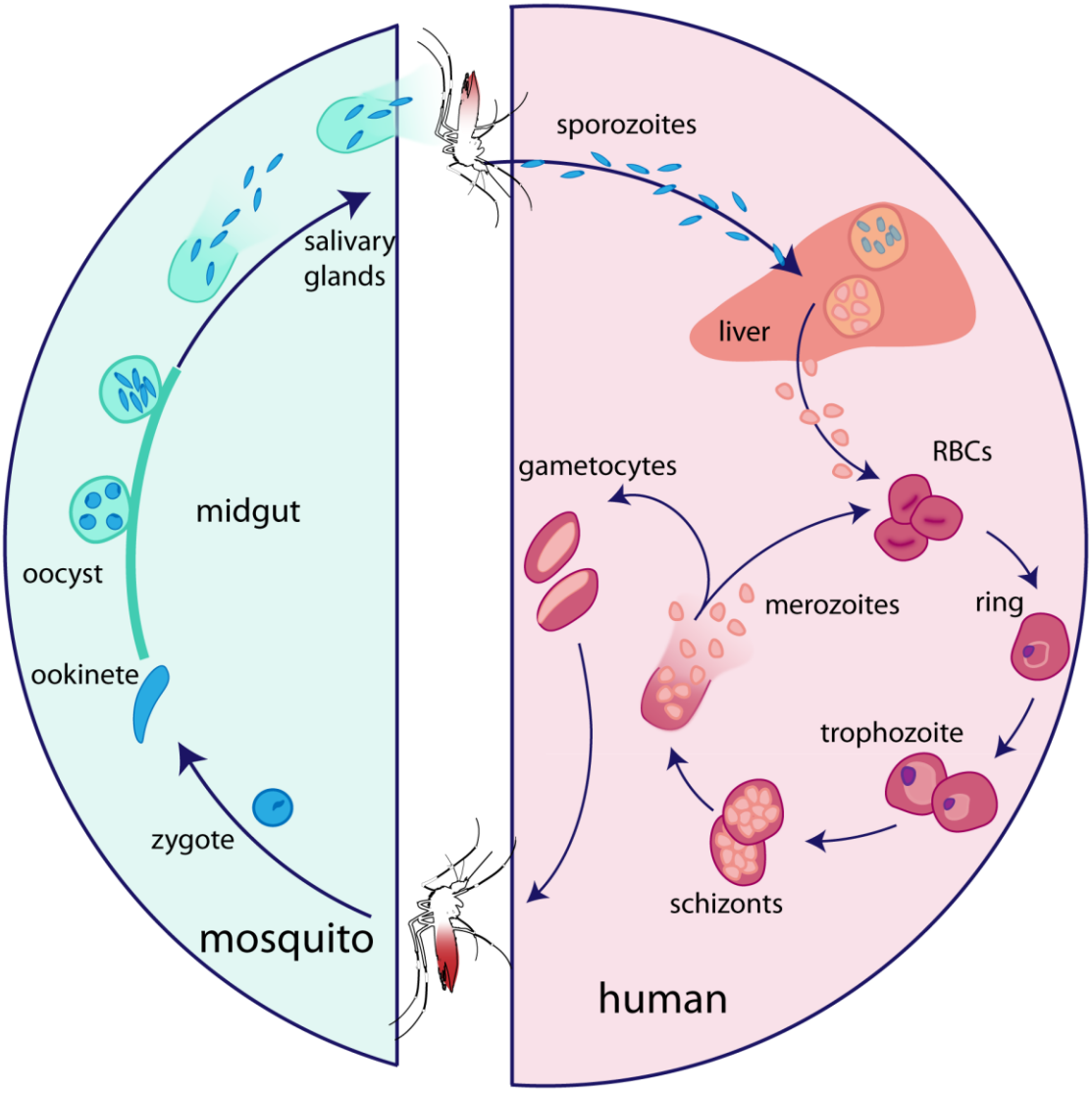


Figure 1.1 | The *Plasmodium falciparum* Life Cycle.

Mature merozoites then egress from the liver cells, re-enter the bloodstream, and invade erythrocytes to begin the intraerythrocytic stages of the parasite life cycle, known as the blood stages. After invasion, the intraerythrocytic parasites advance through the ring, trophozoite and schizont stages before ultimately forming daughter merozoites[2]. The infected erythrocyte then ruptures, releasing mature merozoites into the bloodstream, where they rapidly invade new erythrocytes, beginning another iteration of the cycle. These

repeated cycles of invasion, replication, and egress are responsible for all of the clinical symptoms of malaria, including periodic febrile episodes, coinciding with erythrocyte rupture and merozoite egress, which are characteristic of malaria.

During this process, a small proportion of the blood stage parasites undergo gametogenesis, forming male and female gametocytes in the bloodstream. The gametocytes are ingested and passed back into the invertebrate host when the infected human is bitten by another mosquito, thus beginning another round of the parasite life cycle.

1.1.2 Effector Export in *Plasmodium falciparum* Pathogenesis

While most intracellular pathogens export a limited repertoire of effector proteins to co-opt existing host-cell metabolic machineries, the malaria-causing *Plasmodium falciparum* parasite exports more than 10% of its proteome into its host, the human red blood cell, which the parasite inhabits and reproduces within during the blood stages of its life cycle[2, 3]. The hundreds of proteins in the *P. falciparum* exportome extensively remodel host erythrocytes, creating the infrastructure needed to import nutrients, export waste, and evade splenic clearance of infected erythrocytes[4]. The export of these hundreds of proteins is complicated by the fact that the malaria parasite conceals itself inside a parasitophorous vacuole (PV) derived from invagination of the host cell plasma membrane during invasion[5] (**Fig. 1.2**). Following secretion into the PV, proteins destined for export must be unfolded and transported across the PV membrane (PVM) into the host cell in an ATP-dependent process[6].

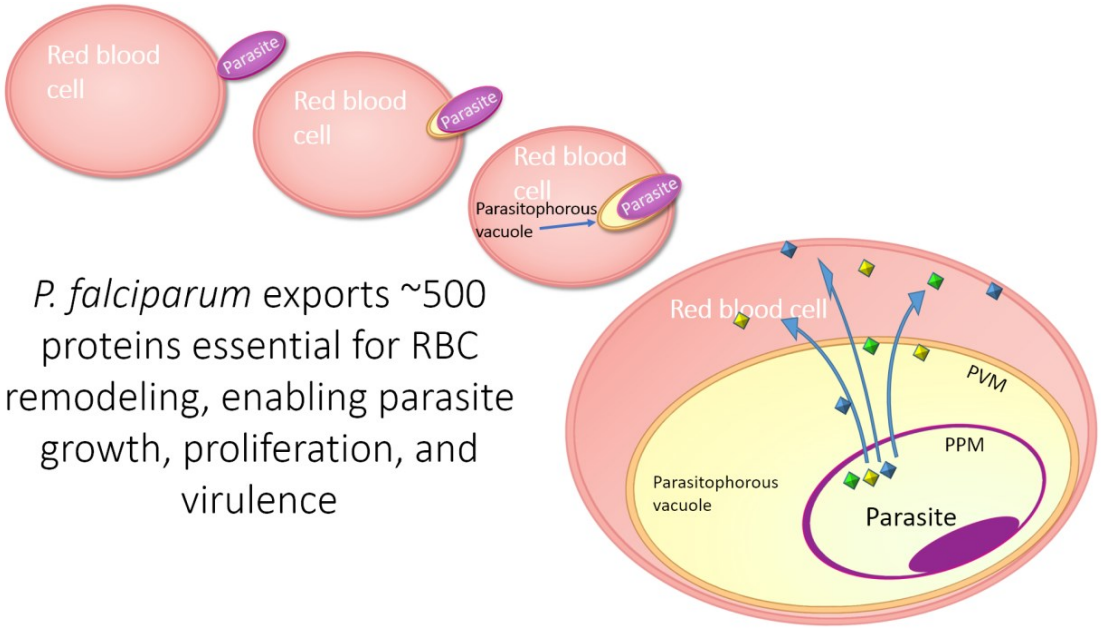


Figure 1.2 | Formation of the parasitophorous vacuole. Invagination of the erythrocyte plasma membrane during invasion leads to the formation of a parasitophorous vacuole (PV) around the parasite. Effector proteins are secreted into the PV, where they are then unfolded and transported across the PV membrane (PVM).

1.1.3 The PEXEL Export Signal Sequence

In 2004, two independent groups identified the existence of a 5-residue motif, RxLxE/Q/D that appeared to be common to all exported proteins known at the time[7, 8]. Subsequent experiments demonstrated that this sequence was necessary for successful targeting of the attached protein to the host cytosol. Dubbed the PEXEL motif, it was found that this 5-residue "host-targeting" motif was conserved across the *Plasmodium* species, generally located ~35 residues downstream of an ER signal peptide. The PEXEL motif was used to

predict additional exported proteins, resulting in the identification of ~400 proteins in the *P. falciparum* proteome that also bore the same 5-residue motif[7-9].

In subsequent work, the PEXEL was determined to be a recognition site for cleavage by plasmepsin V, an aspartic protease localized in the *P. falciparum* ER. Cleavage of the PEXEL motif C-terminal to the leucine residue by plasmepsin V, followed by acetylation of the resulting new N-terminus, was found to be essential for effector protein export[10-13]. The arginine and leucine residues in the motif have been found to be required for cleavage, while the E/Q/D residue appears to be required for translocation across the PVM[9, 14].

Intriguingly, a growing number of exported effector proteins have recently been identified that lack a PEXEL motif[15]. In fact, a consensus export signal sequence has yet to be identified among such proteins identified thus far. Dubbed PEXEL-negative exported proteins, or PNEPs, there seems to be little in common between these proteins, although many of them seem to contain a transmembrane domain near the N-terminus. Both soluble and membrane protein PNEPs have now been identified, both with and without a classical ER signal peptide[16, 17].

In total, this points toward a *P. falciparum* exportome of >500 proteins, equivalent to more than 10% of the *P. falciparum* proteome.

1.1.4 Genetic and Biochemical Evidence for Mechanism of Export

As effector export is essential to *P. falciparum* survival and pathogenesis during the blood stages of the parasite life cycle, the mechanism of effector protein export has been a subject of great interest in the field. In 2009, de Koning-Ward *et al*. identified a five-membered

protein complex, dubbed the *Plasmodium* Translocon of Exported Proteins, which they hypothesized to form an effector protein translocating machine at the PVM[18]. In identifying components of the complex, they specifically looked for proteins that were specific to the *Plasmodium* genus, included a power source (likely an ATPase), localized to the apical end of merozoites, but to the PVM in ring-stage parasites, were essential for intraerythrocytic growth, and specifically bound to known exported effector proteins. Using these constraints, they identified the five components of PTEX: HSP101, an ATP-driven protein unfoldases, PTEX150, a novel protein of unknown structure or function which they hypothesized to play a structural role, EXP2, a novel protein of unknown structure or function which they hypothesized might form a transmembrane pore, and two accessory proteins, PTEX88 and TRX2. While the structure of TRX2 has since been solved[19], PTEX88 remains a novel protein of unknown structure, and the exact functions of both proteins in export remain unclear.

Since its discovery in 2009, a body of genetic and biochemical work from multiple groups has shown that HSP101, PTEX150, and EXP2 are essential for protein translocation and parasite survival. Of note, disrupting PTEX by 1) conditionally dissociating HSP101 from PTEX150 and EXP2 [Ref [20]], or 2) knockdown of HSP101 or PTEX150 [Ref [21]] both blocked effector protein export, leading to the accumulation of effector proteins in the lumen of the PV. Intriguingly, these experiments revealed that disrupting PTEX blocked the export of all classes of exported proteins, regardless of whether they were soluble or membrane proteins, PEXELs or PNEPs. This exciting finding suggests that PTEX is the single gatekeeper through which the entire *Plasmodium* exportome must pass to reach the erythrocyte cytosol. No redundant pathways appear to exist.

Furthermore, parasites expressing cargo proteins fused to a conditionally foldable DHFR domain have been used to demonstrate that effector proteins destined for export, be they PEXEL or PNEP, membrane protein or soluble, must be unfolded before being transported across the PVM[6, 16]. Using this system, Mesen-Ramirez *et al.* further showed that in the presence of a stabilizing ligand, cargo-DHFR fusions proteins stalled in the export process and blocked the export of other cargo proteins that lacked a DHFR domain[22]. Although they were able to show that EXP2 co-immunoprecipitates with stalled cargo-DHFR fusion proteins, they were unable to definitively conclude that the cargo was in fact threaded through a transmembrane channel formed by EXP2 in the absence of structural information.

Despite the genetic and biochemical data demonstrating that the PTEX core components were indeed essential for effector export in *P. falciparum*, there was no irrefutable evidence of PTEX translocation activity prior to the work described in the chapters below. In the absence of high resolution structural information, it was not possible to exclude the possibility that the PTEX components acted upstream of the actual translocation event, either as a complex or individually.

1.1.5 Challenges in Structural Biology of *Plasmodium falciparum*

Since the emergence of recombinant DNA techniques and generic peptide purification tags in the late 1980s, the standard approach in structural biology has been to purify tagged proteins overexpressed in heterologous systems such as *E. coli* or *S. cerevisiae*[23-31]. The ability to produce large quantities of pure protein from recombinant systems has been instrumental in the exponential increase in high resolution structures solved over the past

three decades[28]. Unfortunately, many pathogens of high medical relevance are recalcitrant to structural characterization using conventional recombinant approaches. This is particularly so in the case of *Plasmodium falciparum*, where the paucity of high resolution structural and functional information is compounded by the fact that 50% of the *P. falciparum* proteome is novel[32-34], bearing no similarity to existing structures in the PDB. Many promising *P. falciparum* drug targets are membrane proteins, but there are only two unique integral membrane protein structures from *P. falciparum* in the PDB. To address the paucity of structures from *P. falciparum* and other challenging pathogens, we have developed and implemented methodologies for structure determination from challenging endogenous sources, described below.

1.2 Single-Particle Cryoelectron Microscopy

Cryoelectron microscopy (cryoEM) is an imaging technique used to visualize biological samples at high resolution using phase contrast transmission electron microscopy (TEM). Phase contrast TEM is commonly used in materials science to directly image the atomic structure of materials such as metals, semiconductors, nanoparticles, graphene, and carbon nanotubes up to a resolution of 0.5Å. However, the high energy electron beams used to image samples in material science applications are too harsh for imaging biological samples such as viruses, tissue sections, and individual macromolecular complexes such as proteins, which cannot withstand high vacuum and are highly susceptible to radiation damage resulting in bond breakage and the loss of mass.

In phase contrast TEM, as electrons emerge from the electron source, electrons are focused and condensed into a parallel beam using electromagnetic lenses called condenser

lenses (**Fig. 1.3**). As the electron beam passes through the sample, interaction with particles in the sample scatters some of the electrons, altering their phases. On the far side of the sample, the scattered and transmitted electrons are focused by another electromagnetic lens known as an objective lens into an image, which is then magnified by a third electromagnetic lens known as a projection lens. The magnified image can then be captured using a detector or photographic film. The differences in phase of the electrons as they exit the sample are detected as differences in intensity at the detector, giving rise to phase contrast.

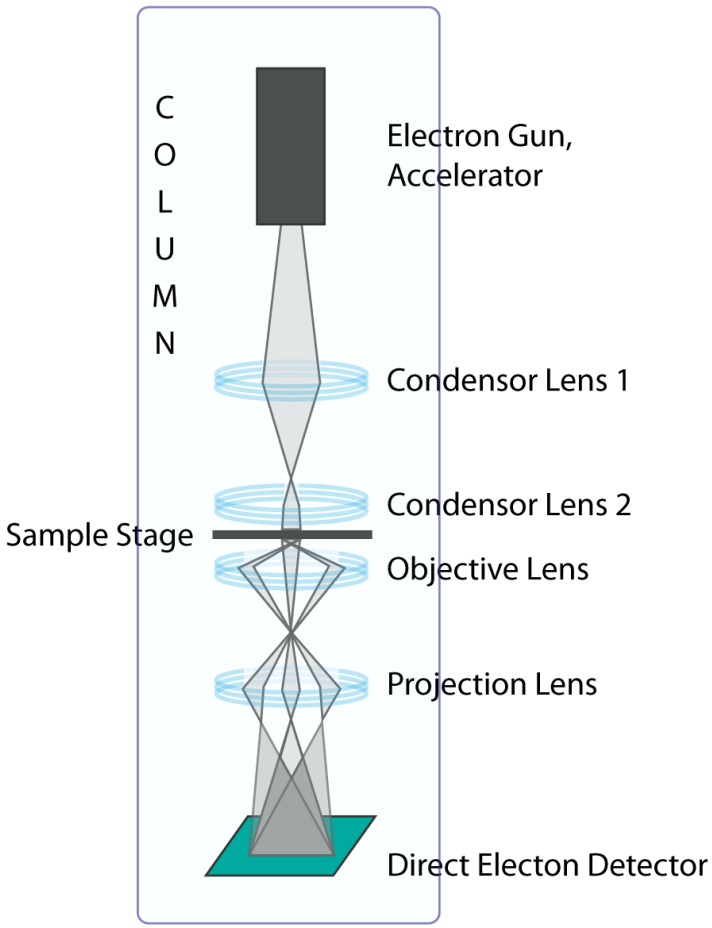


Figure 1.3 | Schematic of a transmission electron microscope.

High resolution cryoEM as it exists today was developed as a technique combining several strategies that together allow for the successful imaging of biological samples at near-atomic resolution using phase contrast TEM, with three key innovations leading to a "resolution revolution" . For their contributions in these three key areas of innovation, Jacques Dubochet, Joachim Frank and Richard Henderson earned the 2017 Nobel Prize in Chemistry. The three key innovations that make cryoEM possible are the following:

1) Preservation of biological samples at cryogenic temperatures (-195°C) in a protective layer of amorphous, vitreous ice.
2) Highly sensitive direct electron detectors capable of single electron counting.
3) Improved computer algorithms for image processing and three-dimensional (3D) reconstruction.

1.2.1 Preservation in a frozen-hydrated state is essential for high resolution structure determination of biological samples by cryoEM

To obtain high resolution images of biological samples, it is essential that biological samples are preserved at cryogenic temperatures in a very thin layer of amorphous or vitreous ice. This protects the samples from the high vacuum in the column of the electron microscope, minimizes the radiation damage from the electron beam, and thus preserves the high resolution information in the resulting images.

To accomplish this, a small volume of as sample, such as a protein complex, suspended in aqueous buffer, is applied to an electron microscopy grid (**Fig. 1.4a**). Typical electron microscopy grids consist of a small disk of copper or gold mesh supporting a layer of amorphous or lacey carbon that is perforated with holes, which can range in diameter

from 0.6-3µm or more (**Fig. 1.4d-f**). The majority of the sample is then blotted away with filter paper, leaving a very thin film of sample and buffer left across the grid (**Fig. 1.4b**). The grid is then plunged rapidly into liquid ethane at -195°C, freezing the sample and buffer so rapidly that the water does not have time to form a crystalline lattice, but rather forms a layer of amorphous or vitreous ice (**Fig. 1.4c, f-g**). The molecules or particles of the protein complex are preserved in a frozen-hydrated state, randomly oriented within the layer of vitreous ice (**Fig. 1.4f-g**).

Vitrification is essential for obtaining high resolution images of the sample, as crystalline ice damages biological samples and scatters electrons, introducing noise that interferes with high resolution information, degrading the image quality.

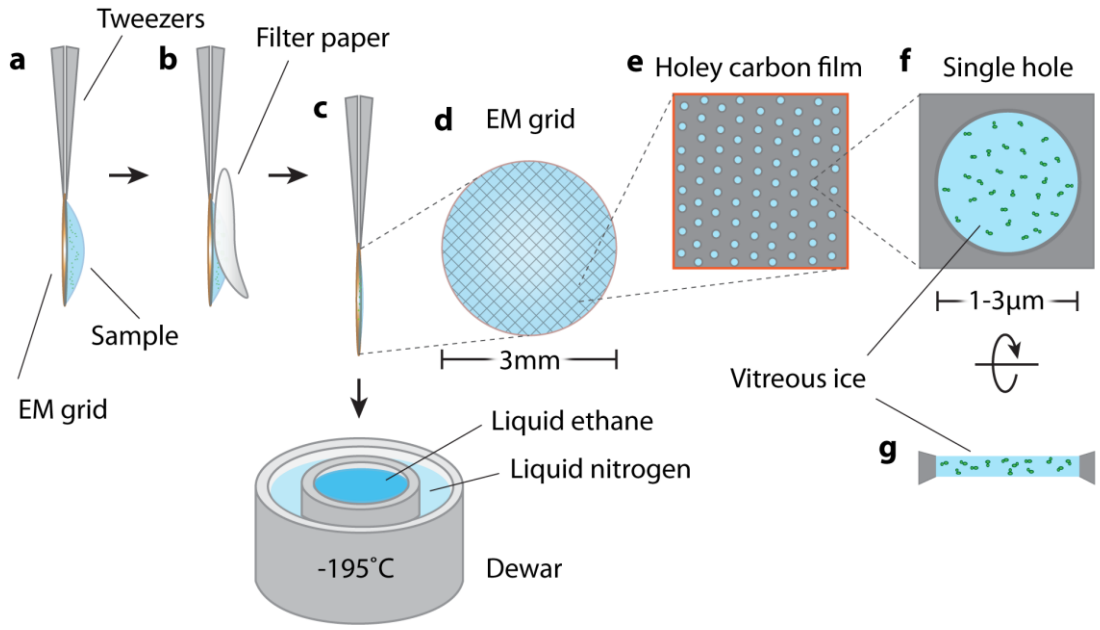


Figure 1.4 | CryoEM sample preparation. a, A small volume of sample is applied to an EM grid. **b,** Excess sample is blotted away. **c,** The grid is plunged into liquid ethane at -195°C, preserving the sample in a layer of vitreous ice. d-g, Sequence of enlarged views of the metal mesh EM grid (d), holey carbon film in a single square of the EM grid (e), top (f) and side (g) views of a single hole from the carbon film, containing the sample suspended in a thin layer of vitreous ice.

1.2.2 The critical role of direct electron detectors in structure determination of biological samples to near-atomic resolution by single-particle cryoEM

The vitrified grid containing the protein particles randomly suspended in the vitreous ice is then inserted into the column of the electron microscope. As the electron beam passes through the sample, the electrons interact with the particles, creating many different two dimensional (2D) projections, corresponding to different views of the randomly oriented protein particles, which are then captured on the detector below.

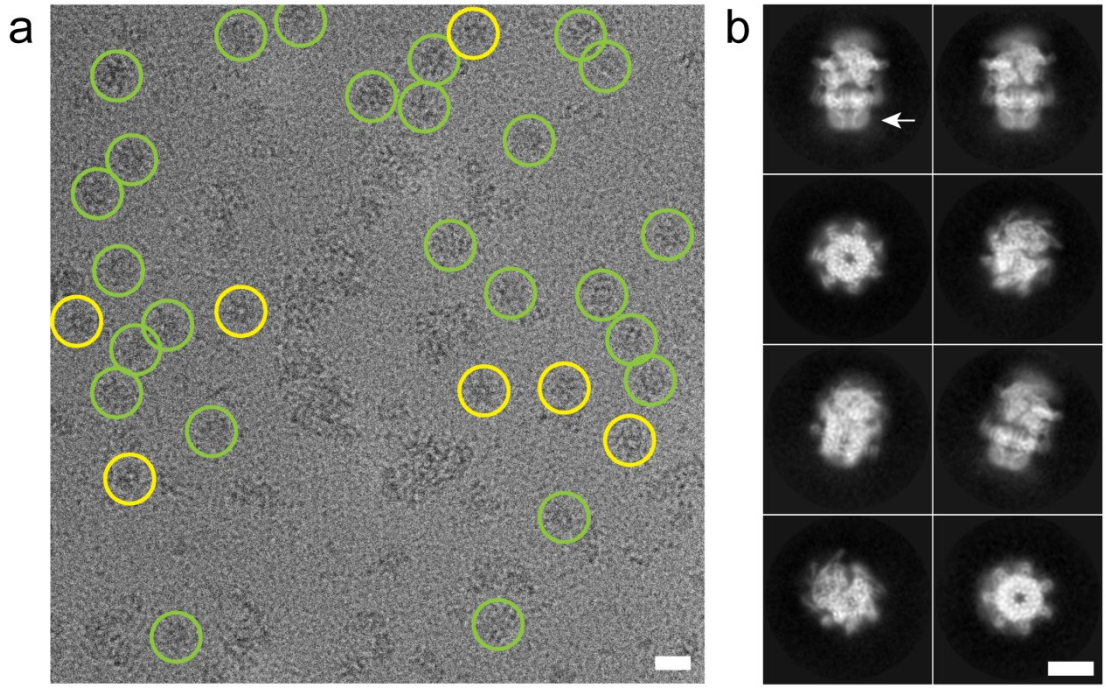


Figure 1.5 | Typical cryoEM micrograph and 2D class averages. Individual 2D projections, referred to as particles, are indicated with circles in the representative cryoEM micrograph in (**a**). 2D classification and averaging of particles from a cryoEM dataset yields class averages corresponding to 2D projections arising from the original sample in multiple orientations. Scalebars are 200Å and 100Å respectively.

Although preserving samples at cryogenic temperatures helps to mitigate radiation damage from the electron beam, biological samples must be imaged using a very low electron dose, as they are very sensitive to radiation damage from the electron beam and cannot withstand the high electron doses typically used for TEM imaging in materials sciences. Low-dose imaging preserves the integrity of biological samples, but produces images with very low contrast (poor to noise).

Prior to 2013, cryoEM images were recorded on either photographic film or indirect electron detectors. Indirect electron detectors rely on scintillators that convert primary electrons hitting the detector to photons. The photons are then recorded on a CCD or CMOS sensor. Unfortunately, the primary electrons scatter in the scintillator, generating photons in a volume exceeding a single pixel. As a consequence, the recording of each primary electron is blurred in the final image recorded on the sensor, and noise is introduced into the final recorded image, resulting in reduced signal to noise and the loss of high resolution details[35]. Because of this, indirect electron detectors have relatively good detective quantum efficiencies (DQEs) at low spatial frequencies, resulting in good contrast, but poor DQE at high spatial frequencies[36]. As such, these detectors struggled to separate signal from noise in the low contrast images yielded by low-dose imaging of biological samples. The images recorded using these detectors were noisy and much of the high resolution information was lost, precluding high resolution structure determination.

Conversely, photographic film exhibits significantly better DQE than indirect electron detectors at high frequencies, but relatively poorer DQE at low spatial frequencies resulting in poor contrast which necessitated imaging at higher defocus values to achieve enough

contrast in the final images for the picking and subsequent alignment of particles during image processing for single-particle reconstruction[36].

The implementation of direct electron detectors capable of single electron counting, which exhibit significantly higher DQEs than either photographic film or indirect electron detectors across both high and low spatial frequencies, played a pivotal role in bringing about the recent "resolution revolution" in cryoEM[37], although the first atomic structure of a biological macromolecule solved by cryoEM was achieved by Zhou and colleagues, using photographic film in a 300kV FEI Titan Krios cryo-electron microscope in 2010[38]. The vastly superior DQE of the new direct electron detectors make them significantly better at separating signal from noise in low dose images[35, 39-41]. The ability to capture images retaining high resolution information at unprecedented low doses, coupled with extremely fast frame rates, enabled the recording of multiple frames over a total exposure of several seconds. The resulting frames could then be aligned to correct for motion, or drift, of the sample during the exposure, thereby minimizing motion-induced blurring[37, 42]. As such, the images recorded on the direct electron detectors were less noisy, and much more high resolution information was retained, enabling routine structure determination of biological samples to near-atomic resolution and precipitating an exponential increase in the number of near-atomic resolution cryoEM structures in the years since they were first implemented in cryoEM in 2013[43].

1.2.3 Key software packages in the single-particle cryoEM resolution revolution

The images or micrographs containing the 2D projections of the sample must then be processed and analyzed to yield a high resolution three dimensional (3D) reconstruction of the original particle that gave rise to the 2D projections. To accomplish this, each individual 2D projection, often referred to as a particle, is selected and extracted from each micrograph, generally yielding hundreds of thousands of 2D particles. These hundreds of thousands of particles are first classified and sorted into self-similar views. The particles in each class are then averaged together to yield high resolution 2D class averages, each corresponding to a unique view of the original 3D sample particle. 2D classes corresponding to "junk" particles such as ice contamination or aggregates are discarded. Various algorithms can then be used to reconstruct a 3D volume from the particles contained within the remaining "good" 2D class averages, yielding the original 3D structure that gave rise to the 2D projections.

The advent of high resolution image collection enabled by direct electron detection and single electron counting introduced a need for new image processing software capable of accommodating the increase in resolution and sheer quantity of cryoEM images produced by the new detectors. In particular, there was a need for algorithms capable of classification and alignment of particle images at unprecedented accuracy – down to near-atomic resolution. The rise of a multitude of new image processing softwares and software suites not only addressed these needs, but also introduced an unprecedented level of automation to the process, with an added focus on user-friendly interfaces, making cryoEM image processing more accessible and routine to both seasoned electron microscopists and

newcomers to the field. A key contribution from Li *et al.*, MotionCor[37] and its successor, MotionCor2[42], are the industry standards for alignment and correction of beam-induced motion in micrograph movie frames.

Of the many software packages now available for single-particle image processing of cryoEM data, two are heavily used in the work described below, RELION[44, 45] and cryoSPARC[46]. The first of the new image processing software packages, and the most widely used today, RELION implements a Bayesian maximum likelihood approach to classification and refinement. cryoSPARC, a relative newcomer to the field, employs stochastic gradient descent (SGD) and branch-and-bound maximum likelihood optimization algorithms to enable unsupervised *ab initio* 3D classification and 3D refinement at unprecedented speeds, using a fraction of the computational power required by other software packages.

1.2.4 Structural Studies from Challenging Endogenous Sources Enabled by CryoEM

In the work described in the following chapters, we demonstrate that by leveraging the recent innovations in cryoEM described above, we are able to accommodate the lower yields and heterogeneity of samples enriched directly from endogenous sources, *via* either tag-free methods or epitope tags introduced in endogenous loci using CRISPR-Cas9 gene editing.

We present below the near-atomic resolution structures of the unique malarial translocon PTEX, purified directly from *P. falciparum* parasites in multiple functional states[47]. These structures are the first near-atomic resolution cryoEM structures of a protein

isolated directly from an endogenous source using an epitope tag inserted into the endogenous locus with CRISPR-Cas9 gene editing.

Furthermore, we present a structural proteomics method whereby protein complexes are enriched directly from the cellular milieu using tag-free methods and identified by imaging and structure determination using CryoEM and mass spectrometry. As a proof of principle, we used this approach to study *P. falciparum*, an organism that has proven recalcitrant to traditional structural biology approaches. By directly imaging components of the parasite cell lysate, we obtained near-atomic resolution (3.3Å) structures of multiple protein complexes implicated in the pathogenesis of malarial parasites, from a single cryoEM dataset.

The work described below demonstrates that the recent dramatic advances in cryoEM have opened the door for near-atomic resolution structure determination of a vast number of biological systems that were previously intractable. While high resolution structural study has until recently been restricted to proteins that are amenable to expression in recombinant systems and capable of either forming well-ordered crystals (for X-ray crystallography) or tumbling rapidly in solution (for nuclear magnetic resonance), the combination of cryoEM and CRISPR gene-editing now opens the door to near-atomic structure determination of virtually any protein from any organism, regardless of size or complexity. Additionally, without the need for finding constructs or mutants that are more stable or provide better crystal contacts, proteins can observed in their native, biologically relevant states. Furthermore, cryoEM has the added benefit that it is possible to achieve multiple high resolution structures of several different conformational states of a single

protein complex, or even several structures of completely unrelated protein complexes from a single cryoEM dataset.

Together, the CRISPR-based "top down" approach and the tag-free "bottom up" approach to structure determination from endogenous sources presented here represent the future of structural biology and cryoEM: direct visualization of protein complexes as they exist in the cellular milieu at near-atomic resolution.

1.3 Thesis Outline

This thesis is organized into four chapters as follows:

Chapter 1 provides the background to, and motivations behind, structural studies of protein complexes from *P. falciparum,* with a particular focus on the mechanism of effector export pathway. It also summarizes the objectives and main results of the thesis.

Chapter 2 describes the atomic resolution structures of PTEX, a unique malarial translocon, purified directly from the human malaria parasite *P. falciparum* in multiple functional states.

Chapter 3 describes a structural proteomics method we have developed, whereby protein complexes are enriched directly from the cellular milieu using tag-free methods and identified by imaging and structure determination using CryoEM, mass spectrometry, and *cryoID*, a program we developed that is able to semi-autonomously identify the proteins in medium to near-atomic resolution cryoEM maps without the need for any prior knowledge of their identities or primary sequences.

Finally, Chapter 4 summarizes the work presented in Chapters 2 and 3 and discusses the direction of future work.

1.4 References

1 WHO. World malaria report 2018. Report No. ISBN 978-92-4-156565-3, 210 (2018).

2 N. J. Spillman, J. R. Beck &D. E. Goldberg. Protein Export into Malaria Parasite-Infected Erythrocytes: Mechanisms and Functional Consequences. *Annu Rev Biochem* **84**, 813-841, doi:10.1146/annurev-biochem-060614-034157 (2015).

3 A. F. Cowman, J. Healer, D. Marapana &K. Marsh. Malaria: Biology and Disease. *Cell* **167**, 610-624, doi:10.1016/j.cell.2016.07.055 (2016).

4 T. F. de Koning-Ward, M. W. Dixon, L. Tilley &P. R. Gilson. Plasmodium species: master renovators of their host cells. *Nat Rev Microbiol* **14**, 494-507, doi:10.1038/nrmicro.2016.79 (2016).

5 K. Lingelbach &K. A. Joiner. The parasitophorous vacuole membrane surrounding Plasmodium and Toxoplasma: An unusual compartment in infected cells. *J Cell Sci* **111**, 1467-1475 (1998).

6 N. Gehde *et al.* Protein unfolding is an essential requirement for transport across the parasitophorous vacuolar membrane of Plasmodium falciparum. *Mol Microbiol* **71**, 613-628, doi:10.1111/j.1365-2958.2008.06552.x (2009).

7 M. Marti, R. T. Good, M. Rug, E. Knuepfer &A. F. Cowman. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**, 1930-1933, doi:10.1126/science.1102452 (2004).

8 N. L. Hiller *et al.* A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* **306**, 1934-1937, doi:10.1126/science.1102737 (2004).

9 J. A. Boddey &A. F. Cowman. Plasmodium Nesting: Remaking the Erythrocyte from the Inside Out. *Annu Rev Microbiol* **67**, 243-269, doi:10.1146/annurev-micro-092412-155730 (2013).

10 J. A. Boddey *et al.* An aspartyl protease directs malaria effector proteins to the host cell. *Nature* **463**, 627-U652, doi:10.1038/nature08728 (2010).

11 H. H. Chang *et al.* N-terminal processing of proteins exported by malaria parasites. *Mol Biochem Parasit* **160**, 107-115, doi:10.1016/j.molbiopara.2008.04.011 (2008).

12 A. R. Osborne *et al.* The host targeting motif in exported Plasmodium proteins is cleaved in the parasite endoplasmic reticulum. *Mol Biochem Parasit* **171**, 25-31, doi:10.1016/j.molbiopara.2010.01.003 (2010).

13 I. Russo *et al.* Plasmepsin V licenses Plasmodium proteins for export into the host erythrocyte. *Nature* **463**, 632-636, doi:10.1038/nature08726 (2010).

14 J. A. Boddey, R. L. Moritz, R. J. Simpson &A. F. Cowman. Role of the Plasmodium export element in trafficking parasite proteins to the infected erythrocyte. *Traffic* **10**, 285-299, doi:10.1111/j.1600-0854.2008.00864.x (2009).

15 A. Heiber *et al.* Identification of new PNEPs indicates a substantial non-PEXEL exportome and underpins common features in Plasmodium falciparum protein export. *PLoS Pathog* **9**, e1003546, doi:10.1371/journal.ppat.1003546 (2013).

16 C. Gruring *et al.* Uncovering common principles in protein export of malaria parasites. *Cell Host Microbe* **12**, 717-729, doi:10.1016/j.chom.2012.09.010 (2012).

17 S. Haase *et al.* Sequence requirements for the export of the Plasmodium falciparum Maurer's clefts protein REX2. *Mol Microbiol* **71**, 1003-1017, doi:10.1111/j.1365-2958.2008.06582.x (2009).

18 T. F. de Koning-Ward *et al.* A newly discovered protein export machine in malaria parasites. *Nature* **459**, 945-949, doi:10.1038/nature08104 (2009).

19 M. Peng, D. Cascio &P. F. Egea. Crystal structure and solution characterization of the thioredoxin-2 from Plasmodium falciparum, a constituent of an essential parasitic protein export complex. *Biochem Biophys Res Commun* **456**, 403-409, doi:10.1016/j.bbrc.2014.11.096 (2015).

20 J. R. Beck, V. Muralidharan, A. Oksman &D. E. Goldberg. PTEX component HSP101 mediates export of diverse malaria effectors into host erythrocytes. *Nature* **511**, 592-595, doi:10.1038/nature13574 (2014).

21 B. Elsworth *et al.* PTEX is an essential nexus for protein export in malaria parasites. *Nature* **511**, 587-+, doi:10.1038/nature13555 (2014).

22 P. Mesen-Ramirez *et al.* Stable Translocation Intermediates Jam Global Protein Export in Plasmodium falciparum Parasites and Link the PTEX Component EXP2 with Translocation Activity. *PLoS Pathog* **12**, e1005618, doi:10.1371/journal.ppat.1005618 (2016).

23 <EMBOJ-1984-Munro & Pelham. Use of peptide tagging to detect proteins expressed from cloned genes - deletion mapping functional domains of Drosophila hsp70.pdf>.

24 J. Field *et al.* Purification of a RAS-responsive adenylyl cyclase complex from Saccharomyces cerevisiae by use of an epitope addition method. *Mol Cell Biol* **8**, 2159-2165 (1988).

25 E. Hochuli, H. Dobeli &A. Schacher. New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. *J Chromatogr* **411**, 177-184 (1987).

26 T. P. Hopp *et al.* A Short Polypeptide Marker Sequence Useful for Recombinant Protein Identification and Purification. *Bio-Technol* **6**, 1204-1210, doi:DOI 10.1038/nbt1088-1204 (1988).

27 S. Munro &H. R. Pelham. Use of peptide tagging to detect proteins expressed from cloned genes: deletion mapping functional domains of Drosophila hsp 70. *EMBO J* **3**, 3087-3093 (1984).

