Published in final edited form as:
*Nat Methods*. 2020 January ; 17(1): 79–85. doi:10.1038/s41592-019-0637-y.

# Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu

**Chi-Min Ho**[1,2,3,10], **Xiaorun Li**[3,4,10], **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 the Microscale, School of Life Sciences University of Science and Technology of China Hefei Anhui 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 in St. Louis, 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

[10]These authors contributed equally: Chi-Min Ho, Xiaorun Li.

## 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

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

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; JAW performed the mass spectrometry; 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 the Phenix tool *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, JAW, and DEG helped edit the paper; all authors approved the paper.

Competing Interests

The authors declare no competing interests.

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. As a proof of principle, we applied this approach to the malaria parasite *Plasmodium falciparum*, an organism that has resisted conventional structural biology approaches, to obtain atomic models of multiple protein complexes implicated in intraerythrocytic survival of the parasite. Our approach is broadly applicable for determining structures of undiscovered protein complexes enriched directly from endogenous sources.

## Introduction

The recent "resolution revolution" in cryo electron microscopy (cryoEM)[1-9] has opened the door for high-resolution structure determination of a vast number of previously intractable biological systems. There is no need for crystallization, as samples for cryoEM are preserved in a frozen-hydrated state, randomly oriented within a 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 with cryoEM. 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[10], 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. However, this approach introduces an intriguing challenge: If we obtain a near-atomic (3.0–4.0 Å) resolution cryoEM map of a protein complex from a heterogeneous sample, how do we identify the protein(s) this map represents?

In instances where an unidentified protein is crystallized, the high purity (low complexity) of the sample and the high resolutions (1.5–2.5Å) of crystallographic density maps make identification of the unknown protein trivial[11-14]. However, identifying protein(s) in cryoEM maps from heterogeneous mixtures enriched directly from endogenous sources is extremely challenging, due to the large pool of potential candidates and the varying local resolutions and low overall resolutions of 3.0–4.0Å (compared to crystallography) typical of routinely achievable cryoEM maps. A systematic approach for identifying a complex from typical near-atomic (3.0–4.0Å) cryoEM maps would be transformative for cryoEM and systems biology, opening the door for structure determination of novel, unidentified molecules and complexes enriched directly from endogenous sources.

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. 1). This workflow employs our program, *cryoID*, to semi-autonomously identify proteins in cryoEM maps at better than 4.0Å resolution 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[15]. By directly imaging components of the parasite cell lysate, we obtained multiple near-atomic resolution structures of protein complexes implicated in the pathogenesis of malarial parasites, from a single cryoEM dataset. We then used *cryoID* to unambiguously assign side chains and identify the complex, enabling atomic model building.

## Results

### 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 of endogenous sources (Fig. 1a).

**Step 2: Sample evaluation.** We then assess the complexity of each fraction by SDS-PAGE and negative stain EM (Fig. 1b-c, Supplementary Figure 1).

**Step 3: Mass spectrometry.** Promising fractions containing uniform particles in negative stain EM (Supplementary Figure 1) 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. 1d).

**Step 4: cryoEM imaging.** Each promising fraction is imaged on a Titan Krios cryo-electron microscope, generally yielding datasets containing several distinct protein complexes in each image (Fig. 1e). To deconvolute mixtures of several distinct protein complexes within a single dataset and resolve them into multiple three dimensional (3D) structures, we leverage cryoSPARC's ability to perform unsupervised *ab initio* 3D classification and refinement[7], 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 can 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. 1f). As some amino acid side chains can look quite similar in 3.0–4.0Å cryoEM maps, we incorporated a certain degree of error tolerance into *cryoID* by introducing our novel, "degenerate" six-letter code that clusters the 20 amino acid residues into six simplified groups, based on the similarity of their side chain densities in typical cryoEM density maps (Fig. 2).

### cryoID

*cryoID*, a key component of the above workflow, determines the unique identity of the protein(s) in a near-atomic resolution (better than 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 side chain densities varies with map quality and local resolution(s), which can fluctuate widely across a cryoEM map; 2) Small and medium size residues can be difficult to differentiate accurately even in promising regions of the map, as their shapes are often less distinctive and can be context dependent. Aspartate and asparagine, for example, which have side chains of a similar size and shape and are often difficult to distinguish without prior knowledge of the primary sequence.

We designed *cryoID* to emulate strategies used by human modelers in *de novo* model-building to overcome these challenges (detailed description in Supplementary Text 1), resulting in a workflow consisting of four main tasks (Fig. 1f). ***1-Selection:*** Identifying high resolution segments of the map with a continuous backbone and clearly distinguishable side chain densities. ***2- Prediction:*** Automatically tracing the polypeptide backbone for each map segment and semi-automatically predicting the identities of the side chains for each residue in the segment, yielding a predicted primary sequence for the segment. ***3-Simplification:*** Translating both the cryoEM map segment sequences and the pool of candidate protein sequences into our novel simplified "six-letter" code (Fig. 2). By using the six-letter code, we introduce a redundancy that imparts a degree of tolerance for errors made by *cryoID* during sequence prediction and eliminate the need for *cryoID* to differentiate between side chains of similar size and shape, which are often indistinguishable in typical cryoEM density maps. ***4-Searching:*** Performing a customized BLASTP[16] search of the entire pool of candidate proteins using the previously predicted primary sequences for the cryoEM map segments as queries, yielding an expectation, or E-value, that indicates the similarity between the query sequence and the candidate protein (Fig. 3). The E-values for each query against a single candidate protein are combined to yield a composite E-value, which indicates the likelihood of the candidate protein being the correct protein. These four tasks are accomplished by two *cryoID* sub-programs, ***get_queries*** (**Selection** and **Prediction**) and ***search_queries*** (**Simplification** and **Searching**). Both sub-programs can be executed automatically with default parameters. Upon completion of the *get_queries* sub-program, *cryoID* launches the external visualization program Coot, which allows users to inspect, select, and, if necessary, correct obvious mistakes (*e.g.*, mistaking G for R) in the queries generated by get_queries (Supplementary Fig. 2). (This manual intervention could potentially be eliminated in the absence of obvious mistakes.)

We show, in a series of benchmarking experiments described below, that it is possible to uniquely identify the protein in a cryoEM map by using multiple simplified sequences of sufficient length from a single map to search a large (100s to 100,000s) pool of candidates.

## Benchmarking *cryoID* using simulated data

Parameters such as the number of queries, length of each query, and number of errors in each query influence the ability of *cryoID* to arrive at a unique answer. We 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, 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 1. To cover a wide range of different sequences with varying amino acid compositions and achieve a statistical significance (p-value < 0.01), we tested each condition (*m,n*) 1000 times. To achieve this, we randomly generated multiple unique sets of queries from a full length *P. falciparum* protein sequence for each condition (*m,n*), then used *cryoID* to simplify and search each set of queries against a pool of 880 full length *P. falciparum* protein sequences, also simplified. For each set of queries, *cryoID* sorted protein candidates in the pool by their composite E-values and monitored their % identity with the queries (Supplementary Fig. 3). We repeated this process with additional full length *P. falciparum* proteins until we had tested 1000 unique sets of queries for each condition (*m,n*).