28 R. PDB. *PDB Statistics: Overall Growth of Released Structures Per Year*, <https://www.rcsb.org/stats/growth/overall> (2018).

29 J. Porath, J. Carlsson, I. Olsson &G. Belfrage. Metal Chelate Affinity Chromatography, a New Approach to Protein Fractionation. *Nature* **258**, 598-599, doi:DOI 10.1038/258598a0 (1975).

30 A. H. Rosenberg *et al.* Vectors for selective expression of cloned DNAs by T7 RNA polymerase. *Gene* **56**, 125-135 (1987).

31 D. B. Smith &K. S. Johnson. Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. *Gene* **67**, 31-40 (1988).

32 M. J. Gardner *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498-511, doi:10.1038/nature01097 (2002).

33 N. Hall *et al.* A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307**, 82-86, doi:10.1126/science.1103717 (2005).

34 A. P. Waters. Genome-informed contributions to malaria therapies: feeding somewhere down the (pipe)line. *Cell Host Microbe* **3**, 280-283, doi:10.1016/j.chom.2008.04.005 (2008).

35 G. M. R. N. Clough, A. I. Kirkland. in *Journal of Physics: Conference Series* Vol. 522 (IOP Publishing, 2013).

36 Y. Cheng. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **161**, 450-457, doi:10.1016/j.cell.2015.03.049 (2015).

37 X. Li *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* **10**, 584-590, doi:10.1038/nmeth.2472 (2013).

38 X. Zhang, L. Jin, Q. Fang, W. H. Hui &Z. H. Zhou. 3.3 A cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. *Cell* **141**, 472-482, doi:10.1016/j.cell.2010.03.041 (2010).

39 G. McMullan, S. Chen, R. Henderson &A. R. Faruqi. Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy* **109**, 1126-1143, doi:10.1016/j.ultramic.2009.04.002 (2009).

40 G. McMullan, A. T. Clark, R. Turchetta &A. R. Faruqi. Enhanced imaging in low dose electron microscopy using electron counting. *Ultramicroscopy* **109**, 1411-1416, doi:10.1016/j.ultramic.2009.07.004 (2009).

41 G. McMullan *et al.* Experimental observation of the improvement in MTF from backthinning a CMOS direct electron detector. *Ultramicroscopy* **109**, 1144-1147, doi:10.1016/j.ultramic.2009.05.005 (2009).

42 S. Q. Zheng *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* **14**, 331-332, doi:10.1038/nmeth.4193 (2017).

43 EMBL-EBI. *EMDB statistics*, <https://www.ebi.ac.uk/pdbe/emdb/statistics_main.html/> (2018).

44 S. H. W. Scheres. A Bayesian View on Cryo-EM Structure Determination. *J Mol Biol* **415**, 406-418, doi:10.1016/j.jmb.2011.11.010 (2012).

45 S. H. W. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519-530, doi:10.1016/j.jsb.2012.09.006 (2012).

46 A. Punjani, J. L. Rubinstein, D. J. Fleet &M. A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-+, doi:10.1038/Nmeth.4169 (2017).

47 C. M. Ho *et al.* Malaria parasite translocon structure and mechanism of effector export. *Nature* **561**, 70-+, doi:10.1038/s41586-018-0469-4 (2018).

Chapter 2

Malaria Parasite Translocon Structure and Mechanism of Effector Export

Chi-Min Ho[1,2,3], Josh R. Beck[4,5], Mason Lai[2], Yanxiang Cui[6], Daniel E. Goldberg[4], Pascal F. Egea[1,3,*], Z. Hong Zhou[1,2,6,*]

[1] The Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA

[2] Department of Microbiology, Immunology, & Molecular Genetics, University of California, Los Angeles, CA 90095, USA

[3] Department of Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA

[4] Departments of Medicine and Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110, USA

[5] Department of Biomedical Sciences, Iowa State University, Ames, IA 50011, USA

[6] California NanoSystems Institute, University of California, Los Angeles, CA 90095, USA

* Correspondence and request for materials should be addressed to Z.H.Z. (Hong.Zhou@UCLA.edu, for cryoEM and atomic modeling) and P.F.E. (PEgea@mednet.ucla.edu, for biochemistry).

2.1 Abstract

The putative *Plasmodium* Translocon of Exported Proteins (PTEX) is essential for transport of malarial effector proteins across a parasite-encasing vacuolar membrane into host erythrocytes, but the mechanism of this process remains unknown. Here we show PTEX is a *bona fide* translocon by determining near-atomic resolution cryoEM structures of the endogenous PTEX core complex of EXP2, PTEX150 and HSP101, isolated from *Plasmodium falciparum* in the *engaged* and *resetting* states of endogenous cargo translocation with CRISPR/Cas9-engineered epitope tags. EXP2 and PTEX150 interdigitate to form a static, funnel-shaped pseudo-sevenfold symmetric protein-conducting channel spanning the vacuolar membrane. Tethered above this funnel, the spiral-shaped AAA+ HSP101 hexamer undergoes a dramatic compaction that allows three of six tyrosine-bearing pore loops lining the HSP101 channel to dissociate from the cargo, resetting the translocon for the next threading cycle. Our work reveals the mechanism of *P. falciparum* effector export, enabling structure-based design of drugs targeting this unique translocon.

2.2 Introduction

Malaria has devastated major civilizations since the dawn of humanity and remains a significant burden to our society, responsible for nearly half a million deaths annually[1]. This infectious disease is caused by *Plasmodium* parasites, which invade and reproduce within human erythrocytes, inducing the clinical symptoms of malaria[2, 3]. These parasites export hundreds of effector proteins to extensively remodel host erythrocytes, which have limited capacity for biosynthesis[4-6]. Collectively known as the exportome, these proteins create the infrastructure necessary to import nutrients, export waste, and evade splenic clearance of infected erythrocytes[7]. Most of these proteins bear a 5-residue motif, the *Plasmodium* Export Element (PEXEL)[8-10]. The malaria parasite conceals itself inside a parasitophorous vacuole (PV) derived from invagination of the host cell plasma membrane during invasion[11] (**Fig. 2.1**).

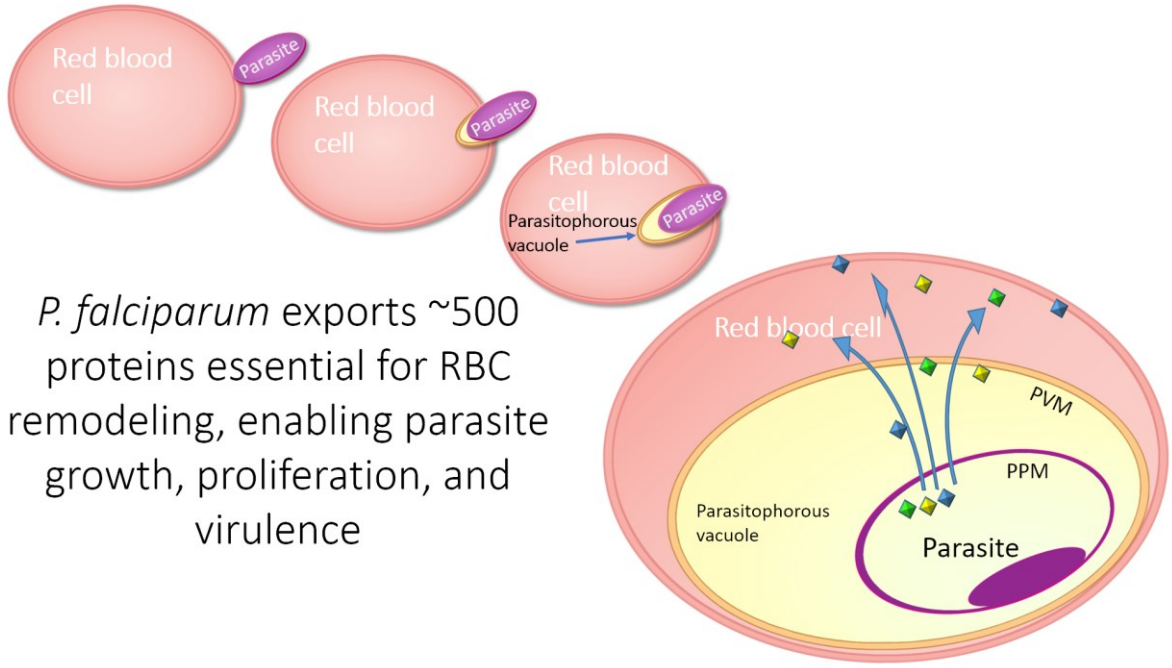


Figure 2.1 | Formation of the parasitophorous vacuole. Invagination of the erythrocyte plasma membrane during invasion leads to the formation of a parasitophorous vacuole (PV) around the parasite. Effector proteins are secreted into the PV, where they are then unfolded and transported across the PV membrane (PVM).

Following secretion into the PV, proteins destined for export are unfolded and transported across the PV membrane (PVM) into the host cell in an ATP-dependent process[12, 13]. To accomplish this, it was proposed that the parasite has evolved a unique membrane protein complex, the *Plasmodium* Translocon of Exported Proteins (PTEX)[14]. PTEX is the only known point of entry to the host cell for exported proteins and an attractive drug target, as disrupting PTEX blocks delivery of key virulence determinants, inducing parasite death[15, 16].

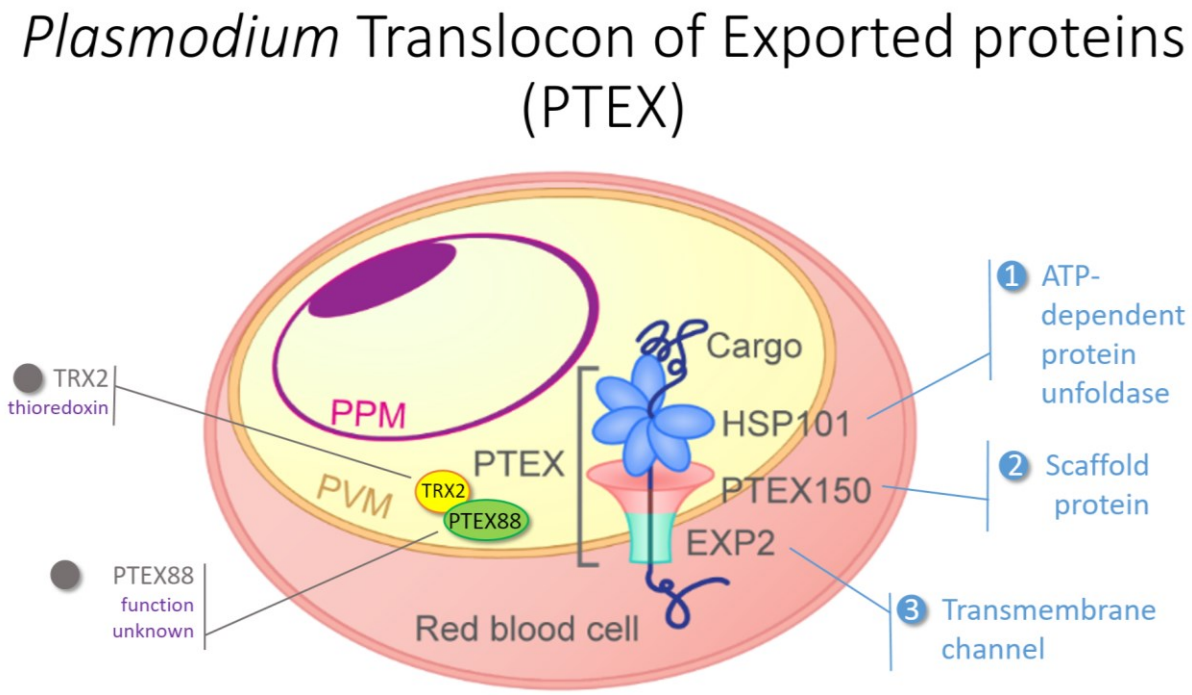


Figure 2.2 | Diagram of a parasite-infected human erythrocyte. PPM, parasite plasma membrane; PVM, parasitophorous vacuole.

PTEX was suggested to be a >1.2MDa membrane protein complex with a core composed of the HSP101 ATPase and two novel proteins, PTEX150 and EXP2 (**Fig. 2.2**)[14, 17]. HSP101 belongs to the Class 1 Clp/HSP100 family of AAA+ ATPases, PTEX150 has no known homologs beyond the *Plasmodium* genus, and EXP2 is a PVM protein[14, 18] conserved among vacuole-

dwelling apicomplexans[19]. All three core components are essential for protein export and parasite survival[15, 16, 20]. A model of PTEX-mediated translocation was proposed in which HSP101 unfolds and threads proteins through an oligomeric EXP2 transmembrane channel spanning the PVM, with PTEX150 playing a structural role between EXP2 and HSP101[14-17]. However, without structural information, the global architecture of PTEX, the stoichiometry of its components, and direct evidence for the proposed molecular mechanism have proven elusive.

In this study, we purify PTEX directly from the human malaria parasite *P. falciparum* and determine near-atomic resolution cryoEM structures of the complex in multiple functional states. Our atomic models reveal the architecture and mechanism of this unique translocon and pave the way for development of novel therapeutics against this promising new malarial drug target.

2.3 Results

2.3.1 Architecture of the PTEX core complex

To purify PTEX from *P. falciparum*, we used CRISPR/Cas9 editing to introduce a 3xFLAG epitope tag on the endogenous HSP101 C-terminus (**Fig. 2.3a-c**) and purified the endogenously assembled PTEX core complex directly from parasites cultured in human erythrocytes (**Fig. 2.3d, Fig. 2.4**).

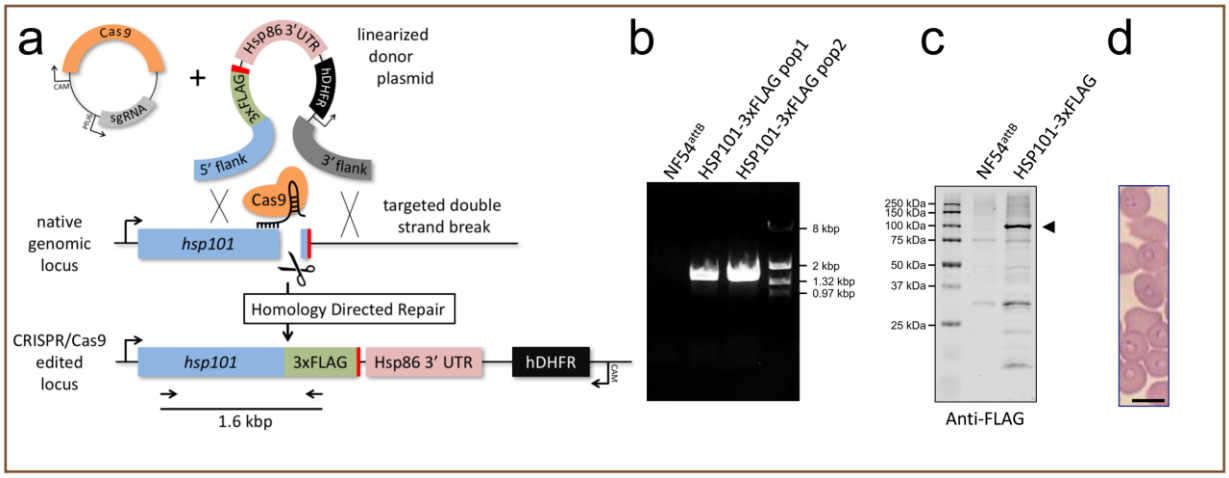


Figure 2.3 | Generation of HSP101-3xFLAG parasites. a, Schematic showing strategy for endogenous tagging of *P. falciparum hsp101* with 3xFLAG using CRISPR/Cas9 editing. Diagnostic PCR primers and expected amplicon following successful integration are shown. sgRNA, single guide RNA; UTR, untranslated region; CAM, calmodulin promoter; PfU6, *P. falciparum* U6 promoter; hDHFR, human dihydrofolate reductase. **B,** Diagnostic PCR with genomic DNA template from NF54^{attB} parent or two independent populations of HSP101-3xFLAG parasites. Kb, kilobase pairs. The experiment was performed one time. **C,** Western blot of NF54^{attB} and HSP101-3xFLAG parasites probed with mouse-anti-FLAG M2 antibody (Sigma) and goat-anti-mouse IRDye 680 secondary (Li-cor). Arrowhead indicates full-length HSP101-3xFLAG (predicted molecular weight 102.9 kDa after signal peptide cleavage). kDa, kilodaltons. Data represent two independent experiments. **D,** Giemsa staining of parasite-infected human erythrocytes from which PTEX was purified. Scale bar: 5um. For source data, see Supplementary Figure 3.

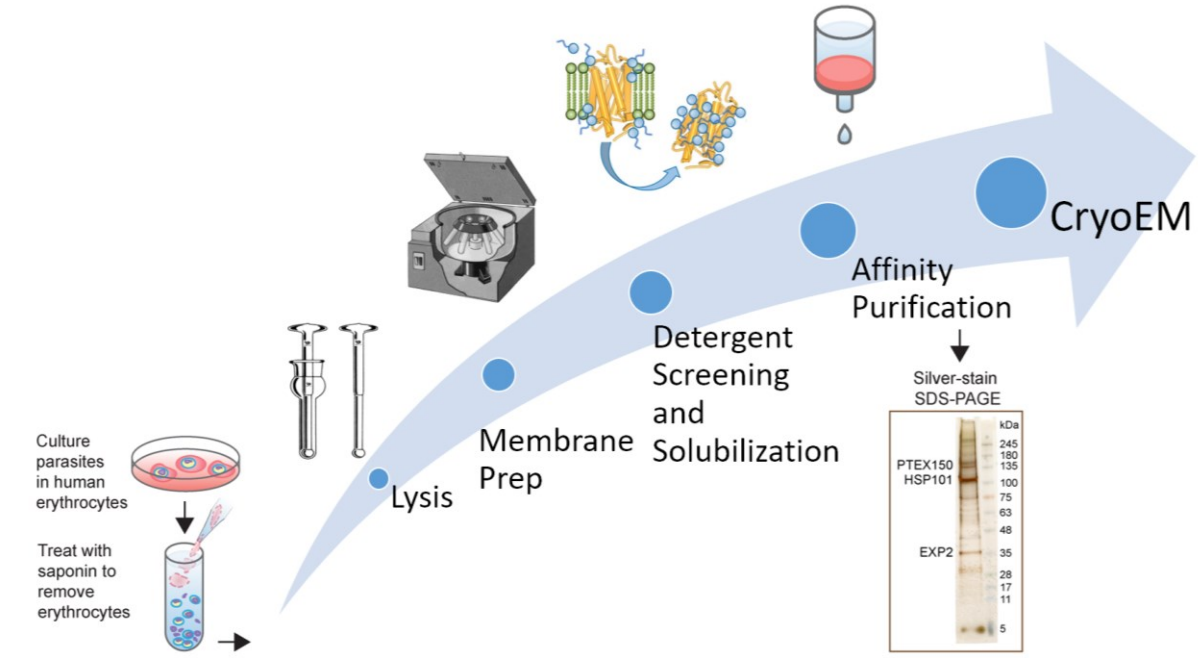
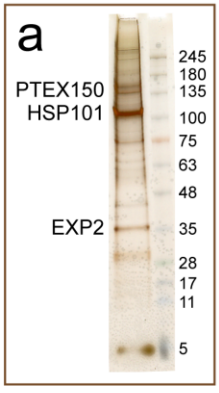


Figure 2.4 | PTEX purification workflow. Workflow illustrating protocol developed for purifying endogenous PTEX complex from blood-stage *P. falciparum* parasites cultured in human erythrocytes.



b

	Accession	Seq coverage	Seq count	Spec count	NSAF	EMPAI	Description
1	PF3D7_1116800	55.5	72	228	0.0750967	2.5892193	product=heat shock protein 101 (HSP101)
2	PF3D7_0917900	48.3	44	101	0.04622617	2.040885	product=heat shock protein 70 (HSP70-2)
3	PF3D7_1436300	58	58	135	0.04056941	2.8018937	product=translocon component PTEX150 (PTEX150)
4	PF3D7_1471100	40.8	13	34	0.03535177	1.5585859	product=exported protein 2 (EXP2)
5	PF3D7_1006200	59.8	7	7	0.019522185	2.9627802	product=DNA/RNA-binding protein Alba 3 (ALBA3)

Figure 2.5 | Analysis of purified PTEX. a, Silver stained SDS-PAGE gel of the FLAG-purified PTEX sample. Identities of the bands labeled EXP2, PTEX150, and HSP101 were confirmed by tryptic digest LC-MS. **b,** Tryptic digest liquid chromatography-mass spectrometry (LC-MS) analysis of the FLAG-purified PTEX sample. The PTEX core components are among the five most abundant species detected in the purified sample. For gel and blot source data, see Supplementary Figure 1.

Figure 2.6 | Experimentally determined secondary structure elements and detected mass-spec fragments mapped to the primary sequences of the three PTEX proteins. For EXP2 (**a**), PTEX150 (**b**), and HSP101 (**c**), secondary structure elements are shown as tubes (helices), lines (loops), and arrows

(strands) above the corresponding sequence and are colored as in Fig. 2a, 3a, and 4a. In the sequences shown below, residues resolved in our structures are colored according to protein colors in Fig. 1c-f: EXP2 (mint), PTEX150 (salmon) and HSP101 (cornflower). Signal peptide residues are colored gold. All residues in the mature proteins that are not resolved in our structures are shown in grey. The 3xFLAG residues at the C-terminus of HSP101 are colored green. Peptides detected in tryptic digest LC-MS/MS analysis of the purified PTEX sample are shown as black lines below the corresponding sequences. Arrowheads above the EXP2 sequence indicate truncations sites described in this work and in Garten *et al.*[20] immediately before (Δ222-287, red arrowhead) and after (Δ234-287, green arrowhead) the assembly strand. Arrowheads above PTEX150 sequence indicate previously described truncation sites[21] (Δ847-993, red arrowhead; Δ869-993, green arrowhead).

CryoEM analysis yielded two distinct conformations of PTEX particles, one extended (195Å) and the other compact (175Å) (**Fig. 2.9**). Endogenous cargo polypeptide densities are visible in the central pore of HSP101 in both structures (**Fig. 2.10, 2.36**).

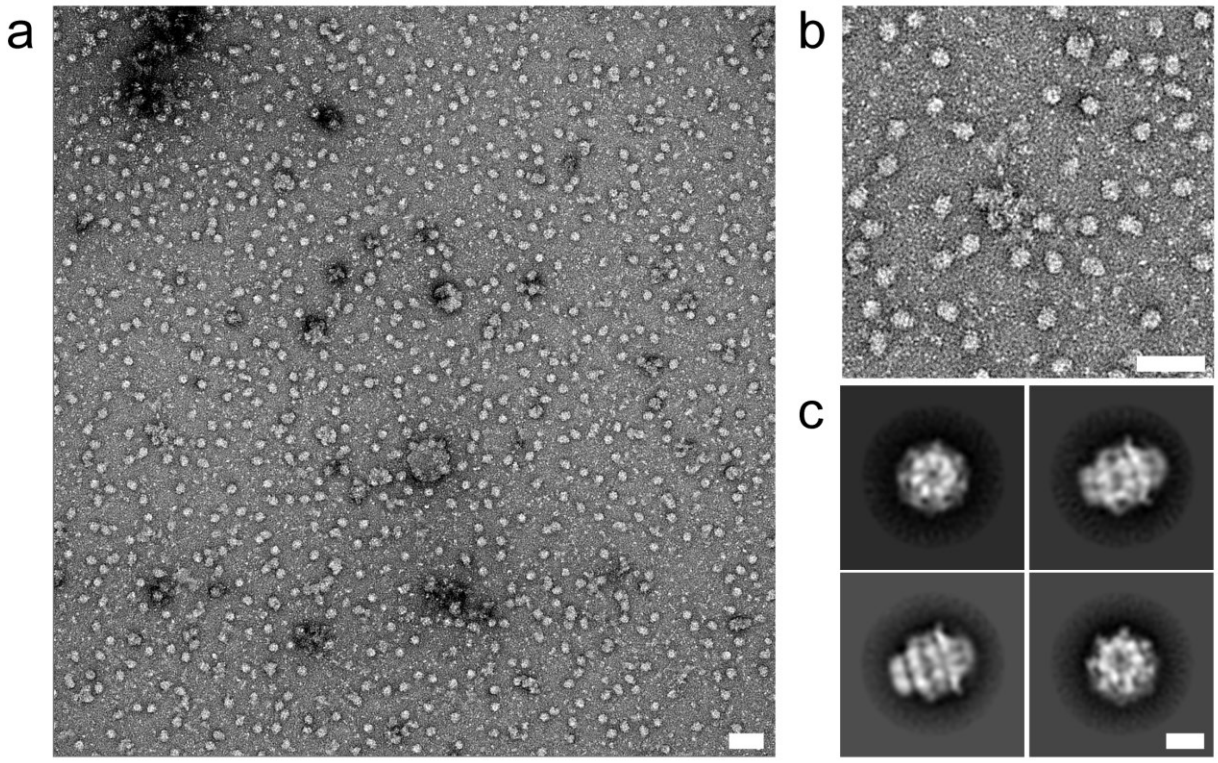


Figure 2.7 | Negative stain analysis of purified PTEX. Representative negative stain micrograph (**a**), enlarged portion of micrograph (**b**), and two-dimensional class averages (**c**) of the PTEX core complex in multiple orientations. Scale bars are 700Å, 700Å, and 100Å respectively.

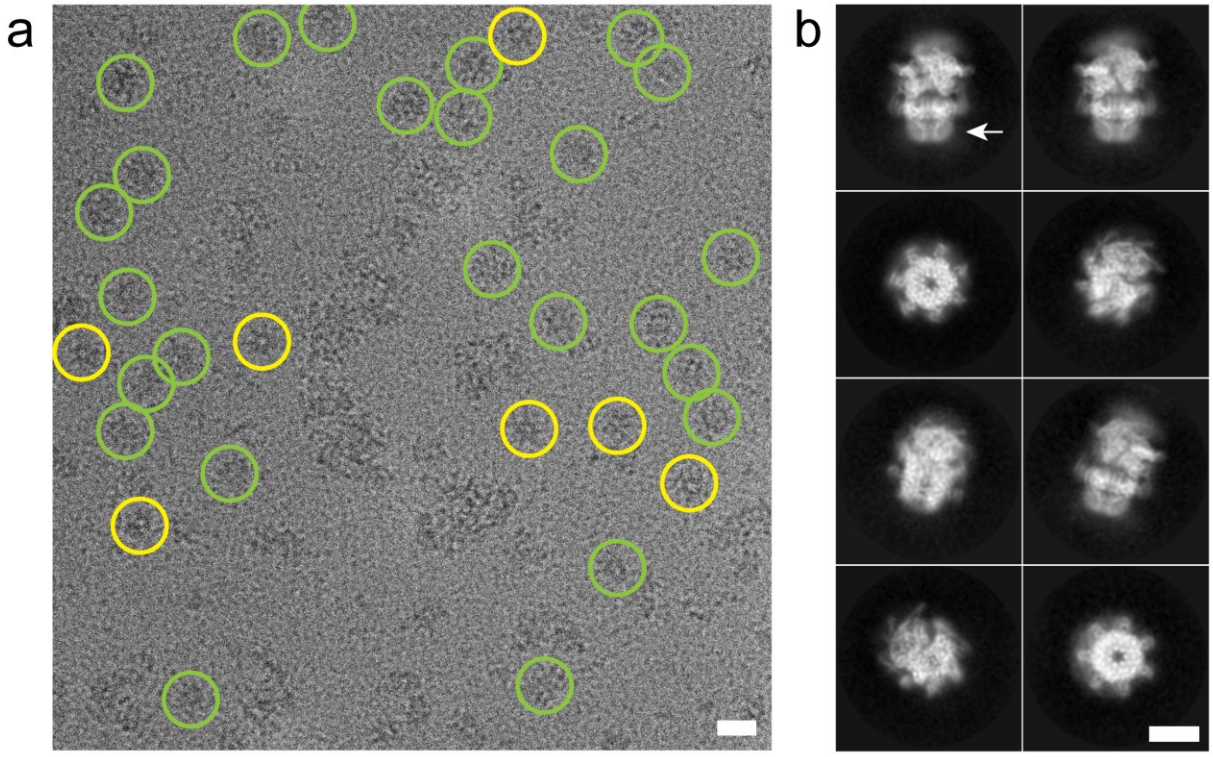


Figure 2.8 | CryoEM analysis of purified PTEX. Representative cryoEM micrograph (**a**) and two-dimensional class averages (**b**) of the PTEX core complex in multiple orientations. Arrow in upper left panel of (**b**) indicates the detergent belt, which is visible as a less-dense (dimmer) halo surrounding the denser (brighter) densities of the alpha helices visible in the TMD in side views. Scale bars are 200Å and 100Å, respectively.

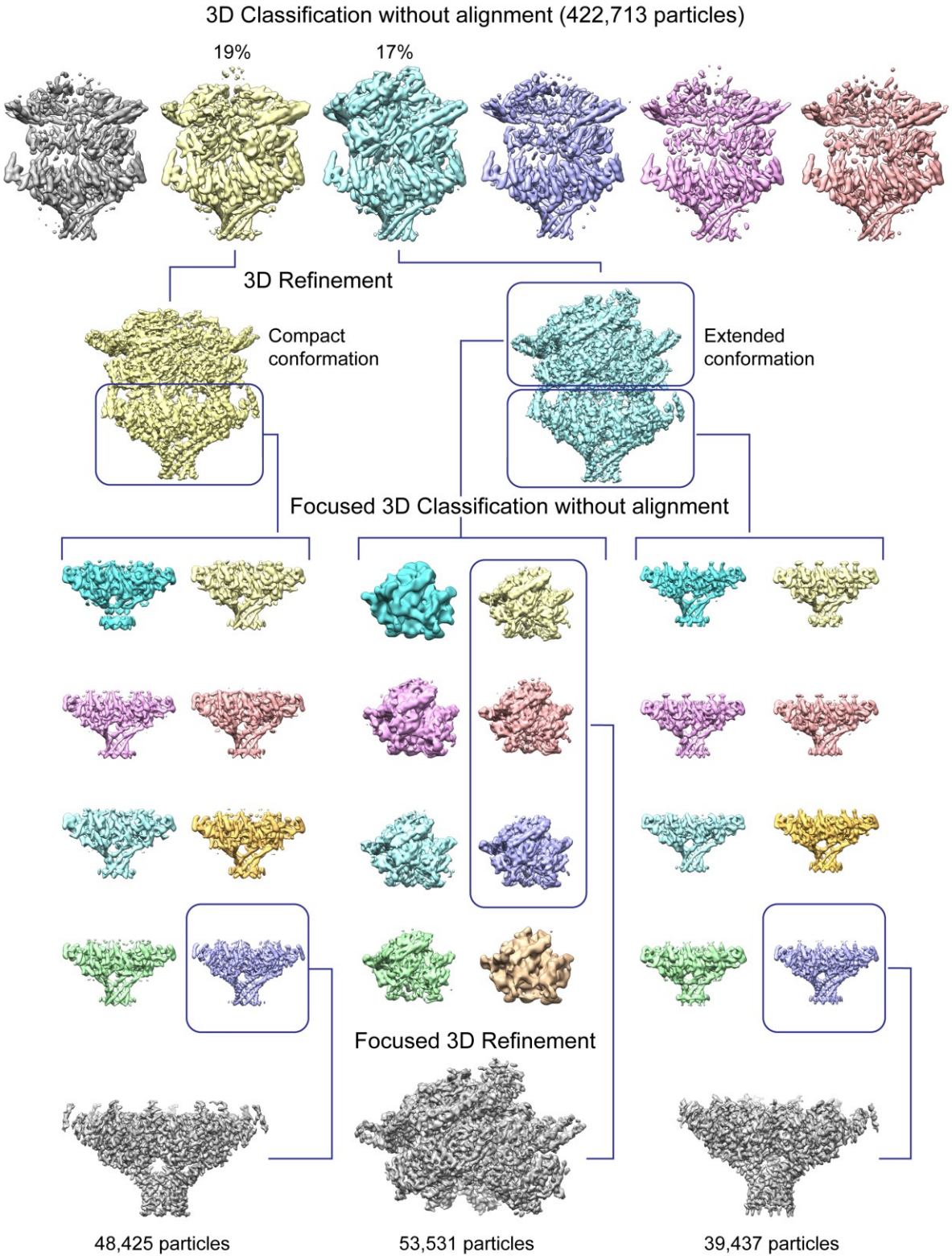


Figure 2.9 | Overview of 3D Image Processing Workflow. Illustration of workflow for 3D classification, and focused classification and refinement. Maps are displayed at higher thresholds where the detergent belt is not visible for clarity, to avoid obscuring details of the transmembrane helices.

Based on differences in the arrangement of HSP101 subunits relative to the cargo between the two conformations, we designated them as the *engaged* and *resetting* states, respectively (**Fig. 2.10**).

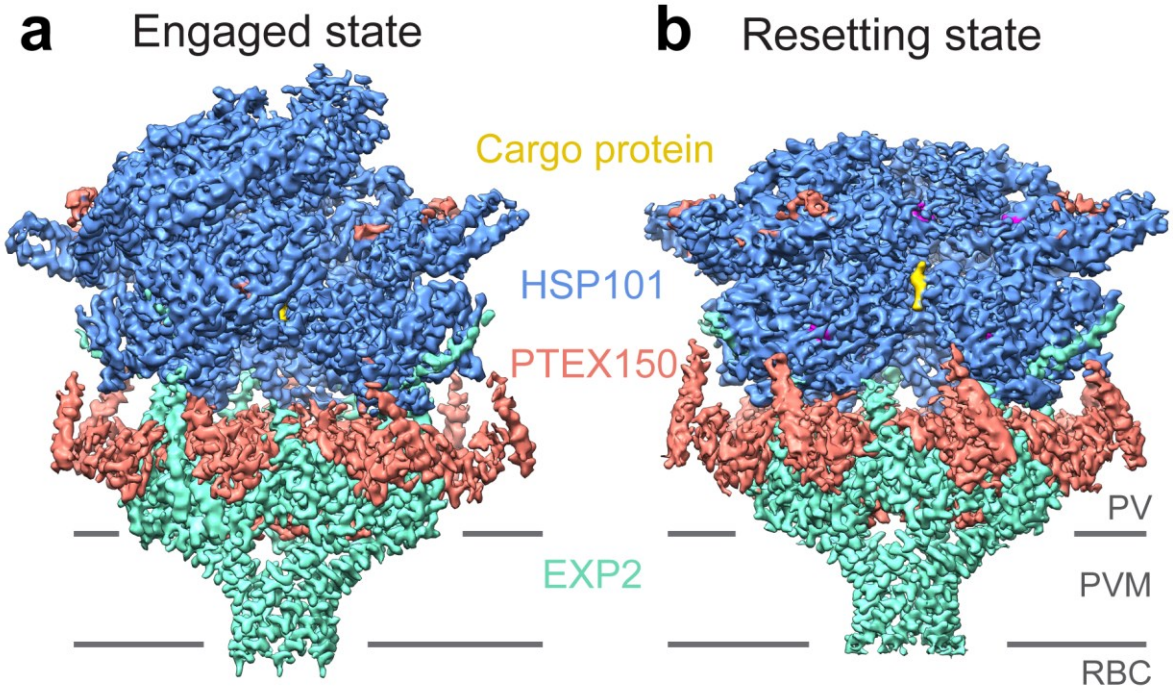


Figure 2.10 | CryoEM maps of the PTEX core complex in the *engaged* (a) and *resetting* (b) states. Horizontal lines represent the PVM bilayer, estimated based on the detergent belt density, visible at lower thresholds (see Extended Data Fig. 7).

Both maps are at near-atomic resolution, varying from 3-3.6Å in the transmembrane © and core regions to 5-8Å in the periphery (**Fig. 2.11**). Clear sidechain densities throughout most regions of both maps (**Fig. 2.10, 2.12; Supplementary Videos 1-2**) enabled us to build *de novo* atomic models of the three constituent proteins for both conformational states (**Fig. 1d-e**), each containing 20 subunits with 6,898 amino acid residues modeled. All subunits were built independently, as conformations varied between subunits.

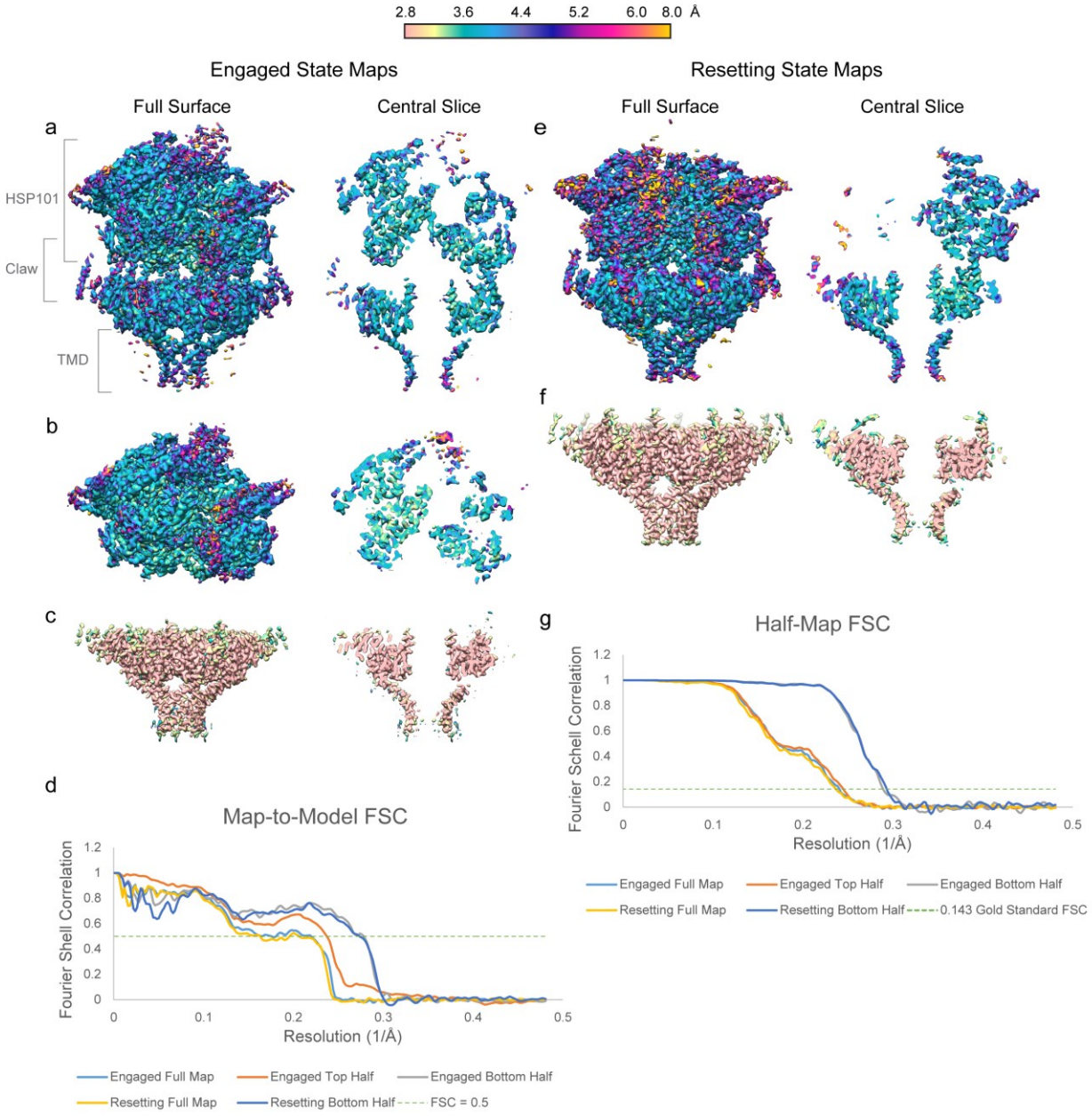


Figure 2.11 | Resolution assessments of the two PTEX states. a-c,e-f Local resolution evaluations of the full PTEX map (**a**) and the focus-refined maps of the upper/hexameric (**b**) and lower/heptameric (**c**) halves of PTEX in the *engaged* state, and the full PTEX map © and the focus-refined map of the lower/heptameric (**f**) half of PTEX in the *resetting* state, calculated by Resmap[22] and colored according to resolution. Maps are displayed at higher thresholds where the detergent belt is not visible for clarity, to avoid obscuring details of the transmembrane helices. **D,** Global resolution assessment of the *engaged* and *resetting* state maps as measured using the "Gold-standard" Fourier shell correlation (FSC) curves generated by RELION[23, 24] by comparison of two independently refined "half-maps". **g,** Map-to-model FSC curves demonstrating the degree of correlation between the refined PTEX models and the experimental cryoEM maps for the *engaged* and *resetting* states.

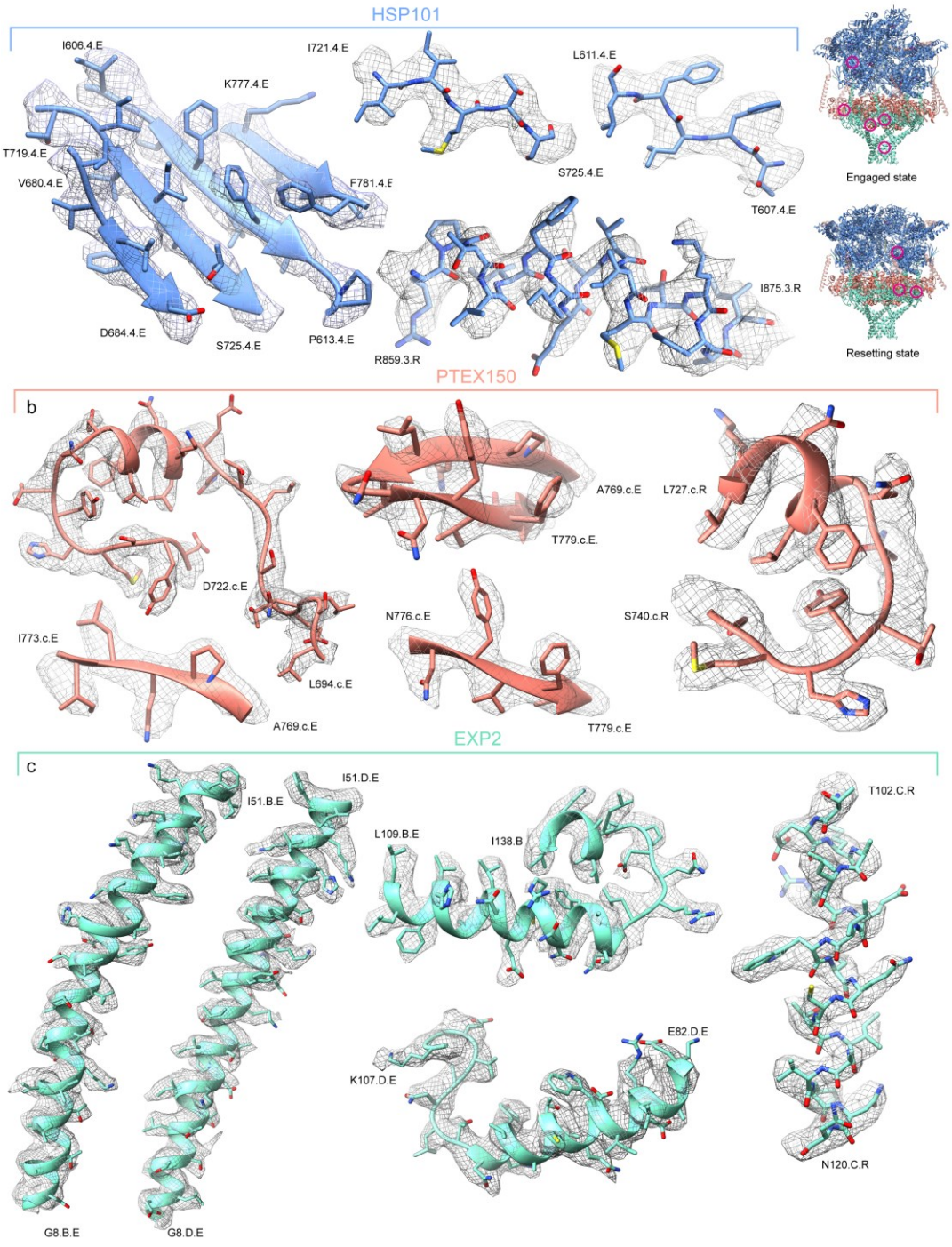


Figure 2.12 | Representative regions of cryoEM density and atomic models. Additional cryoEM densities (mesh) superposed with our atomic models for HSP101 (**a**), PTEX150(668-823) (**b**), and EXP2 (**c**). Displayed regions correspond to areas circled in magenta on guide figures (inset, upper right), and are colored as in guide figure: HSP101 (cornflower blue), PTEX150(668-823) (salmon), EXP2 (mint). Terminal residues for each segment are labeled with the amino acid, residue number, protomer, and state.

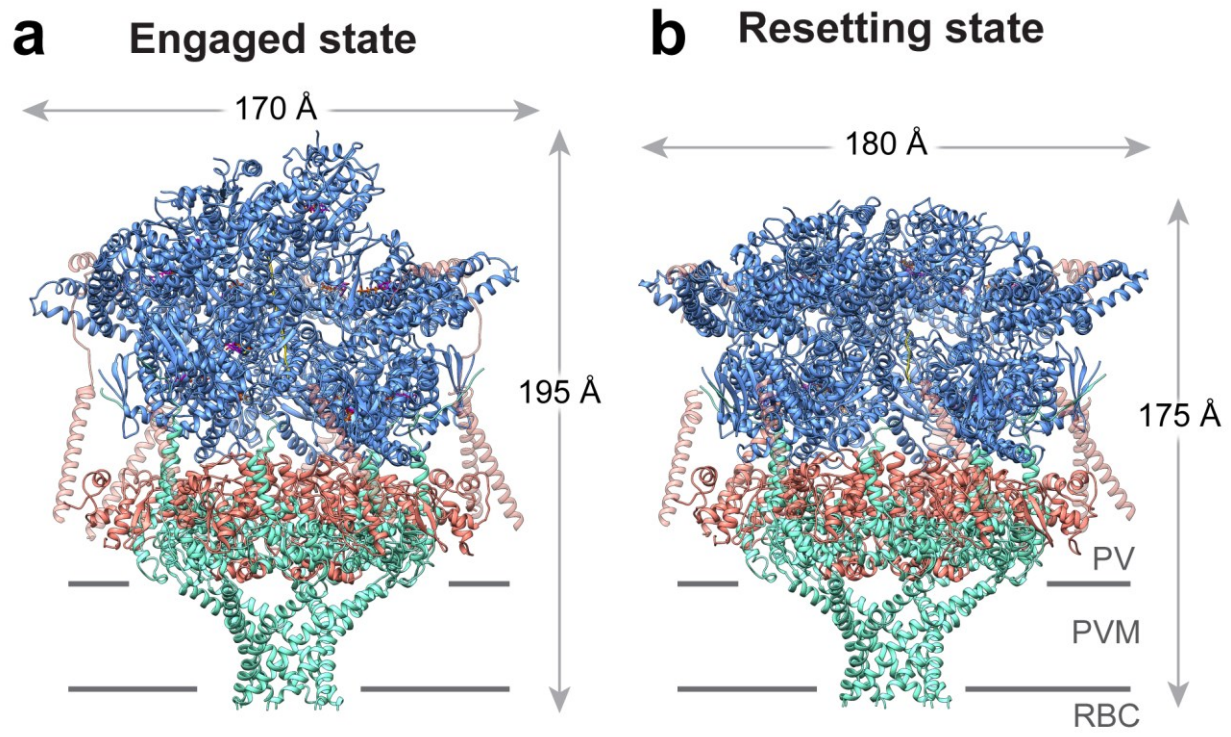


Figure 2.13 | Atomic models of the PTEX core complex in the engaged (**a**) and resetting (**b**) states. Horizontal lines represent the PVM bilayer, estimated based on the detergent belt density, visible at lower thresholds

Both structures reveal PTEX to be a tripartite membrane protein complex with a 6:7:7 stoichiometry and a calculated mass of 1.6Mda, composed of a hexameric HSP101 protein-unfolding motor tethered to a PVM-spanning, pseudosymmetric funnel formed by seven protomers of EXP2 interdigitating with seven protomers of PTEX150 (**Fig. 1d-k, Supplementary Video 3**).

Two transiently associated[25] accessory proteins, PTEX88 and TRX2[14], are not observed in our

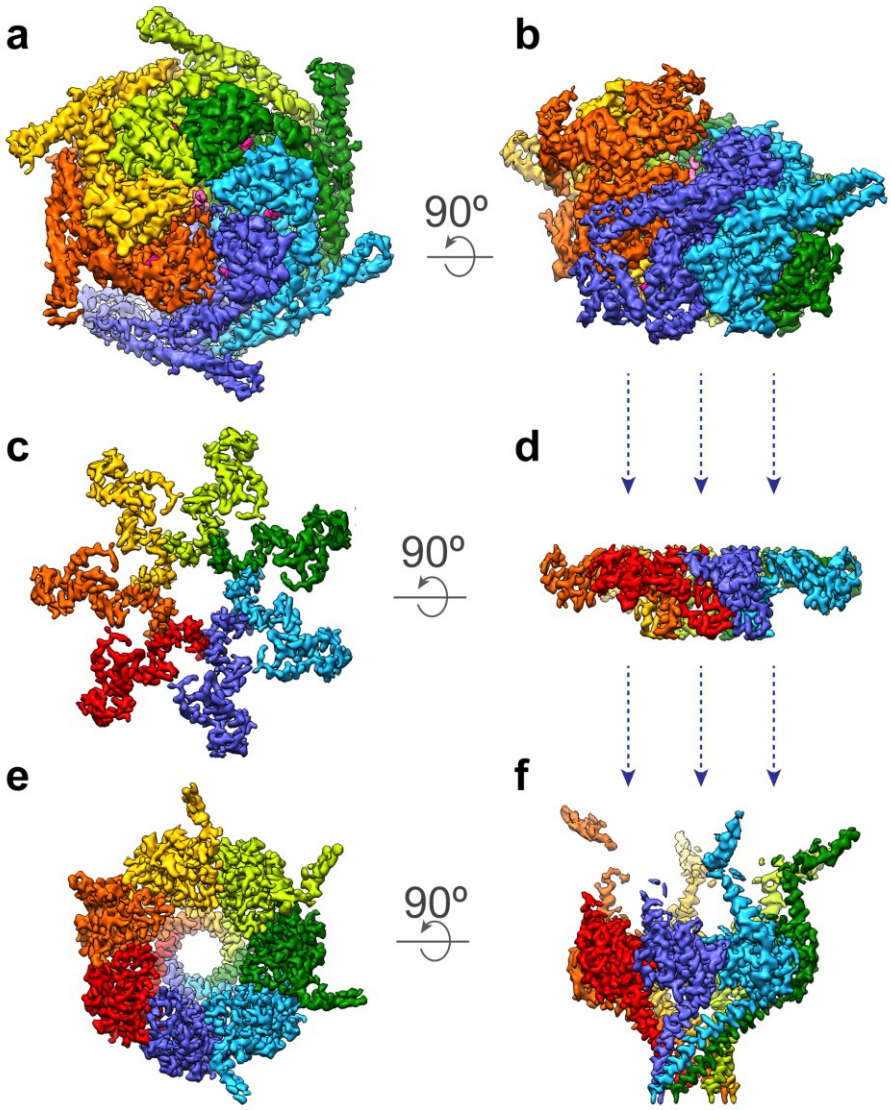


Figure 2.14 | Architecture and stoichiometry of PTEX. Top and side views of the HSP101 (**a-b**), PTEX150 (**c-d**), and EXP2 (**e-f**) cryoEM maps, coloured by protomer.

structures. At the PVM, each EXP2 monomer contributes a single TM helix to form a sevenfold (C7)-symmetric protein-conducting channel spanning the membrane (**Fig. 1j-k**). Six HSP101 protomers are tethered atop the PTEX150/EXP2 funnel in a hexameric right-handed spiral, with a gap between the bottom-most and top-most protomers (**Fig. 1f-g, Supplementary Video 4**). The HSP101 hexamer is oriented such that a single unbroken channel extends from the top of the

HSP101 hexamer to the bottom of the heptameric EXP2 transmembrane pore (**Fig. 1l-n, Extended Data Fig. 2d**). The most constricted point along the channel occurs in HSP101, measuring 4Å and 10Å in diameter, in the *engaged* and *resetting* states, respectively (**Fig. 1l**).

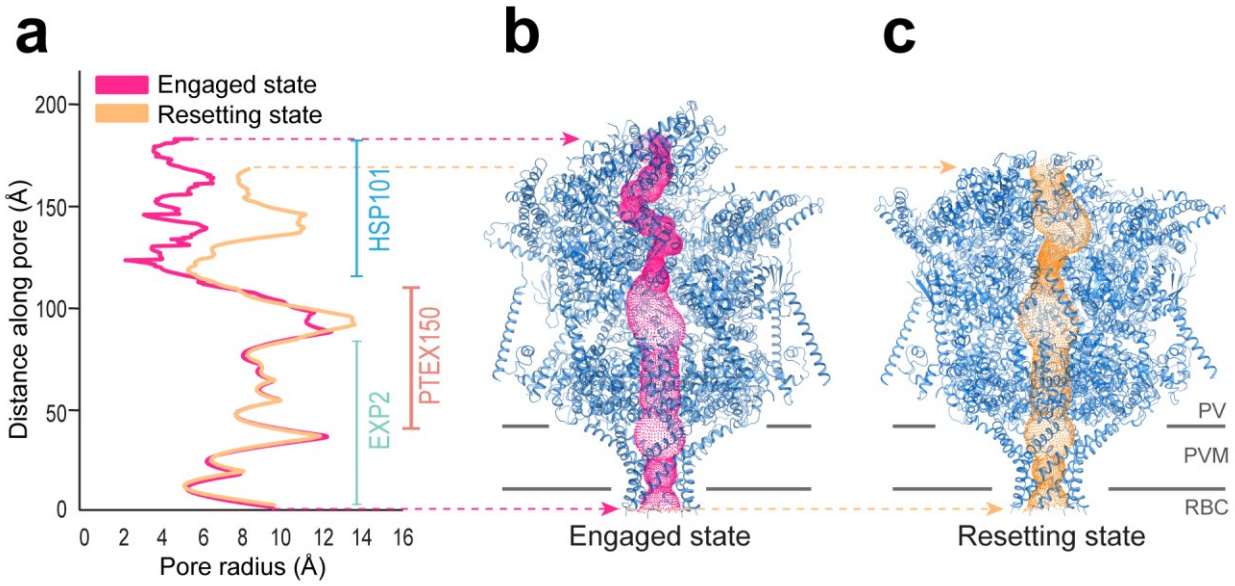


Figure 2.15 | The central channel of the PTEX complex. Pore radius (**a**) and protein-conducting channel (**b-c**) calculated using HOLE[60].

The seventh EXP2 and PTEX150 protomers are situated beneath the gap between HSP101 protomers 1 and 6, accommodating the remarkable symmetry mismatch between the asymmetric HSP101 hexamer and the pseudo-sevenfold-symmetric PTEX150/EXP2 tetradecamer (**Fig. 1f-k, Extended Data Fig. 2e-j**).

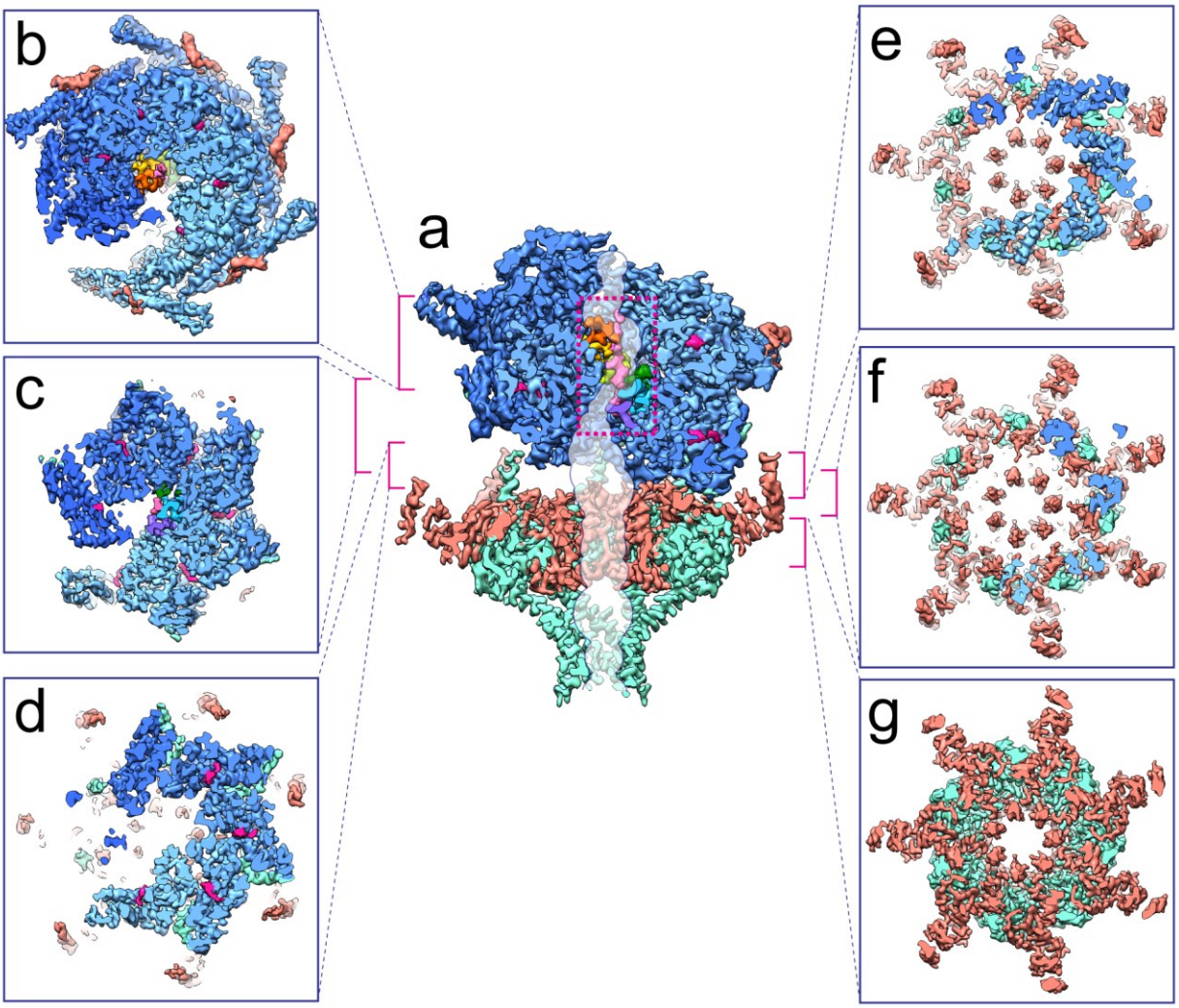


Figure 2.16 | Cross sections of the PTEX structure reveal details of the symmetry mismatch between HSP101, PTEX150, and EXP2. a, Side view of the bisected *engaged* state PTEX cryoEM map. The protein-conducting channel, calculated using HOLE[26], is shown superimposed over the bisected map in translucent white with a navy outline. The HSP101 NBD2 pore loop densities are colored by HSP101 protomer, and the cargo density is colored pink. **B-g,** The transition from the assymetric HSP101 spiral to the C7 pseudosymmetric PTEX150(668-823)-EXP2 heptamer is depicted using a series of cross sections taken perpendicular to the central axis of the translocon, spanning the area of symmetry mismatch. The section of the translocon corresponding to each cross-sectional image is indicated with a brackets in (**a**).

Analyses of our PTEX150 and EXP2 structures with four commonly used structural similarity search programs[27-30] revealed no consistent structural similarities to any known proteins, including

the pore-forming toxin Hemolysin E (HlyE), with which EXP2 was previously speculated to share structural homology[14]. Below, we describe the structural details of the individual proteins in the *engaged* state, followed by a comparison of the two states that suggests a mechanism of translocation.

2.3.2 EXP2 forms a heptameric protein-conducting channel across the PVM

Residues G27-S234 of EXP2 are well resolved in our structure, accounting for 80% of the mature protein (Extended Data Fig. 6a). EXP2 is a single-pass transmembrane protein consisting of a kinked 60Å-long N-terminal TM helix followed by a globular body domain and ending in an assembly domain composed of a linker helix followed by the assembly strand (Fig. 2a-b). The body domain contains five helices (B1-5), stabilized by an intraprotomer C113-C140 disulfide bond (Fig. 2c).

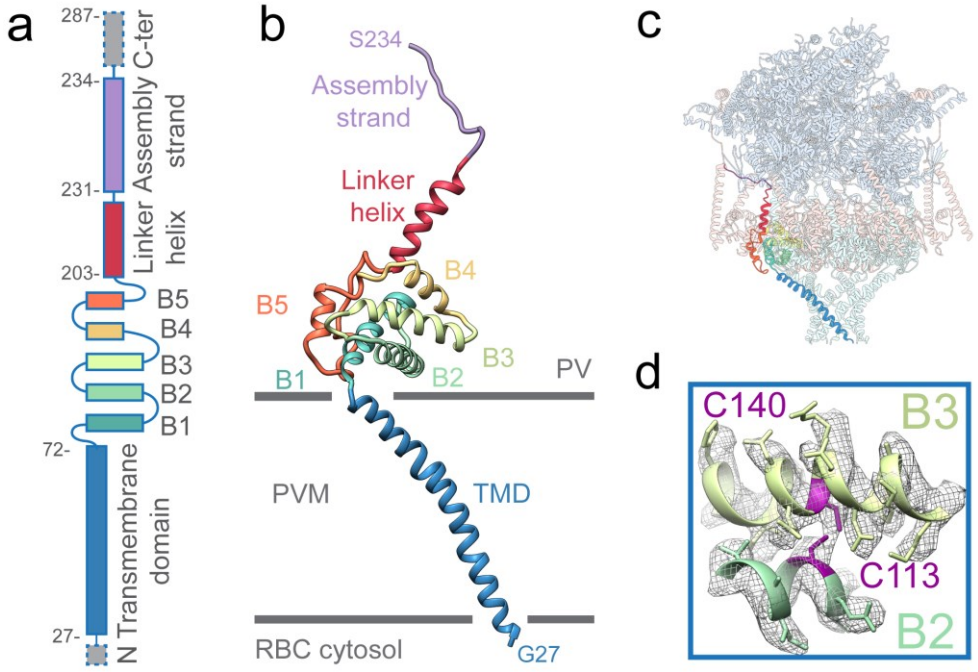


Figure 2.17 | The EXP2 monomer. Linear schematic (**a**) and ribbon diagram (**b**) of the EXP2 monomer in the *engaged* state. Dashed gray boxes represent unmodeled regions. **c,** one EXP2

monomer (coloured) within the PTEX complex. **d,** Density (mesh) and model of C113-C140 disulfide bond.

Seven EXP2 protomers (labeled A-G) oligomerize to form a funnel-shaped C7-pseudosymmetric 216kDa heptamer spanning the PVM (Fig. 2d-e). The TMD and body helices B1-3 are symmetric throughout all seven protomers (Extended Data Fig. 3a-b). This symmetry is broken by inter-protomer conformational variations in body helices B4-5 and the assembly domain, which stretch upwards in some protomers to maintain contacts with the asymmetric HSP101 hexamer situated above the EXP2 funnel. This variation is most pronounced in EXP2 protomers F and G (Extended Data Fig. 3a-b).

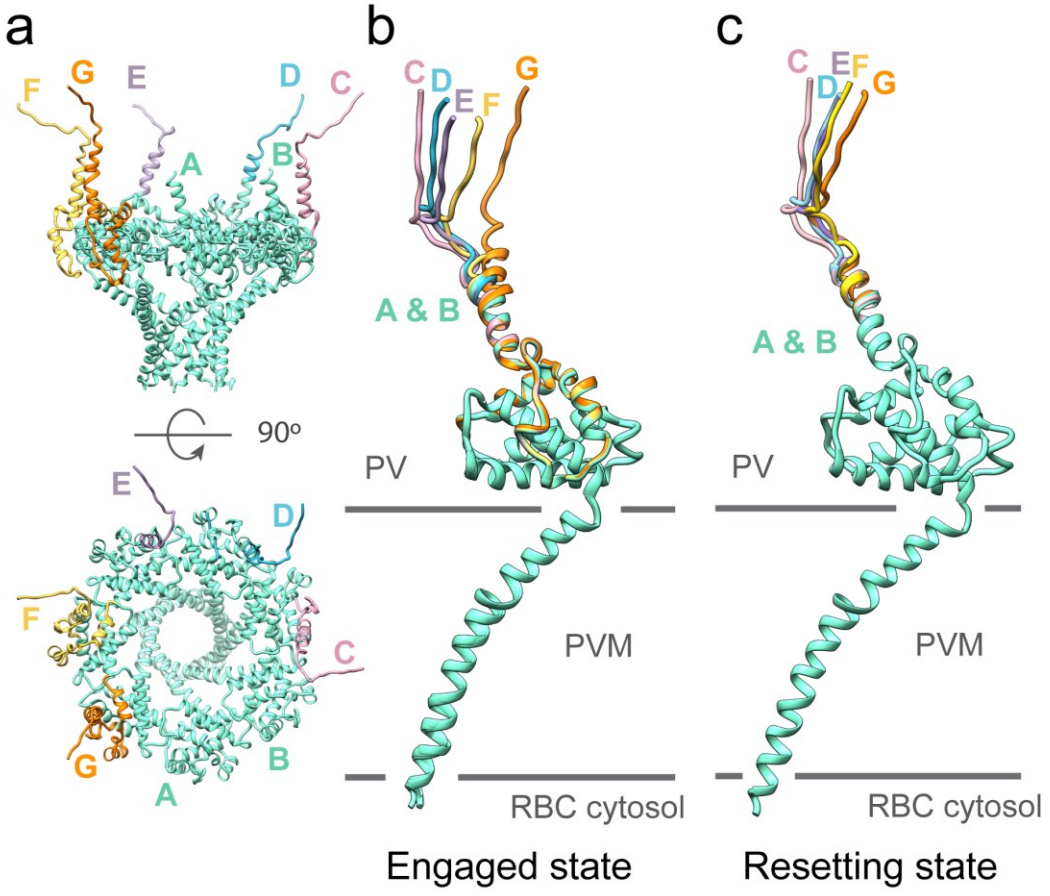


Figure 2.18 | The EXP2 heptamer. a, Side and top views of the EXP2 heptamer in the *engaged* state. Symmetric portions that remain constant between protomers are colored in mint. Portions

that vary between protomers are colored and labeled by protomer. **b-c,** Superposition of the seven EXP2 protomers, labeled A-G, in the *engaged* (**b**) and *resetting* (**c**) states, colored as in (**a**).

In the EXP2 heptamer, the amphipathic TM helices twist slightly around each other, creating a 37Å-long C7-symmetric protein-conducting channel that spans the PVM and forms the stem of the funnel (Fig. 2d-e). The membrane-facing surface of the EXP2 channel is coated with hydrophobic residues, while the inner surface is lined with charged and polar residues, creating an aqueous pore (Fig. 2e).

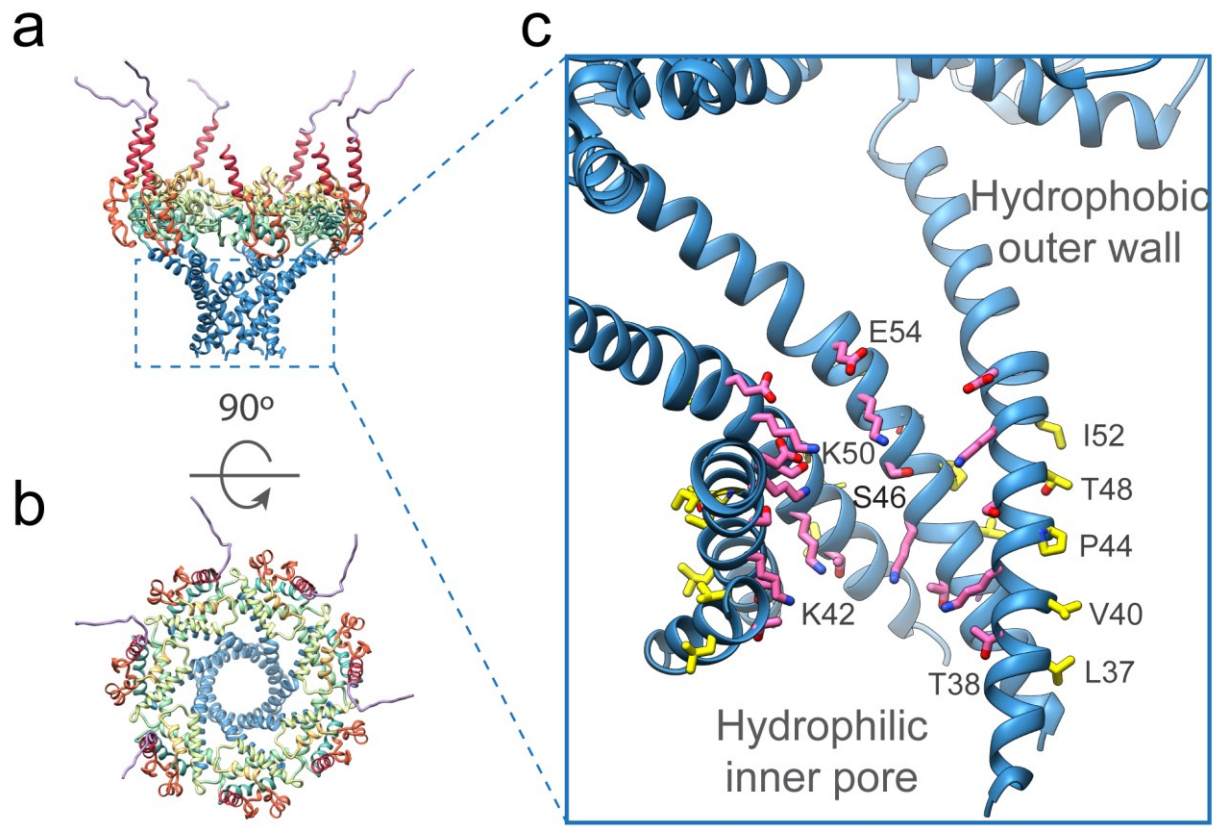


Figure 2.19 | The EXP2 transmembrane protein-conducting pore. Side (**b**) and top (**c**) views of the EXP2 heptamer, coloured as in **(Fig. 2.x)**. **c,** Cutaway of the EXP2 transmembrane channel with hydrophilic residues (pink) lining the inner protein-conducting pore and hydrophobic residues (yellow) on the outer, membrane-facing surface.

The body domains, positioned in a wider ring atop the transmembrane channel on the vacuolar face of the PVM, form the mouth of the funnel. This orientation is consistent with previous analyses of EXP2 topology[14, 20]. Furthermore, a detergent belt is clearly visible in 2D class averages and density maps (Extended Data Fig. 7-8), defining the residues in the TMD that would be buried in the PVM.

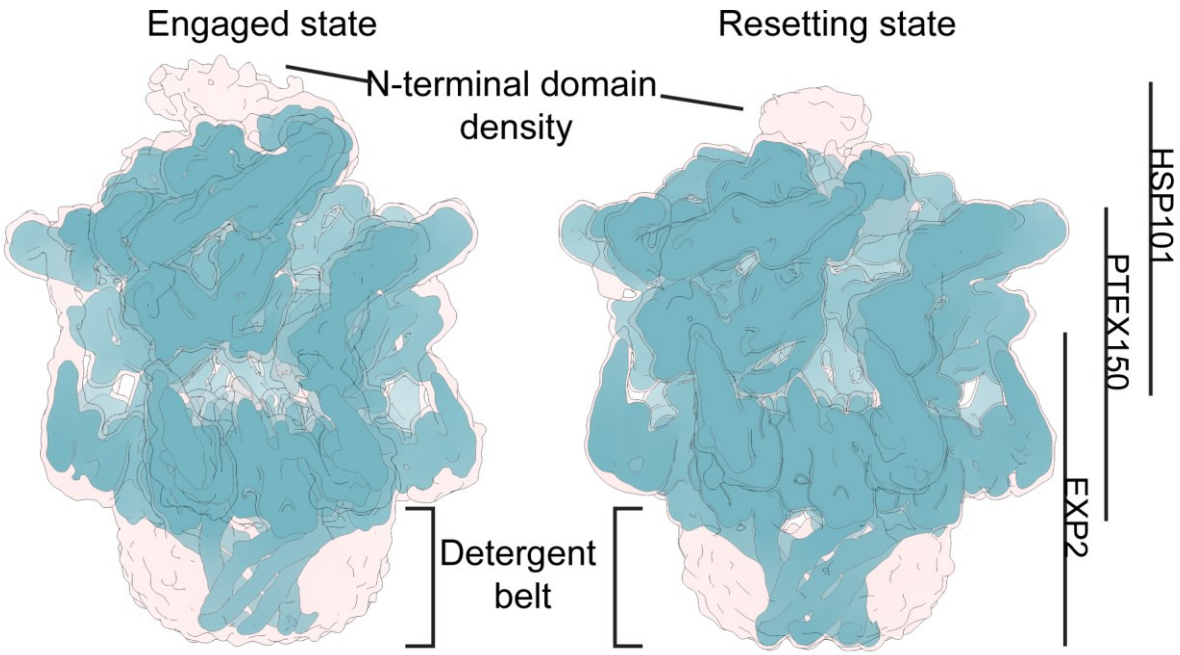


Figure 2.20 | Lower resolution details of the PTEX maps. *Engaged* state (left) and *resetting* state (right) maps were low-pass filtered to 6Å to improve clarity of low resolution details, and are shown overlaid, at two different thresholds to improve visibility of the detergent belt and the poorly-resolved N-terminal domains of HSP101 (teal, higher threshold; peach, lower threshold).

A ring of positively charged residues where the stem meets the mouth of the funnel is positioned to interact with the negatively charged phosphates of the membrane surface (Extended Data Fig. 8a).

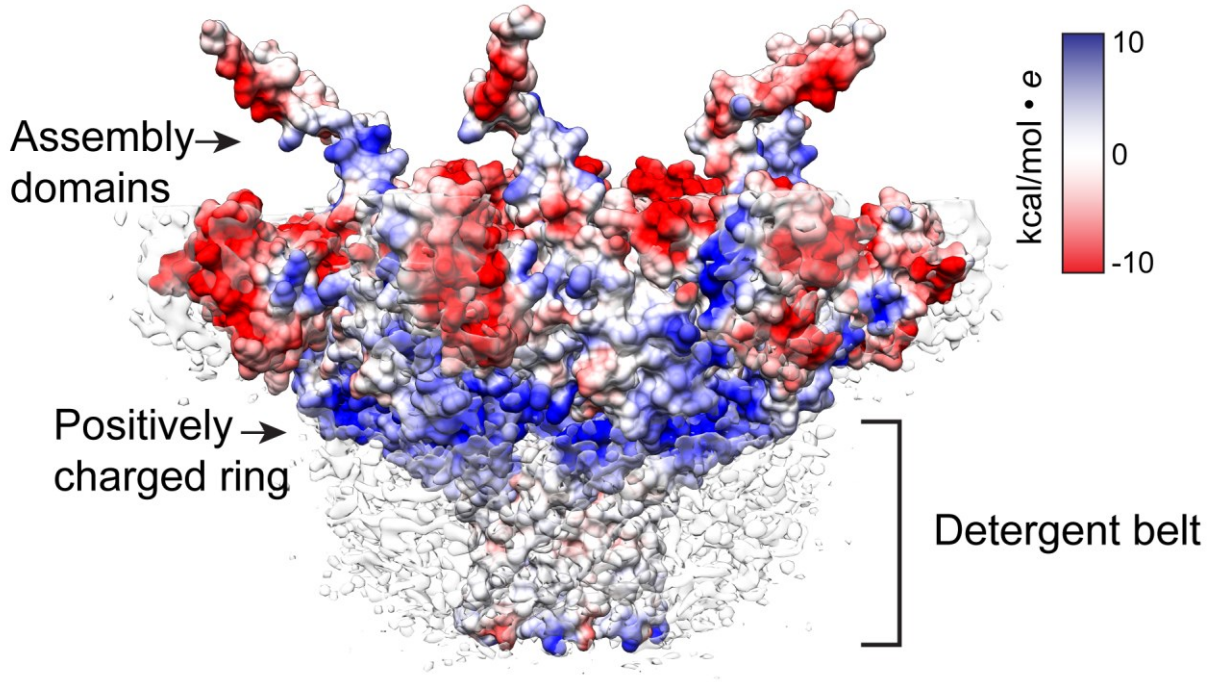


Figure 2.21 | Details of the PTEX map surrounding the detergent belt. The *engaged* state PTEX150/EXP2 heptamer, displayed in surface representation and colored by electrostatic potential. The bottom half of the full *engaged* state density map is superimposed, showing the location of the detergent belt in relation with the EXP2 TMD. A ring of positively charged residues is clearly visible directly above where the PVM surface would normally lie.