Observing the efficacy of each combination of *m* and *n* from this simulated dataset allowed us to empirically determine the optimal query conditions (*m,n*) under which *cryoID* can correctly identify the unique protein matching the queries (Table 1). 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 (Supplementary Fig. 3b-d). For suboptimal query conditions (*m,n*) in which the query sequences were not long or numerous enough, *cryoID* identified multiple protein matches that exhibited 100% identity with the queries, making it impossible to obtain a unique protein ID (Table 1, Supplementary Fig. 3a).

For *cryoID* runs under optimal query conditions (*m,n*), we observe that there is always a clear "gap" in % identity between the correct protein candidate (100% identity with queries) and the next closest matching candidate (Supplementary Fig. 3b-d). This gap increases as *m* or *n* increase. This observation agrees with the theoretical estimation that, given a query set (*m,n*), the probability of achieving high identity with a candidate in the pool by chance can be approximated by $(\frac{1}{6})^{m*n}$. As such, the potential for false positive matches decreases exponentially as *m* and *n* increase. Therefore, we reasoned that it would still be possible for *cryoID* to determine a unique protein ID from queries containing a limited number of errors, provided the number of errors remain less than the number of differences between the next closest matching candidate and the error-free queries.

By running each condition (*m,n*) enough times to achieve statistical significance ($\alpha < 0.01$), 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 10%, 20%, 30%, and 40% errors in the queries.

These results are displayed in Supplementary Table 1, which serves as a reference table for deciding the optimal number and length of queries users should aim for. The table describes the minimum average query length (*n*) required for *cryoID* to identify the correct protein using a given number of queries (*m*), as well as the maximum percentage of errors that can be tolerated for each query condition (*m,n*). The table is based on our protein candidate pool, containing 880 proteins (~750,000 amino acid residues).

## 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, following our workflow detailed in Figure 1f.

We selected two published cryoEM structures within our target resolution range (3.0–4.0Å) from the EMDB, a 3.4Å structure of human gamma-secretase (EMD-3061)[17], and a 3.6Å structure of the *D. melanogaster* NOMPC mechanotransduction channel (EMD-8702)[18]. 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 the cytosolic and transmembrane regions of gamma-secretase. *cryoID* successfully identified the correct protein in the NOMPC map from both a more limited candidate pool consisting of the 3,500 proteins in the *D. melanogaster* proteome, as well as a much larger candidate pool containing all ~560,000 reviewed proteins in the UniProt database (Supplementary Table 2). c*ryoID* also successfully identified the correct proteins in both regions of the human gamma-secretase map from which we generated query sets, from both a candidate pool consisting of the 20,397 proteins in the *H. sapiens* proteome, as well as against the entire ~560,000 reviewed proteins in the UniProt database (Supplementary Table 3-4). For regions of the map corresponding to the other two components of gamma-secretase, *cryoID* generated 1–2 short queries that were insufficient for unambiguous identification of the corresponding component, due to poor map quality. In partial identification cases like this, it is likely that these queries belong to an unidentified component of the complex that is flexibly attached, and only the region that is anchored firmly to the complex is resolved to high resolution. If the flexibly attached component is ordered, performing focused classification and refinement techniques to improve the local resolution of the cryoEM map in this region before applying *cryoID* to identify the unknown component may help. Indeed, in the case of human gamma-secretase, applying *cryoID* to subsequently improved cryoEM maps[19] yielded the successful identification of the previously unidentified components. If this approach fails, the flexibly attached component may be intrinsically disordered. In this case, the pool of potential candidates can be narrowed down to a short-list consisting only of proteins that contain intrinsically disordered domains and are known to interact with the components of the complex previously identified by *cryoID*.

Thus, using published cryoEM maps from the EMDB, we determined that *cryoID* can consistently determine the correct protein(s) in cryoEM maps from entire proteomes ranging anywhere from 1000–20,000 proteins in size. We further found that for the two published cryoEM maps used here, cryoID was capable of identifying the correct protein(s) from a much larger candidate pool consisting of the entire ~560,000 proteins in the UniProt database.

## Application of endogenous structural proteomics workflow to *P. falciparum*

We 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[20-22], bearing no similarity to existing structures in the PDB. 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[10, 23].

We enriched for protein complexes ranging from 100kDa to ≥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 (Supplementary Figure 1) yielded multiple near-atomic resolution cryoEM density maps at an overall resolution ranging from 3.2–3.6Å, including two unknown protein complexes, as well as two distinct conformations of the 20S proteasome that exhibit marked differences from the previously published *P. falciparum* 20S proteasome structure[24, 25] (Fig. 4, Supplementary Figure 4). Mass spectrometry identified a candidate pool of 773 proteins in the fraction (Supplementary Table 5).

We analyzed each of the unknown maps in *cryoID* following the workflow detailed in Figure 1f, generating a single query set per map. In each case, *cryoID* successfully identified the correct protein in the map from the candidate pool, enabling us to build atomic models of the two protein complexes.

**Cross-validation against Pre-existing Crystal Structure of the *P. falciparum* M18 Aspartyl Aminopeptidase.—**The protein in the first 3.2Å cryoEM map was identified to be *P. falciparum* M18 aspartyl aminopeptidase, a 788kDa homo-dodecameric complex with tetrahedral symmetry (Fig. 4a, Supplementary Table 6, Supplementary Video 1). Our *de novo* structure agrees extremely well with the previously reported X-ray crystallographic structure of this complex[26] (Fig. 5a), with both the regulatory (residues 1–92 and 307–577) and catalytic (residues 92–306) domains clearly visible in all subunits (Fig. 5b). As such, the previously published crystal structure serves as a gold standard validation of our method.

**Structure of *P. falciparum* Glutamine Synthetase Reveals New Structural Features Unique to *Plasmodium*.—**The protein in the second 3.2Å 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. 4e, Supplementary Table 7, Supplementary Video 2). This enzyme catalyzes the condensation of glutamate and ammonia into glutamine in an ATP-dependent manner. 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 (either $Mg^{2+}$ or $Mn^{2+}$)[27].

Our *de novo* atomic model from the cryoEM map is similar throughout most of the structure (RMSD 1.5Å) to the previously published atomic model of a close homolog from *S. enterica*, solved by X-ray crystallography to 2.67Å resolution[28] (Supplementary Fig. 5). In particular, the three substrate-binding pockets in the active site are well-conserved (Fig. 5c-e).

However, we do observe one major difference between the two structures (Fig. 5c-e): 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 (Fig. 5c). In the corresponding location in our structure from *P. falciparum*, there is an extra 50-residue insertion here that forms a long loop that folds down along the outside of the structure, in the opposite direction from the active site (Fig. 5e).

**Structures of *P. falciparum* 20S Proteasome Reveal New Conformational States.—**Analysis of the most abundant particles in our dataset yielded two distinct cryoEM maps at 3.48Å and 3.67Å, respectively, which *cryoID* successfully identified to be the *P. falciparum* 20S proteasome in two distinct conformations (Supplementary Fig. 4). All 28 subunits were well-resolved in the 3.48Å structure of the first conformation. However, the β4 and β5 subunits in the 3.67Å structure of the second conformation are disordered, and thus only visible when the map is filtered to lower resolution, suggesting that these subunits are flexible (Supplementary Fig. 4g,h). As such, we named this second structure the 20S proteasome-$\beta4\beta5_{\text{flex}}$. Comparison of our structures with the previously published *P. falciparum* 20S proteasome structure[24, 25] revealed significant differences in both of our new structures, the most striking of which is the disordered β4 and β5 subunits in the 20S proteasome-$\beta4\beta5_{\text{flex}}$ conformation structure.