2.3.3 The PTEX150(S668-D823) heptamer acts as an adaptor between HSP101 and EXP2

Of the 993 residues in PTEX150, S668-D823 are well resolved in our structure and form a hook with a shaft (Fig. 3a-b). The hook domain consists of three short helices (H1-3) joined by several long loops. Directly N-terminal and C-terminal to the hook domain, the shaft is composed of proximal and distal shaft domains (Fig. 3a-b).

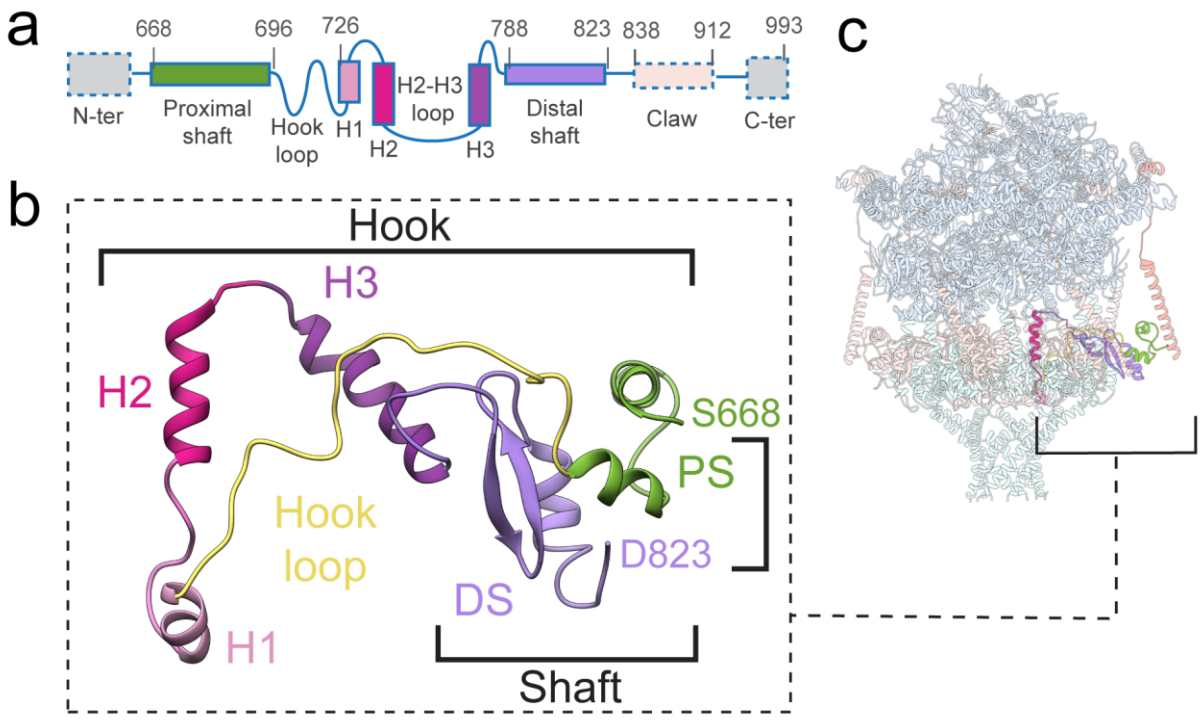


Figure 2.22 | The PTEX150(668-823) monomer. Linear schematic (**a**) and ribbon diagram (**b**) of the PTEX150(668-823) monomer in the *engaged* state. Dashed gray boxes represent unmodeled regions. PS, proximal shaft; DS, distal shaft. **c,** one PTEX150(668-823) monomer (coloured) within the PTEX complex.

The remaining 80% of PTEX150, not visible in our structures, is predicted to be intrinsically disordered (average disorder tendency of 0.83 in IUPred[31, 32], with scores above 0.5 indicating disorder) (**Fig. 2.22, 2.24**), unlike the rigid structured core of PTEX150 (S668-D823) (average disorder tendency score of 0.42, indicating ordered structure) (**Fig. 2.23**), suggesting that this 80% of the protein is too mobile to be observed and may be flexibly arranged outside the stable PTEX core.

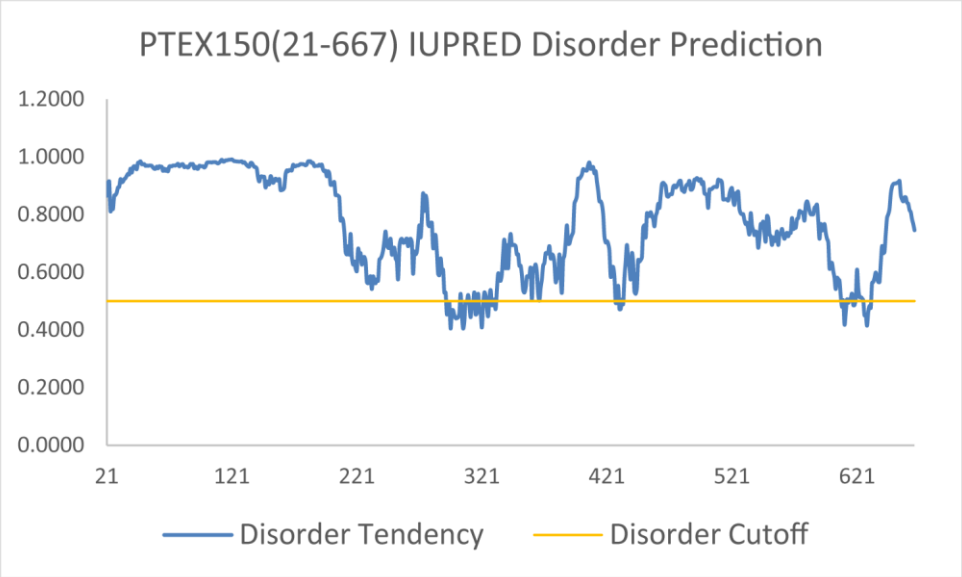


Figure 2.23 | IUPRED analysis of the PTEX150 N-terminus. IUPRED predicts the N-terminal 667 residues of PTEX150 to be highly disordered.

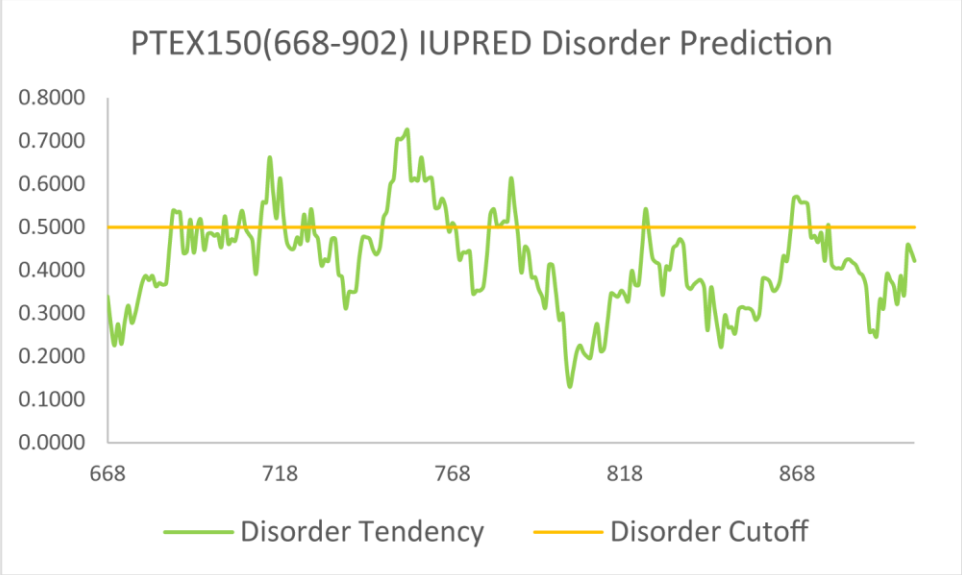


Figure 2.24 | IUPRED analysis of the PTEX150(668-902). IUPRED predicts PTEX150 residues 668-902, of which residues 668-823 are clearly visible in our structure, to be highly disordered.

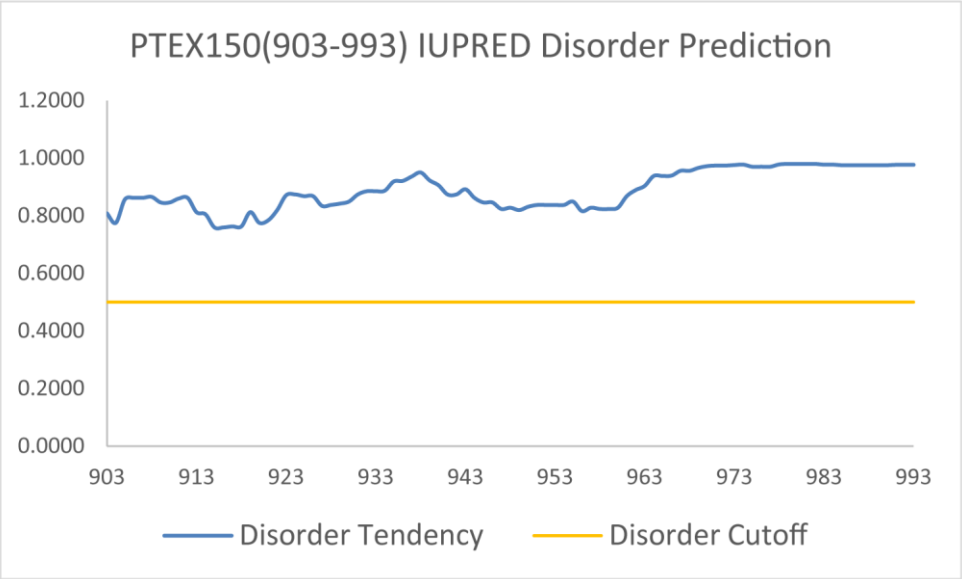


Figure 2.25 | IUPRED analysis of the PTEX150 C-terminus. IUPRED predicts the C-terminal 90 residues of PTEX150 to be highly disordered.

Seven PTEX150(S668-D823) hooks (labeled a-g) oligomerize, forming a flange-shaped C7-pseudosymmetric heptamer that fits into the mouth of the EXP2 channel (**Fig. 2.25**).

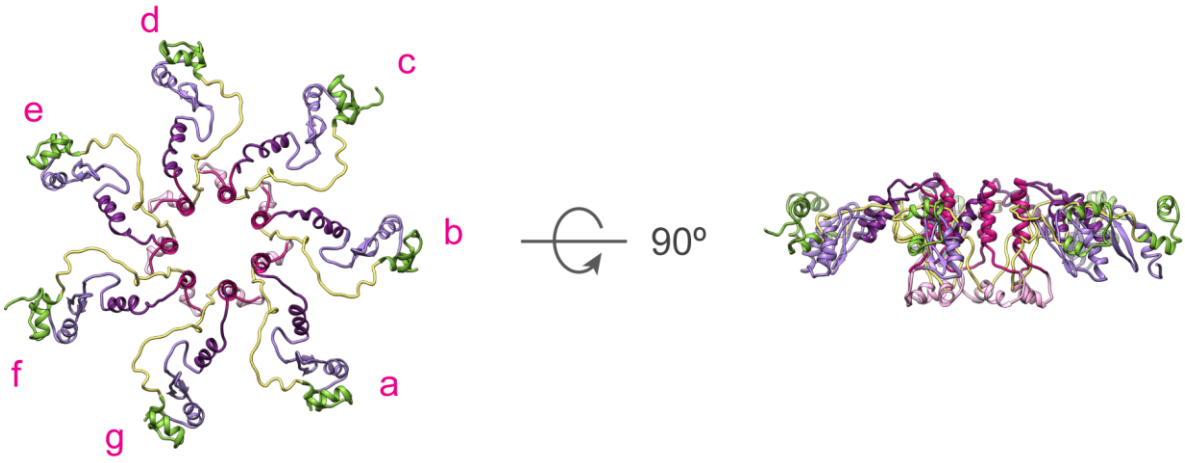


Figure 2.26 | The PTEX150(668-823) heptamer. Top (left) and side (right) views of the PTEX150(668-823) heptamer, colored as in (Fig. 2.21).

Each hook lies in the groove between adjacent EXP2 body domains, and the tip of the hook curls down into the mouth of the EXP2 pore (**Fig. 2.26).**

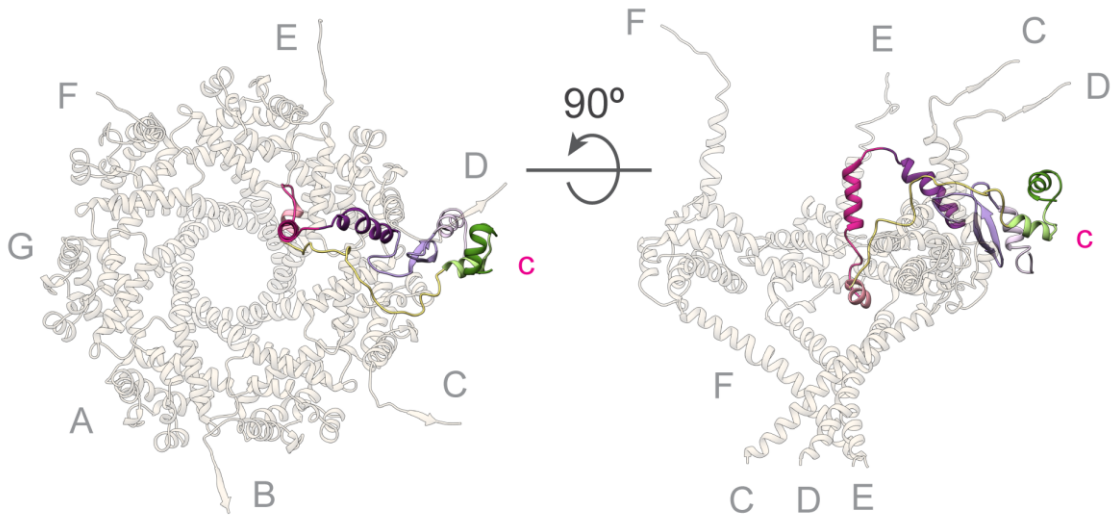


Figure 2.27 | The PTEX150(668-823) monomer and EXP2 funnel. Top (left) and side (right) views of the EXP2 funnel shown with a single PTEX150(668-823) monomer, illustrating how one PTEX150(668-823) monomer hooks into the top of the EXP2 funnel.

A vertical, heptameric ring of H2 helices sits in the mouth of the EXP2 funnel, forming a conduit between the hexameric HSP101 and heptameric EXP2 central pores (**Fig. 2.26**). In this way, PTEX150(S668-D823) serves as an adaptor between HSP101 and EXP2, providing a continuous protected path for unfolded cargo.

2.3.4 Endogenous cargo is observed bound in the channel of the HSP101 protein unfoldase

Class 1 Clp/HSP100 AAA+ ATPases are highly conserved hexameric protein unfoldases associated with diverse functions, which are known to thread polymeric substrates through a central pore[33, 34]. HSP101 is a 598kDa hexamer exemplifying the canonical Class 1 Clp/HSP100 domain architecture[35, 36], with a substrate-binding N-terminal domain (NTD)[37] followed by two AAA+ nucleotide-binding domains (NBD1 and NBD2), each containing a cargo-binding pore loop

(L1 and L2, respectively) that extends into the central pore (Fig. 4a-b). Additionally, HSP101 contains a C-terminal domain (CTD), and a coiled-coil middle domain (MD) insertion in the C-terminal end of NBD1 (Fig. 4a-b).

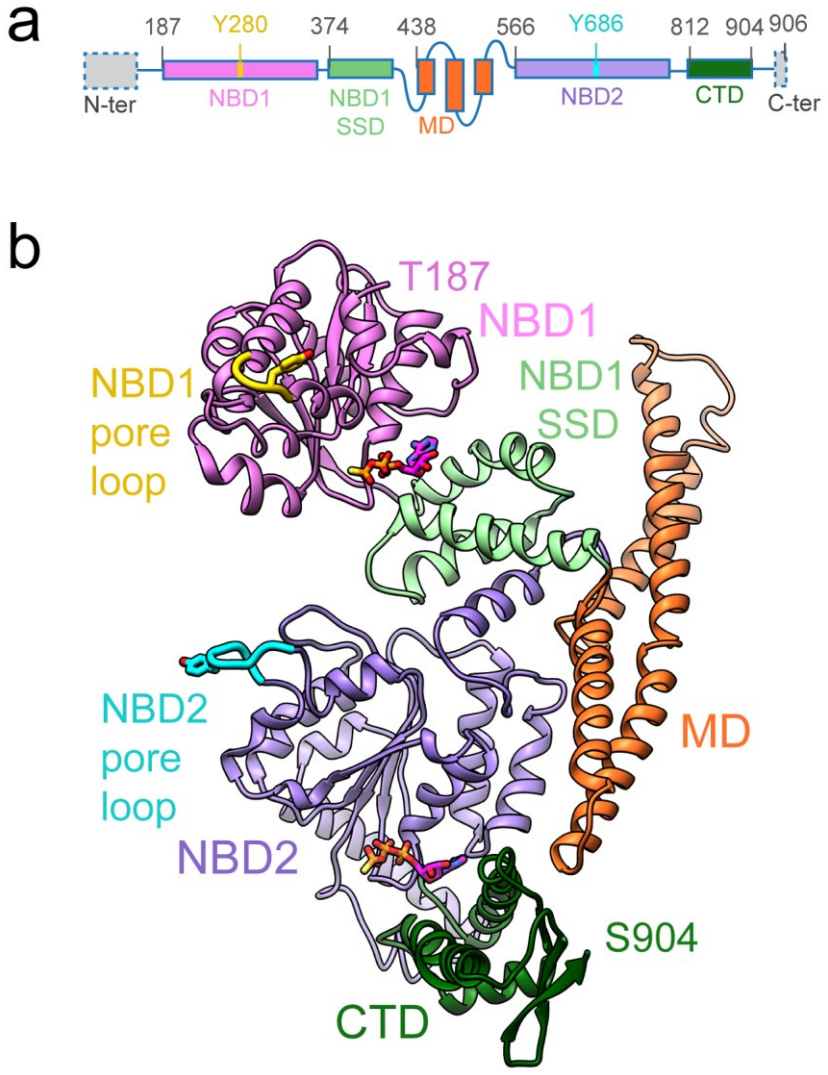


Figure 2.28 | The HSP101 monomer. Linear schematic (**a**) and ribbon diagram (**b**) of the HSP101 monomer in the *engaged* state.

Unlike Class 2 HSP100s [Ref [38]], Class 1 HSP100s form three-tiered hexamers, where the NTDs, NBD1s and NBD2s form the top, middle, and bottom tiers, respectively[35, 36]. In our *engaged* state structure, the NBD1 and NBD2 tiers are arranged in a right-handed ascending spiral[35, 36, 39]

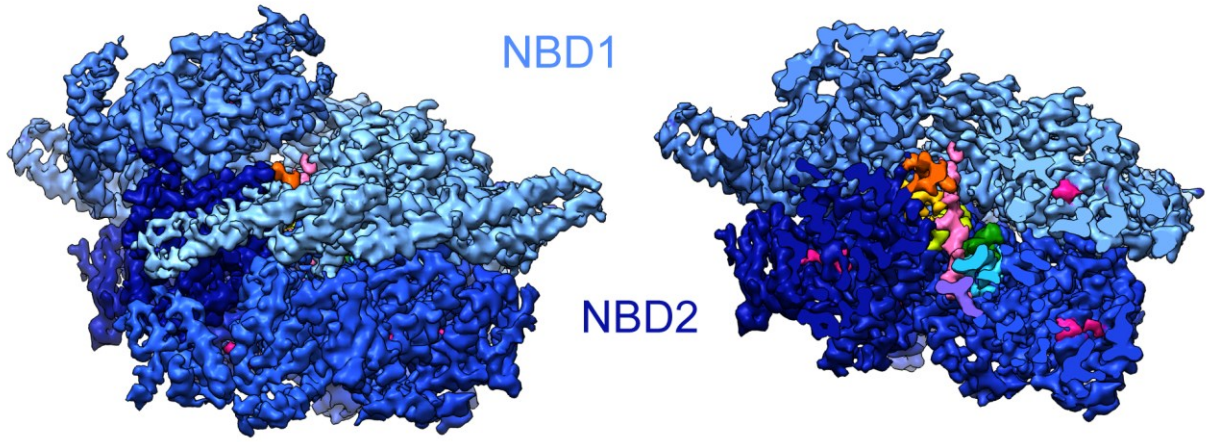


Figure 2.29 | Bisected view of HSP101 cryoEM map reveals endogenous cargo peptide density. Side view of the full (left) and bisected (right) HSP101 hexamer cryoEM map. NBD1 and NBD2 rings are coloured with light (NBD1) and dark (NBD2) blue gradients to emphasize the right-handed spiral shape of the hexamer. In the bisected map, NBD2 pore loop densities are coloured by protomer, ATPγS is colored magenta, and the cargo density is colored light pink.

(Fig. 4c). A layer of weaker density above the NBD1 tier may correspond to the NTDs, which are likely dynamic (Extended Data Fig. 8b). The MDs encircle the upper NBD1 tier. The central pore of the spiral is lined with pore loops bearing tyrosines in a spiral staircase pattern. The tyrosine sidechain densities intercalate with a 45Å-long density clearly visible in the middle of the chaperone pore (Fig. 4c-d, Supplementary Video 5),

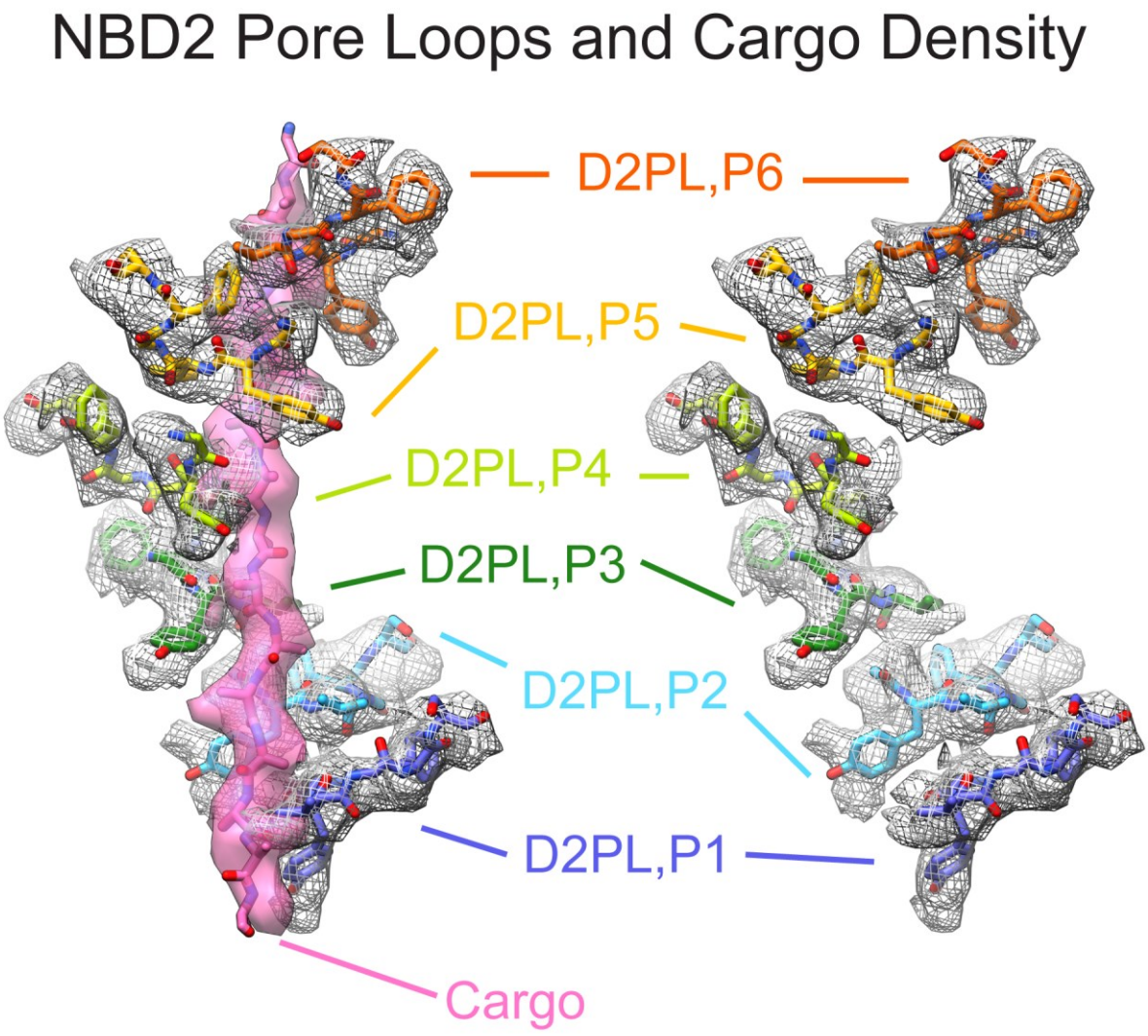


Figure 2.30 | Enlarged side view of the atomic models of the HSP101 NBD2 pore loops and unfolded cargo polypeptide backbone, shown with densities. NBD2 pore loops are colored as in (Fig. 2.28) and labeled by protomer (e. g., D2PL,P1: NBD2 Pore Loop, Protomer 1). Vertical distances between pore loop tyrosines in D2PL,P1-6 are 6.52Å, 6.28Å, 6.38Å, 6.96Å and 6.12Å, respectively.

which closely resembles unfolded cargo polypeptide densities reported in recently published cryoEM structures of homologous HSP100s bound to cargo[36, 38] (Extended Data Fig. 2a-d). The unfolded PTEX cargo polypeptide chain modeled into this 45Å-long density matches very closely

(RMSD of 1.09-1.25Å) with the unfolded cargo polypeptides in these cargo-bound homolog structures (Extended Data Fig. 2a-c).

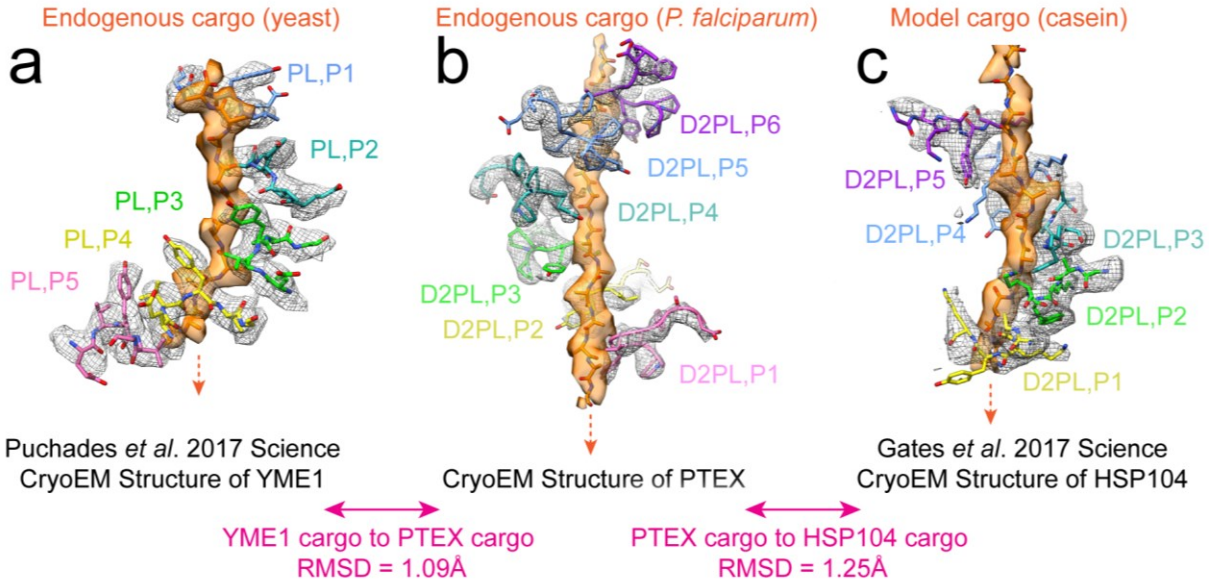


Figure 2.31 | CryoEM densities and atomic models of cargo and pore loops from the near-atomic resolution structures of Clp/HSP100 ATPases YME1[40] (**a**), PTEX HSP101 (**b**), and HSP104[36] (**c**). Tyrosine sidechain densities are clearly visible intercalating with the cargo densities. The modeled *engaged* state PTEX cargo has a calculated RMSD of 1.09Å and 1.25Å to the published YME1 and HSP104 cargo models, respectively. Pore loops are labeled by NBD and protomer (e. g., D2PL,P1: NBD2 Pore Loop, Protomer 1).

2.3.5 Key interactions for PTEX assembly and a potential mechanism for regulation

While the three PTEX components share extensive binding interfaces, we describe only the two most intriguing interactions here. In EXP2 protomers A-F, the assembly strand augments the CTD β-sheet in the HSP101 protomer situated directly above (Fig. 5a-b).

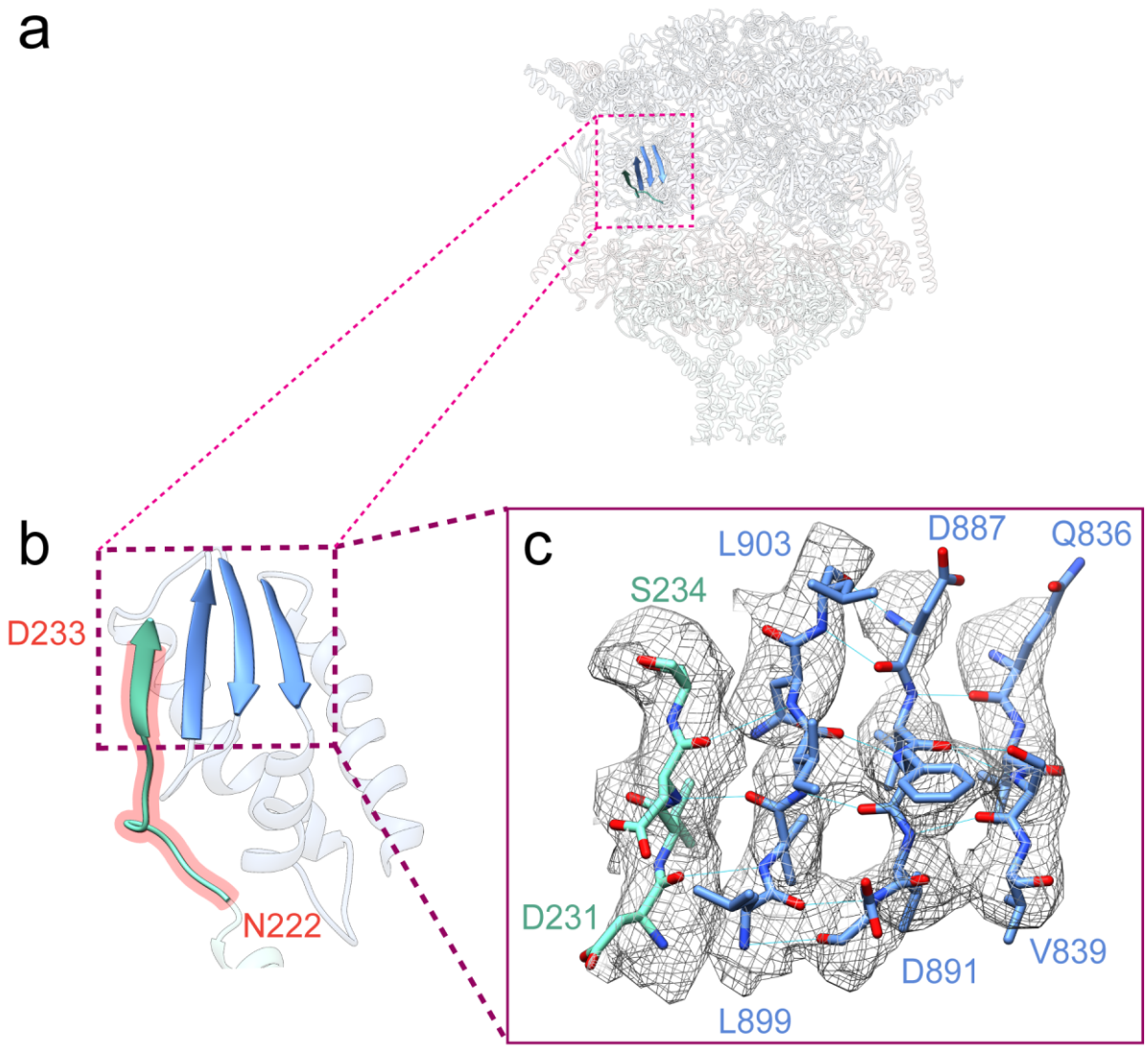


Figure 2.32 | Details of a β-sheet augmentation interaction between the C-termini of EXP2 and HSP101. The location of an intriguing β-sheet augmentation interaction between the C-termini of EXP2 and HSP101 is shown in (**a**), in the context of the full PTEX structure. Ribbon (**b**) and stick (**c**) models of the HSP101 CTD β-sheet augmented by the EXP2 assembly strand, shown with corresponding cryoEM density (mesh). Segment outlined in red was truncated in functional complementation assays.

Protomer G occupies the space beneath the gap between HSP101 protomers 1 and 6 (Fig. 1f-k). This hydrogen bond-mediated interaction tethers the HSP101 hexamer to the transmembrane funnel, positioning the central pore exit directly above the entrance to the PTEX150/EXP2 pore.

We hypothesized that this interaction is essential for assembly of the PTEX core complex, and that the complex must be stably assembled to be active. We tested this using genetic functional complementation in live parasites.

Knockdown of EXP2 produces a lethal defect in parasite growth and export that can be rescued by a mutant version of EXP2 lacking the last 54 residues[20]. Thus, the amino acids immediately following the assembly strand are not essential for PTEX function. However, complementation with a version of EXP2 lacking an additional 12 residues, removing the assembly strand, failed to rescue these phenotypes (Fig. 5c-f). These results demonstrate that the EXP2 assembly strand is critical to PTEX function, consistent with an essential role in docking the HSP101 unfoldase to the EXP2 membrane channel to facilitate translocation.

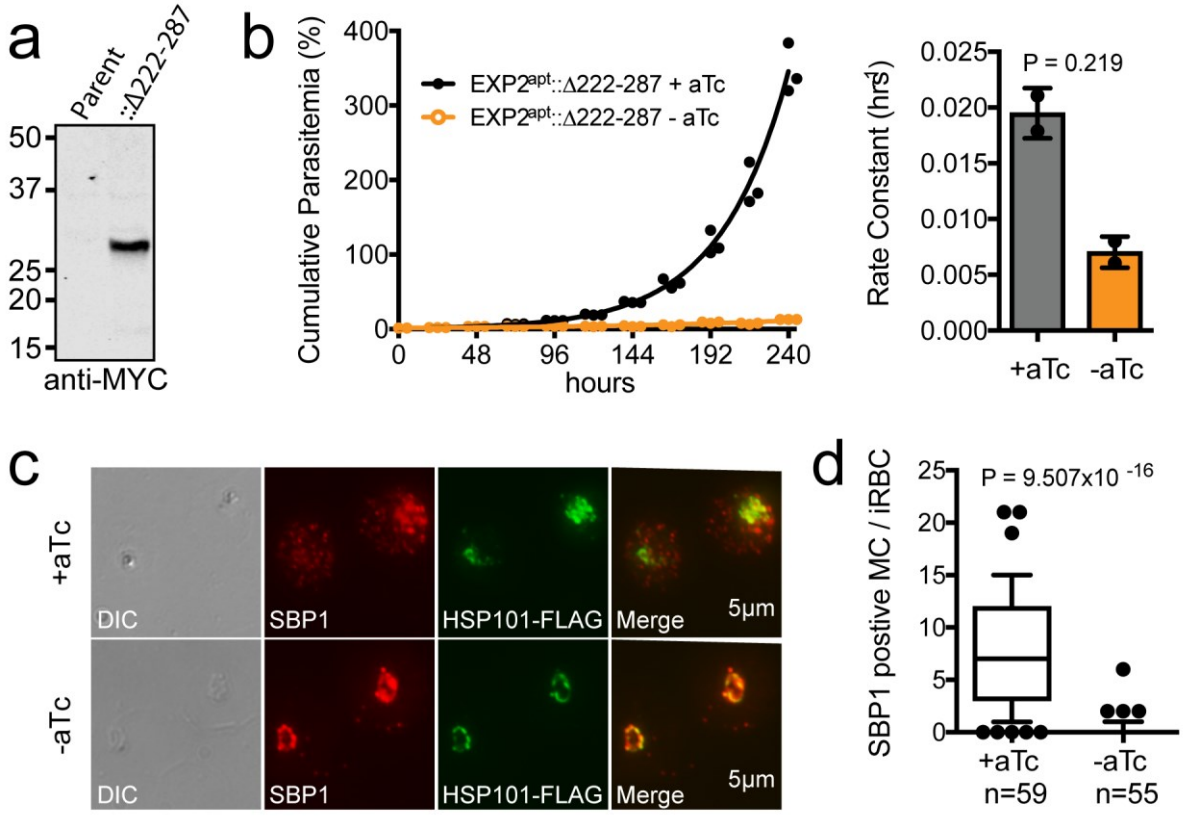


Figure 2.33 | Genetic functional complementation of EXP2 C-terminal truncation mutants. **a,** Western blot of EXP2apt parasites complemented with EXP2Δ222-287-3xMYC (predicted

molecular weight: 27.8 kDa after signal peptide cleavage). For blot source data, see Supplementary Fig. 1. **b,** Growth analysis of EXP2apt::Δ222-287-3xMYC. Parasites were grown with or without aTc to maintain or knockdown endogenous EXP2 expression, respectively. Results from one experiment with three technical replicates are shown. Bar graph shows mean exponential growth rate constant (hr-1) determined from the fit of the two independent experiments and error bars indicate s.d. **c,** Immunofluorescence assay (IFA) detecting exported protein SBP1 and HSP101-3xFLAG (as a PV marker) in EXP2apt::Δ222-287-3xMYC parasites allowed to develop with or without aTc to 24 hr post invasion. Dashed line indicates the traced boundary of the RBC. DIC, differential interference contrast. **d,** Quantification of SBP1 export IFA assays. Data are pooled from two independent experiments, n is the number of individual parasite-infected RBCs. Boxes and whiskers delineate 25th-75th and 10th-90th percentiles, respectively. All P values determined by an unpaired, two-sided t-test. All data shown represent two independent experiments.

A strong, albeit lower resolution claw-shaped density extends from the end of each modeled PTEX150(S668-D832) shaft to the HSP101 MD above, terminating in a three-turn helix resting atop the midpoint of the MD.

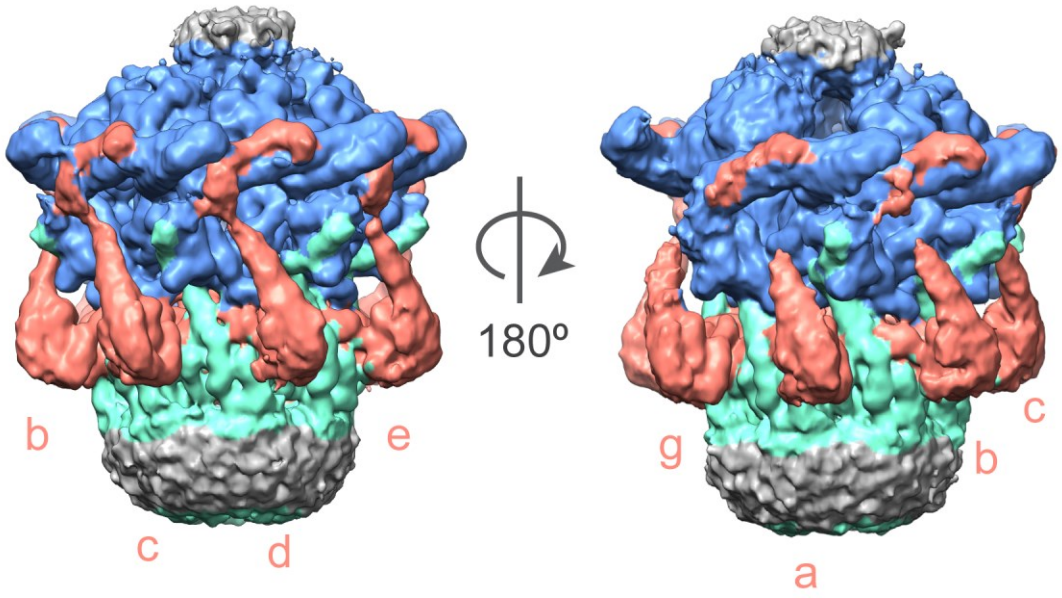


Figure 2.34 | Lower resolution details of the PTEX cryoEM map. *Resetting* state map of PTEX displayed at a lower threshold to show the strong claw-shaped densities extending from the PTEX150(668-823) shaft up to the HSP101 MD.

This helix forms a strong interaction with HSP101 Y488 and Y491 in claws a-e (**Extended Data Fig. 8d-e**) but is not visible in claw f in the *engaged* state. Claw g appears to form an additional interaction with the N-terminal end of the HSP101 protomer 1 MD (Extended Data Fig. 8d). The MD is known to play a critical role in regulating ATPase and unfoldase activities in related HSP100s [Ref [41, 42]], suggesting the importance of this interaction.

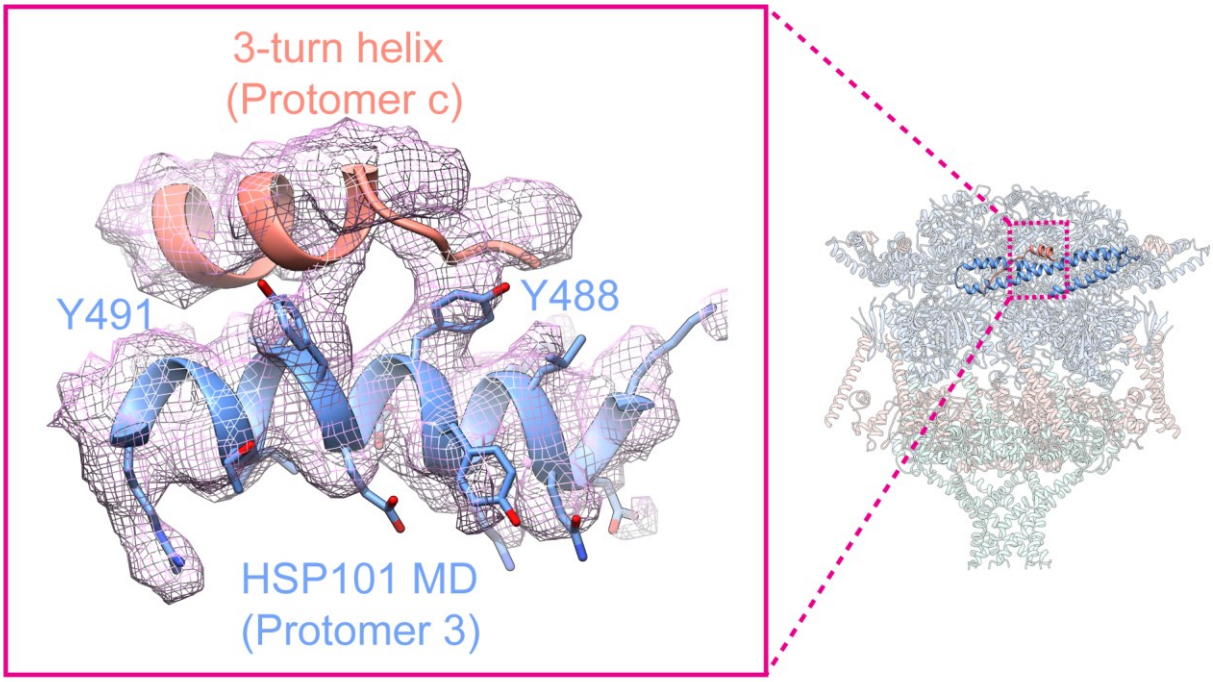


Figure 2.35 | Enlarged view of the interaction between HSP101 Y488 and Y491 and the three-turn helix, shown with corresponding cryoEM density (mesh).

2.3.6 Atomic details of the two observed states suggest a mechanism for translocation

In addition to the above-described *engaged* (195Å) state, a more compact (175Å) *resetting* state of PTEX was also observed. Much of PTEX150(S668-D823) and EXP2 remain unchanged between the *engaged* and *resetting* states, with a dramatic hinge-like swinging motion in the HSP101 hexamer accounting for the 20Å height difference. The TMD and B1-3 helices of EXP2

exhibit C7-symmetry, remaining identical between states (**Supplementary Video 6**). The deviation from C7-symmetry in the B4-5 helices and assembly domain is less pronounced in the *resetting* state (**Extended Data Fig. 3b-c**), likely due to the more planar arrangement of HSP101 protomers. As in the *engaged* state, slight inter-protomer variations in the PTEX150(S668-D823) H2-3 region bridge the gap between EXP2 and HSP101, maintaining a continuous protected path for unfolded cargo proteins.

The spiral staircase of HSP101 tyrosine pore loops in the *engaged* state collapses into a planar "C" shape in the *resetting* state (**Supplementary Video 7**), with a freedom of movement possibly conferred by the gap between HSP101 protomers 1 and 6[35, 36]. Originating at the interface between the NBD2 domains of HSP101 protomers 3 and 4, HSP101 protomers 4-6 swing downwards and outwards, creating a deep vertical cleft through the central pore of the hexamer.

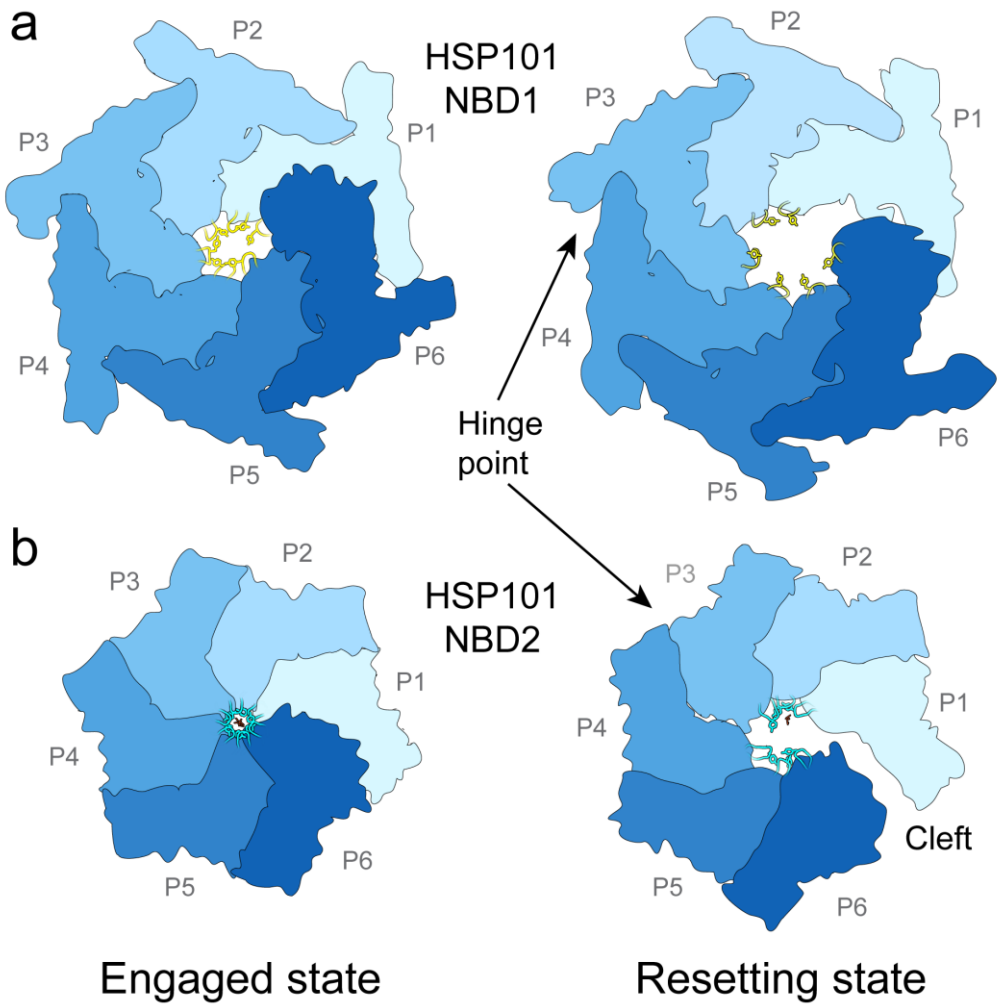


Figure 2.36 | Simplified top views of the HSP101 NBD1 and NBD2 tiers. a-b, Top view of HSP101 NBD1 (**a**) and NBD2 (**b**) in the *engaged* (left) and *resetting* (right) states, shown in simplified surface representation. The hinge point at the interface between HSP101 protomers 3 and 4 is indicated.

This motion pulls the NBD2 loops in protomers 4-6 away from the unfolded cargo (**Fig. 2.35, 2.36, Supplementary Video 6-7**). A shorter (19Å *vs* 45Å), unfolded cargo density remains visible, bound to the NBD2 loops in protomers 1-3, while no peptide density is visible in protomers 4-6 (**Fig. 2.36**).

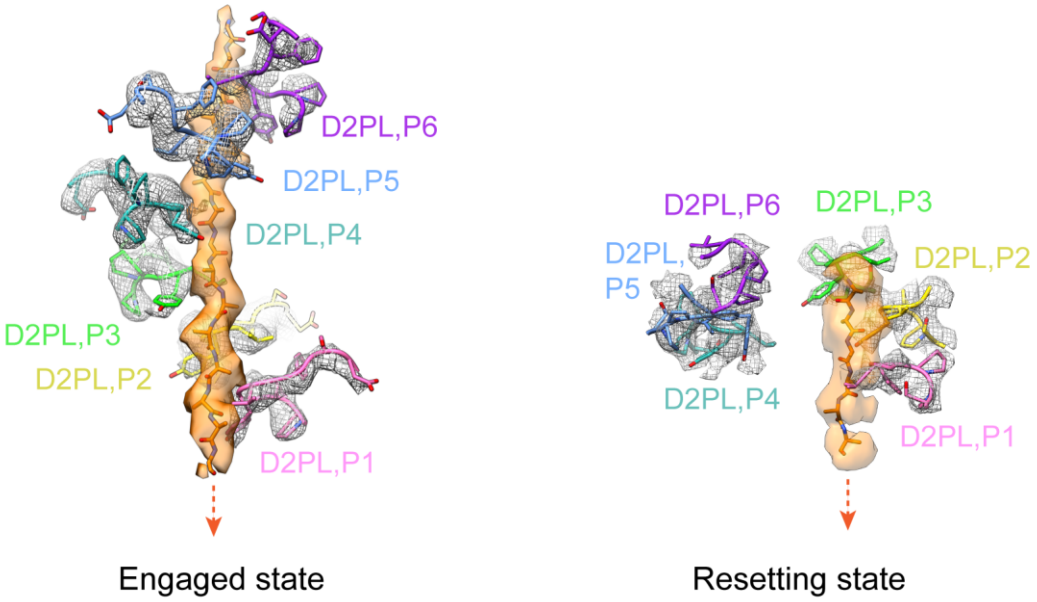


Figure 2.37 | Comparison of the endogenous cargo peptide density between the *engaged* and *resetting* states. Enlarged side view of the atomic models of the HSP101 NBD2 pore loops and unfolded cargo polypeptide backbone in the *engaged* (left) and *resetting* (right) states, shown with corresponding cryoEM densities. Tyrosine sidechain densities are clearly visible intercalating with the cargo densities. The modeled PTEX cargo has a calculated RMSD of 1.09Å and 1.25Å to the published YME1 and HSP104 cargo models, respectively. Pore loops are labeled by NBD and protomer (e. g., D2PL,P1: NBD2 Pore Loop, Protomer 1).

Furthermore, the NBD1 domain of protomer 3 rotates outward, such that the R361 arginine finger remains within 5.2Å of the ATPγS in the protomer 4 NBD1, while the nucleotide in the protomer 4 NBD2 shifts 7.5Å away from the R859 arginine finger in protomer 3 (Extended Data Fig. 3f-g).

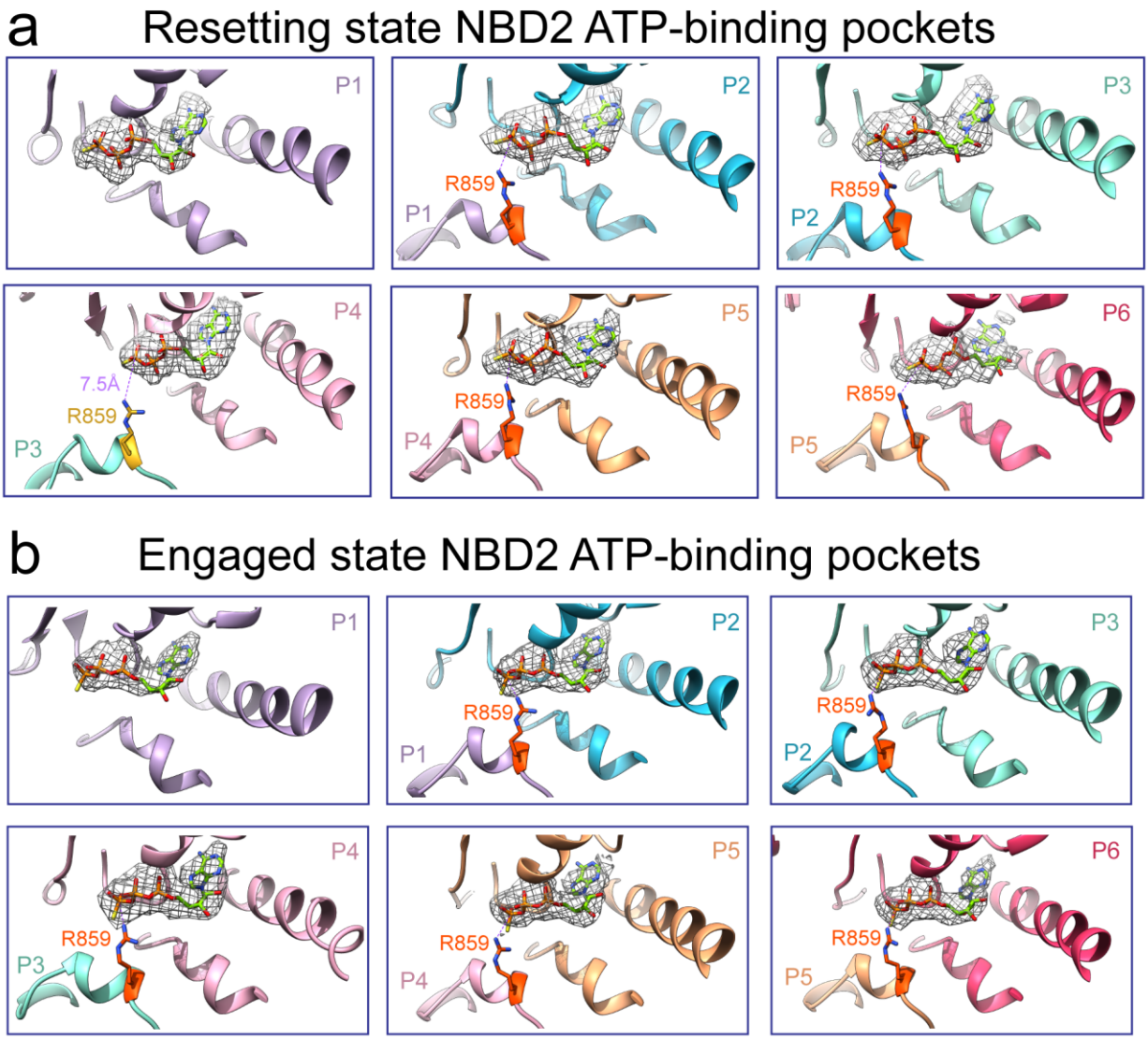


Figure 2.38 | Detailed views of the HSP101 NBD2 ATP binding pockets. a-b, Ribbon diagrams of the *resetting* state (**a**) and *engaged* state (**b**) nucleotide binding pockets are shown for each protomer. ATPγS in each pocket is shown with corresponding cryoEM density (mesh). The R859 arginine finger (sidechain shown in red-orange) is positioned ~3-5.5Å from the phosphorous atom in the γ-phosphate of the ATPγS in the binding pocket of the neighboring protomer in all protomers except R859 in protomer 3 in the *resetting* state (sidechain shown in gold), where the ATPγS bound in the protomer 4 NBD2 nucleotide pocket has shifted ~7.5Å away from the protomer 3 R859 arginine finger.

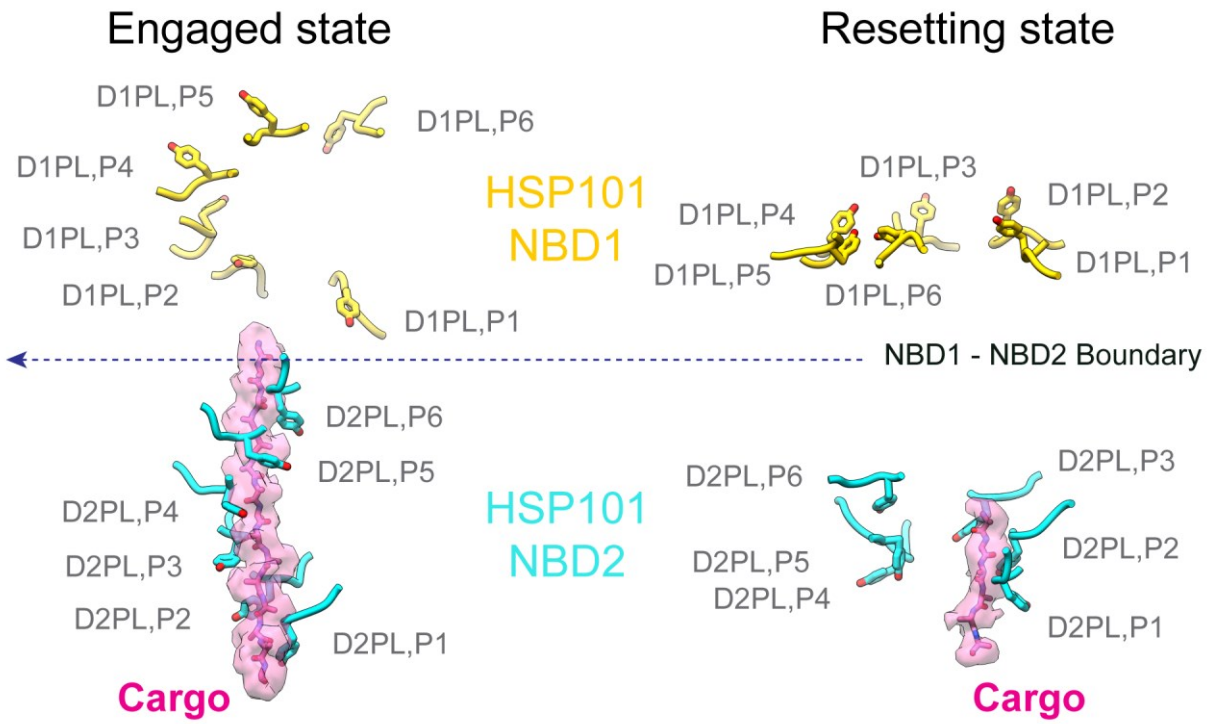


Figure 2.39 | Detailed comparison of the HSP101 cargo-binding site in the *engaged* and *resetting* states. Side views of the HSP101 pore loops with the unfolded cargo peptide backbone models (pink) built into the cryoEM densities (pink) in the *engaged* (left) and *resetting* (right) states. Vertical distances between pore loop tyrosines in consecutive loops are: *engaged* D1PL,P1-6: 9.41Å, 8.61Å, 1.40Å, 3.34Å, 2.28Å; *engaged* D2PL,P1-6: 6.52Å, 6.28Å, 6.38Å, 6.96Å, 6.12Å; *resetting* D1PL,P1-6: 1.75Å, -2.70Å, -1.65Å, -0.78Å, 1.81Å; *resetting* D2PL,P1-6: 5.88Å, 4.56Å, -6.80Å, 2.25Å, 7.88Å.

We propose a PTEX-mediated mechanism of protein translocation via a cyclic process involving at least two discrete states (**Fig. 2.39, Supplementary Video 6-7**), which we have captured by purifying PTEX complexes directly from parasites actively translocating cargo. The pore loops in HSP101 NBD2 form two hands which work together to thread the cargo protein through the central pore. NBD2 loops from HSP101 protomers 1-3 form the passive hand, located closest to the PTEX150(S668-D823)/EXP2 funnel, which stays fixed between states (**Fig. 2.38, 2.39**). NBD2 loops from HSP101 protomers 4-6 form the active hand, which moves along the channel axis (above the passive hand), grasping the unfolding peptide and feeding it through the passive hand.

In the *engaged* state, all six NBD2 pore loops grip the unfolded peptide in the spiral staircase formation (**Fig. 2.38, 2.39**). As the HSP101 hexamer collapses into the *resetting* state, the active hand moves downwards, feeding the newly unfolded peptide through the passive hand, into the PTEX150(S668-D823)/EXP2 funnel below. The passive hand then grips the unfolded peptide, preventing it from slipping back toward the HSP101 apical entrance while the active hand swings outward, releasing the cargo (**Fig. 2.38, 2.39**). Finally, the active hand moves upwards to grasp the unfolding protein further upstream, transitioning back into the *engaged* state. With this elegant cyclic feeding mechanism, the unfolded cargo protein is threaded through the translocon, across the PVM and into the host cell cytosol.

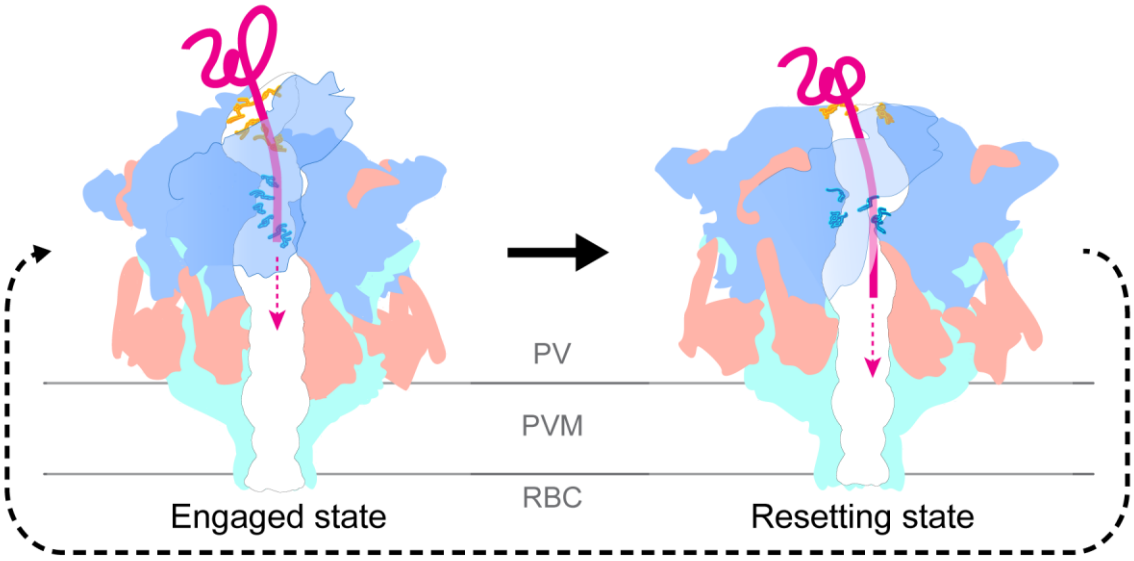


Figure 2.40 | Proposed stepwise feeding mechanism of translocation by PTEX. NBD1 and NBD2 pore loops and cargo are colored as in (**Fig. 2.38**).

2.4 Discussion

The states captured here may be two of several states in the processive phase of translocation. Additional states likely exist for cargo-recognition. Although we did not observe PTEX-free HSP101 oligomers as suggested by Elsworth *et al.*[21], we did observe additional, seemingly cargo-free PTEX complexes (**Fig. 2.9**) which did not refine to better than 7Å, suggesting conformational heterogeneity in the absence of stabilizing cargo interactions. Cargo-PTEX interactions during cargo-recognition may be transient, possibly explaining why we did not observe the HSP101 NTDs or other components potentially required for cargo-recognition.

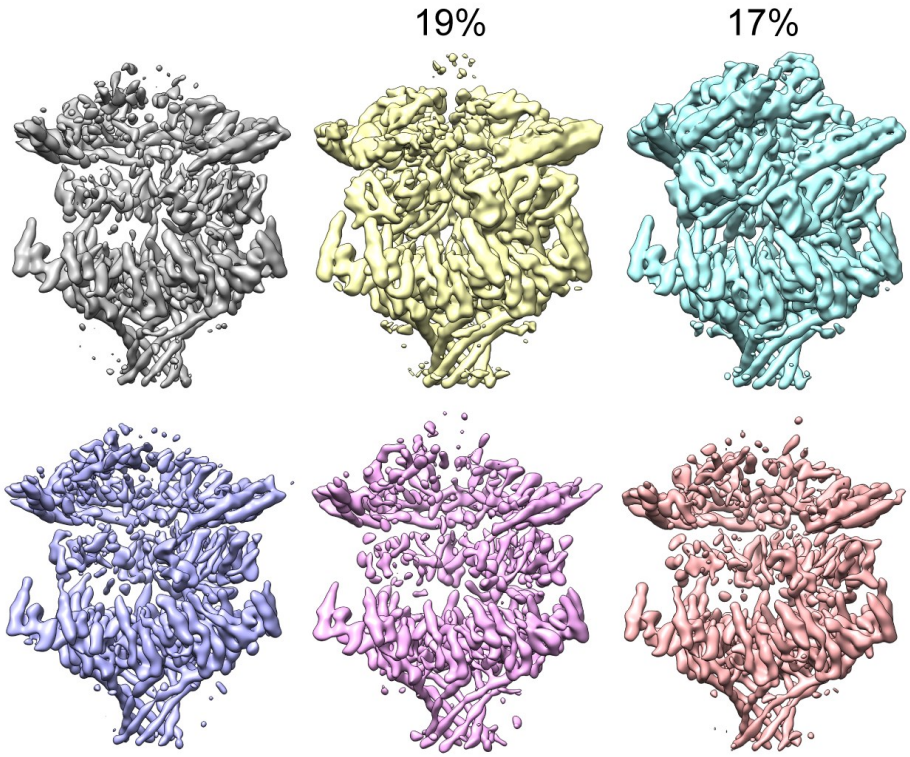


Figure 2.41 | Additional 3D classes may correspond to other states. 3D classification in RELION revealed additional, seemingly cargo-free PTEX states (grey, purple, pink, and salmon) in addition to the two cargo-bound states (yellow, mint) that gave rise to our two near-atomic resolution structures. The density in HSP101, particularly around the cargo-binding site, is weak

and appears not to be well-ordered in the cargo-free PTEX states, suggesting conformational heterogeneity in the absence of stabilizing cargo interactions.

Without these details, the mechanisms for cargo-recognition and subsequent refolding after translocation remain unclear, although some evidence suggests involvement of exported parasite chaperones[43] or co-opted host chaperonins[44]. Interestingly, based on secondary structure prediction and PTEX150 truncation experiments[21], PTEX150 residues D838-F912 may occupy the claw (PTEX150 D838-E873) and three-turn helix (PTEX150 S884-F912) densities that remain unassigned in our structures.

Our work demonstrates the advantages of obtaining structures of challenging protein complexes in functionally relevant states by imaging samples purified directly from endogenous sources. Direct observation of the native PTEX core provides compelling evidence that this complex, comprising EXP2, PTEX150 and HSP101, is a *bona fide* translocon embedded in the PVM that serves as the gateway for the malaria parasite exportome. In addition to establishing the role of EXP2 as the membrane-spanning pore of PTEX and providing insight into the mechanism of this essential protein translocating machine, our structures reveal a unique interaction between the EXP2 assembly domain and the HSP101 C-terminal domain which is indispensable for PTEX function. These highly sought-after atomic structures of PTEX provide exciting possibilities for designing a new class of drugs inhibiting this essential gatekeeper of the malarial exportome.

2.5 Acknowledgements

This research was supported in part by grants from National Institutes of Health (R21AI125983 to P.F.E., R01GM071940/AI094386/DE025567 to Z.H.Z. and K99/R00 HL133453 to J.R.B.). P.F.E. is the Alexander and Renee Kolin Endowed Chair in Molecular Biology and Biophysics. CM.H. acknowledges funding from the Ruth L. Kirschstein National Research Service Award (AI007323). We thank the UCLA Proteome Research Center for assistance in mass spectrometry and acknowledge the use of instruments in the Electron Imaging Center for Nanomachines supported by UCLA and grants from NIH (S10RR23057, S10OD018111 and U24GM116792) and NSF (DBI-1338135 and DMR-1548924). We thank Anthony W. P. Fitzpatrick for input on cryoEM aspects of the project and Judy Su for helping with Figure 1a

2.6 Data availability

The atomic models and the cryoEM density maps are deposited to the Protein Data Bank and the Electron Microscopy Data Bank, under the accession numbers 6E10, 6E11, EMD-8951 and EMD-8952.

2.7 Competing interests

The authors declare no competing interests.

2.8 Materials and Methods

Cells

P. falciparum strain NF54^{attB} [Ref [45]] was used exclusively in the study. De-identified, IRB-exempt expired RBCs were obtained from the blood bank at the St. Louis Children's Hospital. PCR amplified regions from the NF54attB genome were found to match the genome sequence for 3D7, a sub clone of NF54. The prescence of the cg6 localized attB sequence was verified by successful Bxb1-mediated integration at that site.

Parasite culture and genetic modification for PTEX purification

P. falciparum culture was performed as described with the exception that RPMI was supplemented with 0.5% Albumax I [46]. All plasmid construction was carried out by Infusion cloning (Clontech) unless otherwise noted. Integration of a 3xFLAG fusion at the endogenous HSP101 C-terminus was accomplished with CRISPR/Cas9 editing. A Cas9 target site was chosen just upstream of the *hsp101* stop codon (TAATAGTAAAGCTAAAAACT) and the guide RNA seed sequence was synthesized as a sense and anti-sense primer pair (sense shown) 5'-TAAGTATATAATATTTAATAGTAAAGCTAAAAACTGTTTTAGAGCT
AGAA -3', annealed and inserted into the *BtgZI* site of the plasmid pAIO [47], resulting in the plasmid pAIO-HSP101-CT-gRNA1. A 5' homology flank (up to but not including the stop codon) was amplified from *P. falciparum* NF54^{attB} genomic DNA using primers 5'-GACGCGAGGAAAATTAGCATGCATCCTTAAGGAGATTCTGGTATGCCACTTGGTTC-3' and 5'- CTGCACCTGGCCTAGGGGTCTTAGATAAGTTTATAACTAAGTTTTTAGCTTTACTATT-3', incorporating a synonymous shield mutation in the protospacer adjustment motif of the gRNA target site within the *hsp101* coding sequence. A 3' homology flank

(beginning 3 bp downstream of the stop codon) was amplified using primers 5'-CACTATAGAACTCGAGAATTACGCATATATATATATATATATATATATAACATGGGTTG-3' and 5'-GAACCAAGTGGCATACCAGAATCTCCTTAAGGATGCATGCTAATTTTCCTCGCGTC-3'F. The flank amplicons were assembled in a second PCR reaction using primers 5'-CACTATAGAACTCGAGAATTACGCATATATATATATATATATATATATAACATGGGTTG-3' and 5'-CTGCACCTGGCCTAGGGGTCTTAGATAAGTTTATAACTAAGTTTTTAGCTTTACTATT-3' and inserted between *XhoI* and *AvrII* in pPM2GT[46]. The GFP tag between *AvrII* and *EagI* in this vector was then replaced with sequence encoding a 3xFLAG tag using the primer 5'-CTTAGTTATAAACTTATCTAAGACCCCTAGGGACTACAAGGACGACGACGACAAGGATTATAAAGATGATGATGATAAAGATTATAAAGATGATGATGATAAATGACGGCCGCGTCGAGTTATATAATATATTTATG-3' and a QuikChange Lightning Multi Site-Directed Mutagenesis kit (Agilent), resulting in the plasmid pPM2GT-HSP101-3xFLAG. This plasmid was linearized at the *AflII* site between the 3' and 5' homology flanks and co-transfected with pAIO-HSP101-CT-gRNA1 into *P. falciparum* NF54^{attB} parasites[45]. Selection with 10 nM WR99210 was applied 24 hours after transfection. Once parasites returned from selection, integration at the intended site was confirmed by PCR with primers 5'-CGAAAACTTTTATGGTATTAATATAACAG-3' and 5'-CCTTGTCGTCGTCGTCCTTG-3' and a clonal line was isolated by limiting dilution.

For PTEX purification, HSP101-3xFLAG parasites were synchronized by serial treatment with 5% w/v D-sorbitol and then expanded while shaking to increase singlet invasion events and

maintain synchrony. For each preparation, ~$2x10^{10}$ parasite-infected erythrocytes were collected at the ring stage (typically ~500 mls of 2% hematocrit culture at ~20% parasitemia). Erythrocytes were lysed in 10x pellet volume of cold phosphate buffered saline (PBS) containing 0.0125% saponin (Sigma, sapogenin content ≥10%) and EDTA-free protease inhibitory cocktail (Roche or Pierce). Released parasites were washed in cold PBS containing EDTA-free protease inhibitory cocktail and washed cell pellets were frozen in liquid nitrogen and stored at -80°C.

Affinity purification of PTEX core complex from parasite pellets

Frozen parasite pellets were resuspended in Lysis Buffer (25mM HEPES pH 7.4, 10mM $MgCl_2$, 150mM KCl, 10% Glycerol) and homogenized using a glass Dounce tissue homogenizer. The membrane fraction was isolated from the homogenized lysate by centrifugation at 100,000*g* for one hour. The membrane pellet was solubilized in Solubilization Buffer (25mM HEPES pH 7.4, 10mM $MgCl_2$, 150mM KCl, 10% Glycerol, 0.4% Triton X-100) and the solubilized membranes were then applied to anti-FLAG M2 Affinity Gel resin (Sigma). The resin was washed extensively in Wash Buffer (25mM HEPES pH 7.4, 10mM $MgCl_2$, 150mM KCl, 10% Glycerol, 0.015% Triton X-100), after which the protein was eluted from the affinity resin with Elution Buffer (25mM HEPES pH 7.4, 10mM $MgCl_2$, 150mM KCl, 2mM ATPγS, 0.015% Triton X-100, 500µg/ml FLAG peptide).

The presence and relative abundance of the three PTEX core components were verified by silver stained SDS-PAGE and tryptic digest liquid chromatography-mass spectrometry (Extended Data Fig. 1d-e). The extremely low yields achievable when purifying PTEX directly from *P. falciparum* parasites prohibited the conventional approach of evaluating sample quality by size exclusion chromatography. Thus, during the iterative process of screening for optimal purification

conditions, sample quality was assessed by negative stain (uranyl acetate) transmission electron microscopy in an FEI TF20 microscope equipped with a TVIPS 16 mega-pixel CCD camera. Briefly, small datasets of ~100,000 particles were collected and 2D class averages were generated in RELION[23, 24] to assess the presence of sufficient numbers of intact PTEX particles yielding "good" class averages exhibiting distinct features. For example, C7 symmetry could be recognized in top views, and the characteristic Clp/HSP100 layers were visible in side views (Extended Data Fig. 6a-c).

Cryo Electron Microscopy

3µl aliquots of purified PTEX core complex were applied to glow-discharged lacey carbon grids with a supporting ultrathin carbon film (Ted Pella). Grids were then blotted with filter paper and vitrified in liquid ethane using an FEI Vitrobot Mark IV or a home-made manual plunger. CryoEM grids were screened in an FEI Tecnai TF20 transmission electron microscope while optimizing freezing conditions.

Higher resolution cryoEM images were collected on a Gatan K2-Summit direct electron detector in counting mode on an FEI Titan Krios at 300kV equipped with a Gatan Quantum energy filter set at a 20 eV slit width. Fifty frames were recorded for each movie at a pixel size of 1.04Å at the specimen scale, with a 200 ms exposure time and an average dose rate of 1.2 electrons per $Å^2$ per frame, resulting in a total dose of 60 electrons per $Å^2$ per movie. The final dataset consists of a total of 25,000 movies recorded in four separate sessions.

Image processing and 3D reconstruction

Frames in each movie were aligned, gain reference-corrected and dose-weighted to generate a

micrograph using MotionCor2 [Ref [48]]. Aligned and un-dose-weighted micrographs were also generated and used for contrast transfer function (CTF) estimation using CTFFIND4 [Ref [49]] and PTEX particle picking by hand and using Gautomatch[50].

1,508,462 particles were extracted from 19,752 micrographs and initially binned by a factor of 2. After two rounds of reference-free two-dimensional (2D) classification in RELION, 422,713 particles were selected as "good" particles from distinct 2D class averages representing different views of the PTEX core complex. These particles were then used in a one-class *ab initio* reconstruction followed by homogeneous refinement in CryoSPARC[51], yielding a 4.8Å *ab initio* 3D map.