## Discussion

We have presented a workflow for a bottom-up approach to endogenous structural proteomics that uses cryoEM in combination with mass spectrometry to obtain near-atomic resolution structures of native, untagged protein complexes enriched directly from endogenous sources. We demonstrated that *cryoID*, an essential component of the workflow, can successfully identify the protein(s) in better than 4.0Å resolution cryoEM maps of unidentified protein complexes without prior knowledge of the primary sequence(s). This workflow allows us to determine the structures of complexes that are difficult or impossible to obtain using conventional recombinant methods. As a proof of principle, we have shown the successful use of this workflow to obtain near-atomic resolution (3.2Å) structures of multiple protein complexes enriched directly from unmodified *P. falciparum* parasites.

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"[29, 30]. The current state-of-the-art for direct visualization of biological structures *in vivo* involves cryo electron tomographic (cryoET) imaging and reconstruction of thin sections, called lamella, of intact cells, created using a focused ion-beam scanning electron microscope (FIB-SEM)[31-33]. Although cryoFIB milling can be used to overcome the fundamental limitation in the thickness of samples amenable to imaging by cryoET, both cryoET and cryoFIB milling are laborious, technically challenging, and far from routine, typically yielding at most 10 lamellae per day, each encompassing a mere 1–2$\mu m^2$ of imageable area per cell. As such, obtaining enough lamella tomograms to accumulate the half-million particles of a macromolecular complex needed to achieve near-atomic resolution is currently impractical.

The bottom-up endogenous structural proteomics approach presented here represents an immediate step toward direct visualization of native protein complexes as they exist in the cellular milieu at near-atomic resolution. This workflow allows us to determine the structures of complexes that are difficult or impossible to obtain using conventional recombinant methods. By using minimally disruptive, tag-free techniques that avoid over-purification, we are able to enrich large multi-protein assemblies as they exist in vivo. There are an enormous number of organisms that have yet to be fully explored, and the majority of extant structures even from well-studied organisms are derived from recombinant systems. As such, notwithstanding challenges in dealing with low abundance or flexible species, our approach opens the door for structural study of endogenous protein complexes, enabling direct observation at near-atomic resolution of previously unidentified novel protein complexes captured in multiple native conformational states (*e.g.,* changes in binding partners or cycling through sub-complexes like the spliceosome[34-40]) and during various stages of biological processes (*e.g.,* parasite life cycles, progression of cancers or cardiovascular diseases).

## Materials and Methods

### Parasite culture

*P. falciparum* cultures were prepared as described previously[10].

### Sucrose gradient fractionation of *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 (Supplementary Table 5). 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[8, 9] to assess the presence of sufficient numbers of intact particles yielding class averages exhibiting distinct features. For example, various symmetries could be recognized in top and side views (Fig. 1e, Supplementary Figure 1e).

## Mass Spectrometry

Selected fractions of interest were digested using trypsin as previously described[41]. Digested samples were then fractionated online using reversed-phase chromatography and analyzed by tandem mass spectrometry on a Q-Exactive mass spectrometer[42]. Data were analyzed on the IP2 software platform, which utilizes ProLuCID for database searching and DTASelect for filtering with decoy-database estimated false discovery rates[43, 44]. The proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the MassIVE partner repository with the dataset identifier PDX014263.

## Cryo Electron Microscopy

3μl aliquots of fractionated lysate 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. CryoEM grids were screened in an FEI Tecnai TF20 transmission electron microscope to optimize 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 [45]]. Aligned and un-dose-weighted micrographs were also generated and used for contrast transfer function (CTF) estimation using CTFFIND4 [Ref [46]], for particle picking by hand, and for particle picking using Gautomatch[47].

328,435 particles were extracted from 2,514 micrographs. After several rounds of reference-free two-dimensional (2D) classification in RELION, junk particles were excluded. 47,159 particles belonged to “good” 2D class averages that exhibited clear secondary structure features and resembled typical[24] 20S proteasome top and side views. Upon finer classification, two distinct side views and two distinct top views emerged, suggesting the presence of multiple native conformations. All 47,159 particles were subjected to an unsupervised single-class *ab initio* 3D reconstruction followed by a two-class heterogenous refinement using C1 symmetry in CryoSPARC. The resulting two reconstructions were then individually subjected to homogenous refinement using C2 symmetry in CryoSPARC, yielding two distinct structures at a final overall resolution of 3.48Å (from 24,788 particles) and 3.67Å (from 16,588 particles) respectively (Supplementary Figure 4).

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[7], yielding two 3.2Å *ab initio* 3D maps.

Further refinement of the particles in any of the four abovementioned structures in RELION failed to yield any improvement in resolution.

## Manual Model Building and Refinement

Map interpretation was performed with UCSF Chimera[48] and COOT[49]. *P. falciparum* protein sequences were obtained from the National Center for Biotechnology Information (NCBI)[50] and the PlasmoDB[51] 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 [52]] 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 of the monomer 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[53] 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[13].

All figures and videos were prepared with UCSF Chimera, Pymol[54], and Resmap[55]. Molprobity was used to validate the stereochemistry of the final models.

## Building *cryoID*

We developed the Python-based *cryoID* program (Fig. 1f) using the scientific Python development environment Spyder. *cryoID* consists of two main subprograms, *get_queries* and *search_pool.* The subprogram *get_queries* 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 side chain densities. It then automatically traces the polypeptide backbone for each map segment and semi-automatically predicts the identities of the side chains for each residue in the segment, yielding a predicted primary sequence for the segment. The subprogram *search_pool* performs the Simplification and Search functionalities of *cryoID*, translating both the cryoEM map segment sequences (*i.e.,* queries) and all the primary sequences of the pool of candidate proteins into a simplified "six-letter" code and performing a modified blast search of the entire pool of candidate proteins using the predicted cryoEM map segment sequences as queries. Both *get_queries* and *search_pool* sub-programs of *cryoID* are detailed below.

We also developed a graphical user interface (GUI) for *cryoID* using the cross-platform *Qt* GUI toolkit and its Python binding, *PyQt*. The two subprograms can be accessed either *via* the GUI, or from the command line. *cryoID* is an open source program under the MIT license, available for download at github (https://github.com/EICN-UCLA/cryoID).