The original 422,713 "good" particles were then aligned in a 3D refinement in RELION using the 4.8Å CryoSPARC map as an initial reference. All subsequent image-processing steps were performed using RELION. After this refinement, the particles were unbinned, their centers recalculated and used to re-extract particles from the original micrographs without binning. The newly extracted, unbinned particles were then aligned with a second 3D refinement yielding a ~4.5Å reconstruction.

An exhaustive, iterative search of classification and refinement conditions was used to sort out different conformations and further improve resolution (Extended Data Fig. 9). Briefly, upon further sorting using 3D-classification without alignment, we identified two homogenous particle subsets corresponding to the *engaged* and *resetting* states (Extended Data Fig. 9). Particles in the two subsets were refined separately, yielding full maps with overall resolutions of 4.16Å and 4.23Å, respectively.

Focused 3D classification without alignment followed by focused refinement was used to

further improve the resolution of mobile regions of the structure in both states. C7 symmetry was applied in the focused 3D classification and refinement steps of the heptameric halves, comprising EXP2 and PTEX150, yielding a 3.4Å *engaged* state map and a 3.5Å *resetting* state map (Extended Data Fig. 3,9). The same procedure, except with C1 symmetry, was applied to the hexameric half of the *engaged* state, yielding a 4.09Å map (Extended Data Fig. 3,9). This last step was also applied to the hexameric half of the *resetting* state, but did not yield improvements in resolution. Further efforts of focused 3D classification and refinement of individual HSP101 protomers, individual claws, and HSP101 N-terminal domain densities in the two states did not ultimately yield improvements in resolution in either state.

Model Building and Refinement

Map interpretation was performed with UCSF Chimera[52] and COOT[53]. *P. falciparum* protein sequences were obtained from the National Center for Biotechnology Information (NCBI)[54] and the PlasmoDB[55] protein databases. PHYRE2 [Ref [56]] secondary structure predictions were used as an aid for initial manual sequence registration. Models for a single monomer of HSP101, PTEX150, and EXP2 in the *engaged* state were all built *de novo.* This first model for each protein monomer was then placed into the density maps of other protomers to aid de novo modeling of subsequent protomers. Individual protomers in the complex were then manually remodeled to ensure a close fit between densities and models. The same process was repeated for the *resetting* state. Manual refinement targeting protein geometry alone was done primarily along the periphery and flexible regions of the complex (*e.g.*, the MDs of HSP101). While their densities and backbone traces were visible, we were unable to model the claw with its connected three-turn helix, nor one of the 12 MD loops in the *resetting* state (Fig. 5g,h). The three-turn helix displayed a few bulky

side chains interacting with the MD of HSP101, however the lack of backbone connection to our atomic model of the complex and the limited visibility of smaller side chains in this region have made sequence assignment challenging.

Manual refinement targeting both protein geometry and fit with the density map was used primarily in the core regions where resolution was higher and noise was minimal. Rotamers were fit manually in COOT and improved using the 'Back-rub Rotamers' setting. The resulting models for the complexes were subjected to the phenix.real_space_refine program in PHENIX[57]. Following this step, Molprobity[58] reported less than ideal clash scores and map-to-model cross-correlation. To improve the geometry and fit, manual adjustments were made to protein geometry and density map fit, with the additional step of using Molprobity[58] clash dots and sphere-refinement in COOT.

The complex was then broken into three portions: (1) symmetric regions of EXP2 and PTEX150, (2) HSP101, and (3) the full PTEX complex. These model segments were fed back to phenix.geometry_minimization in PHENIX and then to phenix.real_space_refine using simulated annealing and global minimization applying Emsley's Ramachandran restraints[53]. Following another round of manual checks and improvements, all models were subjected to phenix.real_space_refine with default settings one last time.

All figures and videos were prepared with UCSF Chimera, Pymol[59], and Resmap[22]. Molprobity was used to validate the stereochemistry of the final models.

Genetic complementation

For expression of a complementing second copy of truncated EXP2, the *exp2* coding sequence up to codon position 221 was amplified with primers 5'-

CGAATAAACACGATTTTTTCTCGAGATGAAAGTCAGTTATATATTT TCCTTTTTTTTGTTATTCTTCG-3' and 5'-AATCAACTTTTGTTCGCTAGCTTTCTTTG ATTCCATAGATTTCAATTTCTCTTCC-3' and inserted into the plasmid pyEOE-attP-EXP2-3xMYC[20] between *XhoI* and *NheI*, resulting in the plasmid pyEOE-attP-EXP2Δ222-287-3xMYC. This plasmid was co-transfected with pINT[45] into EXP2^{apt}::HSP101-3xFLAG conditional knockdown parasites[20] at the mature schizont stage using a Nucleofector 2b and Basic Parasite Nucleofector kit 2 (Lonza). Selection with 2μM DSM1 [Ref [60]] was applied 24 h post transfection (in addition to 2.5 μg/ml Blasticidin S and 1 μM anhydrotetracycline (aTc) for maintenance of endogenous EXP2 translational control by the aptamer system) to facilitate integration into the attP site engineered in the benign *cg6* locus through integrase mediated attB x attP recombination. Following return from selection, parasites were cloned by limiting dilution, and expression of EXP2Δ222-287-3xMYC was confirmed by western blot.

Parasite growth assays

EXP2^{apt}::Δ222-287 parasites were extensively washed to remove aTc and plated with or without 1μM aTc in triplicate at an initial parasitemia of 1%. Media was changed every 48 h and 1:1 subculture was performed every other day beginning on day 4 to avoid culture overgrowth. Parasitemia (percent of total red blood cells (RBCs) infected) was measured every 24 h by flow cytometry on a FACSCanto (BD Biosciences) by nucleic acid staining of cultured RBCs with PBS containing 0.8 μg/ml acridine orange. Cumulative parasitemias were back calculated based on the subculture schedule and data were fit to an exponential growth equation to determine rate constants using Prism (Graphpad).

Quantification of protein export

For evaluation of protein export by immunofluorescence assay (IFA), mature schizonts were purified on a magnetic column and allowed to invade fresh, uninfected RBCs with shaking for 3 hours before treatment with 5% w/v D-sorbitol to destroy unruptured schizonts. Pulse invaded cells were plated with or without 1 μM aTc and allowed to develop 24 h post invasion. Thin smears of infected RBCs were briefly air dried and immediately fixed in ice cold acetone for 2 minutes. After fixation, samples were blocked for 30 minutes in PBS+3% BSA followed by incubation for one hour with primary antibody solutions containing mouse anti-FLAG M2 mAb (detecting HSP101-3xFLAG to mark the PVM) and rabbit anti-SBP1. After washing, secondary antibody incubation was carried out for one hour with Alexa Fluor anti-mouse 488 and anti-rabbit 594 IgG antibodies (Life Technologies), each diluted 1:2000. After final washing, coverslips were mounted over each sample using Pro-long antifade Gold with DAPI (Life Technologies). Images were collected with an ORCA-ER CCD camera (Hamamatsu) using AxioVision software on an Axio Imager.M1 microscope (Zeiss) with a 100x oil immersion objective using the same exposure times for each image (300 ms for SBP1-594, 150 ms for FLAG-488). Ten images were acquired for each condition using the DAPI channel for field selection to avoid bias. Images were then analyzed using Volocity 6.3 (PerkinElmer). The border of each single-infected erythrocyte was traced using the DIC channel as a guide to define a region of interest (ROI). The PVM was marked using the "find objects" measurement tool for the HSP101-3xFLAG-488 channel (automatic threshold setting with threshold offset set to -30% and minimum object size set to 0.5 μm^2). Individual Maurer's clefts were identified using the "find spots" measurement tool for the SBP1-594 channel (offset minimum spot intensity set to 40% and brightest spot within radius set to 0.5 μm). All spots within the PVM object boundary were then removed using

the “subtract” measurement tool and the number and fluorescent intensity of the remaining spots in each ROI were collected. Data were pooled from two independent experiments and plotted with Prism.

Antibodies

The following primary antibodies were used for IFA and western blot: mouse anti-FLAG mAb clone M2 (Sigma) (IFA: 1:500, WB 1:500); rabbit polyclonal anti-SBP1 [Ref [61]] (IFA: 1:500); mouse anti-cMYC mAb 9E10 (ThermoFisher) (WB: 1:300).

2.9 Supplementary Information

Extended Data Table 1: Cryo-EM data collection, refinement and validation statistics

	PTEX *Engaged* Full (EMDB-8951) (PDB 6E10)	PTEX *Engaged* Top (EMDB-8951) (PDB 6E10)	PTEX *Engaged* Bottom (EMDB-8951) (PDB 6E10)	PTEX *Resetting* Full (EMDB-8952) (PDB 6E11)	PTEX *Resetting* Top (EMDB-8952) (PDB 6E11)	PTEX *Resetting* Bottom (EMDB-8952) (PDB 6E11)
Data collection and processing						
Magnification	×105,000	×105,000	×105,000	×105,000	n/a	×105,000
Voltage (kV)	300	300	300	300	n/a	300
Electron exposure (e–/Å²)	60	60	60	60	n/a	60
Defocus range (μm)	-1.5 to -4.0	-1.5 to -4.0	-1.5 to -4.0	-1.5 to -4.0	n/a	-1.5 to -4.0
Pixel size (Å)	1.04	1.04	1.04	1.04	n/a	1.04
Symmetry imposed	C1	C1	C7	C1	n/a	C7
Initial particle images (no.)	1,508,462	1,508,462	1,508,462	1,508,462	n/a	1,508,462
Final particle images (no.)	72,866	53,531	39,437	78,499	n/a	48,425
Map resolution (Å) FSC threshold	4.09 0.143	4.16 0.143	3.5 0.143	4.23 0.143	n/a	3.4 0.143
Map resolution range (Å)	3.2-7.5	3.0-7.0	2.8-3.6	3.2-7.5	n/a	2.8-3.4
Refinement						
Initial model used (PDB code)	n/a	n/a	n/a	n/a	n/a	n/a
Model resolution (Å) FSC threshold	4.58 0.5	4.23 0.5	3.59 0.5	4.84 0.5	n/a	3.67 0.5
Model resolution range (Å)	4.58	4.23	3.59	4.84	n/a	3.67
Map sharpening *B* factor (Å²)	-180	-180	-170	-180	n/a	-160
Model composition Non-hydrogen atoms Protein residues Ligands	 57,352 6,838 12	 57,352 6,838 12	 57,352 6,838 12	 57,401 6,826 12	 57,401 6,826 12	 57,401 6,826 12
B factors (Å²) Protein Ligand	n/a	n/a	n/a	n/a	n/a	n/a
R.m.s. deviations Bond lengths (Å) Bond angles (°)	 0.008 1.311	 0.009 1.262	 0.008 0.925	 0.008 1.300	 0.007 1.332	 0.006 0.895
Validation MolProbity score Clashscore Poor rotamers (%)	 1.95 9.74 1.04	 1.81 7.35 0.13	 1.53 4.01 0.45	 2.03 10.44 0.47	 1.96 7.46 0.05	 1.64 5.48 0.33
Ramachandran plot Favored (%) Allowed (%) Disallowed (%)	 93.96 5.53 0.51	 94.01 5.76 0.23	 94.99 4.69 0.33	 91.91 7.65 0.44	 89.98 9.58 0.44	 94.98 5.08 0.00

2.10 References

1 WHO. World malaria report 2017. Report No. ISBN: 978 92 4 156552 3, 196 (2017).

2 L. H. Miller, H. C. Ackerman, X. Z. Su &T. E. Wellems. Malaria biology and disease pathogenesis: insights for new treatments. *Nat Med* **19**, 156-167, doi:10.1038/nm.3073 (2013).

3 A. F. Cowman, J. Healer, D. Marapana &K. Marsh. Malaria: Biology and Disease. *Cell* **167**, 610-624, doi:10.1016/j.cell.2016.07.055 (2016).

4 N. J. Spillman, J. R. Beck &D. E. Goldberg. Protein Export into Malaria Parasite-Infected Erythrocytes: Mechanisms and Functional Consequences. *Annu Rev Biochem* **84**, 813-841, doi:10.1146/annurev-biochem-060614-034157 (2015).

5 J. A. Boddey &A. F. Cowman. Plasmodium Nesting: Remaking the Erythrocyte from the Inside Out. *Annu Rev Microbiol* **67**, 243-269, doi:10.1146/annurev-micro-092412-155730 (2013).

6 J. M. Przyborski, B. Nyboer &M. Lanzer. Ticket to ride: export of proteins to the Plasmodium falciparum-infected erythrocyte. *Mol Microbiol* **101**, 1-11, doi:10.1111/mmi.13380 (2016).

7 T. F. de Koning-Ward, M. W. Dixon, L. Tilley &P. R. Gilson. Plasmodium species: master renovators of their host cells. *Nat Rev Microbiol* **14**, 494-507, doi:10.1038/nrmicro.2016.79 (2016).

8 M. Marti, R. T. Good, M. Rug, E. Knuepfer &A. F. Cowman. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**, 1930-1933, doi:10.1126/science.1102452 (2004).

9 N. L. Hiller *et al.* A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* **306**, 1934-1937, doi:10.1126/science.1102737 (2004).

10 A. Heiber *et al.* Identification of new PNEPs indicates a substantial non-PEXEL exportome and underpins common features in Plasmodium falciparum protein export. *PLoS Pathog* **9**, e1003546, doi:10.1371/journal.ppat.1003546 (2013).

11 K. Lingelbach &K. A. Joiner. The parasitophorous vacuole membrane surrounding Plasmodium and Toxoplasma: An unusual compartment in infected cells. *J Cell Sci* **111**, 1467-1475 (1998).

12 I. Ansorge, J. Benting, S. Bhakdi &K. Lingelbach. Protein sorting in Plasmodium falciparum-infected red blood cells permeabilized with the pore-forming protein streptolysin O. *Biochem J* **315**, 307-314, doi:DOI 10.1042/bj3150307 (1996).

13 N. Gehde *et al.* Protein unfolding is an essential requirement for transport across the parasitophorous vacuolar membrane of Plasmodium falciparum. *Mol Microbiol* **71**, 613-628, doi:10.1111/j.1365-2958.2008.06552.x (2009).

14 T. F. de Koning-Ward *et al.* A newly discovered protein export machine in malaria parasites. *Nature* **459**, 945-949, doi:10.1038/nature08104 (2009).

15 J. R. Beck, V. Muralidharan, A. Oksman &D. E. Goldberg. PTEX component HSP101 mediates export of diverse malaria effectors into host erythrocytes. *Nature* **511**, 592-595, doi:10.1038/nature13574 (2014).

16 B. Elsworth *et al.* PTEX is an essential nexus for protein export in malaria parasites. *Nature* **511**, 587-+, doi:10.1038/nature13555 (2014).

17 H. E. Bullen *et al.* Biosynthesis, Localization, and Macromolecular Arrangement of the Plasmodium falciparum Translocon of Exported Proteins (PTEX). *J Biol Chem* **287**, 7871-7884, doi:10.1074/jbc.M111.328591 (2012).

18 D. Johnson *et al.* Characterization of Membrane-Proteins Exported from Plasmodium-Falciparum into the Host Erythrocyte. *Parasitology* **109**, 1-9, doi:Doi 10.1017/S0031182000077696 (1994).

19 D. A. Gold *et al.* The Toxoplasma Dense Granule Proteins GRA17 and GRA23 Mediate the Movement of Small Molecules between the Host and the Parasitophorous Vacuole. *Cell Host Microbe* **17**, 642-652, doi:10.1016/j.chom.2015.04.003 (2015).

20 M. N. Garten, A. S.; Niles, J. C.; Zimmerberg, J.; Goldberg, D. E.; Beck, J. R. . EXP2: a dual function channel protein in the malaria parasite vacuolar membrane. *Nature Microbiology*.

21 B. Elsworth *et al.* Proteomic analysis reveals novel proteins associated with the Plasmodium protein exporter PTEX and a loss of complex stability upon truncation of the core PTEX component, PTEX150. *Cell Microbiol* **18**, 1551-1569, doi:10.1111/cmi.12596 (2016).

22 A. Kucukelbir, F. J. Sigworth &H. D. Tagare. Quantifying the local resolution of cryo-EMEM density maps. *Nat Methods* **11**, 63-+, doi:10.1038/Nmeth.2727 (2014).

23 S. H. W. Scheres. A Bayesian View on Cryo-EM Structure Determination. *J Mol Biol* **415**, 406-418, doi:10.1016/j.jmb.2011.11.010 (2012).

24 S. H. W. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519-530, doi:10.1016/j.jsb.2012.09.006 (2012).

25 S. A. Chisholm *et al.* The malaria PTEX component PTEX88 interacts most closely with HSP101 at the host-parasite interface. *FEBS J*, doi:10.1111/febs.14463 (2018).

26 O. S. Smart, J. G. Neduvelil, X. Wang, B. A. Wallace &M. S. P. Sansom. HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *J Mol Graph Model* **14**, 354-&, doi:Doi 10.1016/S0263-7855(97)00009-X (1996).

27 L. Holm &L. M. Laakso. Dali server update. *Nucleic Acids Res* **44**, W351-355, doi:10.1093/nar/gkw357 (2016).

28 J. F. Gibrat, T. Madej &S. H. Bryant. Surprising similarities in structure comparison. *Curr Opin Struct Biol* **6**, 377-385 (1996).

29 M. Wiederstein, M. Gruber, K. Frank, F. Melo &M. J. Sippl. Structure-based characterization of multiprotein complexes. *Structure* **22**, 1063-1070, doi:10.1016/j.str.2014.05.005 (2014).

30 E. Krissinel &K. Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D* **60**, 2256-2268, doi:10.1107/S0907444904026460 (2004).

31 Z. Dosztanyi, V. Csizmok, P. Tompa &I. Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433-3434, doi:10.1093/bioinformatics/bti541 (2005).

32 Z. Dosztanyi, V. Csizmok, P. Tompa &I. Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**, 827-839, doi:10.1016/j.jmb.2005.01.071 (2005).

33 E. C. Schirmer, J. R. Glover, M. A. Singer &S. Lindquist. HSP100/Clp proteins: A common mechanism explains diverse functions. *Trends Biochem Sci* **21**, 289-296, doi:Doi 10.1016/S0968-0004(96)10038-4 (1996).

34 P. I. Hanson &S. W. Whiteheart. AAA+ proteins: have engine, will work. *Nat Rev Mol Cell Biol* **6**, 519-529, doi:10.1038/nrm1684 (2005).

35 C. Deville *et al.* Structural pathway of regulated substrate transfer and threading through an Hsp100 disaggregase. *Sci Adv* **3**, e1701726, doi:10.1126/sciadv.1701726 (2017).

36 S. N. Gates *et al.* Ratchet-like polypeptide translocation mechanism of the AAA+ disaggregase Hsp104. *Science* **357**, 273-279, doi:10.1126/science.aan1052 (2017).

37 A. P. AhYoung, A. Koehl, D. Cascio &P. F. Egea. Structural mapping of the ClpB ATPases of Plasmodium falciparum: Targeting protein folding and secretion for antimalarial drug design. *Protein Sci* **24**, 1508-1520, doi:10.1002/pro.2739 (2015).

38 C. Puchades *et al.* Structure of the mitochondrial inner membrane AAA+ protease YME1 gives insight into substrate processing. *Science* **358**, doi:10.1126/science.aao0464 (2017).

39 A. L. Yokom *et al.* Spiral architecture of the Hsp104 disaggregase reveals the basis for polypeptide translocation. *Nature Structural & Molecular Biology* **23**, 830-837, doi:10.1038/nsmb.3277 (2016).

40 C. Puchades *et al.* Structure of the mitochondrial inner membrane AAA plus protease YME1 gives insight into substrate processing. *Science* **358**, 609-+, doi:10.1126/science.aao0464 (2017).

41 F. Seyffer *et al.* Hsp70 proteins bind Hsp100 regulatory M domains to activate AAA plus disaggregase at aggregate surfaces. *Nature Structural & Molecular Biology* **19**, 1347-+, doi:10.1038/nsmb.2442 (2012).

42 N. Lipinska *et al.* Disruption of Ionic Interactions between the Nucleotide Binding Domain 1 (NBD1) and Middle (M) Domain in Hsp100 Disaggregase Unleashes Toxic Hyperactivity and Partial Independence from Hsp70. *J Biol Chem* **288**, 2857-2869, doi:10.1074/jbc.M112.387589 (2013).

43 S. Kulzer *et al.* Plasmodium falciparum-encoded exported hsp70/hsp40 chaperone/co-chaperone complexes within the host erythrocyte. *Cell Microbiol* **14**, 1784-1795, doi:10.1111/j.1462-5822.2012.01840.x (2012).

44 S. Batinovic *et al.* An exported protein-interacting complex involved in the trafficking of virulence determinants in Plasmodium-infected erythrocytes. *Nat Commun* **8**, 16044, doi:10.1038/ncomms16044 (2017).

45 S. H. Adjalley *et al.* Quantitative assessment of Plasmodium falciparum sexual development reveals potent transmission-blocking activity by methylene blue. *P Natl Acad Sci USA* **108**, E1214-E1223, doi:10.1073/pnas.1112037108 (2011).

46 M. Klemba, W. Beatty, I. Gluzman &D. E. Goldberg. Trafficking of plasmepsin II to the food vacuole of the malaria parasite Plasmodium falciparum (vol 164, pg 47, 2004). *Journal of Cell Biology* **164**, 625-625, doi:DOI 10.1083/jcb.2004021616447 (2004).

47 N. J. Spillman, J. R. Beck, S. M. Ganesan, J. C. Niles &D. E. Goldberg. The chaperonin TRiC forms an oligomeric complex in the malaria parasite cytosol. *Cell Microbiol* **19**, doi:ARTN e12719 10.1111/cmi.12719 (2017).

48 S. Q. Zheng *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* **14**, 331-332, doi:10.1038/nmeth.4193 (2017).

49 A. Rohou &N. Grigorieff. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol* **192**, 216-221, doi:10.1016/j.jsb.2015.08.008 (2015).

50 K. Zhang. *Gautomatch: a GPU-accelerated program for accurate, fast, flexible and fully automatic particle picking from cryo-EM micrographs with or without templates* (2016).

51 A. Punjani, J. L. Rubinstein, D. J. Fleet &M. A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-+, doi:10.1038/Nmeth.4169 (2017).

52 E. F. Pettersen *et al.* UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).

53 P. Emsley, B. Lohkamp, W. G. Scott &K. Cowtan. Features and development of Coot. *Acta Crystallogr D* **66**, 486-501, doi:10.1107/S0907444910007493 (2010).

54 N. R. Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**, D7-D19, doi:10.1093/nar/gkv1290 (2016).

55 C. Aurrecoechea *et al.* PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* **37**, D539-D543, doi:10.1093/nar/gkn814 (2009).

56 L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass &M. J. E. Sternberg. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858, doi:10.1038/nprot.2015.053 (2015).

57 P. D. Adams *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D* **66**, 213-221, doi:10.1107/S0907444909052925 (2010).

58 V. B. Chen *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D* **66**, 12-21, doi:10.1107/S0907444909042073 (2010).

59 Schrodinger, LLC. *The PyMOL Molecular Graphics System, Version 1.8* (2015).

60 S. M. Ganesan *et al.* Yeast dihydroorotate dehydrogenase as a new selectable marker for Plasmodium falciparum transfection. *Mol Biochem Parasit* **177**, 29-34, doi:10.1016/j.molbiopara.2011.01.004 (2011).

61 T. Blisnick *et al.* Pfsbp 1, a Maurer's cleft Plasmodium falciparum protein, is associated with the erythrocyte skeleton. *Mol Biochem Parasit* **111**, 107-121, doi:Doi 10.1016/S0166-6851(00)00301-7 (2000).

Chapter 3

Structural Proteomics of the Malaria Parasite Plasmodium falciparum

Chi-Min Ho[1,2,3], Xiaorun Li[3,4], Mason Lai[2,3], Thomas C. Terwilliger[5], Josh R. Beck[6,7], James Wohlschlegel[8], Daniel E. Goldberg[6], Anthony W. P. Fitzpatrick[9], Z. Hong Zhou[1,2,3]*

[1]The Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA

[2]Department of Microbiology, Immunology, & Molecular Genetics, University of California, Los Angeles, CA 90095, USA

[3]California NanoSystems Institute, University of California, Los Angeles, CA 90095, USA

[4]Hefei National Laboratory for Physical Sciences at Microscale, University of Science and Technology of China, Hefei, Anhui 230026, China

[5]Los Alamos National Laboratory and the New Mexico Consortium, Los Alamos, NM 87544, USA

[6]Departments of Medicine and Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110, USA

[7]Department of Biomedical Sciences, Iowa State University, Ames, IA 50011, USA

[8]Department of Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA

[9]Zuckerman Institute, Columbia Medical School, New York, NY, USA

* Correspondence and request for materials should be addressed to Z.H.Z. (Hong.Zhou@UCLA.edu).

3.1 Abstract

X-ray crystallography and recombinant protein production have enabled an exponential increase in atomic structures, but often require non-native constructs involving mutations or truncations, and are challenged by membrane proteins and large multi-component complexes. We present here a “bottom-up” endogenous structural proteomics approach whereby near-atomic resolution cryoEM maps are reconstructed ab initio from unidentified protein complexes enriched directly from the endogenous cellular milieu, followed by identification and atomic modeling of the proteins. The proteins in each complex are identified using cryoID, a program we developed to identify proteins in ab initio cryoEM maps from the sequences in the proteome. As a proof of principle, we applied this approach to the malaria parasite Plasmodium falciparum, an organism that has resisted traditional structural biology approaches, to obtain multiple atomic models of protein complexes implicated in the life cycle of the parasite. Our approach opens the door to atomic structures of previously intractable biological systems.

3.2 Introduction

Since the emergence of recombinant DNA techniques and generic peptide purification tags in the late 1980s, the standard approach in structural biology has been to purify tagged proteins overexpressed in heterologous systems such as *E. coli* or *S. cerevisiae*[1-8]. The ability to produce large quantities of pure protein from recombinant systems has been instrumental to the exponential increase in high resolution structures solved over the past three decades, the vast majority of which were obtained using X-ray crystallography and nuclear magnetic resonance (NMR)[9]. These structures have transformed much of our understanding of the molecular mechanisms underlying many fundamental processes in the cell. However, both techniques require large quantities of pure protein at relatively high concentrations. As such, structural studies were biased toward proteins that were amenable to expression in recombinant systems and either capable of forming well-ordered crystals (for X-ray crystallography) or small enough for NMR. In addition, mutations or truncations were often required to produce proteins that satisfied these requirements, leading to uncertainty about how closely the resulting structures reflect their true, biologically relevant states *in vivo*. Furthermore, challenging systems such as membrane proteins and large multi-component complexes (particularly those involving nucleic acids, lipids, or transiently stable intermediate states and interactions) are under-represented in the current body of solved structures, as they are recalcitrant to expression in recombinant systems, often precluding structure determination by X-ray crystallography or NMR.

However, the recent "resolution revolution" in cryoelectron microscopy (cryoEM) has opened the door for high resolution structure determination of a vast number of previously intractable biological systems[10-19]. There is no need for crystallization, as samples for cryoEM are preserved in a frozen-hydrated state, randomly oriented within the layer of vitreous ice. Without

the need to introduce mutations or truncations that provide better crystal contacts, it is possible to observe proteins in native or near-native, biologically relevant states. Moreover, with a dramatically reduced requirement for both quantity and homogeneity of samples for cryoEM, we are no longer restricted to systems that can be produced in large quantities at high purity. In fact, cryoEM has the added advantage that it is possible to achieve multiple high resolution structures of several different conformational states of a single protein complex[20], or even several structures of completely unrelated protein complexes in the same sample, from a single cryoEM dataset.

By leveraging the latest cutting-edge innovations in cryoEM, it is now possible to accommodate the low yields and heterogeneity of samples enriched directly from endogenous sources. This approach introduces an intriguing challenge. If we obtain a near-atomic (3.0-4.0 Å) resolution cryoEM map of a molecule or complex from a heterogeneous sample, how do we identify the molecule or complex this map represents? In the past, proteins have occasionally been crystallized accidentally, requiring a close examination of side chain density along the path of the main chain of the molecule to guess the sequence and identify what molecule was crystallized[21]. In other cases, sequence errors have been identified from crystallographic density maps[22]. However, identifying molecules or complexes in cryoEM maps is challenging, due to the lower overall resolutions (3.0-4.0Å) and varying local resolutions typical of current/routinely achievable cryoEM maps.

There are no existing programs capable of autobuilding atomic models from cryoEM density maps without primary sequence information. Even with primary sequence information, currently existing autobuilding programs cannot match the performance of an experienced human modeler, especially at resolutions worse than 3.5Å. An autobuilding program capable of handling typical cryoEM maps ranging from 2.5-4.0Å in resolution, given the primary sequence, would

represent a significant step forward. An autobuilding program capable of doing so without the primary sequence would be transformative for the field, bringing cryoEM closer to the level of accessibility and automation that is the standard in X-Ray crystallography.

To address this challenge, we have developed a targeted "bottom-up" endogenous structural proteomics approach whereby protein complexes are enriched directly from the cellular milieu and identified by imaging and structure determination using mass spectrometry and near-atomic resolution cryoEM density maps reconstructed ab initio (**Fig. 3.1**). This workflow employs our program, cryoID, to semi-autonomously identify proteins in 3.0Å – 4.0Å resolution cryoEM maps without any prior knowledge of the sequence(s). As a proof of principle, we have applied this approach to P. falciparum, an organism that has proven recalcitrant to traditional structural biology approaches. By directly imaging components of the parasite cell lysate, we obtained near-atomic resolution structures of multiple protein complexes implicated in the pathogenesis of malarial parasites, from a single cryoEM dataset. We then used cryoID to unambiguously assign sidechains and identify the complex, enabling atomic model building.

3.3 Results

3.3.1 Workflow

Our workflow consists of the following five steps, starting from raw cell lysates and potentially yielding atomic models of many native macromolecular complexes:

Step 1: Endogenous purification. We use sucrose gradient fractionation to enrich protein complexes from raw cell lysates from endogenous sources (**Fig. 3.1a**).

Step 2: Sample evaluation. We then assess the complexity of each fraction by SDS-PAGE and negative stain EM (**Fig. 3.1b-c**).

Step 3: Mass spectrometry. Promising fractions containing uniform particles in negative stain EM are analyzed by tryptic digest liquid chromatography-mass spectrometry (LC-MS), yielding a pool of all proteins present in each fraction, usually ~1000-2000 (**Fig. 3.1d**).

Step 4: cryoEM imaging. High resolution images of each promising fraction are collected on a Titan Krios cryo-electron microscope, generally yielding datasets containing several distinct protein complexes in each image (**Fig. 3.1e**). To deconvolute mixtures of several distinct protein complexes within a single dataset and resolve them into multiple three dimensional (3D) structures, we take advantage of cryoSPARC's[17] ability to perform unsupervised *ab initio* 3D classification and refinement, given a mixture of particles from multiple distinct protein complexes.

Step 5: Protein identification and modeling. We have developed a semi-automated program, *cryoID*, which is used to identify the protein(s) in each cryoEM map obtained in Step 4, using only the cryoEM density, from the pool of potential candidates detected in the sample by mass spectrometry in Step 3 (**Fig. 3.1f)**. As some amino acid sidechains can look quite similar in cryoEM maps at 3.0-4.0Å resolution, we incorporated a certain degree of error tolerance into *cryoID via* the use of a simplified, "degenerate" six-letter code that clusters the 20 amino acid residues into 6 simplified groups, based on the similarity of their side-chain densities in typical cryoEM density maps.

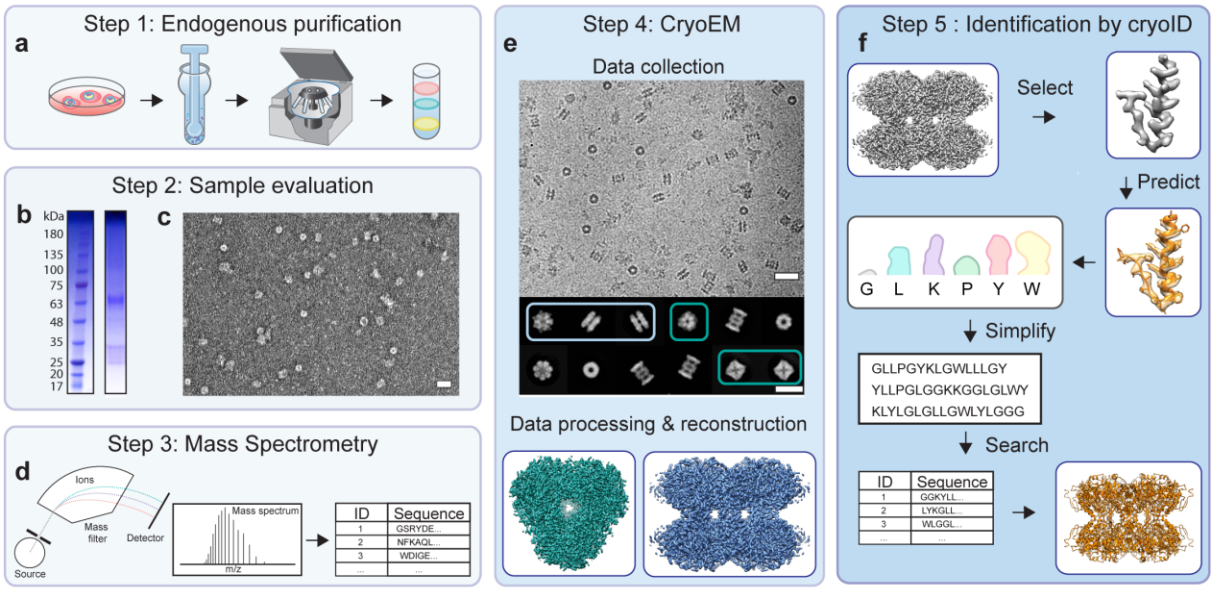


Figure 3.1 | Endogenous structural proteomics workflow. a, Protein complexes are enriched by sucrose gradient fractionation. **b-c,** Fractions are evaluated by SDS-PAGE (**b**) and negative stain electron microscopy (**c**). **d,** Mass spectrometry is identifies a list of all proteins in each fraction. **e,** cryoEM analysis yields near-atomic resolution cryoEM maps. **f,** The proteins in the cryoEM maps are identified using *cryoID*.

3.3.2 *cryoID*

cryoID, a key component of the above workflow, determines the unique identity of the protein(s) in a near-atomic resolution (3.0-4.0Å) cryoEM density map from a pool of candidates (either full proteome(s) from Uniprot, or a list of possible proteins identified by mass spectrometry), using only the information contained within the cryoEM density map.

There are two main challenges in *de novo* modeling into cryoEM maps at 3.0-4.0Å resolution, for both human modelers and automated modeling programs: 1) distinguishability of sidechain densities varies depending on map quality and local resolution(s), which can fluctuate widely across a given cryoEM map, 2) small and medium size residues can be difficult to distinguish accurately even in “promising” regions of the map, as some residues can look quite

similar. Aspartate and asparagine, for example, which have sidechains of a similar size and shape, can often be difficult to distinguish without prior knowledge of the primary sequence.

To overcome these challenges, experienced human modelers start by locating a high resolution region of the map with a number of distinctive, bulky residues, known as markers, interspersed with smaller residues. In typical cryoEM density maps, phenylalanine (F), tyrosine (Y), tryptophan (W), histidine (H), arginine (R), lysine (K) and proline (P) generally have distinctive, easier to recognize 3D shapes, and are commonly used as markers. Conversely, smaller and medium size residues are often harder to distinguish in typical cryoEM density maps, as their shapes are often less distinctive and can be context dependent. In order to uniquely identify the section of the primary sequence corresponding to a given segment of a map, human modelers rely on the pattern of markers (the markers and the spacing between markers) as well as contextual clues, taking into account the quality of the map in the surrounding area, as well as prior knowledge of protein sequence "rules". Once a promising region has been identified, experienced human modelers will then look through the primary sequence for a continuous string of residues that would satisfy the pattern in the map. When going through the primary sequence looking for a match, they weight correct positioning of markers more heavily than placement of smaller and medium size residues. (*i.e.*, incorrect positioning of markers is penalized more heavily). Once the portion of the primary sequence corresponding to a segment of the map has been identified, human modelers have a known starting point and can then confidently proceed with assigning and modeling the remaining sidechains throughout the rest of the map.

Given the variability in resolution and quality of the density across most cryoEM maps, modeling into the entire map without prior knowledge of the primary sequence would be extremely challenging. However, most maps with an overall resolution of 3.0-4.0Å have at least a few regions

of higher resolution, often in the core of the protein complex, where dense protein packing interactions keep everything locked down and relatively stable. Given a map of an unknown protein at 3.0-4.0Å resolution, an experienced human modeler would most likely build into these higher resolution regions to obtain several short sequences, which can then be used as queries in BlastP to identify the protein and obtain the primary sequence.

We designed *cryoID* to emulate these strategies used by experienced human modelers. cryoID accomplishes this by performing four main tasks (**Fig. 3.2).** ***1-Selection:*** Identifying one or more high resolution segments of the map with a continuous backbone and clearly distinguishable sidechain densities. ***2- Prediction:*** Automatically tracing the polypeptide backbone for each map segment and semi-automatically predicting the identities of the sidechains for each residue in the segment, yielding a predicted primary sequence for the segment. ***3-Simplification:*** Translating both the cryoEM map segment sequences and all the primary sequences of the pool of candidate proteins into a simplified "6-letter" code. ***4-Searching:*** Performing a customized BLASTP[23] search of the entire pool of candidate proteins using the predicted cryoEM map segment sequences as queries.

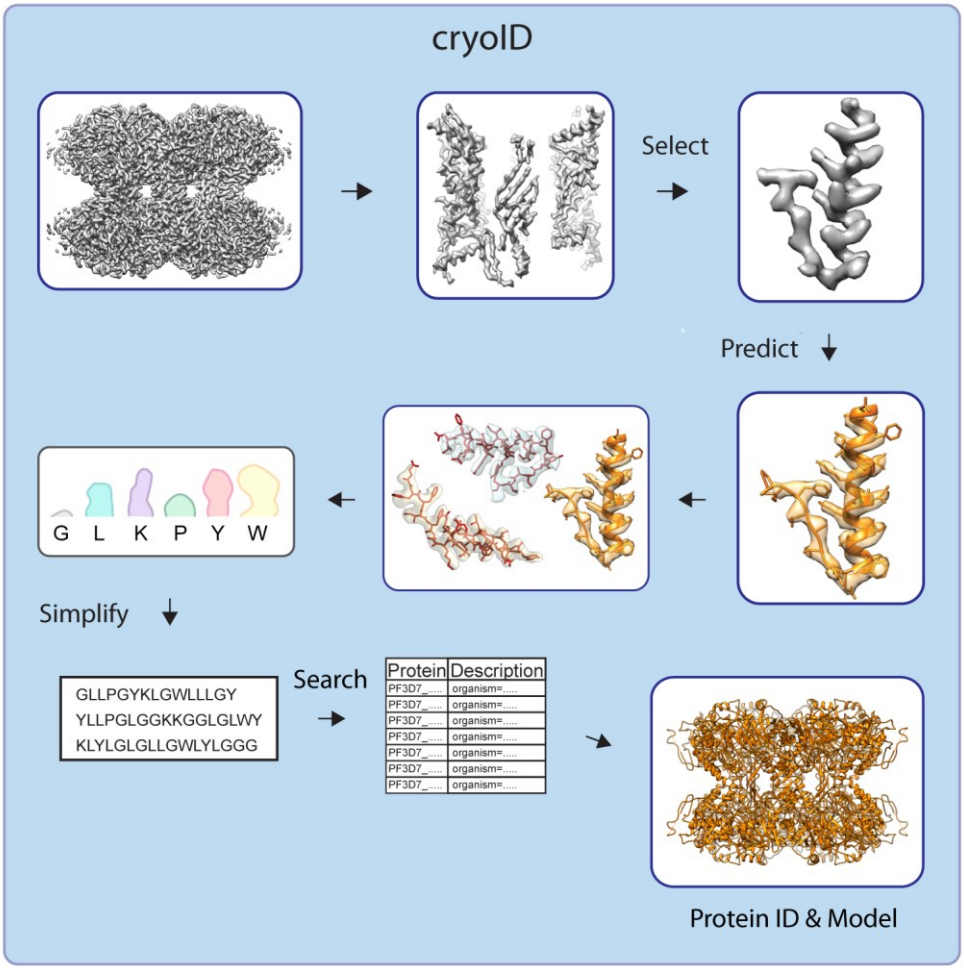


Figure 3.2 | Graphical overview of *cryoID*. The four main functions of *cryoID* are shown: Selection, Prediction, Simplification, and Searching.

3.3.3 Selection. First, to make things as easy as possible, *cryoID* locates the highest resolution regions of the map, with a preference for regions containing a number of markers. Doing this allows *cryoID* to build only into the most promising segments of a given cryoEM map, simplifying the problem. This strategy allows *cryoID* to overcome the first key challenge of *de novo* modeling into 3.0Å-4.0Å resolution cryoEM maps: variability in map quality and resolution.

3.3.4 Prediction. Once a few good regions have been identified, cryoID predicts the identity of each residue in each segment using segment_from_map, a new Phenix subfunction we developed. This tool builds a peptide backbone into the segment and then predicts the identity of each individual side chain in the segment, based on the density. In this way, cryoID is capable of predicting the sequence of short, high resolution segments of a cryoEM map. At this point, cryoID provides the user with the option to further improve the sequence prediction by manually inspecting the modeled sequence and correcting any obvious errors via the cryoID GUI (**Fig. 3.3**).

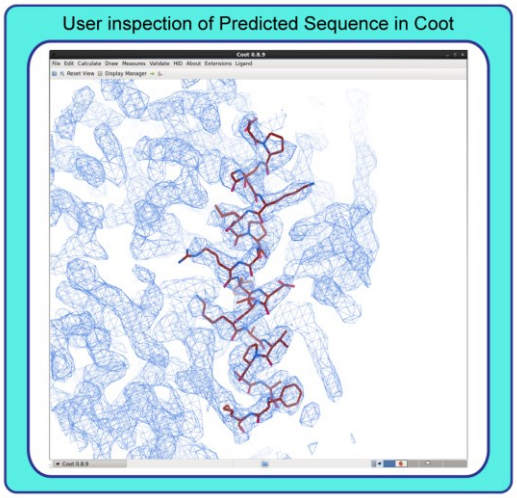


Figure 3.3 | Manual inspection in Coot *via* the *cryoID* GUI.

3.3.5 Simplification. Errors introduced at this stage are accommodated by using a simplified, "degenerate" six-letter code that groups the 20 amino acid residues into 6 simplified groups, based on the similarity of their side-chain densities in typical cryoEM density maps (**Fig. 3.4**).

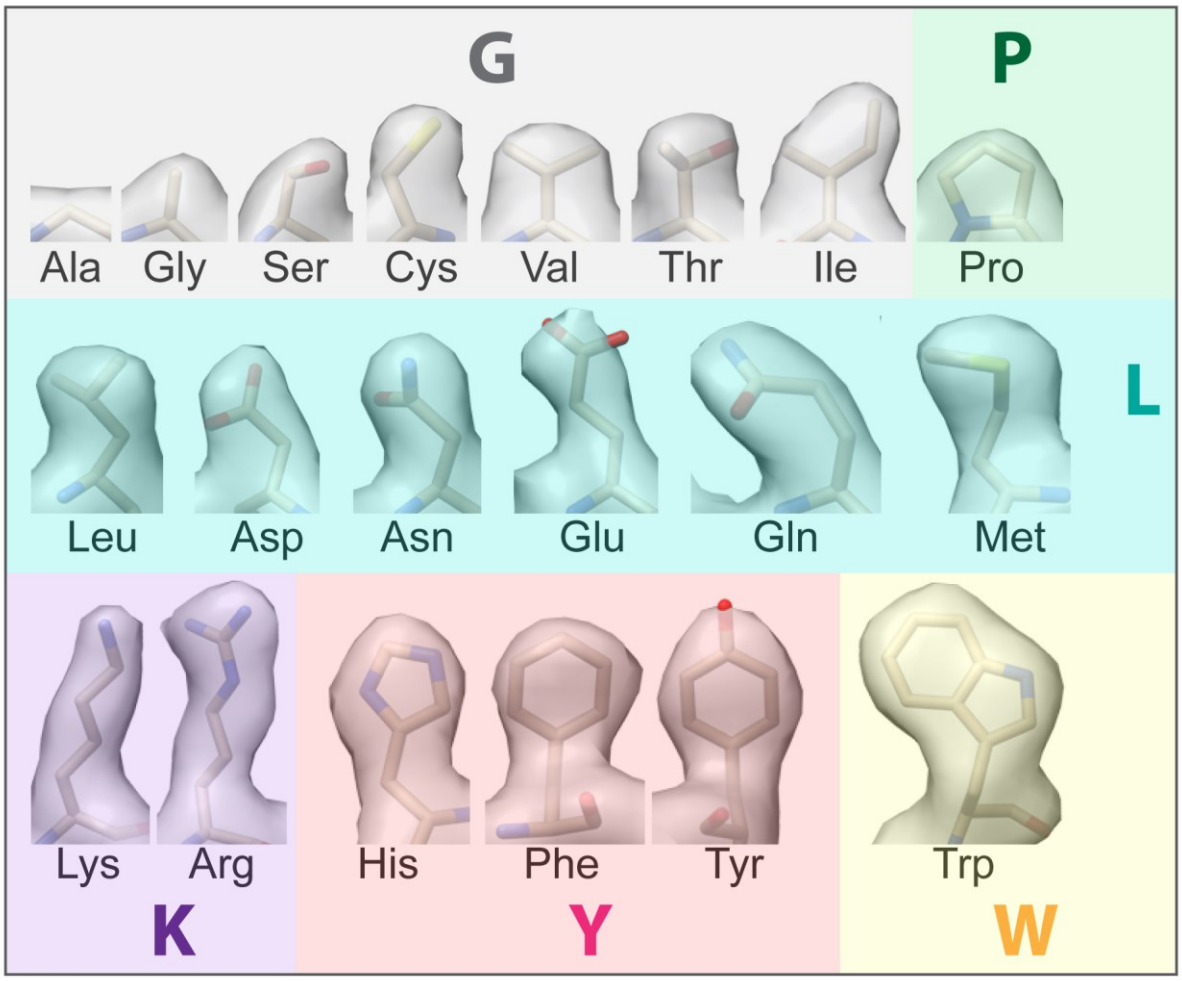


Figure 3.4 | Simplified 6-Letter Code. The 20 amino acid residues are clustered into 6 simplified groups, based on the similarity of their side-chain densities in typical cryoEM density maps. One residue from each group is chosen as the representative of the entire group (denoted by the large colored single letter label in each group shown here). These representatives for each group are used during subsequent searching operations in *cryoID*.

For example, the segment of the map shown in **Fig. 3.5**, DKKAREYANDALKF, is translated into the following simplified sequence: LKKGKLYGLLGLKY. By using the simplified 6-letter code, we introduce a redundancy that imparts a certain amount of tolerance for errors made by *cryoID* during the sequence prediction step.

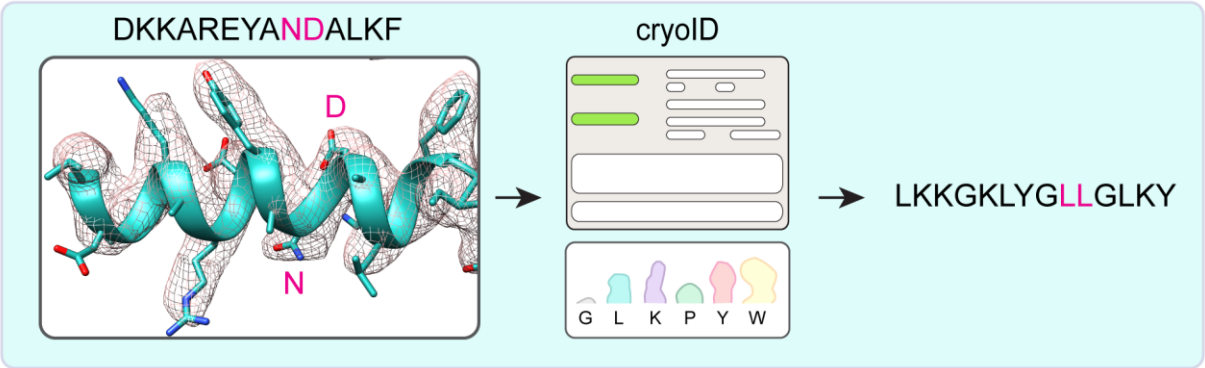


Figure 3.5 | Simplification into the 6-Letter Code. *cryoID* predicts the identity of each residue in the density on the left, and then simplifies the resulting sequence of the entire segment into the degenereate 6-letter code shown on the right.

In the example shown in **Fig. 3.5**, the densities for N and D look very similar, however, there is no need for *cryoID* to be able to differentiate between them, since they both fall within the "L" group and will both be translated into the same letter in the simplified code. In fact, *cryoID* doesn't need to be able to distinguish between any of the members within a given group – as long as it can differentiate between members of different groups. This means that *cryoID* can tolerate errors made during the sequence prediction process provided the incorrect residue still falls within the same group as the correct residue. By using this simplified code, we eliminate the need for *cryoID* to differentiate between small and medium sized side0chains which are often difficult to distinguish in typical cryoEM density maps. The use of the simplified 6-letter code thereby allows *cryoID* to overcome the second key challenge of *de novo* modeling into 3.0Å-4.0Å resolution cryoEM maps: difficulty in accurately distinguishing between small and medium size residues, even in "promising" regions of the map.

Once the primary sequence for the entire protein is simplified as well, we found that this simplified sequence still only occurs once in the entire protein. We then demonstrated, in a series of benchmarking experiments described below, that it is possible to uniquely identify the protein

in a cryoEM map by using multiple such simplified sequences from a single cryoEM map to search a large (100s to 100,000s) pool of potential candidates.

3.3.6 Searching. Searching by *cryoID* is a fully automated process. After generating several simplified query sequences from the cryoEM density map, *cryoID* simplifies the pool of candidate proteins and searches for a protein containing segments that match all of the query sequences (**Fig. 3.6**). Using a customized BLASTP algorithm (**Fig. 3.6b**), *cryoID* aligns each query sequence with the simplified sequence of each candidate protein in the pool, determining an expectation, or E value, which indicates the similarity between the query sequence and the candidate protein. The E values for each query against a single candidate protein are then combined to yield a composite E value, which indicates the likelihood of the candidate protein being the correct protein. *cryoID* then ranks the candidates by their composite E values, with lower E values representing higher likelihood of being the correct protein.

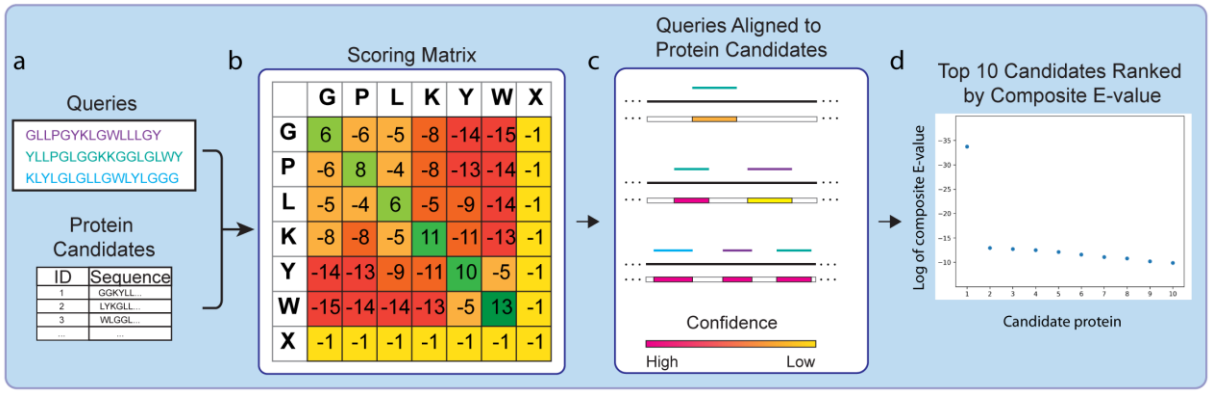


Figure 3.6 | Searching in *cryoID*. *cryoID* runs alignments of the simplified query sequences obtained from the cryoEM maps against each protein in the pool of candidate proteins (also simplified), which were previously identified by mass spectrometry. **b,** The alignment scoring matrix used by *cryoID* is a modified version of the standard BLASTP scoring matrix, which has been modified for use with sequences that have been translated into the simplified 6-letter code used in *cryoID*. **c, d,** *cryoID* uses the composite E-value of each protein candidate against the query set to make a unique proteinID.

3.3.7 Benchmarking *cryoID* using simulated data

Parameters such as the number of queries, length of each query, and the number of errors in each query influence the ability of *cryoID* to arrive at a single unique answer. We first determined the optimal range for each parameter, using simulated datasets.

To determine the minimum number (m) of query sequences of a given length (n), from a single protein, that are required for *cryoID* to arrive at a unique answer, we varied m from 1-10 query sequences, and n from 8-100 residues, as illustrated in **Table 3.1**. To generate each query set, we randomly selected m number of sequences, each containing n number of residues, from one full length protein sequence randomly selected out of a pool of 800 *P. falciparum* proteins identified by mass spectrometry in a sample generated from *P. falciparum* parasite lysate using the workflow illustrated in **Fig. 3.1**. In order to cover a wide range of different sequences with different amino acid compositions and achieve a statistical significance of α=0.01, we tested each condition (m,n) 1000 times with 1000 different sets of queries. We derived the query sets from the full length sequences of 5 different proteins from the pool to ensure efficient random query generation. We then used *cryoID* to simplify and blast each set of queries against the full pool of 800 *P. falciparum* proteins, also simplified. For reach run, *cryoID* sorts protein candidates in the pool by the composite E value and monitors their % identity with the queries.

Once the efficacy of all the combinations of m and n were tabulated (**Table 3.1**), we used our simulated data to empirically define the optimal query conditions (m,n) under which *cryoID* will correctly identify the single unique protein matching the query. For some query conditions (m,n), the query sequences were not long or numerous enough to obtain a single unique protein ID. For query sets falling under these suboptimal query conditions, *cryoID* identified multiple protein candidates that exhibit 100% identity with the queries. We defined optimal query

conditions (*m,n*) as those for which *cryoID* consistently identified only a single protein that exhibited 100% identity with the queries.

Number of queries (*m*)	Query sequence length (*n*)														
	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	√
2	×	×	×	×	×	×	×	√	√	√	√	√	√	√	√
3	×	×	×	√	√	√	√	√	√	√	√	√	√	√	√
4	×	×	√	√	√	√	√	√	√	√	√	√	√	√	√
5	×	×	√	√	√	√	√	√	√	√	√	√	√	√	√
6	×	√	√	√	√	√	√	√	√	√	√	√	√	√	√
7	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
8	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
9	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
10	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

Table 3.1 | Determining optimal parameters for Searching in *cryoID*. The result for each query condition (*m,n*) shown here represents 1000 tests of that condition with 1000 different sets of randomly generated queries drawn from a candidate pool of 800 *P. falciparum* proteins. Conditions under which *cryoID* was consistently able to identify the correct protein are marked with green checks. Condition under which *cryoID* was unable to arrive at a single unique protein ID are marked with red X's. P(wrong) < 0.001, N = 1000, $\alpha = 0.01$.

For *cryoID* runs under optimal query conditions (*m,n*), we observed that there is always a clear "gap" in % identity between the correct protein candidate (100% identity with queries) and the next closest matching candidate. This gap increases as *m* or *n* increase. This observation agrees with the theoretical estimation that given a query set (*m,n*) and the candidate pool, the possibility of achieving high identity by chance is proportional to ~(1/6)^*n***m* and decreases exponentially. As *m* or *n* increases, the information contained in the query set increases, and the right candidate becomes easier to distinguish (the increasing gap observed) from incorrect candidates matching by chance. In other words, as the length and number of queries increase, the likelihood of a perfect alignment of the query sequences with an incorrect candidate decreases. As such, we reasoned that it would still be possible for *cryoID* to determine a unique protein ID if the queries contain a limited

number of errors, as long as the number of errors remains less than the number of differences between the next closest matching candidate and the error-free queries.

By running each condition (*m,n*) 1000 times (to achieve α=0.01) with 1000 different sets of queries of number *m* and length *n*, we empirically determined the minimum number of differences that can (probabilistically) occur between the correct protein and the next closest match in the pool, thus defining the maximum number of errors in a query set of (*m,n*) that *cryoID* can tolerate. We then used this information to predict the optimal conditions (m,n) for *cryoID*, given an incidence of errors in the queries of 10%, 20%, 30%, and 40%.

These results are displayed in **Table 3.2**, which may serve as a guide for determining the optimal number and length of queries users should aim for. The table describes the minimum average query length required for *cryoID* to identify the correct protein using a given number of queries (*m*). The maximum number of errors per 10 residues that can be tolerated for each combination of *m* and *n* is also indicated. The table is based on our protein candidate pool, which contains 880 proteins and about 750,000 amino acid residues, and may need to be adjusted for candidate pools that are significantly larger.

Number of queries	Query Sequence Length			
1	30	45	75	>100
2	20	30	55	100
3	16	24	32	75
4	15	20	30	65
5	13	19	28	55
6	12	18	27	50
7	12	16	25	45
8	11	16	24	40
9	11	15	22	40
10	11	15	21	40
Tolerable percentage errors (ungapped)	**10%**	**20%**	**30%**	**40%**

Table 3.2 | Minimum query length for correct protein ID by *cryoID*. The maximum number of errors per 10 residues tolerated for each query condition (*m,n*) is indicated.

3.3.8 Validation of *cryoID* using published cryoEM maps from the EMDB

Having determined the optimal range for each parameter, using simulated datasets, we then tested *cryoID* against published experimental cryoEM maps available from the EMDB, using our workflow detailed in **Fig. 1**.

We selected two published structures within our target resolution range (3.0-4.0Å) from the EMDB, a 3.4Å structure of human gamma-secretase[24] (EMD-3061), and a 3.55Å structure of the *Drosophila* NOMPC mechanotransduction channel[25] (EMD-8702). Gamma-secretase is a four-membered intramembrane protease consisting of presenilin, PEN-2, nicastrin and APH-1. NOMPC is a homotetrameric integral membrane protein. We analyzed each map in *cryoID*,

generating a single query set for NOMPC, and two separate query sets for gamma-secretase (one for the cytosolic region, and one for the transmembrane region). *cryoID* successfully identified the correct protein in the NOMPC map both from a more limited candidate pool consisting of the 3,500 proteins in the *D. melanogaster* proteome, and from a much larger candidate pool consisting of the entire ~600,000 proteins in the UniProt database (Fig. x), achieving a single unique protein ID in both instances. *cryoID* also successfully identified two of the four proteins in the human gamma-secretase map from a candidate pool consisting of the 20,397 proteins in the *H. sapiens* proteome, achieving a single unique protein ID for both query sets. Against the entire ~600,000 proteins in the UniProt database, *cryoID* successfully identified the correct protein for one of the query sets (single unique protein ID), but was not able to achieve a single unique protein ID for the second query set (correct answer was ranked 23rd out of ~600,000).

Using published cryoEM maps from the EMDB, we determined that *cryoID* can consistently determine correct protein(s) in cryoEM maps from candidate pools of up to ~24,000 proteins, and in some cases from a much large candidate pool consisting of the entire ~600,000 proteins in the UniProt database.

	Gamma-secretase transmembrane region: APH-1			
	Length	Number of X	Errors	E-value
GGKLXGYGGGLGYXGGXGGYGGL	23	3	1	9.3e-12
PYYYLGGGYGGGXLGLLYGYKGG GY XGGGGK	31	2	5	5.5e-14
LPXYGGGGLGGYGGGGXLGXWGY G	24	3	3	4.4e-09
Compounded E-value	1.1e-32			
Ranking out of human proteome (20.39)	1st			
Ranking out of UniProt Database (~600,7000)	1st			

Table 3.3 | Human gamma-secretase *cryoID* results query set 1.

	Gamma-secretase excellular region: Nicastrin			
	Length	Number of X	Errors	E-value
GGXGPLGGYLGWGXG	15	2	3	3.4e-02
LLYYGGGGPPGGXGGKGGXYGL	21	2	2	3.2e-07
GGGKGGXLGGGGGLXKGP	18	2	2	1.0e-3
Compounded E-value	4.4e-11			
Ranking out of human proteome (20.39)	1st			
Ranking out of UniProt Database (~600,000)	23nd			

Table 3.4 | Human gamma-secretase *cryoID* results query set 2.

	NOMPC transmembrane region: NOMPC			
	Length	Number of X	Errors	E-value
WGGXLYLGGYGGYLLGGGGGGGGLLGGLKGGGYXKG	36	2	6	1.3e-17
LLXGGGKYLGGXGGYGLGYG	20	2	3	2.4e-06
GGXYGGXGGGYGYGLGXGGG	20	3	4	2.0e-03
Compounded E-value	1.1e-26			
Ranking out of *D. melanogaster* proteome (3,500)	1st			
Ranking out of UniProt Database (~600,000)	1st			

Table 3.5 | *Drosophila* NOMPC *cryoID* results.

3.3.9 Application of the endogenous structural proteomics workflow to *P. falciparum*

We then used the challenging organism *P. falciparum* to further test the ability of our entire workflow to yield near-atomic resolution structures of protein complexes enriched directly from endogenous sources. Many pathogens of high medical relevance are recalcitrant to structural characterization using traditional recombinant approaches. This is particularly so in the case of *P. falciparum*, where the paucity of high resolution structural and functional information is compounded by the fact that **50% of the *P. falciparum* proteome is novel**, bearing no known similarity to existing structures in the PDB[26-28]. Many of the most promising *P. falciparum* drug targets are membrane proteins, but there are only two unique integral membrane protein structures from *P. falciparum* in the PDB.

We enriched for protein complexes ranging from 0.1-2.0+ MDa from *P. falciparum* NF54 parasite lysate using sucrose gradient fractionation. Analysis of a single cryoEM dataset collected from the fraction that looked the most promising by SDS-PAGE and negative stain EM yielded

multiple near-atomic resolution cryoEM density maps at an overall resolution of 3.3Å (**Fig. 3.7**). Mass spectrometry identified a candidate pool of 800 proteins in the fraction.

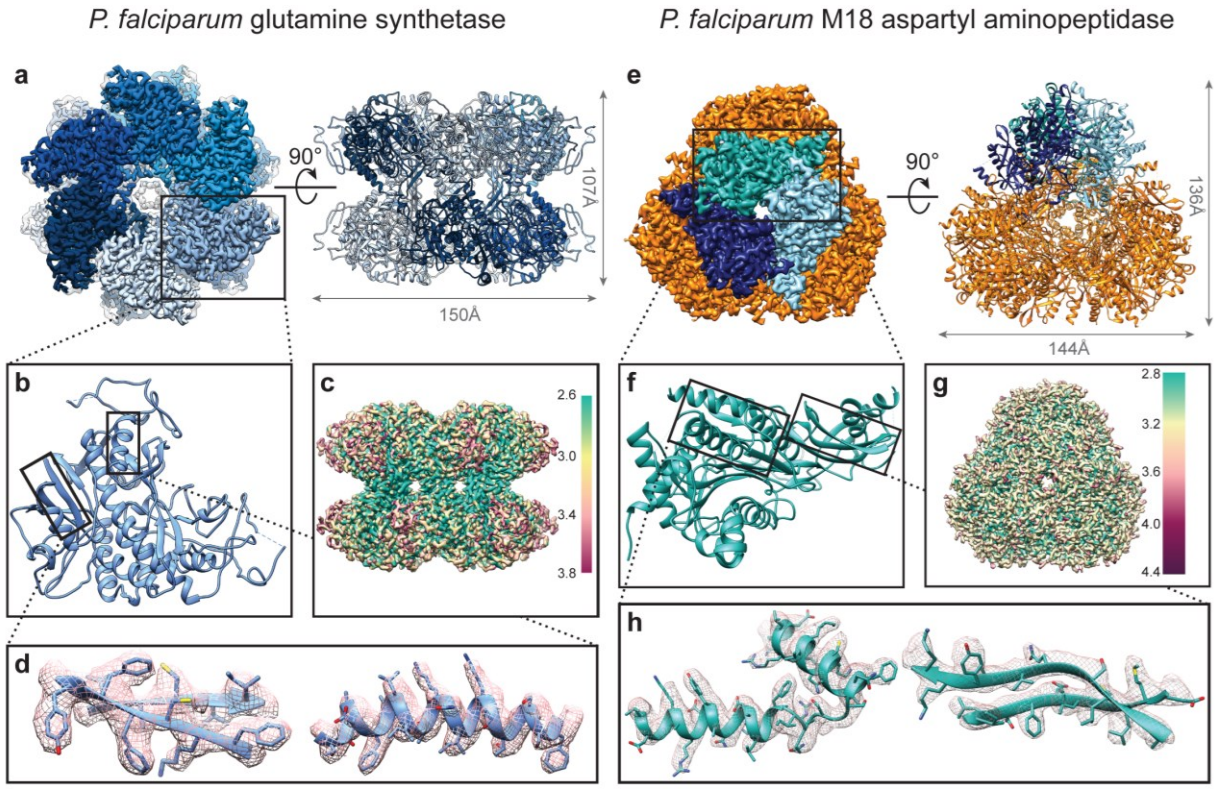


Figure 3.7 | CryoEM structures of proteins enriched directly from P. falciparum parasite lysates. a,e, 3.2Å cryoEM density map and atomic model of P. falciparum glutamine synthetase (**a**) and M18 aspartyl aminipeptidase (**e**). **b,f,** Enlarged view of the Pf glutamine synthetase (**b**) and M18 aspartyl aminopeptidase (**f**) monomer. c,g, Local resolution calculated using Resmap and two unfiltered halves of the reconstruction for Pf glutamine synthetase (**c**) and M18 aspartyl aminopeptidase (**g**). **d,h,** Detailed view of regions boxed in (**b & f**), displayed with corresponding cryoEM density.

We analyzed each map in *cryoID*, following the workflow detailed in **Fig. 3.1f**, generating a single query set from each map. In each case, *cryoID* successfully identified the correct protein in the map from the candidate pool consisting of the 800 proteins in the sample identified by mass spectrometry, enabling us to build atomic models of the two protein complexes.

	Ref6 map: C0H551: glutamine synthetase from *P. falciparum*			
	Length	Number of X	Errors	E-value
LGYGGLLGXGYLKYKKLL	18	1	3	3.6e-08
GGYKLPLGGGGXYLGGGGLGLGGK	23	1	5	4.5e-09
PLGLGLYXLGGKYLGKGGGGGYGKG	25	1	4	3.7e-12
YLGGPYLGGLGGKLKLGXGL	20	1	3	7.3e-09
Compounded E-value	6.4e-35			
Ranking out of candidate protein pool identified by mass spectrometry (800)	1st			

Table 3.6 | Ref6 map *cryoID* results.

	Ref7 map: Q8I2J3: M18 aspartyl aminopeptidase from *P. falciparum*			
	Length	Number of X	Errors	E-value
LGKGYGLGGLXYGXKLGGLYLGGKXLKLLL	30	3	5	9.7e-15
GKYGLLGGGYGXYGGYLLLL	20	1	5	1.7e-06
GGGXGYGGLLYLKKGGGGGGY	21	1	5	6.4e-07
Compounded E-value	7.9e-26			
Ranking out of candidate protein pool identified by mass spectrometry (800)	1st			

Table 3.7 | Ref6 map *cryoID* results.

3.3.10 Cross-validation against Pre-existing Crystal Structure of the *P. falciparum* M18 Aspartyl Aminopeptidase

The protein in the first cryoEM map was identified to be *P. falciparum* M18 aspartyl aminopeptidase, a 788kDa homo-dodecameric complex with tetrahedral symmetry which is thought to play a role in hemoglobin metabolism during the intraerythrocytic stage of the parasite life cycle[29, 30]. Our *de novo* structure agrees extremely well with the previously reported X-ray

crystallographic structure of this complex[29] (**Fig. 3.8**), with both the regulatory (residues 1-92 and 307-577) and catalytic (residues 92-306) domains clearly visible in all subunits (**Fig. 3.7**). As such, the previously published crystal structure serves as a gold standard validation of our method.

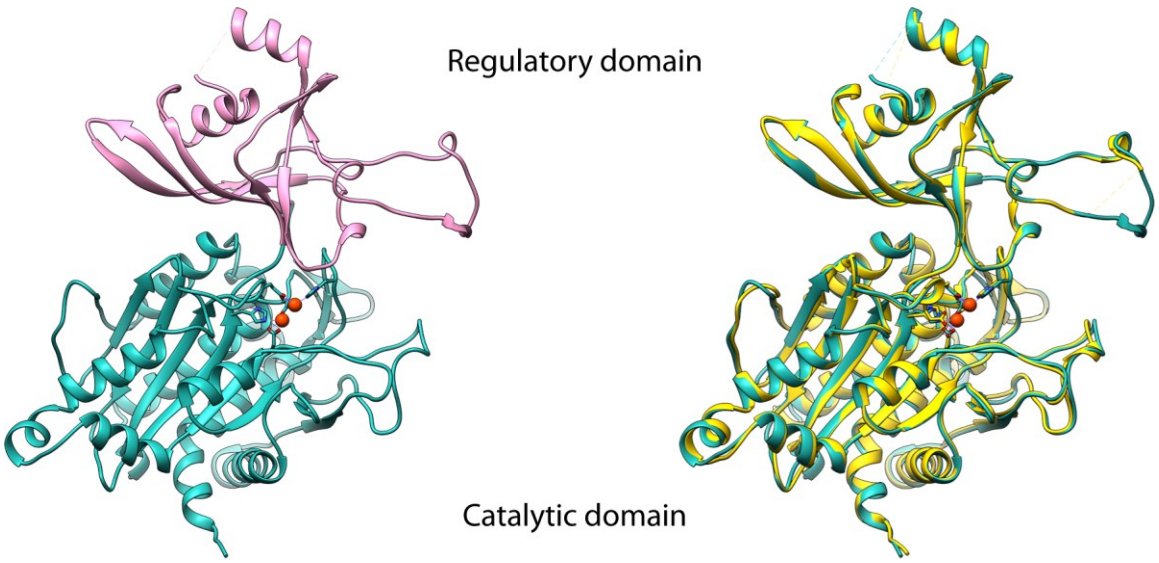


Figure 3.8 | Details of the M18 aspartyl aminopeptidase monomer. A single monomer from our atomic model of the *P. falciparum* M18 aspartyl aminopeptidase (*Pf*M18AAP), solved by cryoEM using our endogenous structural proteomics workflow, is shown on the left, colored to indicate the regulatory (pink) and catalytic (sea green) domains. On the right, our atomic model of *Pf*M18AAP (sea green), solved by cryoEM using our endogenous structural proteomics workflow, is shown superimposed with the previously published structure of *Pf*M18AAP (gold), solved using X-ray crystallography. The structures align with an RMSD of 0.548Å.

3.3.11 Structure of *P. falciparum* Glutamine Synthetase Reveals New Structural Features Unique to *Plasmodium*

The protein in the first cryoEM map was identified to be *P. falciparum* glutamine synthetase, a 759kDa homo-dodecameric complex which adopts a two-tiered ring shape with D6 symmetry (**Fig. 3.7**). This enzyme catalyzes the condensation of glutamate and ammonia into glutamine in an ATP-dependent manner[31-34]. The active site, positioned between adjacent monomers, contains

binding sites for ATP, glutamate, and ammonia, as well as two pockets for the binding of divalent cations[31, 35] (either Mg^{2+} or Mn^{2+}).

Our de novo structure agrees well with structure prediction based on the previously published atomic model of a close homolog, glutamine synthetase from *S. enterica*, solved by X-ray crystallography to 2.67Å resolution[35] (PDB accession code 1FPY). Our de novo atomic model from our cryoEM map is similar to the *S. enterica* glutamine synthetase crystal structure throughout most of the structure (RMSD 1.5Å), particularly in the active site, where the three substrate-binding pockets are well-conserved.

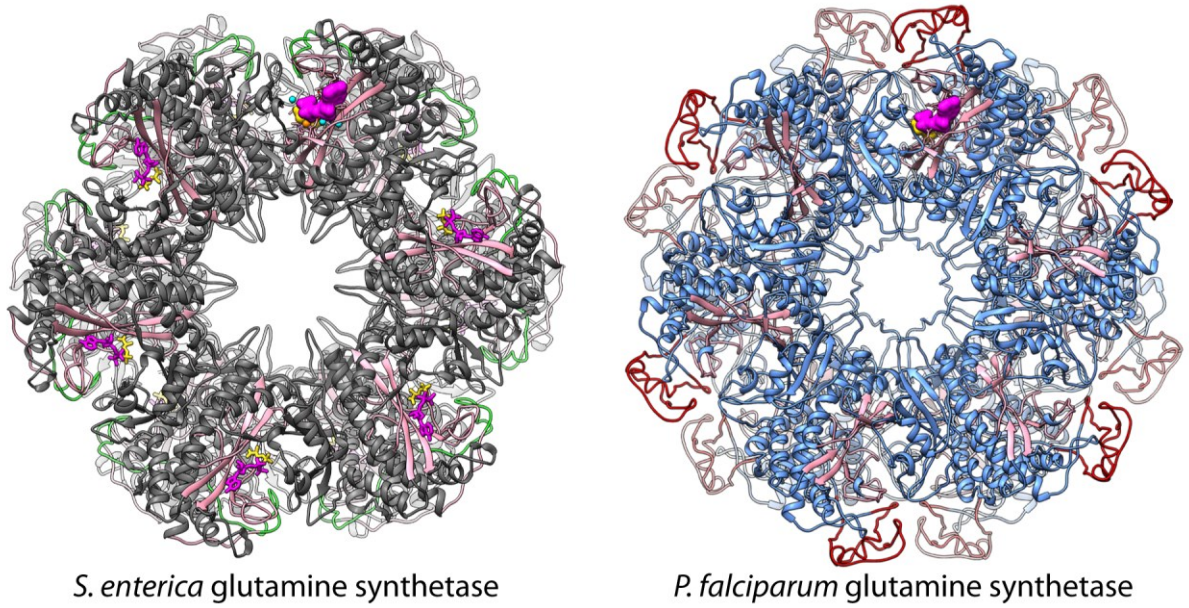


Figure 3.9 | Comparison of glutamine synthetase from *P. falciparum* (by endogenous cryoEM) and *S. enterica* (by X-ray crystallography). The active sites, shown in light pink in both models, is well-conserved. One region in which the two structures diverge is highlighted in red in the *P. falciparum* structure and green in the *S. enterica* structure.

However, we do observe one major difference between the two structures. In the *S. enterica* structure, residues 393-410 form a short loop shaped like a flap that folds partially across the entrance to the active site. In the corresponding location in our structure from *P. falciparum*, there

is an extra 50-residue insertion here that forms a long loop which folds down along the outside of the structure, in the opposite direction from the active site.

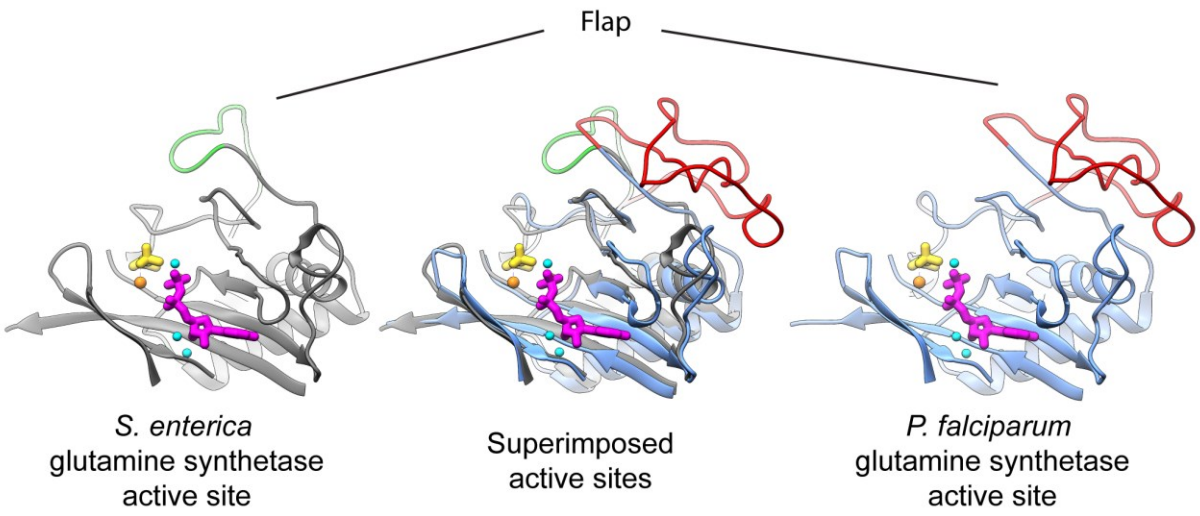


Figure 3.10 | Details of the glutamine synthetase monomer. A single monomer from our atomic model of the *P. falciparum* glutamine synthetase, solved by cryoEM using our endogenous structural proteomics workflow, is shown on the right, colored in cornflower blue. The previously published structure of the *S. enterica* glutamine synthetase, solved using X-ray crystallography, is shown on the left, colored dark grey. ADP (magenta), piperaquine (gold), manganese (orange), and water (baby blue) are visible in the active site. The structures align with an RMSD of 1.5Å and are shown superimposed in the middle. We observed an extra 50-residue insertion in the *P. falciparum* structure (colored red) that is absent in the *S. enterica* structure. This long insertion forms a large flap that curls away from the active site, unlike the shorter flap formed by the corresponding region in the *S. enterica* glutamine synthetase (colored green), which curls toward the active site.