## Selection and Prediction Using *get_queries*

The first *cryoID* subprogram, *get_queries,* generates multiple query sequences from the cryoEM density map. *cryoID* automatically assesses the backbone continuity when generating queries in the following manner: *cryoID equence_from_map* function in the Phenix software package, which automatically builds a model, typically composed of many fragments, into the cryoEM density. First, regular secondary structure (helices, strands) are identified. Then the contour level in the map is initially set to a very high level and is gradually lowered until connections appear between the secondary structure elements. If branching occurs, the process is stopped. This procedure defines the connectivity of a fragment and its ends. Once a fragment is identified, the chain is traced through it by following high density, Cα/Cβ positions are identified from bumps extending from the main chain in the density, and an all-atom model is constructed using Pulchra[56] (an existing fast all-atom reconstruction algorithm). Only the longest fragments are considered in the consequent sequence prediction and analysis (typically 15–50 residues in length). At each residue in each fragment, the most likely amino acid is identified by comparison of side chain density in the cryoEM map at this position with a library of side-chain densities extracted from a large set of cryoEM maps and models. This results in a predicted sequence for each fragment.

Upon completion, two files are generated: a FASTA file containing the query sequences and a corresponding .pdb file containing the atomic coordinates for each sequence.

*get_queries* then calls COOT to open the density map with the pdb file for user inspection, and if necessary, manual deletion of poor segment sequences, and correction of incorrectly assigned residues to the right group (Supplementary Fig. 2) and unidentifiable residues to “X” (which is designated with the residue type “MSE”) using the “mutate residue” tool in COOT. The modified pdb file is then saved and passed on to the second cryoID subprogram as the query set. Experienced users can also directly modify the sequences in the FASTA file, using the segment density as a reference.

The *get_queries* subprogram requires three user-provided inputs: the cryoEM density map filename, high resolution limit for map analysis, and symmetry of 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 *get_queries* subprogram should complete relatively quickly, generally within one hour, depending on factors such as map size and symmetry. For reference, processing of the glutamine synthetase and M18 aspartyl aminopeptidase maps was completed in 5–10min, while processing of the NOMPC and human gamma-secretase maps were completed in ~50min on a single CPU core. The version of Phenix utilized by *cryoID* in all the benchmarking results was Phenix-1.14–3260.

## Simplification and Searching Using *search_pool*

The second *cryoID* subprogram, *search_pool*, performs alignments of the query sets generated by the *get_queries 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 *get_queries cryoID* subprogram. In the latter case, the sequence information is extracted by calling PHENIX.*print_sequence* in the *search_pool* subprogram. 2) a FASTA file containing a list of the proteins in the user-defined candidate protein pool. The *search_pool* subprogram then translates both the query and candidate sequences into the simplified six-letter code (Fig. 2a).

Once this is accomplished, the *search_pool* subprogram calls the widely distributed local alignment search tool BLASTP[16] 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 → P, L-like → Z, K-like → M, 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 (Fig. 3b). 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 adjusting the penalty scores.

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). If the polarity of the query sequence (*i.e.*, N/C termini) is unknown, users can add a "polarity unknown" flag (-r) in the GUI options so that the program will try to align the query against the protein pool in both polarities. 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, 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 composite E-value of the search results of all queries, as defined below:

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

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

Finally, the program sorts all matched proteins based on the composite E-value, and the resulting list is saved in a summary file. The protein candidate with the smallest composite 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 Supplementary Table 1.

In rare instances where the query contains too many errors, the queries are too few, or the length too short, the matched protein with the smallest composite E-value is a false positive. False positive matches can be readily recognized either 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 (3.4A) and the *D. melanogaster* mechanotransduction channel NOMPC, which have global resolutions in our target resolution range (better than 4.0Å).

## *cryoID* and human gamma-secretase

Human gamma-secretase is a four-membered intramembrane protease consisting of presenilin, PEN-2, nicastrin and APH-1. We tested *Get_queries* for on the human gamma-secretase density map (*Homo sapiens,* EMD-3061, protein complex, no symmetry, and reported resolution of 3.4Å) with a symmetry of C1 and several different resolution inputs (3.0Å, 3.2Å and 3.4Å) and found the selected segments to be quite consistent. Running *get_queries* automatically with the resolution input of 3.2Å yielded the best set of query models, and was used for query generation. Upon completion of *get_queries*, we inspected the resulting queries with the density maps in the Coot pop-up window and observed the queries were localized to two distinct 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. *Get_queries* successfully generated pdb files with predicted query sequences for the three segments. We then manually inspected the query models, correcting residues incorrectly assigned by *Get_queries* 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 tested 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 ~560,000 reviewed 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 (Supplementary Table 3-4). Against the much larger ~560,000 protein candidate pool, *cryoID* correctly identified APH-1 (Q96BI3) as the correct protein for the transmembrane region query set (Supplementary Table 3), and Nicastrin as the correct protein for the extracellular region query set (Supplementary Table 4).

### *cryoID* and the NOMPC mechanotransduction channel

The NOMPC mechanotransduction channel from *D. melanogaster* is a homotetrameric integral membrane protein that mediates gentle-touch sensation. We tested *Get_queries* on the NOMPC density map (*D. melanogaster*, EMD-8702, C4 symmetry, reported resolution 3.6Å). The 3.2Å and C4 symmetry input parameters yielded one region in the transmembrane domain with clean, continuous backbone density throughout, from which *Get_queries* produced a set of three query sequences. We manually inspected the query models, correcting residues incorrectly assigned by *Get_queries* and extending the queries on both ends as the density permitted. This yielded the following degenerate sequences, which were then used for searching:

1. WGGXLYLGGYGGYLLGGGGGGGGLLGGLKGGGYXKG,
2. LLXGGGKYLGGXGGYGLGYG,
3. GGXYGGXGGGYGYGLGXGGG.

Using this set of query sequences, we tested 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 ~560,000 reviewed 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) (Supplementary Table 2).

### 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 four maps at 3.2Å, 3.2Å, 3.48Å, and 3.67Å overall resolutions. As a control for *cryoID*, we independently identified the proteins in the two 3.2Å 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 773 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 (Unknown Protein 1), 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 Unknown Protein 1 cryoEM map *de novo* (*i.e.* from scratch, building one residue at a time), using only 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 = 0.55Å).

In the case of our map that was ultimately determined to be glutamine synthetase (Unknown Protein 2), 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 Unknown Protein 2 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 *Get_queries* on the M18 aspartyl aminopeptidase density map, named Unknown Protein 1 (*P. falciparum*, Unknown Protein 1, T symmetry, reported resolution 3.2Å), using a symmetry of T and a high resolution limit of 3.2Å as the initial input parameters. 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 *Get_queries* 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 773 proteins identified in this sucrose gradient fraction by mass spectrometry. *cryoID* correctly identified the corresponding protein for the query set generated from the Unknown Protein 1 cryoEM map, making a unique protein ID of M18 aspartyl aminopeptidase from *P. falciparum* (Q8I2J3) (Supplementary Table 6). 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 0.55Å), serving as a gold standard validation of our workflow and *cryoID*'s performance.

We then tested *Get_queries* on the glutamine synthetase density map, named Unknown Protein 2 (*P. falciparum,* Unknown Protein 2, D6 symmetry, reported resolution 3.2Å), using a symmetry of D6 and a high resolution limit of 3.2Å as the initial input parameters. We tuned the resolution parameter according to local resolution estimates[57] and ultimately found 3.0Å yielded the best results. We then manually inspected the query models, correcting residues incorrectly assigned by *Get_queries* 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 tested the ability of *cryoID* to correctly identify the corresponding protein, from a candidate pool consisting of the 773 proteins identified in this sucrose gradient fraction by mass spectrometry. *cryoID* correctly identified the corresponding protein for the query set generated from the Unknown Protein 2 cryoEM map, making a unique protein ID of glutamine synthetase from *P. falciparum* (C0H551) (Supplementary Table 7). 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 (Fig. 5c-e).