3.4 Discussion

We have presented here a workflow for endogenous structural proteomics, which uses cryoEM in combination with mass spectrometry to obtain near-atomic resolution structures of protein complexes enriched directly from endogenous sources using tag-free methods. We have demonstrated the efficacy of *cryoID*, an essential component of the workflow, in successfully identifying the protein(s) in 3.0-4.5Å resolution cryoEM maps of unidentified protein complexes, without prior knowledge of the primary sequence(s). *cryoID* represents a significant step toward the realization of an autobuilding program capable of automatically modeling into typical (2.5-4.5Å) cryoEM maps of novel, unidentified protein complexes without primary sequence information, which would be transformative for the field. As a proof of principle, we have demonstrated the successful use of this workflow to obtain near-atomic resolution (3.3Å) structures of multiple protein complexes implicated in the pathogenesis of malarial parasites, enriched directly from unmodified *P. falciparum*. We have demonstrated that this endogenous structural proteomics workflow allows us to solve the structures of complexes that are difficult or impossible to recapitulate using traditional recombinant systems. By using minimally disruptive, tag-free techniques that avoid over-purification, we are able to enrich for large multi-protein assemblies as they exist in vivo, caught "in the act", in multiple native conformations, sometimes even bound to native substrates.

The holy grail of structural biology is to determine atomic structures for the entire proteome of a living cell without disrupting the macromolecular complexes from their native environment – commonly referred to as "*in situ* structural biology"[36, 37]. Such pleomorphic cellular structures are reconstructed using cryo electron tomography (cryoET), and resolutions of individual complexes inside the tomograms can be improved using sub-tomographic averaging.

However, there is a fundamental limitation to the thickness of samples that can be imaged by cryoET, as issues such as inelastic and multiple electron scattering events lead to prohibitively low SNR in resulting images, precluding structure determination at atomic resolution. Most cells exceed the desirable 200nm sample thickness for cryoET, previously restricting cryoET imaging to the cell peripheries, or artifact-ridden cryosections of high-pressure frozen cells. In recent years, focused ion-beam scanning electron microscopes (FIB-SEM) have been used to create thin sections, called lamella, of intact cells by milling away material above and below an area of interest in the cell. cryoET imaging of these lamella and subsequent sub-tomographic averaging are used to visualize macromolecular complexes *in situ* at molecular resolution (~2-3nm)[38-40]. However, FIB-SEM milling is laborious and technically difficult, typically yielding at most 10 lamellas per day, each encompassing a mere 1-2μm of imageable area per cell. As such, obtaining enough lamella tomograms to accumulate the half-million particles of macromolecular complexes needed to achieve near-atomic resolution is currently not practical. Therefore, despite exciting progress, this approach is far from routine, and atomic resolution *in situ* cellular structures will likely require years of development to achieve.

The "bottom-up" endogenous structural proteomics approach presented here represents an immediate step toward the ultimate goal of direct visualization of native protein complexes as they exist in the cellular milieu at near-atomic resolution. The methods developed here provide tools for structural study of endogenous protein complexes, enabling identification and *de novo* model-building of novel, previously unidentified protein complexes as they exist in their native environments. This platform will enable direct observation of previously unidentified novel protein complexes and their evolution (changes in binding partners, or cycling through sub-

complexes like the spliceosome) throughout various stages of biological processes (parasite life cycles, progression of cancers or cardiovascular diseases, etc.) at near-atomic resolution.

3.5 Acknowledgements

This research was supported in part by grants from National Institutes of Health (R21AI125983 to P.F.E., R01GM071940/AI094386/DE025567 to Z.H.Z. and K99/R00 HL133453 to J.R.B.). CM.H. acknowledges funding from the Ruth L. Kirschstein National Research Service Award (AI007323). We thank the UCLA Proteome Research Center for assistance in mass spectrometry and acknowledge the use of instruments in the Electron Imaging Center for Nanomachines supported by UCLA and grants from NIH (S10RR23057, S10OD018111 and U24GM116792) and NSF (DBI-1338135 and DMR-1548924).

3.6 Data Availability

The atomic models and the cryoEM density maps are deposited to the Protein Data Bank and the Electron Microscopy Data Bank, under the accession numbers of XXXX, XXXX, EMD-XXXX, and EMD-XXXX, respectively.

3.7 Competing Interests

The authors declare no competing interests.

3.8 Materials and Methods

Parasite culture

P. falciparum culture was performed as described with the exception that RPMI was supplemented with 0.5% Albumax I [41]. Parasites were synchronized by serial treatment with 5% w/v D-sorbitol and then expanded while shaking to increase singlet invasion events and maintain synchrony. For each preparation, ~$2x10^{10}$ parasite-infected erythrocytes were collected at the ring stage (typically ~500 mls of 2% hematocrit culture at ~20% parasitemia). Erythrocytes were lysed in 10x pellet volume of cold phosphate buffered saline (PBS) containing 0.0125% saponin (Sigma, sapogenin content ≥10%) and EDTA-free protease inhibitory cocktail (Roche or Pierce). Released parasites were washed in cold PBS containing EDTA-free protease inhibitory cocktail and washed cell pellets were frozen in liquid nitrogen and stored at -80°C.

Sucrose gradient fractionation *P. falciparum* parasite lysate

Frozen parasite pellets were resuspended in Lysis Buffer (25mM HEPES pH 7.4, 10mM $MgCl_2$, 150mM KCl, 10% Glycerol) and homogenized using a glass Dounce tissue homogenizer. The cytosolic fraction was isolated from the homogenized lysate by centrifugation at 100,000*g* for one hour. The soluble lysate was then fractionated with a 15-40% sucrose gradient.

The presence and relative abundance of large protein complexes of interest were ascertained by silver stained SDS-PAGE and tryptic digest liquid chromatography-mass spectrometry (**Fig. 3.1b-d**). The extremely low yields achievable when purifying protein complexes directly from *P. falciparum* parasites prohibited the conventional approach of evaluating sample quality by size exclusion chromatography. Thus, during the iterative process of screening for fractions containing complexes of interest as well as optimal fractionation conditions,

sample quality was assessed by negative stain (uranyl acetate) transmission electron microscopy in an FEI TF20 microscope equipped with a TVIPS 16 mega-pixel CCD camera. Briefly, small datasets of ~100,000 particles were collected and 2D class averages were generated in RELION[18, 19] to assess the presence of sufficient numbers of intact particles yielding "good" class averages exhibiting distinct features. For example, various symmetries could be recognized in top and side views (**Fig. 3.1e)**.

Cryo Electron Microscopy

3µl aliquots of fractionated lysate were applied to glow-discharged Quantifoil EM grids (Quantifoil). Grids were then blotted with filter paper and vitrified in liquid ethane using an FEI Vitrobot Mark IV. CryoEM grids were screened in an FEI Tecnai TF20 transmission electron microscope while optimizing freezing conditions.

Higher resolution cryoEM images were collected on a Gatan K2-Summit direct electron detector in super-resolution counting mode on an FEI Titan Krios at 300kV equipped with a Gatan Quantum energy filter set at a 20 eV slit width. Fifty frames were recorded for each movie at a pixel size of 1.07Å at the specimen scale, with a 200ms exposure time and an average dose rate of 1.2 electrons per $Å^2$ per frame, resulting in a total dose of 60 electrons per $Å^2$ per movie. The final dataset consists of a total of 2,514 movies.

Image processing and 3D reconstruction

Frames in each movie were aligned, gain reference-corrected and dose-weighted to generate a micrograph using MotionCor2 [Ref [42]]. Aligned and un-dose-weighted micrographs were also generated and used for contrast transfer function (CTF) estimation using CTFFIND4 [Ref [43]] and

particle picking by hand and using Gautomatch[44].

328,435 particles were extracted from 2,514 micrographs. After several rounds of reference-free two-dimensional (2D) classification in RELION, junk particles and classes clearly corresponding to 20s proteasome (47,159 particles) were excluded. 22,596 particles belonging to "good" 2D class averages that exhibited clear secondary structure features but did not resemble proteasomes were then used in an unsupervised four-class *ab initio* 3D reconstruction, followed by separate homogeneous refinements in CryoSPARC[17], yielding two 3.3Å *ab initio* 3D maps. Further refinement of the particles giving rise to each of the two 3.3Å maps in RELION failed to yield any improvement in resolution.

Manual Model Building and Refinement

Map interpretation was performed with UCSF Chimera[45] and COOT[46]. *P. falciparum* protein sequences were obtained from the National Center for Biotechnology Information (NCBI)[47] and the PlasmoDB[48] protein databases. Sequence registration during model building of *P. falciparum* glutamine synthetase and *P. falciparum* M18 aspartyl aminopeptidase was guided by reference to homologs (accession codes 1FPY and 4EME, respectively) as well as PHYRE2 [Ref [49]] secondary structure predictions. For the M18 aspartyl aminopeptidase structure, each residue in the monomer was manually refit in COOT to optimize geometry and fit. For the glutamine synthetase structure, each residue in the monomer was manually traced and built *de novo* in COOT. The model for each protein was then propagated to match the biological assembly and rigid-body fit into the density map.

Manual refinement targeting both protein geometry and fit with the density map was used primarily in the core regions where resolution was higher and noise was minimal. To improve the

geometry and fit, manual adjustments were made to protein geometry and density map fit, using Molprobity[50] clash dots and sphere-refinement in COOT. Rotamers were fit manually in COOT and improved using the 'Back-rub Rotamers' setting. The resulting models for the complexes were subjected to the phenix.real_space_refine program in PHENIX[51].

All figures and videos were prepared with UCSF Chimera, Pymol[52], and Resmap[53]. Molprobity was used to validate the stereochemistry of the final models.

Building the *cryoID* User Interface

We developed the Python-based ProgamX software (**Fig. 3.2**) using the PyQt GUI toolkit and the open source Python development environment Spyder. *cryoID* consists of two main subprograms, *generate_queries_from_map* and *search_candidate_database.* The subprogram *generate_queries_from_map* performs the Selection and Prediction functionalities of *cryoID*, identifying one or more high resolution segments of the map with a continuous backbone and clearly distinguishable sidechain densities and automatically tracing the polypeptide backbone for each map segment and semi-automatically predicting the identities of the sidechains for each residue in the segment, yielding a predicted primary sequence for the segment. The subprogram *search_candidate_database* performs the Simplification and Search functionalities of *cryoID*, Translating both the cryoEM map segment sequences and all the primary sequences of the pool of candidate proteins into a simplified "6-letter" code and performing a modified blast search of the entire pool of candidate proteins using the predicted cryoEM map segment sequences as queries.

The two subprograms can be accessed either from the command line, or *via* the graphical user interface (GUI). *cryoID* is open source and available for download upon request.

Selection and Prediction Using *generate_ queries_from_map*

The first *cryoID* subprogram, *generate_queries_from_map,* generates multiple query sequences from the cryoEM density map. It calls the *phenix.sequence_from_map* function in the PHENIX software package to identify the best regions of the density map. It then predicts the most likely sequence for each of these best regions by matching the side chain density of each residue with those in an ideal (crystallographic) rotamer library. Upon completion, it generates a pdb file containing the atomic coordinates for each of the resulting query sequences.

generate_queries_from_map then calls COOT to open the density map with the pdb file for user inspection, and if necessary, manual correction of incorrectly assigned residues using the "mutate residue" tool in COOT. We recommend a minimum length of 15 residues for each query sequence. For sequences below the recommended 15 residues in length, users may extend on either end of the resulting query sequences if the cryoEM density on either end is of sufficient quality. We use the letter "X" to denote unidentifiable residues, which is designated with the residue type MSE in Coot. The modified pdb file is then saved and passed on to the second *cryoID* subprogram.

The *generate_queries_from_map* subprogram requires three user-provided inputs: the filename, resolution and symmetry for the cryoEM density map, which can be provided either from our GUI or from the command line. For the resolution parameter, one may start with the average global resolution reported by reconstruction programs and then fine-tune this parameter based on the estimated local resolution of the selected regions. The *generate_queries_from_map* subprogram is very fast; for reference, processing of the glutamine synthetase and M18 aspartyl aminopeptidase maps was completed in 5-10min.

Simplification and Searching Using *search_candidate_database*

The second *cryoID* subprogram, *search_candidate_database*, performs alignments of the query sets generated by the *generate_queries_from_map cryoID* subprogram against the full length protein sequences of all the proteins in a user-defined candidate protein pool. The program requires two inputs: 1) a file containing a list of query sequences in either standard fasta format or in the pdb format generated by the *generate_queries_from_map cryoID* subprogram. In the latter case, the sequence information is extracted by calling *phenix.print_sequence* in the *search_candidate_database* subprogram. 2) a file containing a list of the proteins in the user-defined candidate protein pool, either in the form of a standard result file from mass spectrometry, or in the form of a text file containing a list of UniProt protein names. The *search_candidate_database* subprogram then reads in the names of candidate proteins listed in the sequence pool file and retrieves their sequences and length information from the UniProt website *via* the database identifier mapping service. The *search_candidate_database* subprogram then translates both the query and candidate sequences into the simplified six-letter code (Table 1).

Once this is accomplished, the *search_candidate_database* subprogram calls the widely distributed local alignment search tool *Blastp*[23] to search the candidate protein pool for the protein that contains segments matching the query sequences. In preparation for the *Blastp* search, the program first generates a local database from the degenerate candidate protein sequences with *makeblastdb*. The codes for the 6 degenerate categories are selected (G-like → G, P-like → M, L-like → Z, K-like → K, Y-like → Y, W-like → W) based on the PAM30 scoring matrix so that the substitution scoring matrix used has higher bonus scores for matches to P-like/K-like/Y-like/W-like categories and appropriate penalty scores for mismatches depending on the severity of side chain shape dissimilarity between categories (Table 2). By default, the program prohibits gapped alignments (insertions/deletions) during *Blastp* search by setting very high penalty scores for gap

open (-32767) and gap extension (-32767). Experienced users may take advantage of additional arguments through the advanced option input in the GUI. For example, users can choose to include gapped alignments during the *Blastp* search by setting the penalty score as -15/ -3 for gap open/extension and setting the maximum allowable number of gaps.

For each of the query sequences, the program calls a *blastp* search. The following arguments are used for each search to optimize the *blastp* search for short degenerate sequences: "*-task blastp-short -matrix PAM30 -db ./database/dbname -query query_file -out output_name -evalue 1 -comp_based_stats F -dbsize dbsize -searchsp searchsp -word_size 2 -gapopen 32767 -gapextend 32767 -outfmt 7*", where dbsize/searchsp specifies the effective size of the database/search space (in our case we use the actual size). Each search generates a list of sequence segments belonging to the protein candidate pool that match the query with alignment statistics (such as alignment length, % identity, E value, number of mismatches *etc*.). The program then evaluates these matched sequences based on the alignment length and E value: those with very short length (<60% of the query length) are discarded; if gaps are allowed, those with more than the maximum allowable gaps are discarded; and for each matched protein, the one with the smallest E value is selected.

For each matched protein, the program quantifies the quality of the match by calculating a compounded E value of the search results of all queries, as defined below:

$$E_{i,final} = \prod_{j=1}^{N} \min(E_{i,j},1) \bullet l_i$$

where i is the ith protein, j the jth query, N the number of queries, $E_{i,j}$ the E value of the ith protein for query j, and l length factor (*i.e.*, length/1000) of the ith protein. If the polarity of the query

sequence (i.e., N/C termini) is unknown, users can check “polarity unknown” flag in the GUI so that the program will try to align the query against the protein pool in both polarities.

Finally, the program sorts all matched proteins based on the compound E value, and the resulting list is saved in a file called ‘output name_detailed.txt’. The protein candidate with the smallest compounded E value is on the top of the list and should correspond to the correct protein, provided that the queries satisfy the rules as outlined in Table 3.

In rare instances where the query contains too many errors or the queries are too few or the length too short, the matched protein with the smallest compounded E value is a false positive. False positive matches can be recognized during model building or if their abundance in the mass spectrometry results do not agree with their contribution to the particle population in the cryoEM images.

Benchmarking on published structures

We tested *cryoID* against two published experimental cryoEM maps available for the EMDB, the human gamma-secretase complex (3.4A) and the *D. melanogaster* mechanotransduction channel NOMPC, which have global resolutions in our target resolution range (3.0~4.5Å).

***cryoID* and Gamma-secretase**

Human gamma-secretase is a four-membered intramembrane protease consisting of presenilin, PEN-2, nicastrin and APH-1. We tested G*enerate_queries_from_map* for on the human gamma-secretase density map using the following input parameters: (*Homo sapiens,* EMD-3061, protein complex, no symmetry, and reported resolution of 3.4A) with several different resolution inputs (3.0Å, 3.2Å and 3.4Å) and found the selected regions to be quite consistent. The 3.2A input result,

which yielded slightly better query models, were used for query generation. We manually inspected the density maps and found two regions with clean, continuous backbone density throughout, one region in the extracellular domains (referred to as the extracellular region) and one region in the transmembrane domains (referred to as the transmembrane region). The first region contained three segments: two helices and one beta-strand, whose side chain densities were easily distinguishable using the simplified six-letter code. G*enerate_queries_from_map* successfully generated pdb files with predicted query sequences for the three segments. We then manually inspected the query models, correcting residues incorrectly assigned by G*enerate_queries_from_map* and extending the queries on both ends as the density permitted. This yielded the following degenerate sequences, which were then used for searching:

1) GGXGPLGGYLGWGXG
2) LLYYGGGGPPGGXGGKGGXYGL
3) GGGKGGXLGGGGGLXKGP

Selecting and processing segments in the transmembrane region in the same way yielded query sequences for three helical segments:

1) GGKLXGYGGGLGYXGGXGGYGGL,
2) PYYYLGGGYGGGXLGLLYGYKGGGYXGGGGK
3) LPXYGGGGLGGYGGGGXLGXWGYG.

Using these two sets of query sequences, we test the ability of *cryoID* to correctly identify the corresponding protein, first from a more limited candidate pool consisting of the 20.397 proteins in the *H. sapiens* proteome, and then against a much larger candidate pool consisting of the entire ~600,000 proteins in the UniProt database. When given the smaller 20,397 protein candidate pool, *cryoID* correctly identified the corresponding proteins for both of the query sets

generated from the human gamma-secretase cryoEM map, correctly making a unique protein ID of Nicastrin (Q92542) for the extracellular region query set and APH-1 (Q96BI3) for the transmembrane region query set. Against the much larger ~600,000 protein candidate pool, *cryoID* correctly identified the corresponding protein for the transmembrane region query set, correctly making a unique protein ID of APH-1 (Q96BI3), but was unable to make a unique protein ID for the extracellular region query set, ranking Nicastrin, the correct answer, as 23rd out of ~600,000 candidates.

cryoID and the NOMPC mechanotransduction channel

The NOMPC mechanotransduction channel from *Drosophila* is a homotetrameric integral membrane protein that mediates gentle-touch sensation. We tested G*enerate_queries_from_map* on the NOMPC density map using the following input parameters: (*Drosophila*, EMD-8702, C4 symmetry, reported resolution 3.55A). The 3.2Å and C4 symmetry input parameters yielded one region in the transmembrane domain with clean, continuous backbone density throughout, from which G*enerate_queries_from_map* produced a set of three query sequences. We manually inspected the query models, correcting residues incorrectly assigned by G*enerate_queries_from_map* and extending the queries on both ends as the density permitted. This yielded the following degenerate sequences, which were then used for searching:

1) WGGXLYLGGYGGYLLGGGGGGGLLGGLKGGGYXKG,
2) LLXGGGKYLGGXGGYGLGYG,
3) GGXYGGXGGGYGYGLGXGGG.

Using this set of query sequences, we test the ability of *cryoID* to correctly identify the corresponding protein, first from a more limited candidate pool consisting of the 3,500 proteins in

the *D. melanogaster* proteome, and then against a much larger candidate pool consisting of the entire ~600,000 proteins in the UniProt database. In both cases, *cryoID* correctly identified the corresponding protein for the query set generated from the *Drosophila* NOMPC cryoEM map, correctly making a unique protein ID of NOMPC (E0A9E1).

Benchmarking on new experimental structures obtained from *P. falciparum* parasites using the endogenous structural proteomics workflow

We tested *cryoID* against two unpublished experimental cryoEM maps, which we obtained from *P. falciparum* parasite lysates using our endogenous structural proteomics workflow, yielding two maps at 3.3Å resolution. As a control for *cryoID*, we independently identified the proteins in the two maps and built *de novo* atomic models into the two maps by hand. To identify the proteins in each map, we manually sorted through the 800 possible protein candidates identified by mass spectrometry, discarding all proteins that were too low in abundance, and all proteins that had the wrong symmetry, oligomeric state, size, or overall structure based on published atomic models (including atomic models of known homologs from other organisms). After discarding all of the candidates that were obviously wrong, we were left with 5-10 potential candidates. We then compared published structures of the candidates or their homologs against our cryoEM maps until we found a structure for each that appear to fit well in the density.

In the case of our map that was ultimately determined to be M18 aspartyl aminopeptidase (Ref7), the published crystal structure of M18 aspartyl aminopeptidase from *P. falciparum* (PDB accession code 4EME) fit perfectly into our density map. We further confirmed the protein ID by

independently building into our Ref7 cryoEM map *de novo*, using the primary sequence of *P. falciparum* M18 aspartyl aminopeptidase as a guide. Our resulting atomic model matched the previously published crystal structure almost perfectly (RMSD = 0.548Å).

In the case of our map that was ultimately determined to be glutamine synthetase (Ref6), the published structure of a homolog, glutamine synthetase from *S. enterica*, fit well into our density map. In order to test whether our map was truly glutamine synthetase, we independently built into our Ref6 cryoEM map *de novo*, using the primary sequence of *P. falciparum* glutamine synthetase as a guide. Our resulting atomic model matched the *S. enterica* crystal structure well (RMSD = 1.5Å), with the exception of a 50 residue long insertion near the active site.

In the meantime, we tested G*enerate_queries_from_map* on the M18 aspartyl aminopeptidase density map, named Ref7, using the following initial input parameters: (*P. falciparum*, Ref7, T symmetry, reported resolution 3.3Å). We tuned the resolution parameter according to local resolution estimates and ultimately found 3.2Å yielded the best results. We then manually inspected the query models, correcting residues incorrectly assigned by G*enerate_queries_from_map* and extending the queries on both ends as the density permitted. This yielded the following degenerate sequences, which were then used for searching:

1) LGKGYGLGGLXYGXKLGGLYLGGKXLKLLL
2) (GKYGLLGGGYGXYGGYLLLL
3) GGGXGYGGLLYLKKGGGGGGY

Using this set of query sequences, we test the ability of *cryoID* to correctly identify the corresponding protein, from a candidate pool consisting of the 800 proteins identified in this sucrose gradient fraction by mass spectrometry. *cryoID* correctly identified the corresponding protein for the query set generated from the Ref7 cryoEM map, making a unique protein ID of

M18 aspartyl aminopeptidase from *P. falciparum* (Q8I2J3). We confirmed the identification by manually building a *de novo* atomic model into the rest of the map, and then comparing the resulting atomic model with the pre-existing published atomic model of the M18 aspartyl aminopeptidase from *P. falciparum*, solved by X-ray crystallography to 2.6Å resolution (PDB accession code 4EME). Our *de novo* atomic model from our cryoEM map agreed well with the published model (RMSD 0.548Å), serving as a gold standard validation of our workflow and *cryoID*'s performance.

We then tested G*enerate_queries_from_map* on the glutamine synthetase density map, named Ref6, using the following initial input parameters: (*P. falciparum*, Ref7, D6 symmetry, reported resolution 3.3Å). We tuned the resolution parameter according to local resolution estimates and ultimately found 2.8Å yielded the best results. We then manually inspected the query models, correcting residues incorrectly assigned by G*enerate_queries_from_map* and extending the queries on both ends as the density permitted. This yielded the following degenerate sequences, which were then used for searching:

1) LGYGGLLGXGYLKYKKLL
2) GGYKLPLGGGGXYLGGGGLGLGGK
3) PLGLGLYXLGGKYLGKGGGGGYGKG
4) YLGGPYLGGLGGKLKLGXGL

Using this set of query sequences, we test the ability of *cryoID* to correctly identify the corresponding protein, from a candidate pool consisting of the 800 proteins identified in this sucrose gradient fraction by mass spectrometry. *cryoID* correctly identified the corresponding protein for the query set generated from the Ref6 cryoEM map, making a unique protein ID of glutamine synthetase from *P. falciparum* (C0H551). We confirmed the identification by manually

building a *de novo* atomic model into the rest of the map, and then comparing the resulting atomic model with the pre-existing published atomic model of a close homolog, glutamine synthetase from *S. enterica,* solved by X-ray crystallography to 2.67Å resolution (PDB accession code 1F1H). Our *de novo* atomic model from our cryoEM map agreed well with the *S. enterica* glutamine synthetase structure (RMSD 1.5Å) throughout most of the structure, particularly in the active site. However, we do observe one major difference between the two structures. In the *S. enterica* structure, residues 393-410 form a short loop shaped like a flap that folds partially across the entrance to the active site. In our structure from *P. falciparum*, there is an extra 50 residue insertion here that forms a long loop which folds down along the outside of the structure, in the opposite direction from the active site.

3.9 References

1 J. Field *et al.* Purification of a RAS-responsive adenylyl cyclase complex from Saccharomyces cerevisiae by use of an epitope addition method. *Mol Cell Biol* **8**, 2159-2165 (1988).

2 T. P. Hopp *et al.* A Short Polypeptide Marker Sequence Useful for Recombinant Protein Identification and Purification. *Bio-Technol* **6**, 1204-1210, doi:DOI 10.1038/nbt1088-1204 (1988).

3 D. B. Smith &K. S. Johnson. Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. *Gene* **67**, 31-40 (1988).

4 A. H. Rosenberg *et al.* Vectors for selective expression of cloned DNAs by T7 RNA polymerase. *Gene* **56**, 125-135 (1987).

5 E. Hochuli, H. Dobeli &A. Schacher. New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. *J Chromatogr* **411**, 177-184 (1987).

6 <EMBOJ-1984-Munro & Pelham. Use of peptide tagging to detect proteins expressed from cloned genes - deletion mapping functional domains of Drosophila hsp70.pdf>.

7 S. Munro &H. R. Pelham. Use of peptide tagging to detect proteins expressed from cloned genes: deletion mapping functional domains of Drosophila hsp 70. *EMBO J* **3**, 3087-3093 (1984).

8 J. Porath, J. Carlsson, I. Olsson &G. Belfrage. Metal Chelate Affinity Chromatography, a New Approach to Protein Fractionation. *Nature* **258**, 598-599, doi:DOI 10.1038/258598a0 (1975).

9 R. PDB. *PDB Statistics: Overall Growth of Released Structures Per Year*, <https://www.rcsb.org/stats/growth/overall> (2018).

10 Y. Cheng. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **161**, 450-457, doi:10.1016/j.cell.2015.03.049 (2015).

11 X. Li *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* **10**, 584-590, doi:10.1038/nmeth.2472 (2013).

12 H. Liu *et al.* Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks. *Science* **329**, 1038-1043, doi:10.1126/science.1187433 (2010).

13 G. McMullan, S. Chen, R. Henderson &A. R. Faruqi. Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy* **109**, 1126-1143, doi:10.1016/j.ultramic.2009.04.002 (2009).

14 G. McMullan, A. T. Clark, R. Turchetta &A. R. Faruqi. Enhanced imaging in low dose electron microscopy using electron counting. *Ultramicroscopy* **109**, 1411-1416, doi:10.1016/j.ultramic.2009.07.004 (2009).

15 G. M. R. N. Clough, A. I. Kirkland. in *Journal of Physics: Conference Series* Vol. 522 (IOP Publishing, 2013).

16 X. Zhang, L. Jin, Q. Fang, W. H. Hui &Z. H. Zhou. 3.3 A cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. *Cell* **141**, 472-482, doi:10.1016/j.cell.2010.03.041 (2010).

17 A. Punjani, J. L. Rubinstein, D. J. Fleet &M. A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-+, doi:10.1038/Nmeth.4169 (2017).

18 S. H. W. Scheres. A Bayesian View on Cryo-EM Structure Determination. *J Mol Biol* **415**, 406-418, doi:10.1016/j.jmb.2011.11.010 (2012).

19 S. H. W. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519-530, doi:10.1016/j.jsb.2012.09.006 (2012).

20 C. M. Ho *et al.* Malaria parasite translocon structure and mechanism of effector export. *Nature* **561**, 70-+, doi:10.1038/s41586-018-0469-4 (2018).

21 E. Niedzialkowska *et al.* Protein purification and crystallization artifacts: The tale usually not told. *Protein Sci* **25**, 720-733, doi:10.1002/pro.2861 (2016).

22 J. Osipiuk, M. A. Walsh &A. Joachimiak. Crystal structure of MboIIA methyltransferase. *Nucleic Acids Res* **31**, 5440-5448, doi:10.1093/nar/gkg713 (2003).

23 S. F. Altschul, Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. . Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

24 X. C. Bai *et al.* An atomic structure of human gamma-secretase. *Nature* **525**, 212-217, doi:10.1038/nature14892 (2015).

25 P. Jin *et al.* Electron cryo-microscopy structure of the mechanotransduction channel NOMPC. *Nature* **547**, 118-122, doi:10.1038/nature22981 (2017).

26 M. J. Gardner *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498-511, doi:10.1038/nature01097 (2002).

27 N. Hall *et al.* A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307**, 82-86, doi:10.1126/science.1103717 (2005).

28 A. P. Waters. Genome-informed contributions to malaria therapies: feeding somewhere down the (pipe)line. *Cell Host Microbe* **3**, 280-283, doi:10.1016/j.chom.2008.04.005 (2008).

29 K. K. Sivaraman *et al.* X-ray crystal structure and specificity of the Plasmodium falciparum malaria aminopeptidase PfM18AAP. *J Mol Biol* **422**, 495-507, doi:10.1016/j.jmb.2012.06.006 (2012).

30 F. Teuscher *et al.* The M18 aspartyl aminopeptidase of the human malaria parasite Plasmodium falciparum. *J Biol Chem* **282**, 30817-30826, doi:10.1074/jbc.M704938200 (2007).

31 D. Eisenberg, H. S. Gill, G. M. Pfluegl &S. H. Rotstein. Structure-function relationships of glutamine synthetases. *Biochim Biophys Acta* **1477**, 122-145 (2000).

32 A. Ginsburg, J. Yeh, S. B. Hennig &M. D. Denton. Some effects of adenylylation on the biosynthetic properties of the glutamine synthetase from Escherichia coli. *Biochemistry-Us* **9**, 633-649 (1970).

33 W. W. Krajewski *et al.* Crystal structures of mammalian glutamine synthetases illustrate substrate-induced conformational changes and provide opportunities for drug and herbicide design. *J Mol Biol* **375**, 217-228, doi:10.1016/j.jmb.2007.10.029 (2008).

34 S. H. Liaw, I. Kuo &D. Eisenberg. Discovery of the ammonium substrate site on glutamine synthetase, a third cation binding site. *Protein Sci* **4**, 2358-2365, doi:10.1002/pro.5560041114 (1995).

35 H. S. Gill &D. Eisenberg. The crystal structure of phosphinothricin in the active site of glutamine synthetase illuminates the mechanism of enzymatic inhibition. *Biochemistry-Us* **40**, 1903-1912 (2001).

36 S. Asano, B. D. Engel &W. Baumeister. In Situ Cryo-Electron Tomography: A Post-Reductionist Approach to Structural Biology. *J Mol Biol* **428**, 332-343, doi:10.1016/j.jmb.2015.09.030 (2016).

37 M. Beck &W. Baumeister. Cryo-Electron Tomography: Can it Reveal the Molecular Sociology of Cells in Atomic Detail? *Trends Cell Biol* **26**, 825-837, doi:10.1016/j.tcb.2016.08.006 (2016).

38 S. Albert *et al.* Proteasomes tether to two distinct sites at the nuclear pore complex. *Proc Natl Acad Sci U S A* **114**, 13726-13731, doi:10.1073/pnas.1716305114 (2017).

39 J. Mahamid *et al.* Visualizing the molecular sociology at the HeLa cell nuclear periphery. *Science* **351**, 969-972, doi:10.1126/science.aad8857 (2016).

40 S. Mosalaganti *et al.* In situ architecture of the algal nuclear pore complex. *Nat Commun* **9**, 2361, doi:10.1038/s41467-018-04739-y (2018).

41 M. Klemba, W. Beatty, I. Gluzman &D. E. Goldberg. Trafficking of plasmepsin II to the food vacuole of the malaria parasite Plasmodium falciparum (vol 164, pg 47, 2004). *Journal of Cell Biology* **164**, 625-625, doi:DOI 10.1083/jcb.2004021616447 (2004).

42 S. Q. Zheng *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* **14**, 331-332, doi:10.1038/nmeth.4193 (2017).

43 A. Rohou &N. Grigorieff. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol* **192**, 216-221, doi:10.1016/j.jsb.2015.08.008 (2015).

44 K. Zhang. *Gautomatch: a GPU-accelerated program for accurate, fast, flexible and fully automatic particle picking from cryo-EM micrographs with or without templates* (2016).

45 E. F. Pettersen *et al.* UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).

46 P. Emsley, B. Lohkamp, W. G. Scott &K. Cowtan. Features and development of Coot. *Acta Crystallogr D* **66**, 486-501, doi:10.1107/S0907444910007493 (2010).

47 N. R. Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**, D7-D19, doi:10.1093/nar/gkv1290 (2016).

48 C. Aurrecoechea *et al.* PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* **37**, D539-D543, doi:10.1093/nar/gkn814 (2009).

49 L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass &M. J. E. Sternberg. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858, doi:10.1038/nprot.2015.053 (2015).

50 V. B. Chen *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D* **66**, 12-21, doi:10.1107/S0907444909042073 (2010).

51 P. D. Adams *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D* **66**, 213-221, doi:10.1107/S0907444909052925 (2010).

52 Schrodinger, LLC. *The PyMOL Molecular Graphics System, Version 1.8* (2015).

53 A. Kucukelbir, F. J. Sigworth &H. D. Tagare. Quantifying the local resolution of cryo-EMEM density maps. *Nat Methods* **11**, 63-+, doi:10.1038/Nmeth.2727 (2014).

Chapter 4

Conclusion

To address the paucity of structures from *P. falciparum* and other systems that have proven recalcitrant to structural biology approaches dependent on recombinant methods, we have presented here two distinct approaches that leverage recent advances in cryoEM to enable structure determination of protein complexes enriched directly from endogenous sources. Together, these complementary approaches enable high resolution structure determination of a vast number of previously intractable biological systems.

First, employing a targeted, CRISPR-enabled "top down" approach, we determined near-atomic resolution structures of the unique malaria parasite translocon PTEX, which we purified directly from *P. falciparum* parasites in multiple functional states. Our structures are the first near-atomic resolution cryoEM structures of a protein isolated directly from an endogenous source using an epitope tag inserted into the endogenous locus with CRISPR-Cas9 gene editing.

Beyond establishing that PTEX is a *bona fide* translocon and elucidating the mechanism of this essential gatekeeper to the malaria parasite exportome, our structures revealed an interaction between the transmembrane pore (EXP2) and the protein-unfolding motor (HSP101), which we then demonstrated is indispensable for PTEX activity and parasite survival in erythrocytes *in vivo.* While this work represents a significant stride forward in our understanding of effector protein export in malaria parasites, there are many outstanding questions regarding the mechanisms of cargo-recognition and regulation, as well as the potential for rational design of PTEX inhibitors. In addition to further structural studies by single-particle cryoEM, which could shed light on the relationship between ATP-bound state and translocation activity, an *in vitro* translocation activity assay would be a useful tool in answering some of these questions, providing a means for defining

the roles that various accessory proteins may play in cargo recognition and regulation of translocation activity.

Next, to complement the more traditional "top down" approach used to obtain the near-atomic resolution structures of PTEX, we have developed a "bottom up" endogenous structural proteomics method whereby protein complexes are enriched directly from the cellular milieu and identified by imaging and structure determination using cryoEM and mass spectrometry. As a proof of principle, we successfully applied this approach to the study of the *P. falciparum* proteome, which has previously proven recalcitrant expression in recombinant systems, often precluding structure determination by X-ray crystallography or NMR. As an illustration of this point, many of the most promising *P. falciparum* drug targets are membrane proteins, but there are only two unique integral membrane protein structures from *P. falciparum* in the PDB. The paucity of high resolution structural and functional information is compounded by the fact that 50% of the P. falciparum proteome is novel, bearing no known similarity to existing structures in the PDB. By directly imaging components of the parasite cell lysate, we obtained near-atomic resolution (3.3Å) structures of multiple protein complexes implicated in the pathogenesis of malarial parasites, from a single cryoEM dataset.

This work represents a significant step forward toward achieving the ultimate long-term goal in structural biology: directly visualizing protein complexes in action in living cells, at near-atomic resolution. However, in order to elucidate the dynamics of large multi-component molecular machines like the full effector protein export machinery in *P. falciparum*, it will be necessary to capture the details of protein-protein interactions that are too transient to preserve outside of the intact cell. The full effector protein export machinery sits on the host-parasite interface, spanning the PPM, PV, and PVM, but so much information regarding the complex

interactions of the machinery components with each other and with the two membranes is lost when we break the cell open. As such, it remains largely unclear how the full machinery at these two membranes coordinates the delivery of effector proteins across the two membranes and refolds them on the other side. To capture this information, it will be necessary to directly visualize the entire machinery, as it exists in the cell, using super resolution correlative light and electron microscopy (SR-CLEM), focused-ion beam scanning electron microscopy (FIB-SEM), and cryo-electron tomography (CryoET).

The body of work described here addresses a known need for methods that overcome the limitations of structural biology approaches that depend on recombinant systems, opening the door for high resolution structure determination of a vast number of previously intractable biological systems. The use of cryoEM enables us to overcome the lower yields and heterogeneity of samples enriched directly from endogenous sources, making it possible to unambiguously determine near-atomic resolution cryoEM structures of previously intractable protein complexes enriched directly from endogenous sources. Together, the targeted, CRISPR-enabled "top down" approach and the tag-free "bottom up" approach to structure determination from endogenous sources presented above represent the future of structural biology and cryoEM: direct visualization of protein complexes as they exist in the cellular milieu at near-atomic resolution.

fin