Finally, we used *cryoID* to successfully identify the 3.48Å and 3.67Å maps to be the *P. falciparum* 20S proteasome, from a candidate pool containing the entire *P. falciparum* proteome (Supplementary Tables 6 and 7). Furthermore, *cryoID* readily identified and distinguished between individual 20S proteasome subunits, enabling us to fit the previously published *P. falciparum* 20S proteasome model (PDB 5FMG)[24] into each of our two maps. Each residue in the model was then manually refit into each of our two maps in COOT to optimize geometry and fit. We were also able to build a number of previously unmodeled sections in many of the α and β subunits in both structures, thanks to improvements in local resolution or conformational differences in several regions throughout both maps. Comparison of our structures with the previously published *P. falciparum* 20S proteasome structure revealed several major differences, the most striking of which is the disordered β4 and β5 subunits in our 3.67Å structure of the 20S proteasome-β4$\beta5_{flex}$ conformation.

## 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 6PEV, 6PEW, EMD-20333, EMD-20334, respectively. The proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the MassIVE partner repository with the dataset identifier PDX014263.

## Code Availability

*CryoID* is an open source program under the MIT license, available for download at github (https://github.com/EICN-UCLA/cryoID).

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

This research was supported in part by grants from National Institutes of Health (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 resources 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).

## References

1. Cheng Y. Single-Particle Cryo-EM at Crystallographic Resolution. Cell 161, 450–457, doi:10.1016/j.cell.2015.03.049 (2015). [PubMed: 25910205]
2. Li X 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). [PubMed: 23644547]
3. Liu H 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). [PubMed: 20798312]
4. McMullan G, Chen S, Henderson R &Faruqi AR. Detective quantum efficiency of electron area detectors in electron microscopy. Ultramicroscopy 109, 1126–1143, doi:10.1016/j.ultramic.2009.04.002 (2009). [PubMed: 19497671]
5. Clough GMRN, Kirkland AI. in Journal of Physics: Conference Series Vol. 522 (IOP Publishing, 2013).
6. Zhang X, Jin L, Fang Q, Hui WH &Zhou ZH. 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). [PubMed: 20398923]
7. Punjani A, Rubinstein JL, Fleet DJ & Brubaker MA. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat Methods 14, 290-+, doi:10.1038/Nmeth.4169 (2017). [PubMed: 28165473]
8. Scheres SHW. A Bayesian View on Cryo-EM Structure Determination. J Mol Biol 415, 406–418, doi:10.1016/j.jmb.2011.11.010 (2012). [PubMed: 22100448]

9. Scheres SHW. 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). [PubMed: 23000701]
10. Ho CM et al. Malaria parasite translocon structure and mechanism of effector export. Nature 561, 70–75, doi:10.1038/s41586-018-0469-4 (2018). [PubMed: 30150771]
11. Niedzialkowska E et al. Protein purification and crystallization artifacts: The tale usually not told. Protein Sci 25, 720–733, doi:10.1002/pro.2861 (2016). [PubMed: 26660914]
12. Osipiuk J, Walsh MA &Joachimiak A. Crystal structure of MboIIA methyltransferase. Nucleic Acids Res 31, 5440–5448, doi:10.1093/nar/gkg713 (2003). [PubMed: 12954781]
13. Adams PD et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D 66, 213–221, doi:10.1107/S0907444909052925 (2010). [PubMed: 20124702]
14. Porebski PJ, Cymborowski M, Pasenkiewicz-Gierula M &Minor W. Fitmunk: improving protein structures by accurate, automatic modeling of side-chain conformations. Acta Crystallographica Section D-Structural Biology 72, 266–280, doi:10.1107/S2059798315024730 (2016).
15. PDB R. PDB Statistics: Overall Growth of Released Structures Per Year, <https://www.rcsb.org/stats/growth/overall> (2018).
16. Altschul SF, Gish W, Miller W, Myers EW &Lipman DJ. Basic local alignment search tool. J Mol Biol 215, 403–410, doi:10.1016/S0022-2836(05)80360-2 (1990). [PubMed: 2231712]
17. Bai XC et al. An atomic structure of human gamma-secretase. Nature 525, 212–217, doi:10.1038/nature14892 (2015). [PubMed: 26280335]
18. Jin P et al. Electron cryo-microscopy structure of the mechanotransduction channel NOMPC. Nature 547, 118–122, doi:10.1038/nature22981 (2017). [PubMed: 28658211]
19. Zhou R et al. Recognition of the amyloid precursor protein by human gamma-secretase. Science 363, doi:10.1126/science.aaw0930 (2019).
20. Gardner MJ et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419, 498–511, doi:10.1038/nature01097 (2002). [PubMed: 12368864]
21. Hall N 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). [PubMed: 15637271]
22. Waters AP. 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). [PubMed: 18474354]
23. Newby ZE et al. Crystal structure of the aquaglyceroporin PfAQP from the malarial parasite Plasmodium falciparum. Nat Struct Mol Biol 15, 619–625, doi:10.1038/nsmb.1431 (2008). [PubMed: 18500352]
24. Li H et al. Structure- and function-based design of Plasmodium-selective proteasome inhibitors. Nature 530, 233–236, doi:10.1038/nature16936 (2016). [PubMed: 26863983]
25. Xie SC et al. The structure of the PA28–20S proteasome complex from Plasmodium falciparum and implications for proteostasis. Nat Microbiol, doi:10.1038/s41564-019-0524-4 (2019).
26. Sivaraman KK 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). [PubMed: 22709581]
27. Eisenberg D, Gill HS, Pfluegl GM &Rotstein SH. Structure-function relationships of glutamine synthetases. Biochim Biophys Acta 1477, 122–145 (2000). [PubMed: 10708854]
28. Gill HS &Eisenberg D. The crystal structure of phosphinothricin in the active site of glutamine synthetase illuminates the mechanism of enzymatic inhibition. Biochemistry-Us 40, 1903–1912 (2001).
29. Asano S, Engel BD &Baumeister W. 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). [PubMed: 26456135]
30. Beck M &Baumeister W. 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). [PubMed: 27671779]

31. Albert S 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). [PubMed: 29229809]
32. Mahamid J et al. Visualizing the molecular sociology at the HeLa cell nuclear periphery. Science 351, 969–972, doi:10.1126/science.aad8857 (2016). [PubMed: 26917770]
33. Mosalaganti S et al. In situ architecture of the algal nuclear pore complex. Nat Commun 9, 2361, doi:10.1038/s41467-018-04739-y (2018). [PubMed: 29915221]
34. Bai R, Wan R, Yan C, Lei J &Shi Y. Structures of the fully assembled Saccharomyces cerevisiae spliceosome before activation. Science 360, 1423–1429, doi:10.1126/science.aau0325 (2018). [PubMed: 29794219]
35. Zhang X et al. Structures of the human spliceosomes before and after release of the ligated exon. Cell Res 29, 274–285, doi:10.1038/s41422-019-0143-x (2019). [PubMed: 30728453]
36. Liu S et al. Structure of the yeast spliceosomal postcatalytic P complex. Science 358, 1278–1283, doi:10.1126/science.aar3462 (2017). [PubMed: 29146870]
37. Agafonov DE et al. Molecular architecture of the human U4/U6.U5 tri-snRNP. Science 351, 1416–1420, doi:10.1126/science.aad2085 (2016). [PubMed: 26912367]
38. Galej WP et al. Cryo-EM structure of the spliceosome immediately after branching. Nature 537, 197–201, doi:10.1038/nature19316 (2016). [PubMed: 27459055]
39. Haselbach D et al. Structure and Conformational Dynamics of the Human Spliceosomal B(act) Complex. Cell 172, 454–464 e411, doi:10.1016/j.cell.2018.01.010 (2018). [PubMed: 29361316]
40. Nguyen THD et al. Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 A resolution. Nature 530, 298–302, doi:10.1038/nature16940 (2016). [PubMed: 26829225]
41. Kaiser P &Wohlschlegel J. Identification of ubiquitination sites and determination of ubiquitin-chain architectures by mass spectrometry. Methods Enzymol 399, 266–277, doi:10.1016/S0076-6879(05)99018-6 (2005). [PubMed: 16338362]
42. Kelstrup CD, Young C, Lavallee R, Nielsen ML &Olsen JV. Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. J Proteome Res 11, 3487–3497, doi:10.1021/pr3000249 (2012). [PubMed: 22537090]
43. Tabb DL, McDonald WH &Yates JR 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res 1, 21–26 (2002). [PubMed: 12643522]
44. Xu T et al. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. J Proteomics 129, 16–24, doi:10.1016/j.jprot.2015.07.001 (2015). [PubMed: 26171723]
45. Zheng SQ 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). [PubMed: 28250466]
46. Rohou A &Grigorieff N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. J Struct Biol 192, 216–221, doi:10.1016/j.jsb.2015.08.008 (2015). [PubMed: 26278980]
47. Zhang K. Gautomatch: a GPU-accelerated program for accurate, fast, flexible and fully automatic particle picking from cryo-EM micrographs with or without templates (2016).
48. Pettersen EF et al. UCSF chimera - A visualization system for exploratory research and analysis. J Comput Chem 25, 1605–1612, doi:10.1002/jcc.20084 (2004). [PubMed: 15264254]
49. Emsley P, Lohkamp B, Scott WG &Cowtan K. Features and development of Coot. Acta Crystallogr D 66, 486–501, doi:10.1107/S0907444910007493 (2010). [PubMed: 20383002]
50. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44, D7–D19, doi:10.1093/nar/gkv1290 (2016). [PubMed: 26615191]
51. Aurrecoechea C et al. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res 37, D539–D543, doi:10.1093/nar/gkn814 (2009). [PubMed: 18957442]
52. Kelley LA, Mezulis S, Yates CM, Wass MN &Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10, 845–858, doi:10.1038/nprot.2015.053 (2015). [PubMed: 25950237]

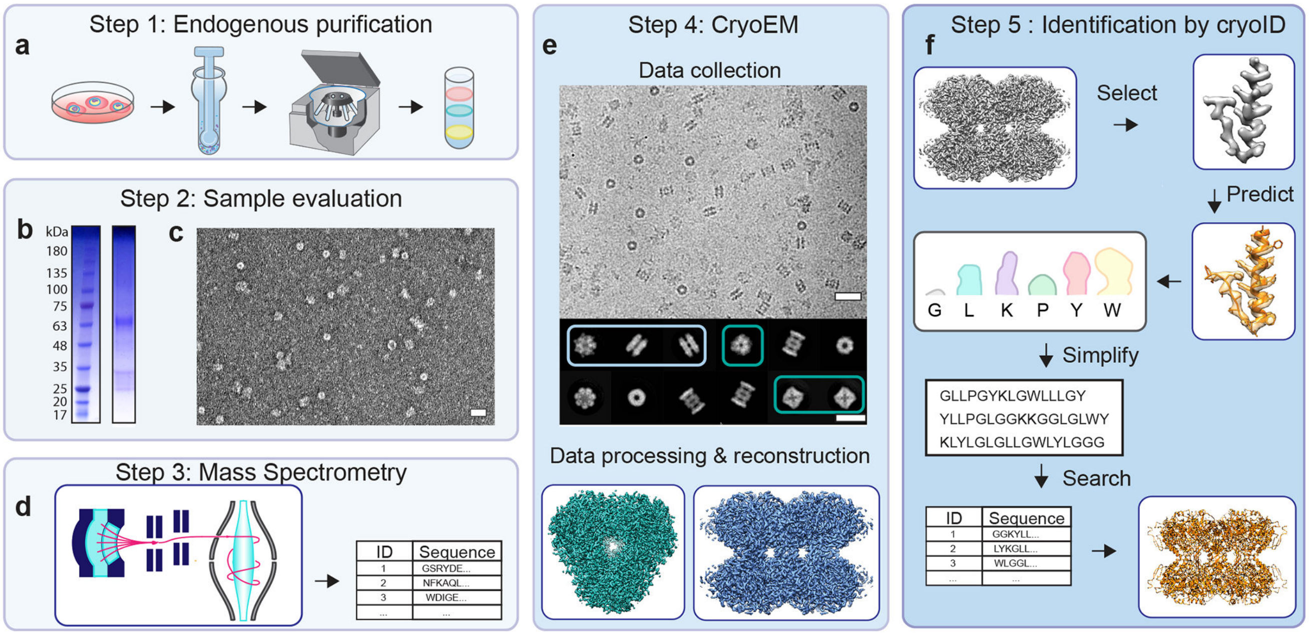
53. Chen VB et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D 66, 12–21, doi:10.1107/S0907444909042073 (2010). [PubMed: 20057044]

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

55. Kucukelbir A, Sigworth FJ &Tagare HD. Quantifying the local resolution of cryo-EMEM density maps. Nat Methods 11, 63-+, doi:10.1038/Nmeth.2727 (2014). [PubMed: 24213166]

56. Rotkiewicz P &Skolnick J. Fast procedure for reconstruction of full-atom protein models from reduced representations. J Comput Chem 29, 1460–1465, doi:10.1002/jcc.20906 (2008). [PubMed: 18196502]

57. Swint-Kruse L &Brown CS. Resmap: automated representation of macromolecular interfaces as two-dimensional networks. Bioinformatics 21, 3327–3328, doi:10.1093/bioinformatics.bti511 (2005). [PubMed: 15914544]


Step 1: Endogenous purification
a
Step 2: Sample evaluation
b
kDa
180
135
100
75
63
48
35
25
20
17
c
Step 3: Mass Spectrometry
d
Step 4: CryoEM
e
Data collection
Data processing & reconstruction
Step 5 : Identification by cryoID
f
Select
Predict
G L K P Y W
Simplify
Search

| ID | Sequence |
|---|---|
| 1 | GSRYDE... |
| 2 | NFKAQL... |
| 3 | WDIGE... |
| ... | ... |

```
GLLPGYKLGWLLLGY
YLLPGLGGKKGGLGLWY
KLYLGLGLLGWLYLGGG
```

| ID | Sequence |
|---|---|
| 1 | GGKYLL... |
| 2 | LYKGLL... |
| 3 | WLGGL... |
| ... | ... |

**Figure 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 identifies a list of all proteins in each fraction. **e,** cryoEM analysis yields near-atomic resolution cryoEM maps. Scale bars, 30 nm (micrograph, top); 10 nm (2D class averages, bottom). **f,** The proteins in the cryoEM maps are identified using *cryoID*.

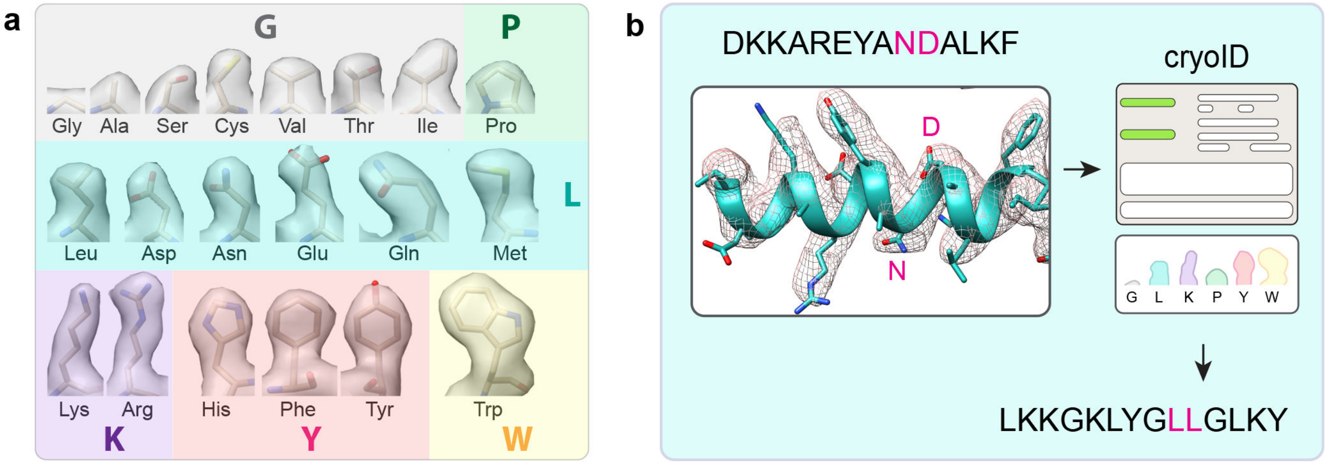
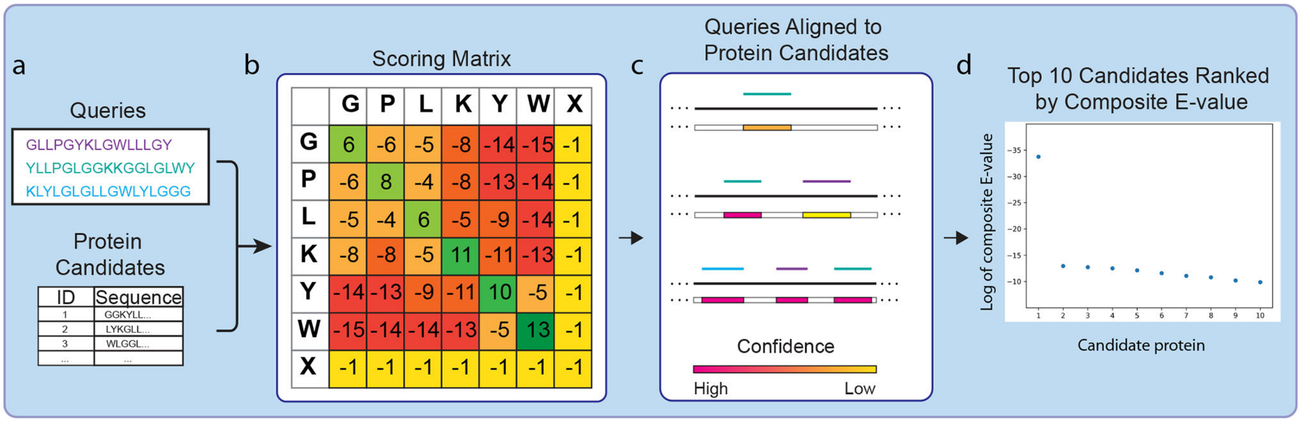


**Figure 2 |. Simplified 6-Letter Code.**

**a,**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. (*i.e.*, G group = small-size side chain density, L group = medium-size side chain density, K group = long and thin side chain density, P group = typical proline side chain density, Y group = long and bulky side chain density) **b,** *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 degenerate 6-letter code shown on the right.



GLLPGYKLGWLLLGY
YLLPGLGGKKGGLGLWY
KLYLGLGLLGWLYLGGG

| ID | Sequence |
|---|---|
| 1 | GGKYLL... |
| 2 | LYKGLL... |
| 3 | WLGGL... |
| ... | ... |

| | G | P | L | K | Y | W | X |
|---|---|---|---|---|---|---|---|
| G | 6 | -6 | -5 | -8 | -14 | -15 | -1 |
| P | -6 | 8 | -4 | -8 | -13 | -14 | -1 |
| L | -5 | -4 | 6 | -5 | -9 | -14 | -1 |
| K | -8 | -8 | -5 | 11 | -11 | -13 | -1 |
| Y | -14 | -13 | -9 | -11 | 10 | -5 | -1 |
| W | -15 | -14 | -14 | -13 | -5 | 13 | -1 |
| X | -1 | -1 | -1 | -1 | -1 | -1 | -1 |

**Figure 3 |. Searching in *cryoID*.**

**a,** *cryoID* runs alignments of the simplified query sequences obtained from the cryoEM maps against each protein in a pool of candidate proteins (also simplified). **b,** We created a customized alignment scoring matrix for *cryoID* by adapting the BLASTP PAM30 scoring matrix to work with the simplified 6-letter code used by *cryoID*. **c, d,** *cryoID* calculates a composite E-value for the alignment between each protein candidate against the query set, and then ranks the candidates by E-value to determine the most likely match to the query set.

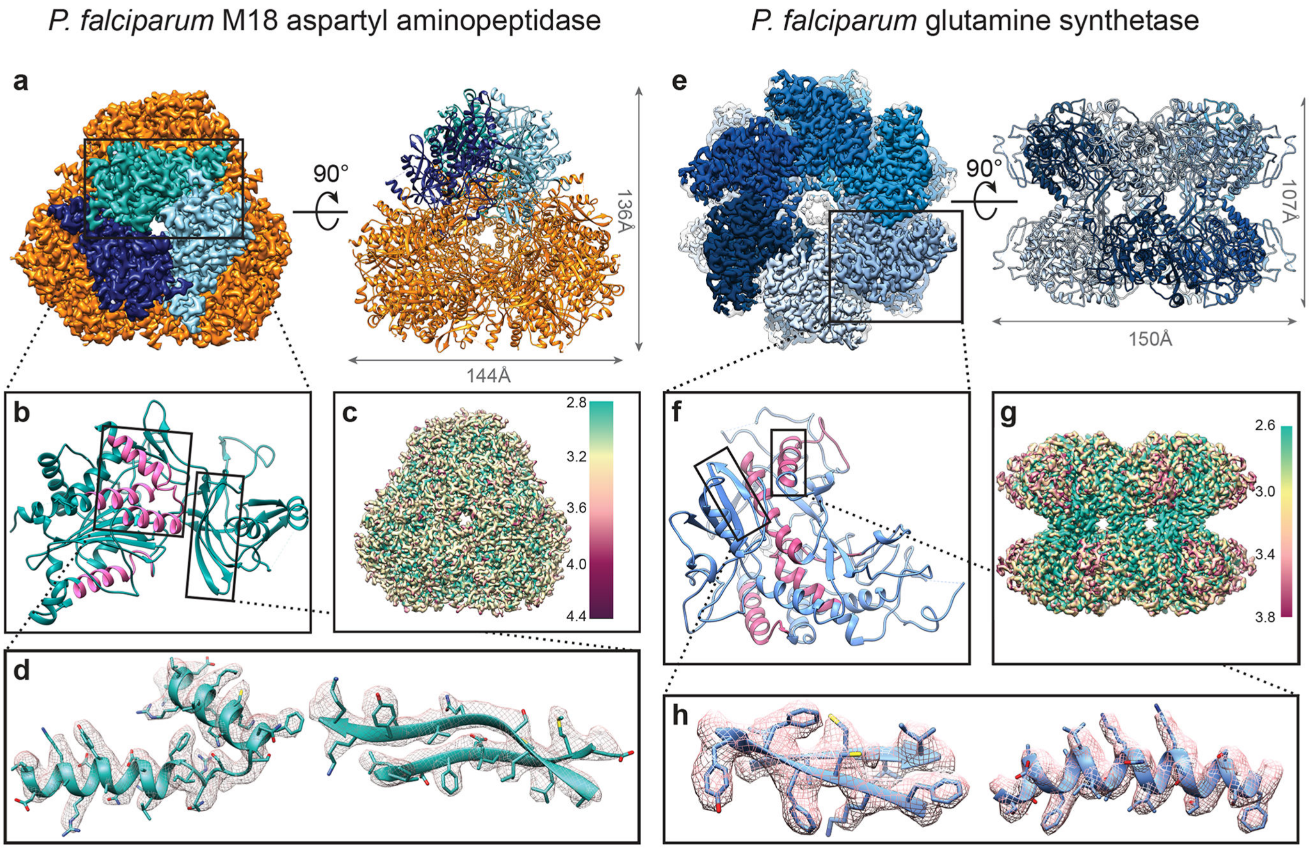

*P. falciparum* M18 aspartyl aminopeptidase
*P. falciparum* glutamine synthetase
a
90°
136Å
144Å
e
90°
107Å
150Å
b
c
2.8
3.2
3.6
4.0
4.4
f
g
2.6
3.0
3.4
3.8
d
h

**Figure 4 |. CryoEM structures of proteins enriched directly from *P. falciparum* parasite lysates.**

**a,e,** 3.2Å cryoEM density map and atomic model of *P. falciparum* M18 aspartyl aminipeptidase (**a**) and glutamine synthetase (**e**). **b,f,** Enlarged view of the *P. faciparum* M18 aspartyl aminopeptidase (**b**) and glutamine synthetase (**f**) monomer. Segments from which the queries for cryoID were generated are highlighted in pink. **c,g,** Local resolution (in Å) calculated using Resmap and two unfiltered halves of the reconstruction for *P faciparum* M18 aspartyl aminopeptidase (**c**) and glutamine synthetase (**g**). **d,h,** Detailed view of regions boxed in (**b & f**), displayed with corresponding cryoEM density.

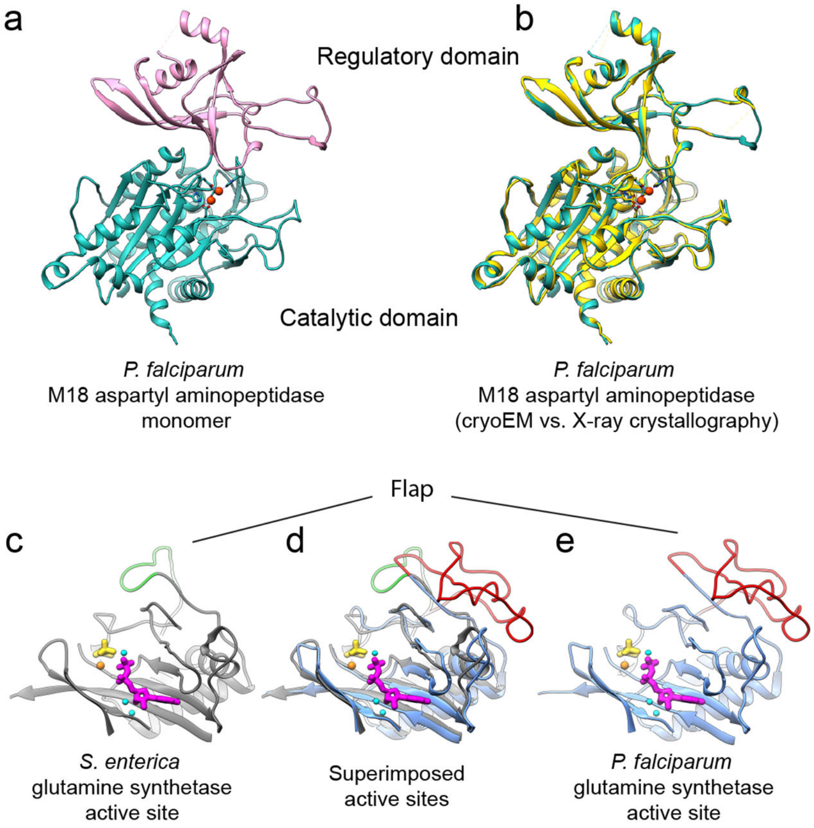


**Figure 5 |. Details of the M18 aspartyl aminopeptidase and glutamine synthetase monomers.**

**a,** 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, colored to indicate the regulatory (pink) and catalytic (sea green) domains. **b,** 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. **c,** The previously published structure of the *S. enterica* glutamine synthetase, solved using X-ray crystallography, colored dark grey. **d,** Our cryoEM structure of the *P. falciparum* glutamine synthetase, colored cornflower blue, is shown superimposed with the *S. enterica* glutamine synthetase crystallographic structure. The two structures align with an RMSD of 1.5Å. **e,** A single monomer from our atomic model of the *P. falciparum* glutamine synthetase, determined by cryoEM using our endogenous structural proteomics workflow, colored in cornflower blue. 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.

### Table 1 |

### Determining optimal parameters for Searching in *cryoID*.

The result for each query condition (m,n) shown here represents 1000 tests of that condition using unique sets of queries randomly generated from full length P. falciparum protein sequences. 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. One-sided statistics test: P(f) < 0.005, N = 1000, p-value < 0.0067, α < 0.01; where P(f) = probability of failure to obtain unique protein ID.

| **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 | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |