Cryo-EM Structure Determination of Crude Isolated Macromolecular Complexes from *P. furiosus*

Major Internship

Molecular and Cellular Life Sciences

Author

Marèl F.M. Spoelstra

(6101909)

Supervisor

Dr. Wenfei Song

Department of Chemistry – Structural Biochemistry – Cryo-EM

Examiners

Prof. dr. Friedrich G. Förster

Department of Chemistry – Structural Biochemistry – Cryo-EM

Dr. Richard A. Scheltema

Department of Chemistry – Biomolecular Mass Spectrometry and Proteomics

---

Educational Institution

Utrecht University

Date

16-05-2022

Universiteit Utrecht

Layman's Summary

Of the three domains of life, the Archaea show similarities with both Bacteria and Eukaryotes, like for transcription and translation. Studying these might shine a light upon the evolution of these important processes in the cell. Organelles in eukaryotes form membrane-enclosed, specialized environments in which specific tasks can take place. In prokaryotes, a similar phenomenon is observed, whereby a specialized environment is enclosed by a protein shell, the encapsulin. These are able among others to store iron. The encapsulins resemble viral capsids, and studying these encapsulins could give some hints about their function, but also about the evolution of viruses.

Furthermore, the archaeal domain comprises some very interesting species that live in extreme conditions. Some can thrive at high temperatures, even up to 100 degrees as in case of the hyperthermophiles. Therefore, their enzymes should remain stable and functional at these conditions. The first forms of life are suggested to have been hyperthermophilic. Studying enzymes of hyperthermophiles could give insight into enzyme evolution, enzyme stability, but also in the light of the global warming it might inform us how different species might deal with the increasing temperatures.

For studying these proteins, we want to determine their structure so we can see how the different amino acids are arranged and which interactions take place. This means we need high resolution at the near-atomic range, which is around the 2-4 Å, with one Ångström being a tenth of a billionth meter. As these particles are much smaller than the wavelength of light, we could not use ordinary light microscopy, but we need another "light" source, with much smaller wavelength, like electrons. To preserve the protein structure, and protect them from the harsh electrons, they are preserved in a thin layer of ice. Imaging results in 2D-projection images of the proteins, which are very noisy. Averaging a lot of 2D-images will average-out the noise, but strengthen the signal of the protein, thus increasing the signal-to-noise ratio (SNR). To obtain a three-dimensional structure of the protein, 2D-images from all different views, e.g. top and side views, need to be combined. For this, it is important to have a high purity sample with a high amount of the protein, to increase the SNR and to collect all the different views.

One of the most challenging tasks is the purification of proteins. That is why we try another approach whereby the cell extract of the hyperthermophile *Pyrococcus furiosus* is slightly purified, so the sample contains some large proteins. The different proteins will show up in our 2D-averages, after which we decide which views belong to which protein and sort them into the different protein groups. After this, the different proteins are processed further separately, to yield a three-dimensional structure of each protein.

This resulted in a large ribosomal subunit resolved to a resolution of 3.22 Å. As we know, the ribosome is made up of both ribosomal RNA, and ribosomal proteins, important for the enzymatic function and the stability, respectively. Our structure indicates the presence of two ribosomal proteins that have additional copies present, something that is not found in bacteria. For a third ribosomal protein, we didn't find the additional copy, although this was described previously, which could be due to the loss of the small subunit.

We also found a 20S proteasome, a protein that can be seen as the trash bin of the cell as it breaks down proteins. However, for this protein we still mis a lot of views and no 3D-image could be rendered, so more images should be collected.

For the encapsulins, we tried to get high purity samples, to be sure we are able to reach high resolution. In the end, we had a pure sample, but at a too low concentration, having not enough particles. This purification has to be repeated, making sure almost no protein will be lost during the different steps of purification.

Furthermore, the ribosome and the proteasome could be easily identified based on morphology, but we still have some interesting 2D-class averages of which we don't know to which protein they belong to. It would be interesting to also analyse our sample by another technique, which could indicate which proteins were present in our sample. Mass spectrometry would be a perfect technique to combine our approach with. This technique can identify proteins based on their mass.

(722 words)

Purification of Large Protein Complexes in *P. furiosus* using Crude Purification Methods

April 29th 2022

Marèl F.M. Spoelstra, BSc*, dr. Wenfei Song, prof. dr. Friedrich G. Förster

Bijvoet Centre for Biomolecular Research – In situ Structural Biology (Förster) at the Structural Biochemistry group, Utrecht University, Utrecht, The Netherlands

Keywords: Macromolecular Complexes, Ribosome, Proteasome, Crude Isolation, Sucrose Gradient, VLPs, Encapsulin, Archaea, *Pyroccocus furiosus*, Cryo-Electron Microscopy indicate the number of words: 9211

ABSTRACT Recent advantages in the cryo-EM field, for both hardware and software, made it able to obtain structures with near-atomic resolution . However, for Single Particle Analysis, protein purification remains the main bottleneck (Bai et al., 2015; Baker, 2018; Danev et al., 2019; Li et al., 2013; Yip et al., 2020). Quite recently, researchers tried a different kind of approach to elucidate high-resolution structures of different macromolecular complexes. For this, crude purification methods were used, whereby the mix of proteins are further purified *in silico* using several rounds of 2D- and 3D-classifications (Ho et al., 2020; Su et al., 2021; Verbeke et al., 2018). For this study, we will validate these kind of crude purification approaches for large macro-complexes found in the hyperthermophilic Archaeon *Pyrococcus furiosus*. Sucrose gradients are used to purify large complexes, like the 50S ribosome and 20S proteasome. We resolved a structure of the 50S ribosome to 3.22 Å resolution. Another large complex we are interested in, and which is purified to high purity, are the Virus Like Particles (VLPs). These are icosahedral structures, resembling viral capsids present in different Bacterial and Archaeal species, including the archaeon *Pyrococcus furiosus* (Akita et al., 2007; Namba et al., 2005). These structures are now known as encapsulins, large protein complexes that form nano-compartments to separate metabolic processes in the cell (Giessen & Silver, 2017). We optimized the purification strategy of these encapsulins to resolve the structure of these encapsulins with its cargo, to see which interactions it has with the shell. For these aims, a combination of biochemistry methods for purification, Cryo-EM and data processing to resolve the structures of these complexes, will be used to resolve these questions

INTRODUCTION

Cryogenic-electron microscopy (cryo-EM) is becoming the main technique for achieving high-resolution structures. With recent advancements like the direct-electron detectors, even near-atomic resolution can be obtained in the range of 2.0-4.0 Å resolution, enabling the building of the amino acid chains in the electron density, reaching similar performances as crystallography (Bai et al., 2015; Baker, 2018; Danev et al., 2019; Li et al., 2013; Yip et al., 2020).

Two commonly used techniques in the cryo-EM field are Single-Particle Analysis (SPA), and cryo-electron tomography (cryoET). For SPA single molecules are purified to a high concentration, and after plunge-freezing many copies of the molecule are available in random orientations. By using computational approaches, a 3D model of the molecule can be reconstructed. For cryoET a tilt-series of a thin sample, like a bacterial cell or for thicker cells a thin slice called a lamella, is obtained during imaging. This will give the different orientations for the same molecule, within it's cellular context, which is then reconstructed to a 3D image using back projection methods. With SPA it is possible to obtain near-atomic resolution, whereas for cryoET the resolution is in the (sub-) nanometer range (Danev et al., 2019).

For SPA, the purification is the major bottleneck in achieving high-resolution structures. Obtaining a pure sample, in a sufficient high concentration, while the particles are still intact, can be highly time-consuming, needing several rounds of optimization for the protocol to find out what will give the best results for a specific protein (Danev et al., 2019). Besides this, you loose information about it's cellular context, and important interaction partners.

Recently, significant improvement in the quality of the data was driven by the development of direct-electron detectors. The improvement of computational algorithms for reconstruction of 3D models from single particles also made it possible to get more out of the data. As processing programs, like Relion®, can deal much better with structural differences between particles during 2D- and 3D-classification, it is possible to deal with more heterogeneities in the sample, therefore needing less pure samples (Bai et al., 2015). Instead of purifying a single molecule, crude purification methods like sucrose gradients are used to obtain enriched protein complexes. Grids are made for these heterogeneous samples, and data is collected on the electron microscope. Besides this, also some of the sample can be analyzed by mass-spectrometry to identify the proteins and reveal possible interesting interaction partners. The micrographs contain a mixture of different protein complexes, which will be sorted into separate classes during several rounds of 2D- and 3D-classification, in principle doing a in silico purification.

We are interested in validating these crude purification approaches for macromolecular complexes in *Pyrococcus furiosus* a hyperthermophilic archaeon which belongs to the Euryachaeota phylum and grows in extreme conditions with temperatures near 100 degrees Celsius (Blumentals et al., 1990). Archaea belong to the prokaryotes, having no membrane enclosed nucleus, but showing distinct molecular features from bacteria. Archaea share some more complex molecular features with eukaryotes, and the origination of the three domains of life could be well understood by studying Archaea (Eme et al., 2017). Processes like transcription, translation, and secretion in Archaea show resemblances with both eukaryotes, as well as bacteria (Bell & Jackson, 1998; Bolhuis, 2004; Schmitt et al., 2020; Wenck & Santangelo, 2020). Hyperthermophiles show extreme enzyme stability and are able to function at high temperatures. As these enzymes are more stable, they can be used to study the function and structure of related, but relative instable eukaryotic proteins (Cavicchioli, 2011). Furthermore, according to current theories, the first forms of life would have been hyperthermophilic. Studying enzymes from the hyperthermophiles might shed light on enzyme evolution, and would also give an understanding about the molecular properties which give them their stability, which might be interesting for protein engineering (Vieille & Zeikus, 2001). Moreover, in the light of global warming, it might also give information about how different species might cope with the increasing temperatures (Barik, 2020).

Researchers already applied these kind of crude isolation approaches to different kind of samples, like the human cell extract, and also to some challenging samples to purify like proteins from the malaria parasite, or membrane proteins. From the human HEK239T cell extract the structures of the 26S proteasome and the HSP60 complex were elucidated (Verbeke et al., 2018). A 3.2 Å structure of crudely isolated glutamine synthetase and M18 aspartyl aminopeptidase from a malaria parasite were solved, for which they used Cryo-EM in combination with mass-spectrometry, along with a program that is able to identify the proteins from ab initio density map (Ho et al., 2020). A similar method was used, named the Build-and-Retrieve (BaR) method, to obtain high resolution structures for both soluble and membrane proteins in *E. coli*. For this, large macromolecular complexes (>100 kDa) were purified in silico with 2D classification, after which they were sorted by preliminary 3D classification and cleaned by several rounds of 2D and 3D classifications. Initial models were built and used as a template to retrieve all particles used for building the density maps. This resulted in the cytochrome b03 structure resolved to 2.20 Å resolution, the OmpF porin channel to 2.54 Å resolution, the succinate-coenzyme Q reductase to 2.50 Å resolution, and OmpC to 2.56 Å resolution. From the raw cell lysate was the catalase-peroxidase, and glutamate decarboxylase structure resolved to 2.17 Å and 2.90 Å resolution, respectively (Su et al., 2021).

We will be applying crude isolation methods to large macromolecular complexes. These would be excellent targets to validate these crude purification methods for obtaining high resolution structures. As these complexes are very large, they provide sufficient contrast for cryo EM in order to identify them on the grids (Henderson, 1995). Some of these complexes, like ribosomes and thermosomes are abundantly present in cells, which will make them easier to purify, identify on the grids, and acquire sufficient copies during imaging. We will use a 20%-60% (w/v) sucrose gradient to isolate different protein fraction. Additionally, we will apply a 30%-60% w/w and w/v sucrose gradients, for which we will combine different protein fractions, and use *in silico* purification.

One of the interesting proteins we encountered are the encapsulins, which were already seen in 2005 as icosahedral structures with electron microscopy inside *Pyroccocus furiosus* (Namba et al., 2005). These icosahedral structures have a spherical shape and a diameter around the 30 nm (Namba et al., 2005). These particles show resemblance with the capsid structure of viruses and are therefore also referred to as Virus-Like Particles (VLPs).
Over the years more of these structures have been studied in different Bacterial and Archaeal species. Their sizes vary from a diameter of 20-24 nm, for the complexes with a triangulation number of T=1, to larger complexes with sizes of 30-32 nm with a T=3 architecture, like the VLPs seen in *Pyroccocus furiosus* (Akita et al., 2007; Giessen & Silver, 2017)(Figure 1A). More recently, also a larger encapsulin was discovered, having a triangulation number of 4, and a diameter of 42 nm, in the Quasibacillus thermotolerans (Giessen et al., 2019).

The encapsulins have a HK97-like fold, a name derived from the fold present in the bacteriophage HK97 capsid protein, but also in other viruses like herpesviruses. Encapsulins show structural homology with the HK97 protomers, which possesses three conserved domains, the peripheral (P) domain, the axial (A) domain, and the elongation loop (E), which is flexible and shows some differences among encapsulins in different species (Akita et al., 2007; Jones & Giessen, 2021; McHugh et al., 2014; Nichols et al., 2017; Sutter et al., 2008) (Figure 1A). It is suggested that the E-loop has a role in defining the triangulation number of the encapsulin shell (Nichols et al., 2017). Furthermore, it is proposed that mutations in the viral capsid accumulated, creating more static capsids with lower triangulation numbers, whereby encapsulins can show some of the viral origins among all three kingdoms of life (Akita et al., 2007; Jones & Giessen, 2021).

It was found out that they belong to the encapsulin family, proteins that form unique nano-compartments, similar in function to the eukaryotic membrane-enclosed organelles, in order to separate different metabolic processes in the cell. By separating these processes in a unique micro-environment, the enzymes involved will function more efficiently, and it minimizes the number of toxic intermediates present in the cytosol (Giessen & Silver, 2017).

Different types of protein complexes are found inside the encapsulins among the different bacterial and archaeal species, and are referred to as the encapsulin cargo proteins. They are targeted into the encapsulin via targeting peptides and have important roles in varying metabolic processes, like the Ferritin-like proteins (FLPs) which are involved in iron mineralization and storage, as are the iron-mineralizing encapsulin-associated Firmicute (IMEFs) cargo proteins (Giessen et al., 2019; Giessen & Silver, 2017). Sequence analysis suggests the presence of the FLP core-cargo protein, and as secondary cargo proteins a nitrite reductase, and a 'DNA-binding proteins from starved cells' (DPSL) cargo protein inside the *Pyroccocus furiosus* encapsulin (Giessen & Silver, 2017)(Figure 1B). A crystal structure is resolved for the encapsulated ferritin (EncFtn), a member of the FLP-superfamily, for *Pyrococcus furiosus*, *H. ochraceum, and R. rubrum*. This was done by expressing a truncated EctFtn in

E.coli. These genes are lacking the encapsulin targeting peptide, or in case of *P. furiosus* which has the encapsulin gene fused to the EctFtn gene, only expressing the EctFtn domain (He et al., 2019)(Figure 1C).

Although a lot is known about the different classes of cargo proteins among species, for *Pyrococcus furiosus* the inside cargo proteins are not completely resolved to high resolution. The function of these encapsulins is not completely clear, as are how the ferritin-like proteins are organized inside the *P. furiosus* encapsulin shell, how it interacts with the capsid, and if other cargo proteins are also present. It would be interesting to study these encapsulins in more detail, as it might present different types of cargo proteins inside the capsid compared to previous results, whereas it is also possible that the *Pyrococcus furiosus* can form different type of encapsulins, depending on the stress condition it encounters.

Studying these encapsulins in *Pyrococcus furiosus* would give more insight in the viral origins in Archaea, but also the other kingdoms of life. Furthermore, it will provide more details about the biological functions of encapsulins and the different cargo proteins.

During this study we will apply crude isolation approaches to obtain high resolution structures of different macromolecular complexes, like the ribosome, but also optimize our crude isolation protocol for specific purification of encapsulins. Different conditions and methods will be tested, using different biochemistry techniques, to find out how we could best purify these from the *P. furiosus* cells, using non-recombinant strategies.

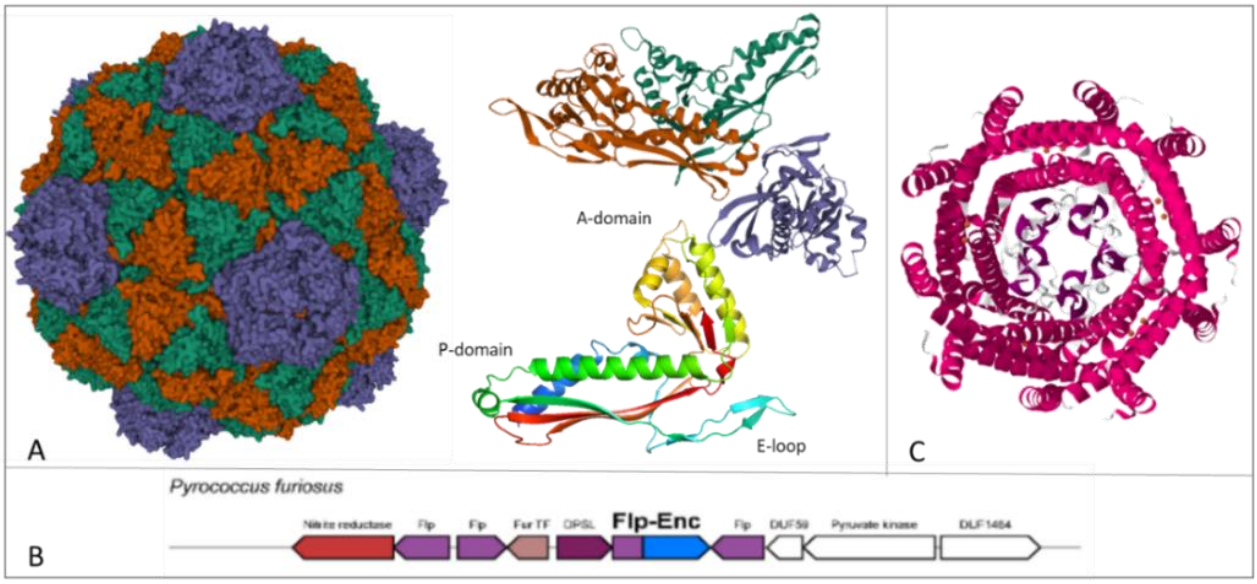


Figure 1. **Known structures for *Pyrococcus furiosus* encapsulin shell and Ferritin-like protein. A)** Overall structure of a *Pyrococcus furiosus* encapsulin capsid on the left, with a single asymmetric subunit on the top-right, and a single protomer on the bottom-right. The A-, B-, and C-subunit are depicted in purple, orange, and green, respectively [PDB: 2E0Z] (Akita et al., 2007). **B)** FLP-encapsulin operon with core-cargo protein FLP, and secondary cargo-proteins as studied by Giessen & Silver (Giessen & Silver, 2017). **C)** Structure of the encapsulated ferritin (EctFtn), belonging to the FLP-superfamily. This protein forms a decamer made up of pentamers of dimers. The alpha-helices are depicted in pink, the beta-sheets are colored purple. In between each dimer interface iron atoms (orange) can be bound [PDB: 5N5E] (He et al., 2019).

RESULTS

In order to validate crude isolation approached in the *P. furiosus*, we will be applying a sucrose gradient. This enables the separation of several large macromolecular complexes into different fractions.

Sucrose Gradient indicates the Presence of Large Complexes

After douncing the cells, the cell lysate was centrifuged at 100 000 rcf for 50 minutes (TLA55 rotor) to separate the soluble proteins from the membrane proteins (Figure 2.1). The soluble proteins were further separated by a 20%-60% (w/w) sucrose gradient. The samples were collected in steps of 200µL, and for each fraction the UV-absorption was measured at 260 nm. The bottom fractions were collected by hand. There are seven discernable absorption peaks which show the presence of different complexes (Figure S1, Table III). Especially clear peaks are found around the wells 24-26, and 31-33. Some clear, but smaller peaks can be seen at wells 19-20 and 39-41 (Figure S1).

Gels show Interesting Samples like Ribosomes and PpsA

For the clear peaks 19-20, 24-26, 31-33, and 39-41 were all samples run on a SDS-PAGE gel, as well as a Native gel. In addition, samples from well 3 and 7 were run, which are the top of the first and second peak, respectively. The gel shows that sample 7 still contains a lot of different proteins, and therefore wouldn't be useful for further analysis.

There are some clear bands in both the SDS-PAGE and native gel for sample 19-20 approximately at the 90 kDa (Figure S2A) and 150 kDa (Figure S2B), respectively. The molecular weight could indicate the presence of phosphoenolpyruvate synthetase (PpsA), a protein that is abundant in *P. furiosus*, and is a homodimer with a molecular weight of 150 kDa, and subunit weight of 92 kDa (Hutchins et al., 2001).

Sample 24-26 gave a high molecular weight band at the native gel, above the 700 kDa (Figure S2B), whereby the SDS-PAGE gel shows a band around the 100 kDa, with a lot of smaller bands at the bottom (Figure S2A).

Even higher molecular weight bands are seen for sample 31-33, at the top of the SDS-PAGE gel (Figure S2A). On the native gel no clear bands are seen, but the protein might remain at the top of the gel. Also this sample shows characteristic bands at the bottom of the SDS-PAGE gel, which is indicative of ribosomes. Both sample 24-26, or 31-33 could contain ribosomes, which will be further verified by negative staining in the next steps.

Sample 39-41 shows only proteins bands on the SDS-gel, around 150 kDa (Figure S2A).

Samples 19-20, 31-33, 24-26 and 39-41 will be dialyzed and checked by negative staining (Figure 2.2a).

Negative Staining Results indicate the Presence of Large Complexes

Negative staining results show only cell debris for the bottom fractions that were collected by hand (Figure S5). Also sample 39-41 doesn't show proteins on the negative staining image, whereas mostly cell debris can be found (Figure S4). For this reason, no cryo-EM images will be taken for these bottom fractions.

On the other hand, samples 31-33 and 19-20 show some promising features, as some large proteins can be seen on the negative staining images, and cryo-EM grids will be made for these samples (Figure 2.2a and Figure S4). Sample 39-41 will still be used for cryo-EM imaging, as it could be possible to find some proteins on these grids. For sample 24-26 no negative staining grids were made, but directly some cryo-EM grids, which will be analyzed in the next steps.

Cryo-EM Grids Show Among Others, Ribosomes and PpsA

Also on the cryo-EM micrographs for sample 39-41 only cell debris can be seen (Figure 2.2a and Figure S4). Likewise, mainly cell debris is found on the micrographs taken for sample 24-26 (Figure S5). These won't be used to collect a dataset.

On the other hand, both sample 31-33 and 19-20 contain proteins. Sample 31-33 is in a relatively pure condition (Figure 2.2a and Figure S4). Sample 19-20, contains a mixture of proteins, were PpsA, thermosomes and 20S proteasomes can be seen (Figure 2.2a and Figure S4). Data sets were collected for both of these samples, whereby only sample 31-33 will be processed for this project.

Data Processing of Sample 31-33 Results in a 3.22 Å Structure of the 50S Ribosome

Around 1000 micrographs were collected for sample 31-33, with a pixel size of 1.04 Å/px. Data was processed using Relion®. Firstly, the micrographs were motion corrected, after which the Contrast-Transfer Functions (CTFs) were estimated, using a minimal and maximal defocus value of 5 000 Å and 30 000 Å, respectively, with a defocus step size of 300 Å. Only micrographs with a good aligned CTF estimation were selected, which have mostly a resolution above the 7.0 Å, resulting in a total of 948 micrographs.

Particles were picked with the AutoPicker in Relion®, using the Laplacian-of-Gaussian function with a log-filter diameter set to particle sizes ranging from 200 Å to 250 Å. This resulted in a number of 126 964 particles being picked. After two 2D-classification rounds in order to clean the data set, a number of 83 149 particles were selected and used to form an initial 3D model. Particles were sorted into four classes, and using a mask diameter of 250 Å. There were then further sorted by a 3D-classification in four classes. No major difference could be seen between the 4 classes, only class 3 contained incomplete or damaged particles, and the other good classes were selected for further processing (Figure 3). The worst class contained about 6.6% of the particles and had an estimated resolution of 22.3 Å. The other particles were almost evenly distributed over the other classes, with a resolution estimate ranging from 12.23 to 12.48 Å. An auto refinement job, with the 1st class of the 3D classification as reference, was run with the resulting 77 658 particles. The refined model had an estimated resolution of 4.33 Å. Applying a mask, based on the refined density, resulted in a 4.27 Å resolution. This model was used to build an tighter mask, which was used for another round of auto refinement, and a postprocessing step resulting in a model with a 3.9 Å resolution according to the gold standard FSC. This density map was further refined using CTF refinement steps in Relion®, to account for anti-symmetrical aberrations like trefoil, and beam tilt, but also symmetrical aberrations like tetrafoil and spherical aberration. Afterwards also per-particle defocus was estimated. No three-fold or four-fold astigmatism seems to be present in our dataset, although there is some

slight beam tilt and axial coma present. Our data contains almost no magnification anisotropy (Figure S6). As final step was Bayesian Polishing applied, so that the final reconstructed density map reached a resolution of 3.22 Å, according to the gold standard FSC (Figure 3). The PDB structure of the 70S ribosome that was previously known was fitted into this density map (Figure 3).

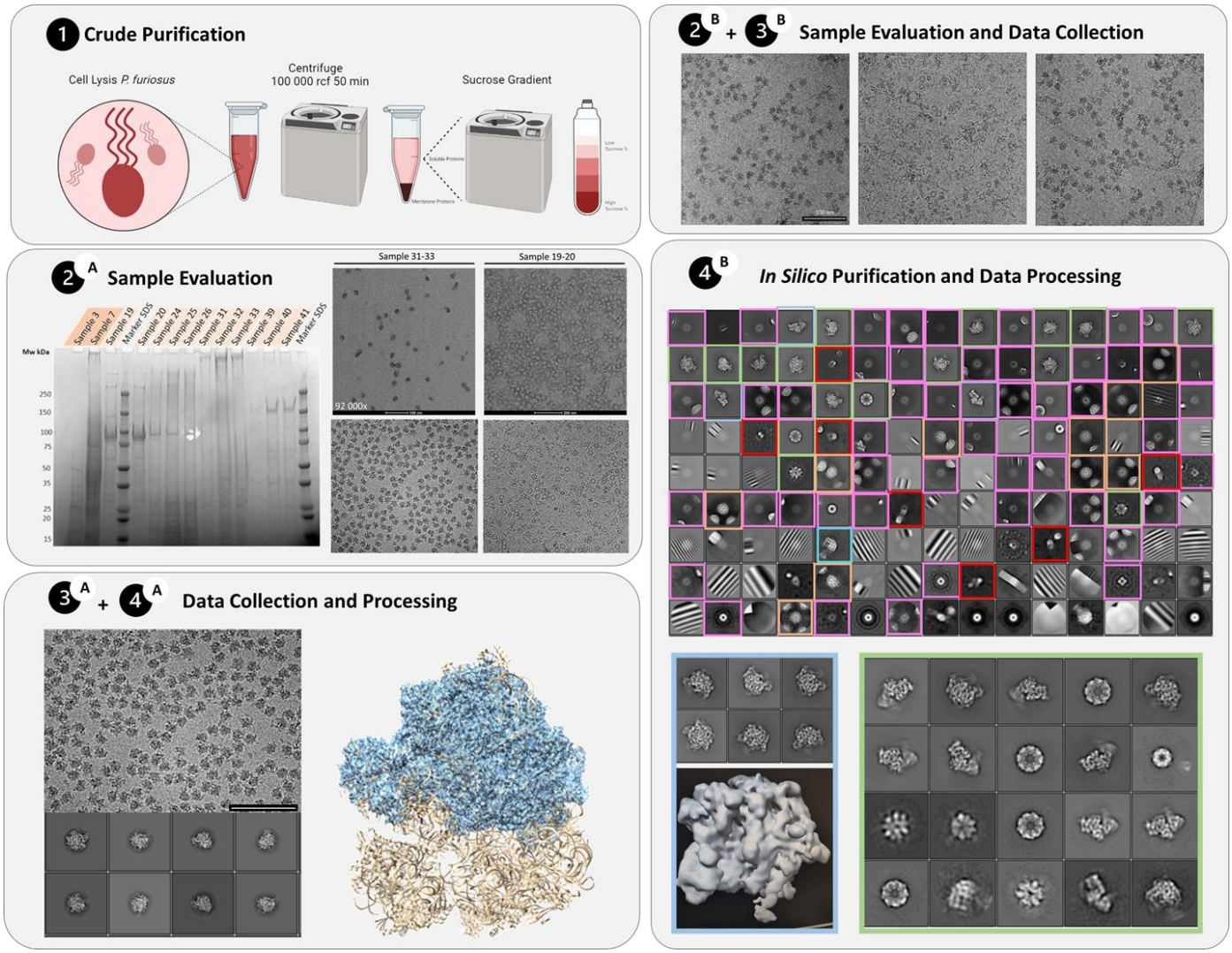


Figure 2**. MGP Crude Purification Protocol.** 1) Sample preparation. *Pyrococcus furiosus* cells were lysed, and centrifuged, after which the broken cells were ultracentrifuged. The supernatant was applied to a sucrose gradient. 2) Sample evaluation. Samples were evaluated using SDS- and Native gels, as well as negative staining. Promising samples were also screened by cryoEM (2a). For additionally run sucrose gradients were samples checked by cryoEM (2b). 3) Data collection. For promising samples was a cryoEM dataset collected, like for sample 31-33, containing the 50S ribosome (3a). For the additional samples were datasets collected and combined, which results in a dataset containing a mix of proteins. AutoPicking parameters were chosen to select all kind of proteins (3b). 4) Data processing. Data was processed in Relion, resulting in a 3.3 Å resolution density map of the 50S proteasome. Some 2D classes and the final density map with the 70S ribosome (4V6U) fitted are shown (4a). For the mix dataset, *in silico* purification was performed using 2D classification to identify and sort the different particles. The 50S ribosome can be easily recognized, outlined in blue. Also, possibly capped 20S proteasomes are present in our dataset, outlined in green. The particles were also sorted based on their size, with in pink particles with a size about 140 Å, in orange a size of 180 Å, and in red 100 Å (4b). Scalebar indicates 100 nm.

Additional Sucrose Gradient Samples Sorted by 2D Classification Show Different Protein Complexes of Different Sizes

Different samples were checked at the cryo-electron microscope, and for samples 23 (30%-60% w/v) and 29 (30%-60% w/w) a data set was collected, with pixel size 1.04 Å/px (Figure 2.2b and 2.3b). These datasets were combined and contain a mix of different large proteins (Figure S7). A total number of 2676 micrographs were motion corrected in Relion®. CTF estimation was then performed and a selection of micrographs was made, based on the resolution estimation, whereby micrographs below 7 Å were discarded. A total of 2499 micrographs were used for selecting particles with the AutoPicker. Different setting for the Laplacian-of-Gaussian AutoPicker were tested on a subset of 16 micrographs, which contain images of the different samples. Using a log-filter diameter with a minimal and maximal value of 150 Å and 300 Å, resulted in the best selection for all types of particles (Figure S7). By using these settings, a total number of 643 490 particles were picked. These particles were extracted with a box size of 440 pixels and rescaling to a size of 100 pixels. Using less binning would take more computational power and memory for the 2D classification, therefore a binning of around 4 was used. In order to sort the different proteins with 2D classifications, a sufficient amount of classes should be chosen. In our case we used 200 classes, with a tau_fudge of 2, and a mask diameter of 400 Å (Figure 2.4b). This 2D classification shows very clearly the presence of al lot of different proteins. Some of them could be recognized directly, like the 50S ribosome, outlined in green. There are also some highly symmetrical classes present, which could be attributed to for example proteasomes or thermosomes. Outlined in blue, are some really large particles, with some flexible or heterogeneous region. These particles resemble a capped 20S proteasome (26S proteasome), which was also previously found by other researchers during crude purification (Verbeke et al., 2018).

The 50S ribosome particles were selected and re-extracted with a box size of 360 px, and binned twice. About 125 424 particles belong to the 50S ribosome.

The particles were sorted in 50 classes using 2D classification and a mask with a diameter of 340 Å, whereby about 121 158 particles were selected from the good classes for building an initial model (Figure 2.4b & Figure S8 & S9).

Further cleaning of the 50S ribosome was performed by 3D classification in 20 classes, with tau_fudge 4. The good classes contain in total a number of 109 175 particles (Figure S9).

As the symmetrical particles could belong to the proteasome, both the 2D classes outlined in blue, and the symmetrical ones are selected and re-extracted (Figure 2.4b). The estimated maximum particle size is around the 260 Å. Therefore a box size of 384 pixels and a binning of 2 were used to extract the 58 031 particles. After 2D classification based cleaning in 50 classes, about 55 551 particles were selected for further processing. Based on the 2D classes, the estimated size would be a bit larger, up to around 270 Å. The particles were for this reason re-extracted with a slightly bigger box-size. Also an initial model was built, which didn't yield a nice model (Figure S10).

The other particles were sorted based on estimated size (Figure 2.4b & Figure S11). The 180 Å sized particles outlined in orange, for the present called "spherical protein", were selected from the 2D classification. This resulted in 17 753 particles belonging to the "spherical protein". An initial model was built using different sizes of masks, with a size of 250 Å offering the best results (Figure S12). The identity of this protein could not been resolved yet, and needs further processing.

Virus Like Particles can be obtained in High Purity

To achieve a high resolution structure of the encapsulins to elucidate the inner density inside the shell, we optimized our crude isolation protocol to obtain pure particles. For this, the P. furiosus cells were broken by using a douncer and centrifuged at 6 000 rpm for 10 minutes. The supernatant containing the broken cells was ultracentrifuged at 39 500 rpm (Ti70 rotor), resulting in three layers, a bottom pellet layer containing mainly cell debris, a cloudy middle layer, and a clear upper layer (Figure 4). This clear upper layer contains the encapsulins, based on previous work done by W. Song.

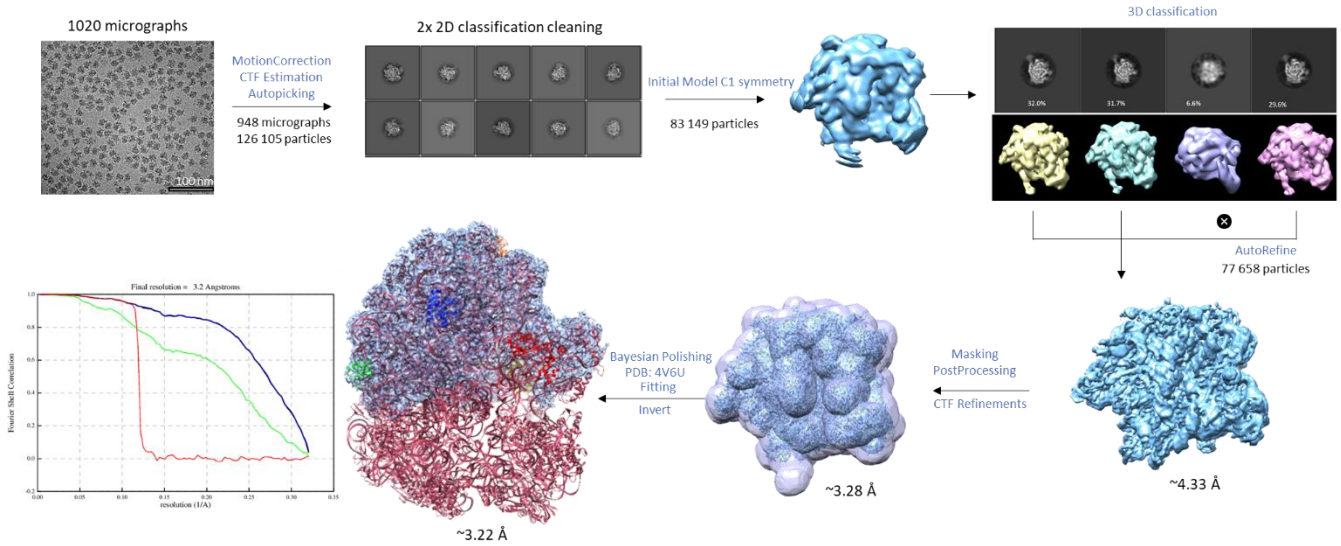


Figure 3. **50S Ribosome Data Processing.** About 1020 micrographs were collected, which were motion corrected and CTF estimated in Relion. After selecting the good micrographs, 948 micrographs were used for AutoPicking, yielding 126 105 particles. The dataset was cleaned using two rounds of 2D classification, after which 83 149 particles were selected to built an initial model. 3D classification was used to further clean the dataset. Class003 was discarded and the other particles were selected for several rounds of AutoRefine, giving a density map with 4.33 Å resolution. Using a mask and several rounds of PostProcessing, and CTF refinements, a map of 3.28 Å was obtained. Into this density map was the 70S ribosome (4V6U) fitted.

Size Exclusion Chromatography results in a Protein Fraction Containing Encapsulins

In order to purify the sample further, the clear upper layer was applied onto a Size Exclusion Column (Hiload Superose© 6). This resulted in four major protein fractions, eluting at approximately 35 mL (B3-B1), 40 mL (C1-C5), 60 mL (C6-D1), and 90 mL (E1-F4), with the wells containing the fraction indicated in brackets (Figure S13). The encapsulins are expected to be present in the third peak, based on previous work done by W. Song. The wells belonging to this peak, C6-D1 are collected and will be applied to a sucrose gradient to purify them further.

A Sucrose Gradient of 10%-50% reveals Encapsulins at Approximately 50% according to Negative Staining

In order to purify the sample further, the clear upper layer was applied onto a 10%-50% (w/v) sucrose gradient (Figure 4). The absorption of each fraction was measured, and shows three main protein fractions (Figure S14). The encapsulins should be present at the last peak, at approximately a sucrose concentration of 40% (w/v), as was already shown by unpublished work of W. Song. For this peak, comprising wells 50-69 and the bottom fractions P1-P7 that are collected by hand, were evaluated using negative staining EM screening, using each evenly numbered well plus sample 69 and the most bottom fraction, P7. Based on the negative staining screening, it could be seen that the virus-like particle could be mostly found in the later fractions, and are rarely seen in samples 50-54, and a few are present in samples 56-68 (Figure S15). However, these samples were diluted the most, and had the highest concentration compared to the others (Table I). Therefore, samples 57-59 are still included for further purification steps, as they possibly contain more encapsulins than would be hypothesized based on the negative staining grids. Each fraction contains 200 µL sample. As the samples will be applied for a second time to size exclusion chromatography, we combined two or three wells together, resulting in samples 57-59, 60-62, 63-65, 66-67, 68-P1, P2-P4, and P5-P7.

A Second Size Exclusion yields Some Pure Encapsulin Samples

After mixing together the sucrose gradient samples, as mentioned above, the different samples were further purified by using a second size exclusion with a Superose® 6 column (Figure 4). This gave a separation into two different peaks for each sample, with the most important one being at approximately 8 mL, as this is were the encapsulins are expected to be (Figure S16 & Table II). Especially the more bottom fractions P2-P4 and P5-P7 show a relatively high peak for this, whereby this first peak at 8 mL is the most prominent one for sample P5-P7 (Figure S16). The wells containing this fraction were collected and imaged by negative staining to evaluate the purity of the sample. Sample $57\text{-}59_{B11\text{-}B10}$ didn't contain a lot of virus-like particles, as did sample $P2\text{-}P4_{B11\text{-}B9}$, $63\text{-}65_{E3\text{-}E4}$, and $66\text{-}67_{B11\text{-}B9}$, although a bit more than $57\text{-}59_{B11\text{-}B10}$. Mainly samples $60\text{-}62_{E3}$ and $60\text{-}62_{E4}$ contain a lot of encapsulins. However, the latter still contains some small molecules in the sample. Some of the other samples are also not completely pure, like sample $60\text{-}62_{E4}$, $68\text{-}P1_{E2\text{-}E3}$, $57\text{-}59_{B11\text{-}B10}$, and $P5\text{-}P7_{B12\text{-}B10}$ (Figure S17). Based on their purity we mixed the samples together, with a mixture of the most pure samples, indicated by a green star, and a mixture with the less pure samples, indicated by the orange stars (Figure S17).

Cryo-EM Screening Indicates a too Low Concentration Sample

Both mixtures were centrifuged to pellet the sample down to concentrate them. The more pure mixture had a final concentration of 0.092 mg/mL, the less pure one of about 0.168 mg/mL. During the screening session for Cryo-EM, the grids showed some encapsulins, but mainly on the carbon edge, almost none were present in the ice. The samples had a too low concentration to obtain good Cryo-EM images, which could be used for collecting a data-set and doing data processing.

Table I. 10%-60% Sucrose Gradient Samples VLP Peak

Sample	Concentration (mg/mL)	Dilution for NS
50	1.08	15x
52	3.83	50x
54	4.84	70x
56	5.13	70x
58	4.32	60x
60	3.14	45x
62	2.13	30x
64	1.24	18x
66	0.57	8x
68	0.31	4.5x
69	0.21	3x
P7	0.43	6x

Table II. Second SEC Samples VLPs

Sample	Wells	Concentration (mg/mL)
57-59	B11-B10	0.02
(1) 60-62	E3	0.11
(2) 60-62	E4	0.13
63-65	E3-E4	0.01
66-67	B11-B9	0.01
68-P1	E2-E3 & $B9_{P5\text{-}P7}$	0.08
P2-P4	B11-B9	0.01
P5-P7	B12-B10	0.21

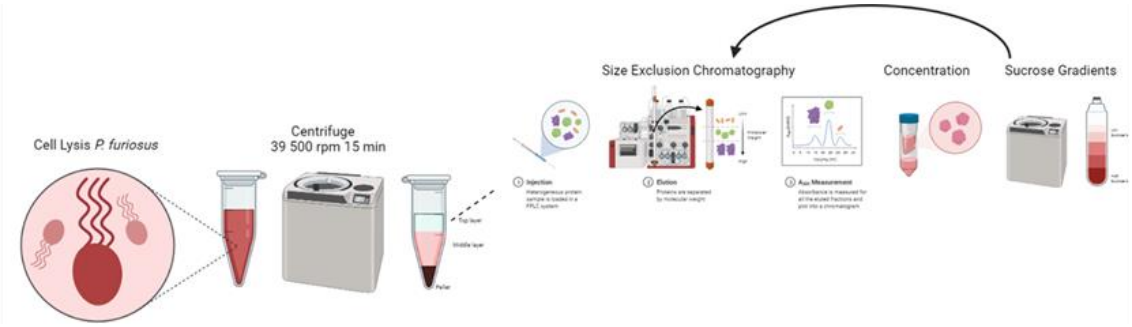


Figure 4. **Optimized Protocol for the Purifiation of Encapsulins.** First the broken cells are centrifuged at 39 500 rpm for 15 minutes. The clear upper layer is purified further by applying a SEC, a 10%-50% sucrose gradient and a second SEC.

DISCUSSION

We have elucidated a 3.22 Å density map of the 50S ribosome in *Pyrococcus furiosus*. Ribosomes consist of a small subunit and a large subunit, 30S and 50S in prokaryotes, and 40S and 60S in eukaryotes, made up of ribosomal proteins and ribosomal RNA. Both bacterial and archaeal ribosomes are made up of 16S, 23S and 5S ribosomal RNAs, whereas eukaryotic ribosomes contain 18S and 25S fragments. About 33 ribosomal proteins are shared between all three domains of life, whereby an additional 34 are shared between eukaryotes and archaea (Londei & Ferreira-Cerca, 2021; Maguire & Zimmermann, 2001). A growing polypeptide-chain, peptidyl-tRNA, will start at the P-site, whereby newly incoming complementary aminoacyl-tRNA will first attach to the A-site. The growing peptide-chain is transferred to the new aminoacyl-tRNA, forming a 1-residue elongated peptidyl-tRNA. The growing peptide chain will then transfer to the P-site, and the deacetylated tRNA will exit the ribosome via the E-site, enabling new aminoacyl-tRNAs to enter the A-site again. During translation, the ribosome undergoes a lot of conformational changes (Maguire & Zimmermann, 2001). For *Pyrococcus furiosus*, there is a 6.6Å resolution structure known for the 70S ribsome (Armache et al., 2013).

Into our 3.2 Å 50S ribosome density we have fitted the molecular model that was previously build using a 6.6 Å density map of the *Pyrococcus furiosus* 70S. One thing that can be immediately noticed, is that we have no density for the L1-stalk (Figure 5). Our initial model showed some density for the L1-stalk, and our final density map shows some part of the L1-stalk, but only when using a low binarization threshold (Figure 6). This stalk aids in guiding the tRNA through the ribosome during translation, and therefore the L1-stalk is highly mobile for its function, which explains why we won't see highly defined density in the map (Trabuco et al., 2010).

However, there are some more ribosomal proteins for which we don't have any density (Figure 5 and Figure 7A,C). One of them is L8e(2), an additional copy of L8e. Archaea contain some additional copies of ribosomal proteins, besides the stalk-proteins, which was not expected as bacteria don't possess these. This shows the intermediate complexity of the archaeal ribosome, and some of the evolutionary changes in the ribosome which could also reveal some of the functions and aspects of the eukaryotic ribosome (Armache et al., 2013). However, when we use a extremely low binarization threshold in Chimera of 0.0250, we can see some density for this ribosomal protein (Figure 7B). This confirms the results of Armache *et al.,* showing an additional copy of L8e.

This is not only the case for L8e(2), also for the other copy S24e we can see no density in our model, even when using lower resolution at the extremely low binarization threshold, there is almost no density for this ribosomal protein (Figure 7C&D).

On the other hand, the additional copy of L14e has some clear density. Especially the top α-helix density can be clearly seen, although the model doesn't fit this perfectly (Figure 7E). The orientation of the L14e(2) model is completely correct for our own model. When focusing the fitting on the L14e(2) in Chimera, we can already see that the helix fits more nicely into our density map. Even some density can be seen for the larger residues, like glutamine and lysine (Figure 7F).

Another ribosomal protein that doesn't show a correct orientation in our density map is the canonical L8e(1) (Figure 7G). Also for this protein we focused the fitting on this protein, resulting in a much better fit into our density map, as can be seen by for example the α-helices (Figure 7H). The canonical L14e(1) protein shows density in our map, although it doesn't fit completely into our density (Figure 7I&J).

Our results confirm the presence of the extra copies of L14e and L8e in the *Pyrococcus furiosus* 50S ribosome, as previously reported by Armache et al., 2012. However, at our resolution, we don't see density for S24e(L). Canonical S24e is normally present on the 30S subunit, and the copy present at the large subunit, might still need the small subunit for binding. During the sucrose gradient we separated both subunits from each other. Also a lower magnesium concentration could account for more easily separating the two subunits during the sucrose gradient steps and losing the small subunit and additionally the S24e(L). Magnesium is known to stabilize the ribosome secondary structure, aiding in the binding of the ribosomal proteins, and a too low concentration (<1 mM) could result in dissociation of the small and large subunit (Akanuma, 2021). Our sample buffer contained a $MgCl_2$ concentration of 3mM, but we could have lost also some magnesium during the purification, for example during the centrifuging steps, or dialysis. Furthermore, the L8e(2) could only be seen at lower binarization levels, and not at the high resolution density

map. This could indicate mobility of the protein, or incomplete occupancy on the 50S ribosomes we have in our sample. On the contrary, the additional L14e shows some more defined density, whereby we can see even some density for the amino acid residues on the helices. The orientation of the model is slightly off, and should be remodeled. This one is still clearly visible in our 3.2 Å density map, which implies less mobility compared to S24e, or higher occupancy. Interestingly, the additional S24e(L) was suggested to be specifically found in Thermococcaceae, which includes the hyperthermophilic species *Pyrococcus* and *Thermococcus* (Armache et al., 2013). This protein might have a role in stabilizing the ribosome in hyperthermophiles, although this is only one hypothesis and should be studied more extensively.

It would be interesting to process our other 50S ribosome dataset from the combined datasets, to see whether these additional ribosomal proteins, L8e(2), L14e(2), S24e(L) are also present in this dataset, or if there is a difference in occupancy. These 50S ribosomes were purified using a 30%-60% (w/v and w/w) sucrose gradient, instead of 20%-60% (w/v). The effect of the sucrose concentration on the separation of the ribosomal proteins could be interesting to study, but also some different salt concentrations and centrifuge speeds could be tested.

Especially, optimizing the purification to obtain also clear densities for L8e(2) and S24e(L) would be interesting, in order to achieve a higher resolution model of the ribosome containing these ribosomal protein to study their interactions and functions.

Besides this, our density map could still be optimized. We see that the model doesn't fit our density perfect (Figure 8). The lysine doesn't have density, which could be due to the high flexibility of the lysine, or a incorrect rotamer. We only run one round of CTF refinements and Bayesian Polishing, while both function best with higher resolution models. As Bayesian Polishing yielded a higher resolution structure, this could give benefits when running an additional CTF refinement. A higher resolution map would benefit the correct modelling. Also, using COOT® instead of Chimera® would improve the modeling.

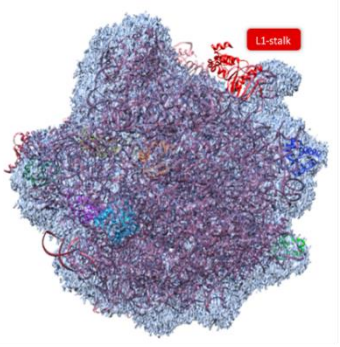


Figure 6. **Top View of the 50S Ribosome at low binarization level**. The top view at a binarization threshold of 0.0160 shows some density for the L1-stalk, which can't be seen at higher resolution due to high mobility.

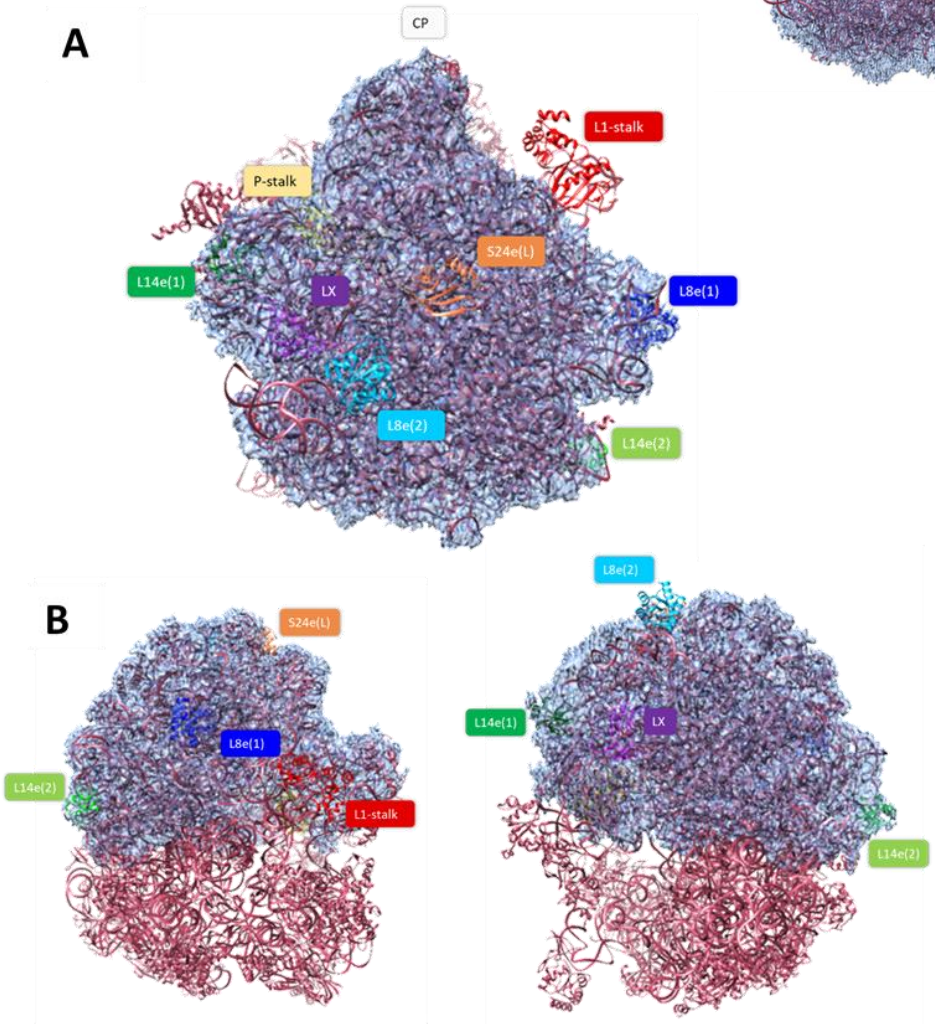
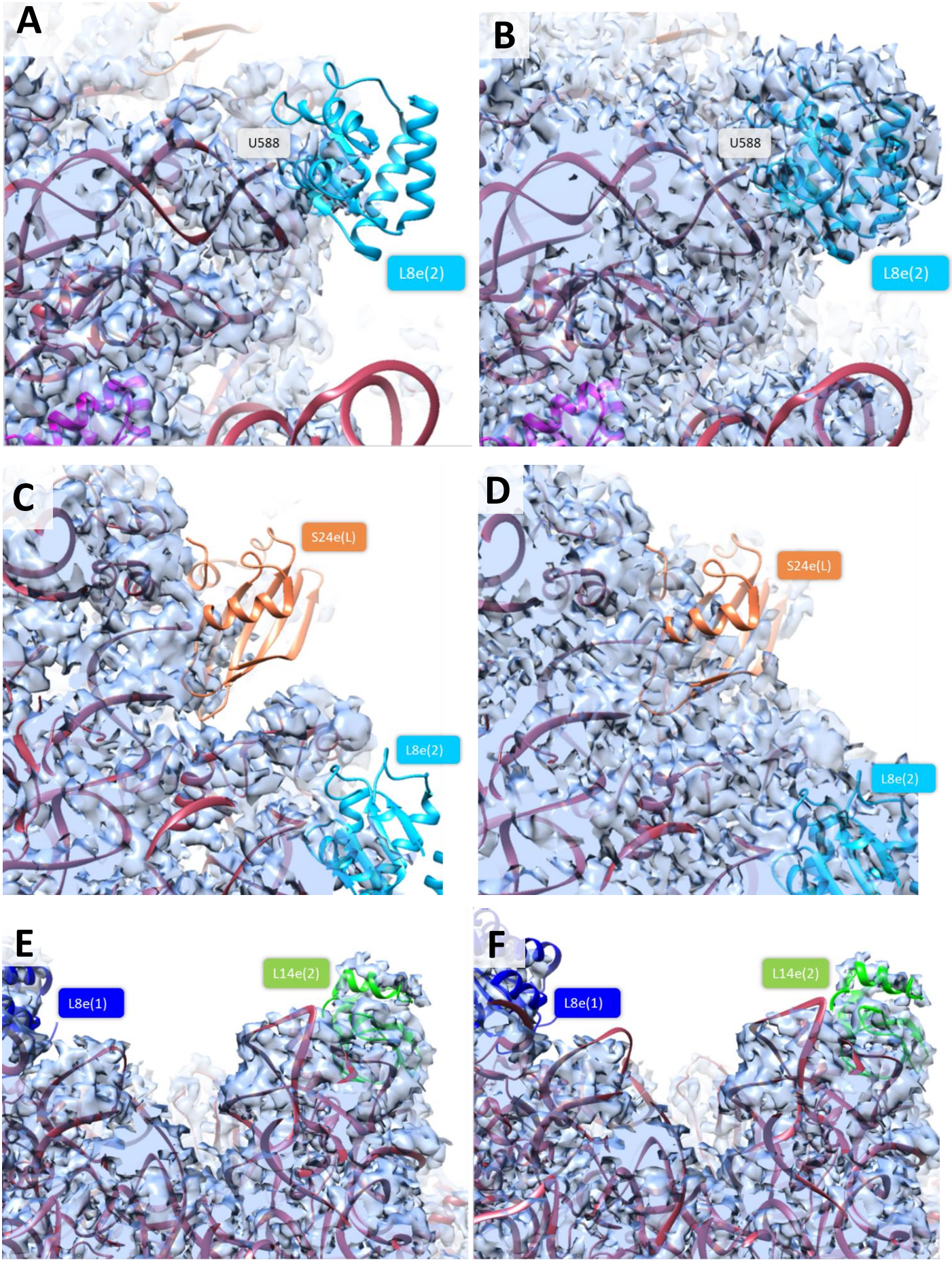


Figure 5. **Top and Side View of the 50S Ribosome and fitted 70S Model**. The top view (A) and two side views (B) of the 50S ribosome. On top is the Central Perturbance (PC), also the L1- and P-stalk are indicated, although we don't see density for these, due to high flexibility. Also the positions of L8e, L14e and their copies L8e(2) and L14e(2) are indicated. The model PDB 4V6U indicates the presence of S24e(L) on the large subunit.


A
U588
L8e(2)
B
U588
L8e(2)
C
S24e(L)
L8e(2)
D
S24e(L)
L8e(2)
E
L14e(2)
L8e(1)
F
L14e(2)
L8e(1)

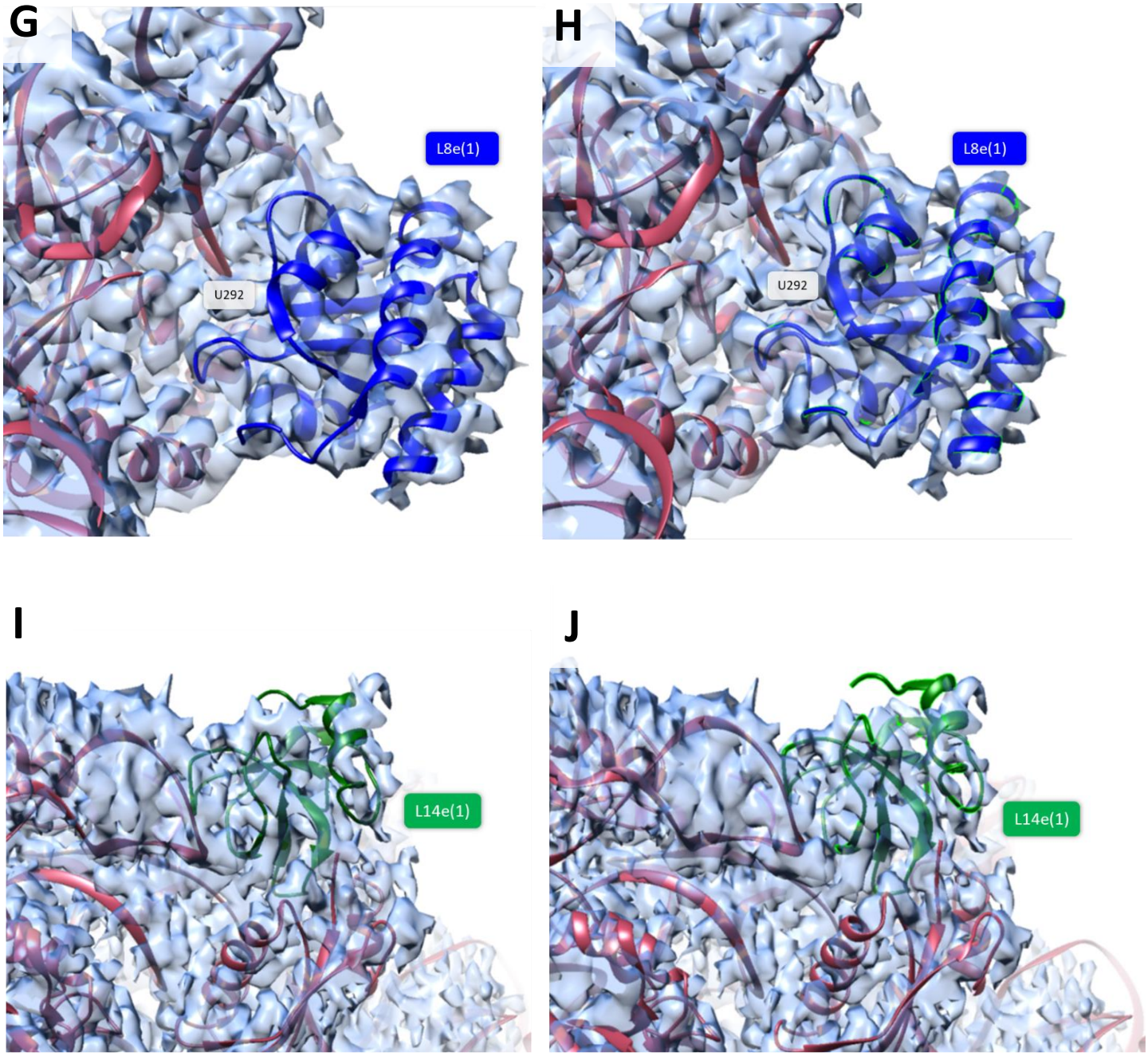


Figure 7. **Ribosomal Proteins L8e(1), L8e(2), L14e(1), L14e(2) and S24e(L).** The additional L8e protein, L8e(2), shows no density in our map at a threshold of 0.0411 (A), but only at lower at 0.0250 (B). S25e(L) shows no density at both the high (C) and low threshold (D). The additional L14e proteins, L14e(2), has a clear density in our own density map, although the orientation is slightly off (E). Fitting the selected L14e(2) into our density shows some improvement in fitting the density (F). Ribosomal protein L8e(1) has also a slightly different position then modeled by PDB 4V6U (G), which is corrected a bit by focusing fitting on this protein (H). Ribosomal protein L14e(1) is also present in our model (I), although the fit is not completely correct, even when using Chimera to fit this protein (J). Images E-J are all at a binarization threshold of 0.0411.

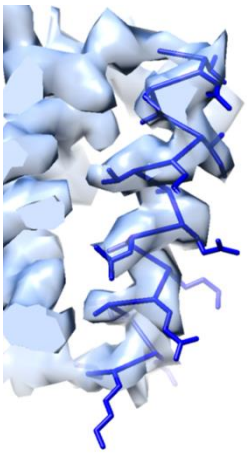


Figure 8. **Zoom-in on a α-Helix of Ribosomal Protein L8e(1).** Zooming in on a α-helix of the L8e(1) protein and showing the side chains, reveals a not perfect fit of the model for our density. Especially the absence of density at the lysine at the bottom is striking. Binarization threshold is 0.0411.

The other dataset contains also a 50S ribosome, present in many copies, which could most likely reach high resolution. Solving this structure would also be interesting. Both 50S Ribosome datasets could be combined, resulting in even more particles that could be used for data processing.

The other dataset has still a lot of proteins that could possibly be identified, like the "spherical protein" with a size of approximately 180 Å, and the smaller proteins about 140 Å. For the capped 20S Proteasome, we need to collect more data, as we miss a lot of views and therefore couldn't obtain a nice initial model, which could be used as reference for the following processing steps. In addition to this, we are not completely sure if the 2D-classes we sorted into the 20S Proteasome group would really be the capped proteasome. Based only on 2D-averages, it is still hard to tell. The highly symmetrical proteins could also belong to thermosomes. For this, sending this sample for mass-spectrometry analysis, could elucidate which proteins are present in our sample. Mass-spectrometry might also deliver some possible candidates for the 180 Å "spherical protein". Furthermore, only about 17 753 particles were grouped together for this protein, which could be a bit too few. Collecting an additional dataset might also be beneficial for this protein.

Although we achieved pure encapsulin samples, we couldn't solve a structure for them by cryo-EM, due to a too low concentration. One of the most plausible reasons for our low concentration, would be the absence of a flow restrictor valve on the injection valve of our ÅKTA machine. This piece normally generates a stable back-pressure, ensuring that no gas bubbles are generated, and also preventing siphoning. In our case, upon injection, some part of our sample went to the waste. Another purification according to our protocol should be performed, so that cryo-EM grids with sufficient amount of particles could be made.

MATERIALS & METHODS

Crude-Isolation

During the purification different centrifuge steps were used. The different fractions are drawn in Figure 1, and the description of the used methods refers to the names used in this figure for each fraction (Figure 9).

Cell Lysis of P. furiosus – freeze-dried *Pyrococcus furiosus* cells were taken out of the -80°C freezer and thawed on ice. To these cells was added 1mL Cell Lysis Buffer Complete [10mM HEPES-NaOH (pH 7.0), 500mM NaCl, 5mM $MgCl_2$, 20% Sucrose (w/v), [+/-] DNAse I (1:10000)(0.1 mg/mL), cComplete protease stop (1/10 tab)]. The cells were lysed by using a 2 mL douncer, and moving up and down for 20 times with pestle A, followed by 20 times with pestle B. The dounced cells were transferred to 2 Eppendorf tubes and the cells were then centrifuged at 4°C, at 10 000 rcf for 15 minutes to separate the broken cells from the unbroken cells. The supernatant (B) was transferred to new Centrifuge tubes, the pellet (C) is frozen, this contains the cell debris. About 20 µL of the supernatant was transferred to another Eppendorf tube to use for SDS-PAGE gels.

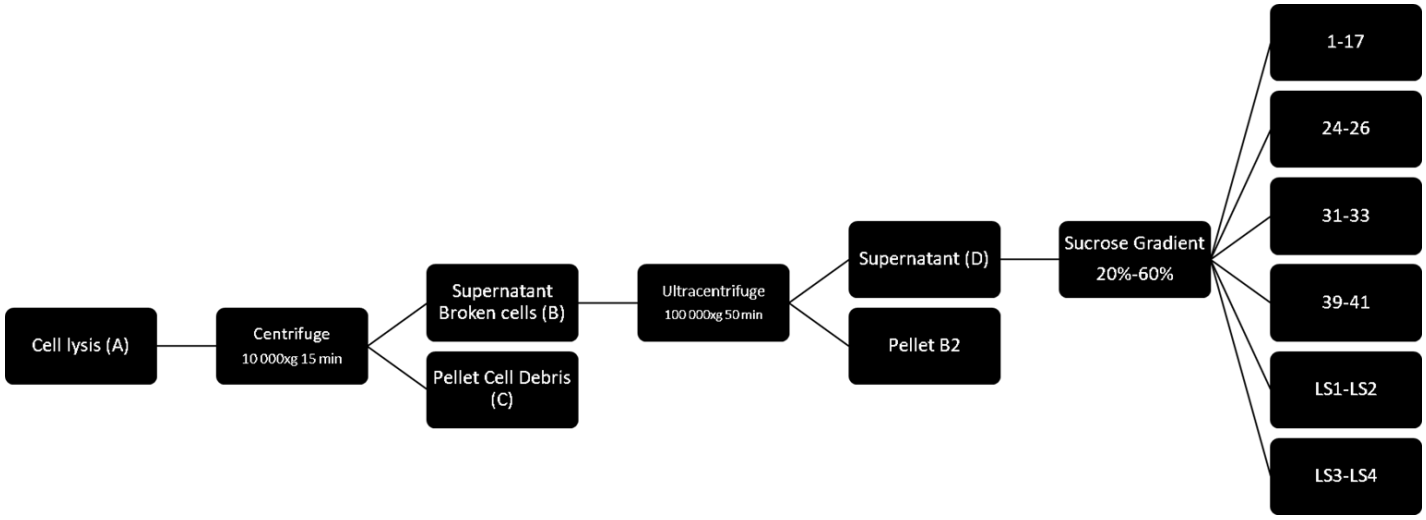


Figure 9. **MGP Purification.** Flow-scheme of the different purification steps used to purify different factions from a Sucrose Gradient. These samples were used for Negative Staining, and 24-26, 31-33, 39-41 also for Cryo-EM imaging. Sample 1-17 is a mix from sample D1 and D2, as is sample 24-26. Other samples contain only fractions from the sucrose gradient of sample D1.

Ultracentrifuge Samples – The supernatant with broken cells (B) was ultracentrifuged with a TLA55 rotor at 100 000 rcf for 50 minutes, at 4°C. After ultracentrifuging, the supernatant containing soluble proteins was transferred to 2 new Eppendorf tubes, and named sample D1 and D2. The pellet with the membrane proteins (B2) was stored in the -80°C freezer.

Sucrose Gradient – The soluble proteins (D) are further purified into different fractions by using a sucrose gradient of 20%-60% [20% Sucrose (w/w) or 60% Sucrose (w/w), 50mM HEPES-NaOH (pH 7.0), 500mM NaCl, 3mM MgCl2, 0.1 mM EDTA (pH 8)], which was ultracentrifuged at 37 000 rpm for 17hrs at 4°C, using the SW-41 Ti rotor. The Piston Gradient Fractionator in combination with the Fraction Collector (FC203B, Gilson) was used to collect different proteins fractions in a 96-wells plate and the UV-absorption at 260 nm (DNA/RNA) and 280 nm (Proteins) was measured, each well containing 200µL. The last fractions at the bottom of the tube (high sucrose concentration) were collected by hand, and stored in the 96-wells plate at wells 12A-12D and named L1-L4.

Table III. 20%-60% Sucrose Gradient

Peak	Sample D1	Top of Peak	Sample D2
1	1-5	3	1-4
2	6-9	7	5-8
3	10-12	10	9-16
4	13-17	14	
5	18-21	19	17-21
6	23-28	25	22-28
7	30-36	32	30-37
8	39-41	40	38-44

SDS-PAGE and Native Gel Analysis – The protein fractions were run on a SDS-PAGE and a Native gel for analysis. Only samples 3, 7, 19, 20, 24, 25, 26, 31, 32, 33, 39, 40, and 41 from sample 1 (Table III) were run on the gels. Samples were prepared by mixing 12 µL sample with 6µL premade 3X SDS Sample Buffer [30% (w/v) glycerol, 6% (w/v) sodium-dodecyl-sulphate, 0.03% bromophenolblue, 187.5 mM Tris pH 6.8]. For the SDS-PAGE a precast 4-15% glycine gel (Bio-Rad®) was used together with 10X Tris/Glycine/SDS buffer (Bio-Rad®) diluted to 1X with MilliQ water [25 mM Tris, 192 mM glycine, 0.1% SDS, pH 8.3]. About 15 µL sample and 5 µL marker (Dual Color Bio-Rad®). The gel was run at constant a constant voltage of 180V until the samples reached the bottom of the gel, about 55 minutes. After running the gel, it was washed for 10 minutes with dH_2O. Staining was performed with premade Coomassie Staining for half an hour, and destained in dH_2O for at least 30 minutes.

Samples were prepared for the native gel by mixing 12 µL of the sample with 3 µL 5x Blue Native Sample Buffer [2.5% Coomassie Brilliant G-250, 50% glycerol, 250mM ε-aminocaprioc acid, 50mM Bis-Tris pH 7.0]. The 4-15% glycine gel (Bio-Rad®) was run on ice in 1x Native Buffer (Bio-Rad®) [25 mM Tris, 192 mM glycine, pH 8.3]. The gel was run at constant and low current of 11 mA for 3 hours, and afterwards stained and destained using the same method as for the SDS-PAGE gel.

A second Native Gel was run for samples 32, 33, 34, 35, 39, 40, 41, and L1. For this 25 µL of sample was mixed with 5 µL 5X Blue Native Sample Buffer, and loaded onto a 7.5% glycine gel. The gel was run for about 1 to 2 hours on ice at 11 mA constant current. Staining and destaining were performed in the same way as described above.

Dialysis – In order to do following purification steps with chromatography methods or doing negative staining, we have to get rid of the high sucrose concentration. This was done by dialyzing the samples. The wells 2-17 from sample D1 and 1-16 from D2 were mixed together and dialyzed with 15 mL Centrifugal Filter Units, MWCO= 100 kDa, till the sucrose concentration had a theoretical concentration of about 0.23%. A centrifugal speed of 3 000 x g was used. Also wells 24-26 from both sample D1 and D2 were mixed and dialyzed till a theoretical sucrose concentration of 0.02% (Table IV). For these theoretical values it was assumed that the sucrose concentration was 60%, the maximum concentration of our sucrose-gradient, and after dialyzing the sucrose concentration should be maximally these theoretical values. Sample 1-17 will be further purified by chromatography techniques. Sample 24-26 had a concentration of 3.61 mg/mL and was further diluted to a concentration of 2.52 mg/mL for cryo-EM imaging.

Also, following wells were mixed together: wells number 39-41, 31-33, 19-20, LS1-LS2, and LS3-LS4. These were dialyzed in a similar way as described above, using 4 mL Centrifugal Filter Units with a Molecular Weight Cut Off of 100 kDa, at 3 000 x g. These were dialyzed till the samples had a theoretical sucrose concentration of 10%. These samples will be imaged with negative staining. Samples 19-20 and 31-33 were diluted to a concentration of 0.7 mg/mL, and 39-41 to 0.8 mg/mL for negative staining (Table V).

Negative Staining – For samples 39-41, 31-33, 19-20, LS1-LS2, and LS3-LS4 negative staining grids were made. For this, 2% Uranyl Acetate (UA), dissolved in water, was centrifuged at maximum speed 11700 x g for 10 minutes, the clear supernatant was transferred to a new tube. Carbon Film Supported Copper Grids, 400-Mesh Copper (100) were used. These were placed on a glass microscope plate covered with parafilm, the carbon layer facing upward and were glow discharged, at 15 mA for 15 seconds. On a parafilm placed at the table, was added for each grid 3 droplets of ddH2O and 3 droplets of 2%-UA. First about 4 µL of protein was added to the grid (carbon side), and removed with filter paper (WhatmanTM Cat No 1001 090) after 1 minute. Then it was washed 3 times with ddH2O, by placing the grid on the water droplets and moving gently; each washing step took about 10 seconds. Then it was washed 2 times for 10 seconds with 2%-Uranyl Acetate, by placing the grid on top of the UA droplets and moving it gently. Afterwards, the last droplet of UA was incubated on the grid for 1 minute.

Cryo-EM grids – Cryo-EM grids were made for samples 24-26, 31-33, 39-41, and 19-20. Sample 31-33 was diluted 3 times, to a total concentration 2.9 mg/mL. A R2/2 copper grid was glow-discharged. Grids were prepared with a VitroBot, at a humidity of 90%, temperature of 20°C, using a wait time of 30 s., a blotting force 0, a blotting time of 4 s., and 3 µL sample.

Image Acquisition – The negative staining grids were imaged using the Talos L120C Electron Microscope, with a 120 kV lanthanum hexaboride (LaB6) electron source, 150 µm C2 aperture, 406 pm pixel size and the use of a BM-Ceta (FEI©) camera. Images were taken at a magnification of 36 000 x, 57 000 x, 92 000 x, or 150k x.

Data Processing 50S Ribosome (31-33) – The data was processed in RELION®. First MotionCorrection and CTF estimation were performed, after which bad micrographs were discarded, resulting in 945 micrographs that are further processed. The particles were picked with the Laplacion-of-Gaussian AutoPicker from RELION®, with a minimum and maximum log-filter diameter of 200 and 250, respectively. The particles were extracted with a boxsize of 600 pixels (px), and rescaled to 200 px, resulting in a pixel size of 3.12 A/px. About 130 000 particles were picked. A 2D classification was performed with these particles, with 50 classes, a mask of 250 Å, a Tau-fudge of 2, and ignoring the CTFs until the first peak. The best classes were selected, resulting in about 84 000 particles that were used for following processing steps.Another 2D classification was run, resulting in discarding about 1 000 particles. A 3D initial model was reconstructed. This was followed by a 3D classification round. The best classes were selected and the particles were extracted using less rescaling, using a boxsize of 300 px, rescaled to 200 px. Then AutoRefinements were run, in combination with 3D classifications. A mask was created for the best AutoRefine modelled and further refined in RELION®, using CTF refinements and Bayesian Polishing.

Table V. Dailysis NS samples

Sample D1	Volume (µL)	Added Volume (µL)	Sucrose%	Volume end (µL)	Concentration (mg/mL)
39-41	350	1750	10	250	1.64
31-33	350	1750	10	250	8.69
19-20	250	1250	10	250	2.90
LS1-LS2	120	600	10	250	0.02
LS3-LS4	250	1000	10	250	0.05

Table IV. Dialysis Sample 24-26

Sample D	Volume (mL)	Added Volume (mL)	Sucrose%	Volume 2 (mL)	Added Volume 2 (mL)	Sucrose%	Volume 3 (mL)	Added Volume 3 (mL)	Sucrose% end
24-26	1.2	13.8	4.80	1	14	0.32	1	14	0.02

Crude Purification Mix Samples SG – Two additional sucrose gradients were run, in a similar manner as described above, using a sucrose concentration of 30%-60% w/v and one with 30%-60% ww. Samples were collected with the fractionator, as described previously. Sample with the top peak at well 23, and 29 for the w/v and w/w sample, respectively were used for making cryo-EM grids, and data collection. The two datasets were combined and used for *in silico* purification.

VLP

During the purification different centrifuge steps were used. The different fractions are drawn in Figure , and the description of the used methods refers to the names used in this figure for each fraction (Figure 10).

Cell Lysis of P. furiosus – freeze-dried *Pyrococcus furiosus* cells were taken out of the -80°C freezer and thawed on ice. To these cells was added 2mL Cell Lysis Buffer Complete [10mM HEPES-NaOH (pH 7.0), 500mM NaCl, 5mM MgCl2, 20% Sucrose (w/v), [+/-] DNAse I (1:10000)(0.1 mg/mL), cComplete protease stop (1/4 tab)]. The cells were lysed by using a 2 mL douncer, and moving up and down for 40 times with pestle A, followed by 50 times with pestle B. The dounced cells were transferred to 2 Eppendorf tubes and the cells were then centrifuged at 4°C, at 6 000 rpm for 10 minutes to separate the broken cells from the unbroken cells. The supernatant was transferred to new Centrifuge tubes, the pellet is frozen, this contains the cell debris.

Ultracentrifuge Samples – The supernatant with broken cells was ultracentrifuged with a Ti70 rotor at 39 500rpm for 15 minutes, at 4°C (Optima XPN 80). After ultracentrifuging, three layers can be seen. On the bottom is a Pellet Layer, in the Middle Layer is a soluble, cloudy layer, and on top a layer which is more transparent compared to the middle layer (Upper Layer). The Upper Layer (UL) was transferred to a new Eppendorf tube, as well is the Middle Layer. Both the Upper and Middle Layer will be used for further purification, the Pellet Layer is stored in the -80°C freezer.

Size Exclusion Chromatography Upper Layer – For separating the protein fractions of the Upper Layer size Exclusion Chromatography (SEC) is performed, using a Superose® 6 Prep Grade HiLoad® column. The column was calibrated with degassed Milli-Q® water overnight for 2 column volumes, whereafter it was calibrated with the Running Buffer A [[10mM HEPES-NaOH (pH 7.0), 500mM NaCl, 5mM MgCl2, 10% Sucrose (w/v)] for 1 column volume. The Upper Layer sample was applied to this column, with a flow rate about 0.45 mL/min, and collected in a 96-deep-well plate, in the serpentine-row fractionation order. Each fraction contains 1.5mL.

Four peaks eluded from the column, B3-B1 (B1), C1-C5 (C2), C6-D1 (D9 & D5), and E1-F4 (E9 & F7), with the top of the peaks in brackets. For the top peaks was 10µL transferred to be used for checking of the sample with negative staining (NS). The encapsulins are expected to be present in sample C6-D1 and E1-F4, and these samples will be used for further purifications steps.

Concentrating C6-D1 and E1-F4 – For better separation, the protein fractions C6-D1 and E1-F4 eluded from the Superose® 6 Column will be loaded unto a sucrose gradient. As these samples are diluted and have too much volume after the AKTA Size Exclusion, they have been concentrated first. For concentrating the samples Centrifugal Filter Units with a Molecular Weight Cut-Off (MWCO) of 100kDa for C6-D1, and a MWCO of 50kDa for E1-F4 were used at a speed of 3 000 x g, till the total volume reached below 500µL. The membranes were washed twice with 50µL of the Cell Lysis Buffer 10% Sucrose [10mM HEPES-NaOH (pH 7.0), 500mM NaCl, 5mM MgCl2, 10% Sucrose (w/v)].

Sucrose Gradient C6-D1 and E1-F4 – The Upper Layer samples, after SEC and concentrating them, will be applied to a sucrose gradient. Sample C6-D1 will be applied to a 10%-50% sucrose gradient, whereas for E1-F4 a sucrose gradient from 10% to 30% will be used. The Piston Gradient Fractionator (BioComp©) was used for making the sucrose gradients. First the 10% sucrose [10% Sucrose (w/v) or 60% Sucrose (w/v), 50mM HEPES-NaOH (pH 7.0), 500mM NaCl, 3mM MgCl2], was added to the two tubes. Then using a syringe, slowly the 30% sucrose (w/v) [30% Sucrose (w/v) or 60% Sucrose (w/v), 50mM HEPES-NaOH (pH 7.0), 500mM NaCl, 3mM MgCl2], was added, by adding it to the bottom of the tube, the same was done for the 50% sucrose (w/v) [50% Sucrose (w/v) or 60% Sucrose (w/v), 50mM HEPES-NaOH (pH 7.0), 500mM NaCl, 3mM MgCl2].The gradient maker used the settings for a 10-30% (w/v) sucrose gradient and for the 10-50% (w/v). The samples were loaded on top of the sucrose gradient, and centrifuged overnight using the Swinging Bucket SW 32 Ti rotor, at a speed of 32 000rom for 18 hrs at 4°C.

The Piston Gradient Fractionator in combination with the Fraction Collector (FC203B, Gilson) was used to collect different proteins fractions in a 96-wells plate and the UV-absorption at 260 nm (DNA/RNA) and 280 nm (Proteins) was measured. Each fraction contains 200µL. The last fractions at the bottom of the tube (high sucrose concentration) were collected by hand. For 10%-30% all pellets, except the last one, were combined. For 10-50 the pellet was collected in steps of 200ul, from P1 to P7.

For the E1-F4 10%-30% Sucrose gradient eluded 2 peaks, from wells 1-52, and 56-end. For sample C6-D1 eluded 4 peaks, the first peak from well 1-30 (20), the other ones from 31-33 (32), from 34-49 (38), and the last one from 50-69 (54). The wells containing the top of the peak are indicated inside brackets.

The encapsulins are suspected to be in the last peak of the C6-D1 sample, and in the pellet of the E1-F4 sample, based on previous work done by W. Song. For each even number of the last peak plus samples 69 and P7, will be used for the concentration was measured. These samples will also be used for negative staining, in order to determine which samples could be used for further purification steps.

Second Size Exclusion Chromatography, UL, SEC, SG 50-69 – Samples SG 50-69, which is the Upper Layer after SEC and the Sucrose Gradient of C6-D1, was used for a second run of size Exclusion Chromatography (SEC), using a Superose® 6 Increase 10/300 GL column. First different mixtures were made, which will all be separately applied to the column. These mixtures were based on negative staining results, as samples below 57 were not pure enough and didn't contain sufficient amounts of encapsulins. Samples 57-59, 60-62, 63-65, 66-67, 68-P1, P2-P4, and P5-P7 were mixed together. The column was calibrated with degassed Milli-Q® water overnight for 2 column volumes, whereafter it was calibrated with the Running Buffer A [[10mM HEPES-NaOH (pH 7.0), 500mM NaCl, 5mM MgCl2, 10% Sucrose (w/v)] for 2 column volumes. The samples were applied to this column, with a flow rate about 0.50 mL/min, and collected in a 96-deep-well plate, in the serpentine-row fractionation order. Each fraction contains 500µL.

Two peaks eluded from the column for each sample, with the first peak eluting around 8mL. This peak is containing most likely the encapsulins, as based on previous work done by W. Song.

The samples were pelleted down by ultracentrifuging them at a speed of 39 500 rpm for 2.5 hrs at 4°C (Ti70 rotor). The concentration was measured for each of the fractions, and negative staining grids were made. The SEC peak fractions from sample 57-59, 60-62, 68-P1 and P5-P7 were mixed, as well as the peaks for 60-62 E3, 63-65, 66-67, and P2-P4, now named sample 1 and 2, respectively. For these two samples cryo-EM grids were made. Sample 2 was applied twice to the grid, each time 3 µL, with a waiting time of 30s, and a blot time of 3 s. For sample 1, two grids were made, in a same manner as described for sample 2, but this time only applied once, and for the other grid thrice.

Image Acquisition – The negative staining grids were imaged using the Talos L120C Electron Microscope, with a 120 kV lanthanum hexaboride (LaB6) electron source, 150 µm C2 aperture, 406 pm pixel size and the use of a BM-Ceta (FEI©) camera. Images were taken at a magnification of 36 000 x, 57 000 x, 92 000 x, or 150k x.

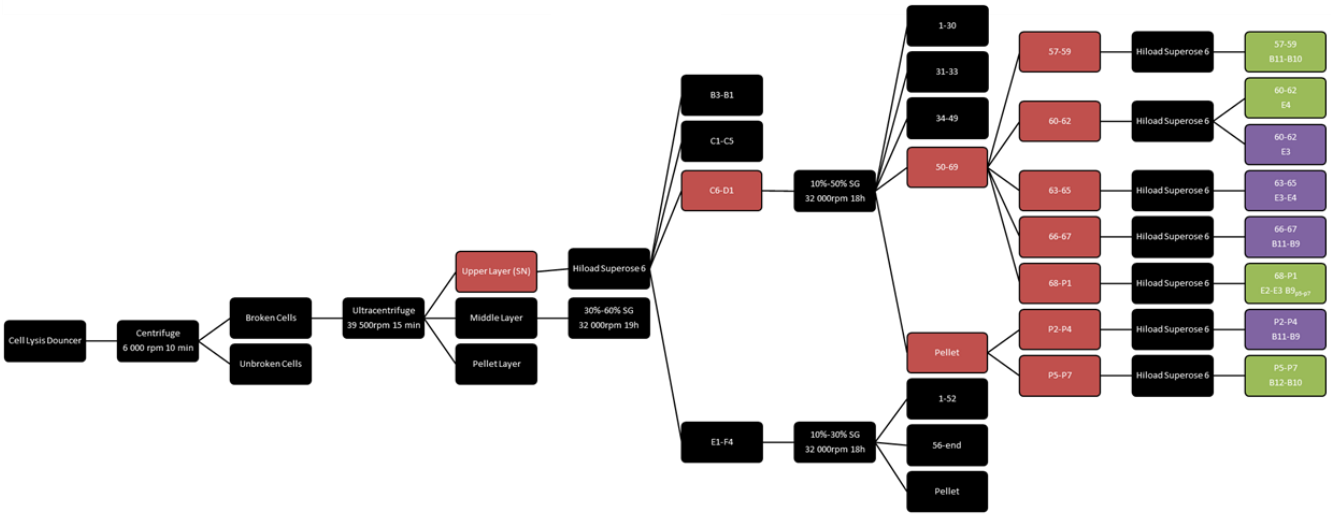


Figure 10. **Encapsulin/VLP Purification.** After cell lysis of the *Pyrococcus furiosus*, the broken cells were centrifuged at 39 500 rpm for 15 minutes. This resulted in three layers, a pellet layer at the bottom, a soluble cloudy middle layer, and a transparent top layer. The top layer was loaded on a Superose6 column, to purify with size exclusion chromatography. The best samples were applied to a Sucrose Gradient. Samples were concentrated and run for another round of SEC. The best samples were mixed together (green), as were the others (purple) to use for cryoEM imaging.

ACKNOWLEDGEMENTS

I would like to thank Wenfei Song, for supervising me during this project, and prof. dr. Friedrich Förster for providing me the opportunity to do this major internship. Furthermore, I would express my appreciation for Menno Bergmeijer for helping me with Relion©. Finally I want to thank the technical staff, including Joke Granneman, Louris Feitsma and Savanne Beeker. Especially Louris for helping with the ÅKTA.

REFERENCES

Akanuma, G. (2021). Diverse relationships between metal ions and the ribosome. *Bioscience, Biotechnology and Biochemistry*, *85*(7), 1582–1593. https://doi.org/10.1093/bbb/zbab070

Akita, F., Chong, K. T., Tanaka, H., Yamashita, E., Miyazaki, N., Nakaishi, Y., Suzuki, M., Namba, K., Ono, Y., Tsukihara, T., & Nakagawa, A. (2007). The Crystal Structure of a Virus-like Particle from the Hyperthermophilic Archaeon Pyrococcus furiosus Provides Insight into the Evolution of Viruses. *Journal of Molecular Biology*, *368*(5), 1469–1483. https://doi.org/10.1016/j.jmb.2007.02.075

Armache, J. P., Anger, A. M., Márquez, V., Franckenberg, S., Fröhlich, T., Villa, E., Berninghausen, O., Thomm, M., Arnold, G. J., Beckmann, R., & Wilson, D. N. (2013). Promiscuous behaviour of archaeal ribosomal proteins: Implications for eukaryotic ribosome evolution. *Nucleic Acids Research*, *41*(2), 1284–1293. https://doi.org/10.1093/nar/gks1259

Bai, X. chen, McMullan, G., & Scheres, S. H. W. (2015). How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, *40*(1), 49–57. https://doi.org/10.1016/j.tibs.2014.10.005

Baker, M. (2018). Cryo-Electron Microscopy Shapes Up. *Nature*, 565–567.

Barik, S. (2020). Evolution of protein structure and stability in global warming. *International Journal of Molecular Sciences*, *21*(24), 1–22. https://doi.org/10.3390/ijms21249662

Bell, S. D., & Jackson, S. P. (1998). Transcription and translation in Archaea: A mosaic of eukaryal and bacterial features. *Trends in Microbiology*, *6*(6), 222–228. https://doi.org/10.1016/S0966-842X(98)01281-5

Blumentals, I. I., Brown, S. H., Schicho, R. N., Skaja, A. K., Costantino, H. R., & Kelly, R. M. (1990). The Hyperthermophilic Archaebacterium, Pyrococcus furiosus Development of Culturing Protocols, Perspectives on Scaleup, and Potential Applications. In *Annals of the New York Academy of Sciences* (Vol. 589, Issue 1). https://doi.org/10.1111/j.1749-6632.1990.tb24254.x

Bolhuis, A. (2004). The archaeal Sec-dependent protein translocation pathway. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1446), 919–927. https://doi.org/10.1098/rstb.2003.1461

Cavicchioli, R. (2011). Archaea - Timeline of the third domain. *Nature Reviews Microbiology*, *9*(1), 51–61. https://doi.org/10.1038/nrmicro2482

Danev, R., Yanagisawa, H., & Kikkawa, M. (2019). Cryo-Electron Microscopy Methodology: Current Aspects and Future Directions. *Trends in Biochemical Sciences*, *44*(10), 837–848. https://doi.org/10.1016/j.tibs.2019.04.008

Eme, L., Spang, A., Lombard, J., Stairs, C. W., & Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, *15*(12), 711–723. https://doi.org/10.1038/nrmicro.2017.133

Giessen, T. W., Orlando, B. J., Verdegaal, A. A., Chambers, M. G., Gardener, J., Bell, D. C., Birrane, G., Liao, M., & Silver, P. A. (2019). *Large protein organelles form a new iron sequestration system with high storage capacity*. https://doi.org/10.7554/eLife.46070.001

Giessen, T. W., & Silver, P. A. (2017). Widespread distribution of encapsulin nanocompartments reveals functional diversity. *Nature Microbiology*, *2*. https://doi.org/10.1038/nmicrobiol.2017.29

He, D., Piergentili, C., Ross, J., Tarrant, E., Tuck, L. R., Logan Mackay, C., McIver, Z., Waldron, K. J., Clarke, D. J., & Marles-Wright, J. (2019). Conservation of the structural and functional architecture of encapsulated ferritins in bacteria and archaea. *Biochemical Journal*, *476*(6), 975–989. https://doi.org/10.1042/BCJ20180922

Henderson, R. (1995). The Potential and Limitations of Neutrons, Electrons and X-Rays for Atomic Resolution Microscopy of Unstained Biological Molecules. *Quarterly Reviews of Biophysics*, *28*(2), 171–193. https://doi.org/10.1017/S003358350000305X

Ho, C. M., Li, X., Lai, M., Terwilliger, T. C., Beck, J. R., Wohlschlegel, J., Goldberg, D. E., Fitzpatrick, A. W. P., & Zhou, Z. H. (2020). Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu. *Nature Methods*, *17*(1), 79–85. https://doi.org/10.1038/s41592-019-0637-y

Hutchins, A. M., Holden, J. F., & Adams, M. W. W. (2001). Phosphoenolpyruvate synthetase from the hyperthermophilic archaeon Pyrococcus furiosus. *Journal of Bacteriology*, *183*(2), 709–715. https://doi.org/10.1128/JB.183.2.709-715.2001

Jones, J. A., & Giessen, T. W. (2021). Advances in encapsulin nanocompartment biology and engineering. In *Biotechnology and Bioengineering* (Vol. 118, Issue 1, pp. 491–505). John Wiley and Sons Inc. https://doi.org/10.1002/bit.27564

Li, X., Mooney, P., Zheng, S., Booth, C. R., Braunfeld, M. B., Gubbens, S., Agard, D. A., & Cheng, Y. (2013). Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods*, *10*(6), 584–590. https://doi.org/10.1038/nmeth.2472

Londei, P., & Ferreira-Cerca, S. (2021). Ribosome Biogenesis in Archaea. *Frontiers in Microbiology*, *12*(July), 1–13. https://doi.org/10.3389/fmicb.2021.686977

Maguire, B. A., & Zimmermann, R. A. (2001). The ribosome in focus. *Cell*, *104*(6), 813–816. https://doi.org/10.1016/S0092-8674(01)00278-1

McHugh, C. A., Fontana, J., Nemecek, D., Cheng, N., Aksyuk, A. A., Heymann, J. B., Winkler, D. C., Lam, A. S., Wall, J. S., Steven, A. C., & Hoiczyk, E. (2014). A virus capsid-like nanocompartment that stores iron and protects bacteria from oxidative stress. *The EMBO Journal*, *33*(17), 1896–1911. https://doi.org/10.15252/embj.201488566

Namba, K., Hagiwara, K., Tanaka, H., Nakaishi, Y., Chong, K. T., Yamashita, E., Armah, G. E., Ono, Y., Ishino, Y., Omura, T., Tsukihara, T., & Nakagawa, A. (2005). Expression and molecular characterization of spherical particles derived from the genome of the hyperthermophilic euryarchaeote Pyrococcus furiosus. *Journal of Biochemistry*, *138*(2), 193–199. https://doi.org/10.1093/jb/mvi111

Nichols, R. J., Cassidy-Amstutz, C., Chaijarasphong, T., & Savage, D. F. (2017). Encapsulins: molecular biology of the shell. In *Critical Reviews in Biochemistry and Molecular Biology* (Vol. 52, Issue 5, pp. 583–594). Taylor and Francis Ltd. https://doi.org/10.1080/10409238.2017.1337709

Schmitt, E., Coureux, P. D., Kazan, R., Bourgeois, G., Lazennec-Schurdevin, C., & Mechulam, Y. (2020). Recent Advances in Archaeal Translation Initiation. *Frontiers in Microbiology*, *11*(5). https://doi.org/10.3389/fmicb.2020.584152

Su, C. C., Lyu, M., Morgan, C. E., Bolla, J. R., Robinson, C. V., & Yu, E. W. (2021). A 'Build and Retrieve' methodology to simultaneously solve cryo-EM structures of membrane proteins. *Nature Methods*, *18*(1), 69–75. https://doi.org/10.1038/s41592-020-01021-2

Sutter, M., Boehringer, D., Gutmann, S., Günther, S., Prangishvili, D., Loessner, M. J., Stetter, K. O., Weber-Ban, E., & Ban, N. (2008). Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural and Molecular Biology*, *15*(9), 939–947. https://doi.org/10.1038/nsmb.1473

Trabuco, L. G., Schreiner, E., Eargle, J., Cornish, P., Ha, T., Luthey-Schulten, Z., & Schulten, K. (2010). The Role of L1 Stalk-tRNA Interaction in the Ribosome Elongation Cycle. *Journal of Molecular Biology*, *402*(4), 741–760. https://doi.org/10.1016/j.jmb.2010.07.056

Verbeke, E. J., Mallam, A. L., Drew, K., Marcotte, E. M., & Taylor, D. W. (2018). Classification of Single Particles from Human Cell Extract Reveals Distinct Structures. *Cell Reports*, *24*(1), 259-268.e3. https://doi.org/10.1016/j.celrep.2018.06.022

Vieille, C., & Zeikus, G. J. (2001). Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. *Microbiology and Molecular Biology Reviews*, *65*(1), 1–43. https://doi.org/10.1128/mmbr.65.1.1-43.2001

Wenck, B. R., & Santangelo, T. J. (2020). Archaeal transcription. *Transcription*, *11*(5), 199–210. https://doi.org/10.1080/21541264.2020.1838865

Yip, K. M., Fischer, N., Paknia, E., Chari, A., & Stark, H. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature*, *587*(7832), 157–161. https://doi.org/10.1038/s41586-020-2833-4

SUPPLEMENTS

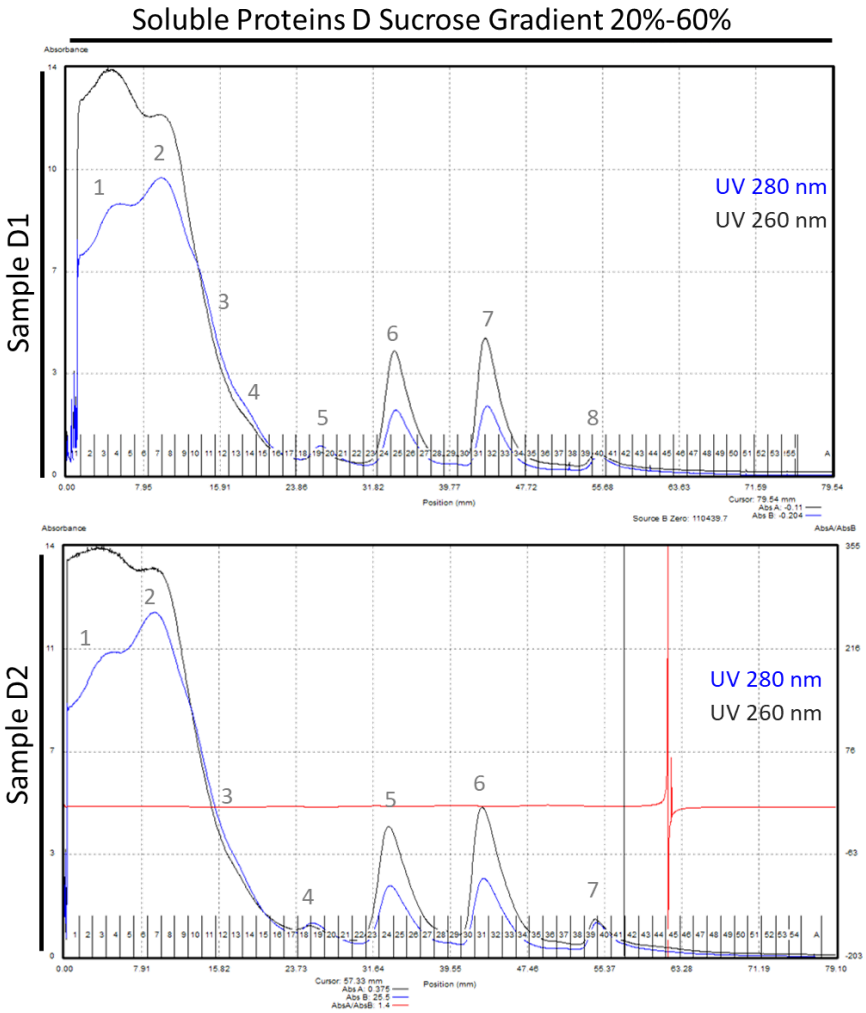


Figure S1. **MGP Purification 20%-60% (w/w) Sucrose Gradient.** After cell lysis, the broken cells were first centrifuged at 100 000 rcf for 50 minutes. The supernatant was applied to a 20%-60% (w/w) sucrose gradient which was ultracentrifuged at 37 000 rpm for 17 hours. The different fractions were collected in steps of 200 µL. UV-absorption was measured at 260 nm for DNA/RNA, black graph, and at 280 nm for proteins, the blue

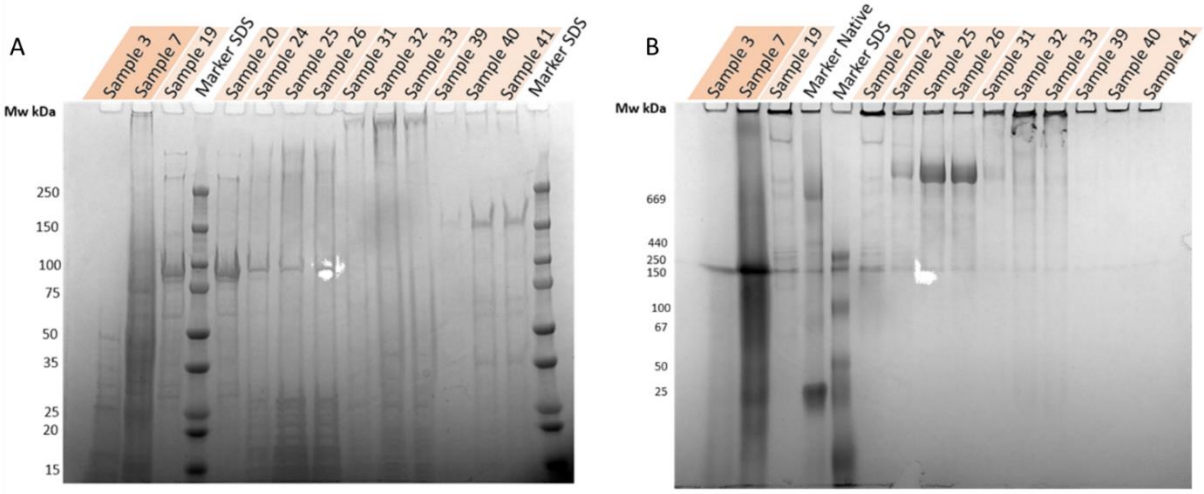


Figure S2. **MGP Purification. Gel results for fractions after a 20%-60% sucrose gradient.** A) SDS-PAGE gel. B) Native gel. Sample 7 still contains a lot of different proteins. Samples 19-20 show bands around the 90 kDa (A), and 150 kDa (B), indicating the possible presence of PpsA. Samples 31-33, and 24-26 show low molecular weight bands at the bottom of the SDS-PAGE gel (A), indicative of ribosomes.

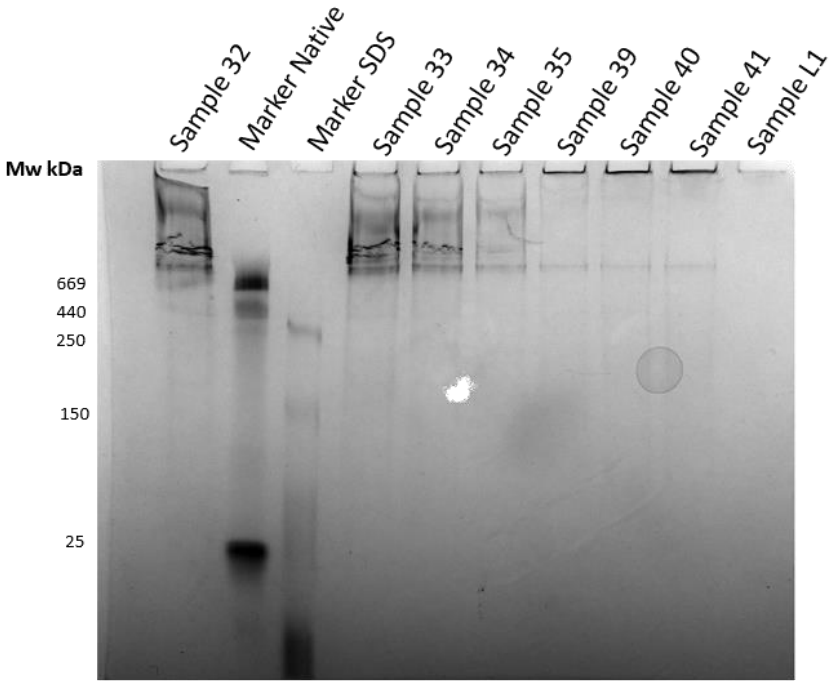


Figure S3. **MGP Purification Additional Native Gel.** Fractions 32-35, and 39-41 and L1 after a 20%-60% (w/w) sucrose gradient were run on an additional native gel

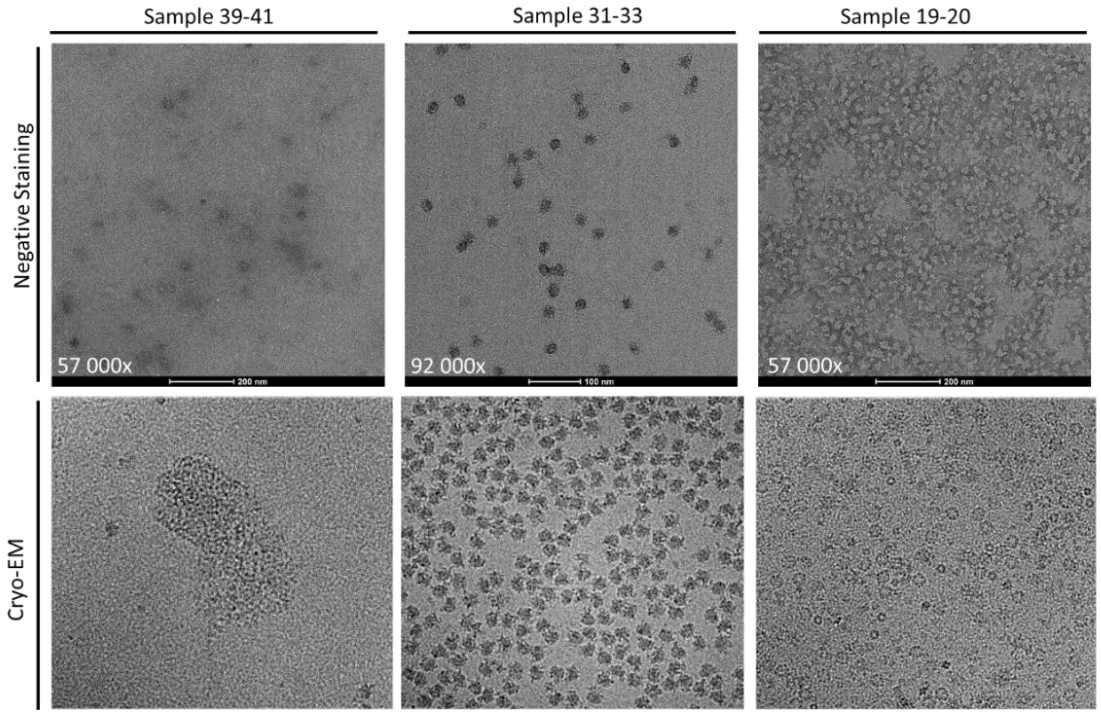


Figure S4. **MGP Purification Negative Staining and CryoEM Images.** Negative staining, on the top, and cryoEM images on the bottom for the samples 39-41 (left), sample 31-33 (middle) and 19-20 (right). Sample 39-41 contains mostly cell debris, as can be seen in both the negative staining and cryoEM images. Sample 19-20 and 31-30 contain a lot of proteins, as can be seen from NS and CryoEM images. Sample 31-33 contains ribosomes, which can be clearly seen in the cryoEM image, whereas sample 19-20 contains PpsA.

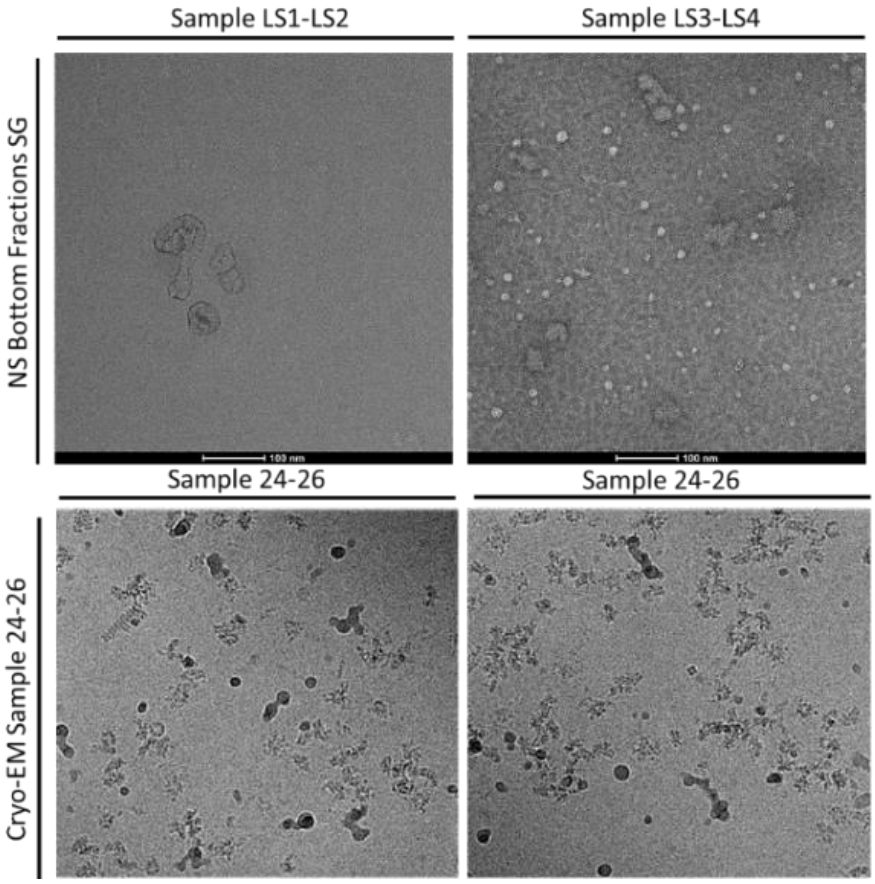


Figure S5. **MGP Purification LS1-4 and Sample 24-26.** Negative staining, on the top, for the bottom fractions of the sucrose gradient 20%-60% w/w that were collected by hand. LS4 is the most bottom layer. From these images can be seen that these samples contain mostly cell debris. On the bottom are cryoEM for the sample 24-26. This sample was directly imaged with cryoEM, and no negative staining grids were made. This sample also contains a lot of cell debris. Both samples weren't further used for further purification and/or processing.

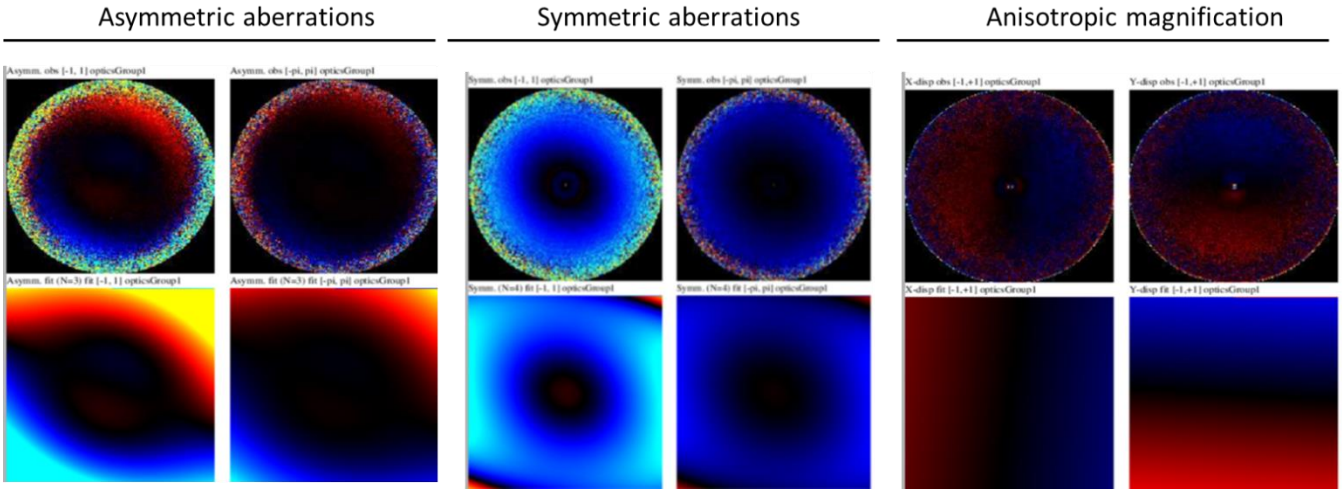


Figure S6**. 50S Ribosome Higher-Order Abberations Refinement Relion.** Accumulated averages for asymmetric abberations, like beam tilt and trefoil, symmetric aberattions, like tetrafoil and spherical abberations, and anisotropic magnification, on the left, middle and right, respectively.

AutoPicker Crude Purification mix

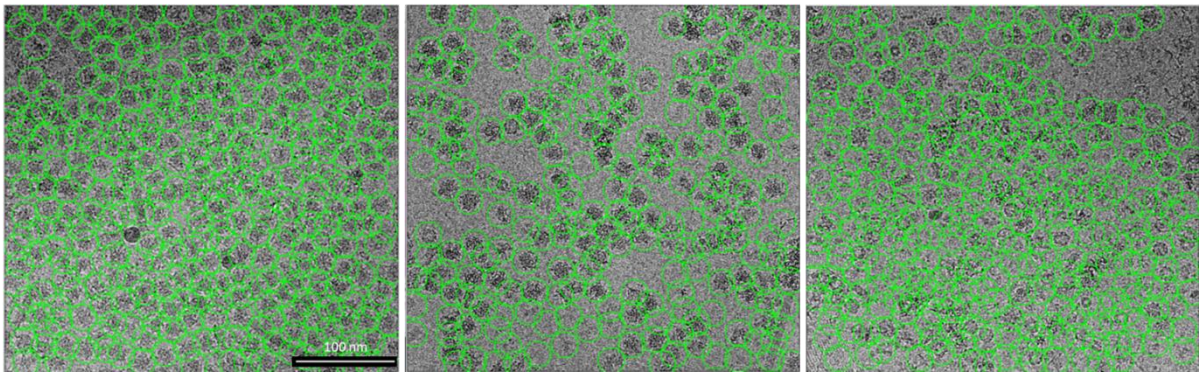


Figure S7**. MGP Crude Purification AutoPicking.** Micrographs containing a mix of different proteins. AutoPicking parameters were chosen to pick all the different kind of particles.

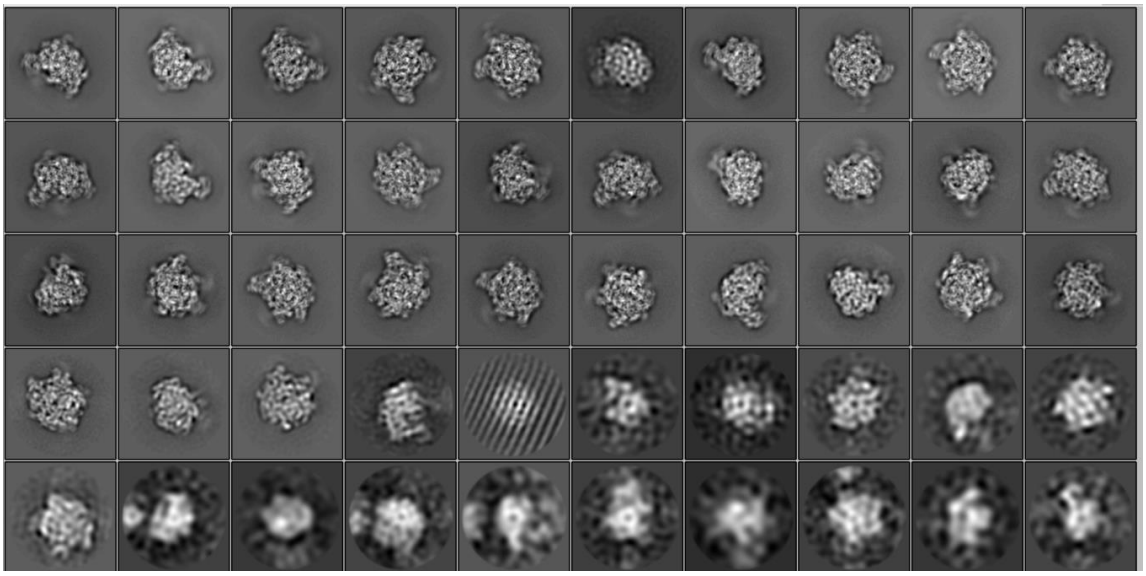


Figure S8**. 2D-Classes for the 50S Ribosome.** The good classes contain 121 158 particles, the bad classes 4266 particles.

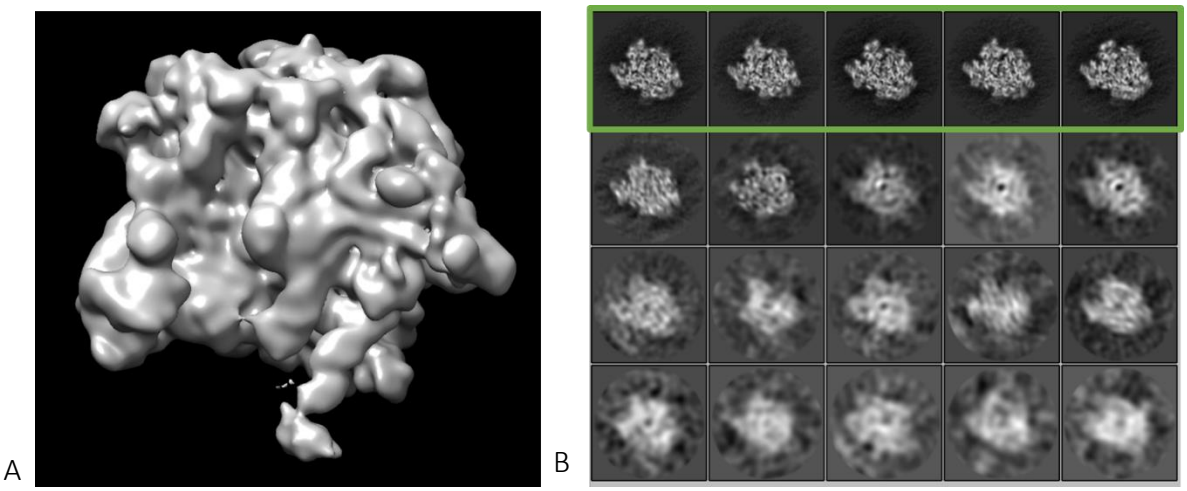


Figure S9**. Initial Model and 3D Classes for the 50S Ribosome of the Mixed Dataset.** A) Initial Model of the 50S Ribosome. B) 3D-classes of the 50S ribosome, with the initial model as reference. The best classes are outlined in green and contain 109 175 particles.

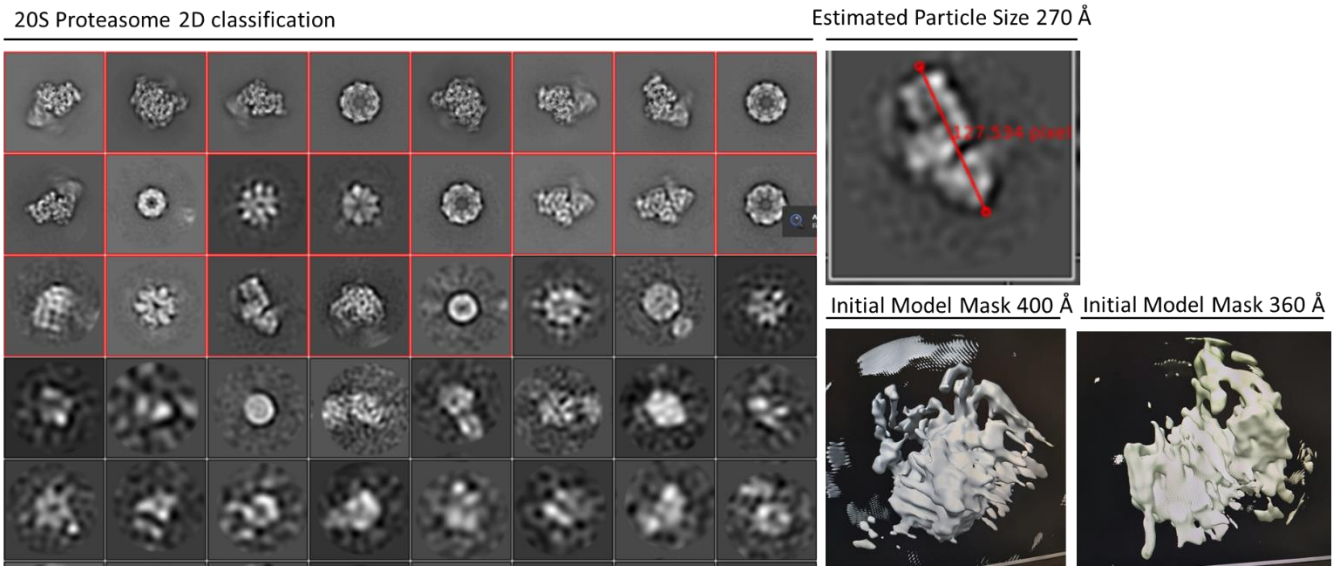


Figure S10**. MGP Crude Purification 20S Proteasome.** 2D classification of particles grouped as 20S proteasome, showing possible capped 20S proteasomes. An estimation of the size was about 270 Å, pixel size is 2.08 Å/px. Initial models didn't give a nice result.

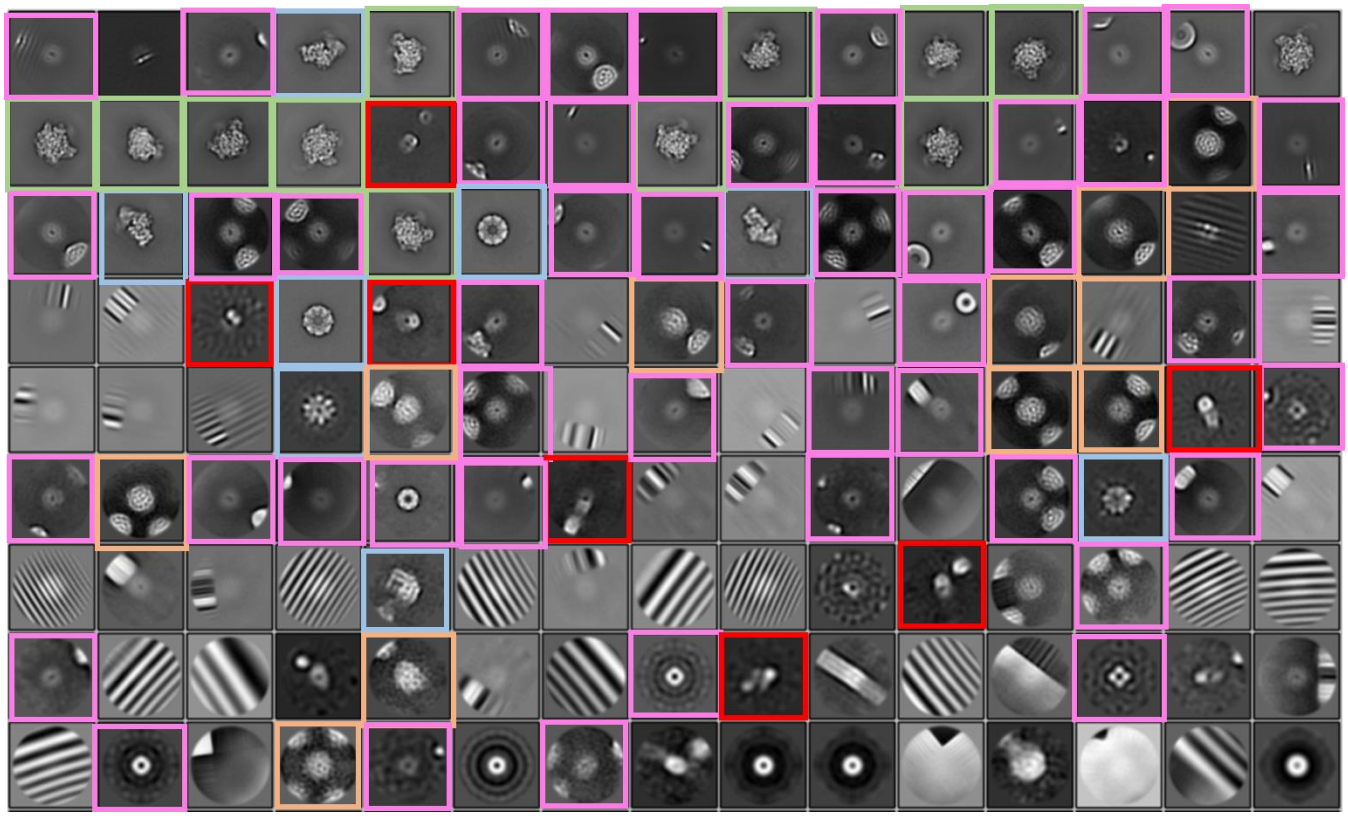


Figure S11. **Enlarged Image of the *in silico* Purification of the Mixed Dataset.** For the mix dataset, *in silico* purification was performed using 2D classification to identify and sort the different particles. The 50S ribosome can be easily recognized, outlined in blue. Also, possibly capped 20S proteasomes are present in our dataset, outlined in green. The particles were also sorted based on their size, with in pink particles with a size about 140 Å, in orange a size of 180 Å, and in red 100 Å (4b).

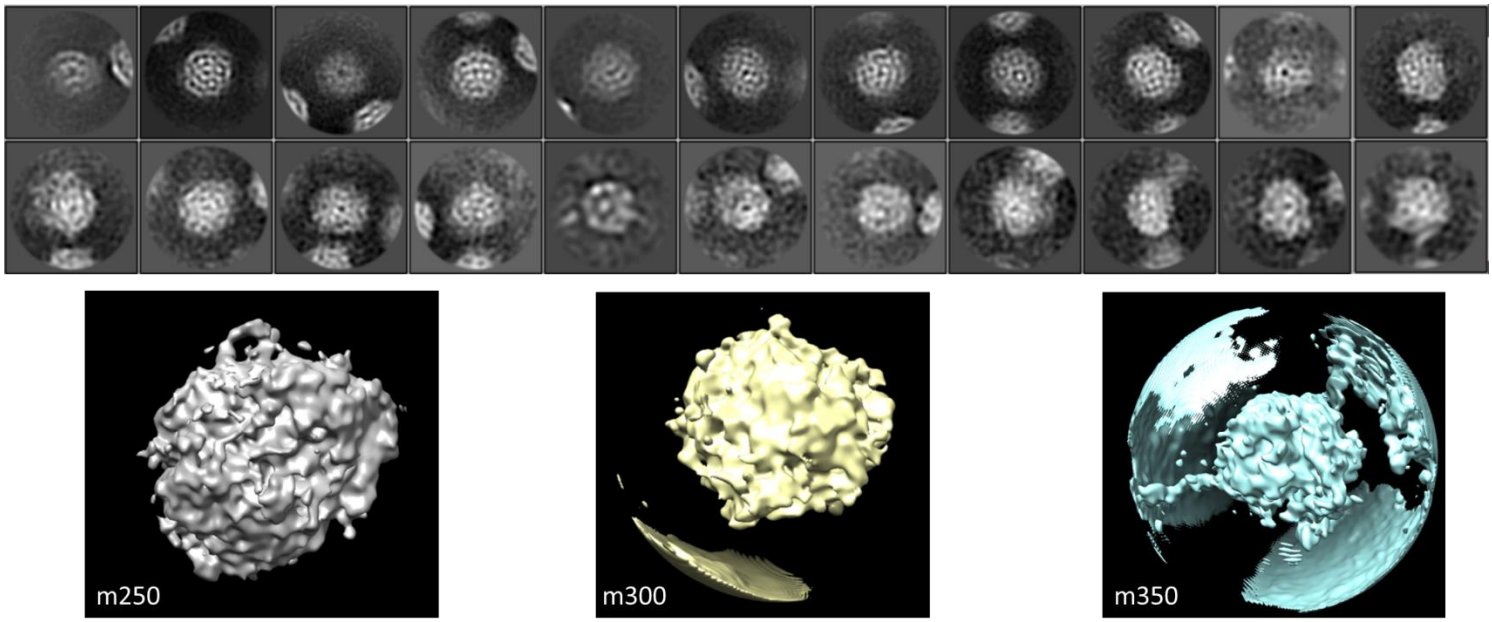


Figure S12.**Classes expected to belong to the "Spherical Protein" are selected from the 200 Classes to be used for building an Initial Model.** The number of particles belonging to the "Spherical Protein" is 17 753. These were then used to built an Initial Model, whereby different mask sizes of 250 Å, 300 Å, and 350 Å were tested. The tightest mask of 250 Å yielded the best result. The identity of this protein is not yet known.

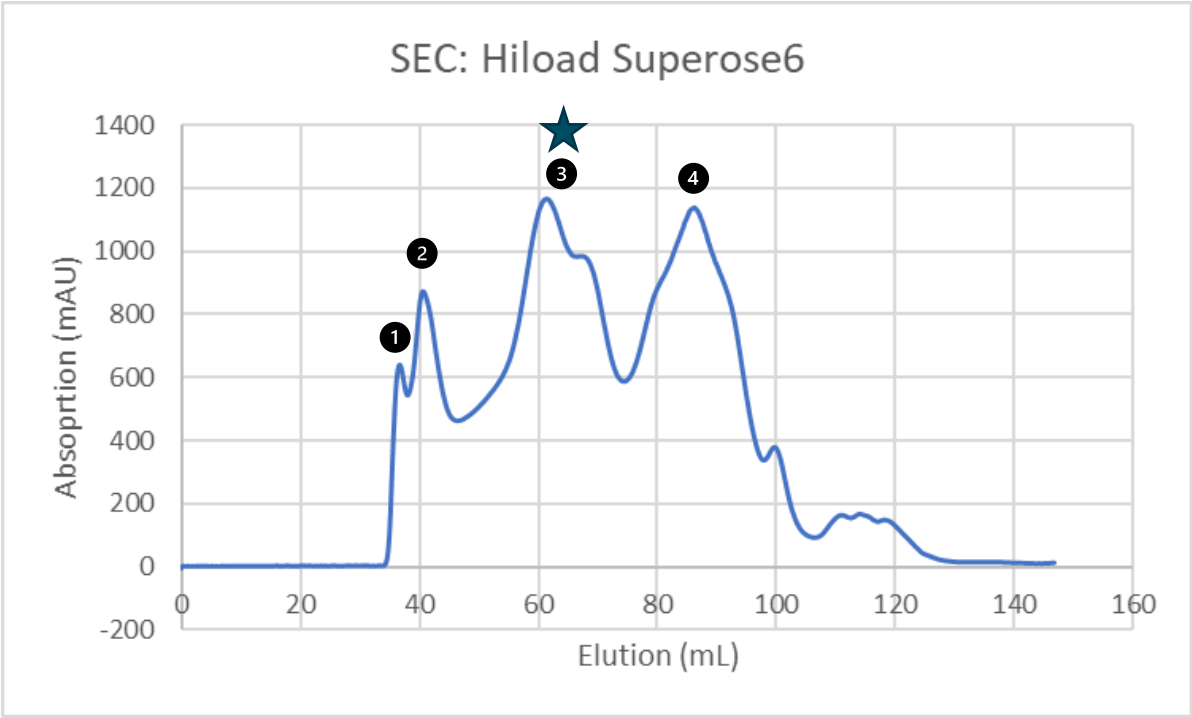


Figure S13. **Size Exclusion Chromatogram of the first SEC of the Upper Layer with Encapsulins Present in the Third Peak.** After the broken cells were ultracentrifuged at 39 500 rpm for 15 minutes, the clear upper layer was applied to a Hiload Superose© 6 column, resulting in the separation of proteins in different fractions. The encapsulins eluted roughly at the 60 mL, wells C6-D1.

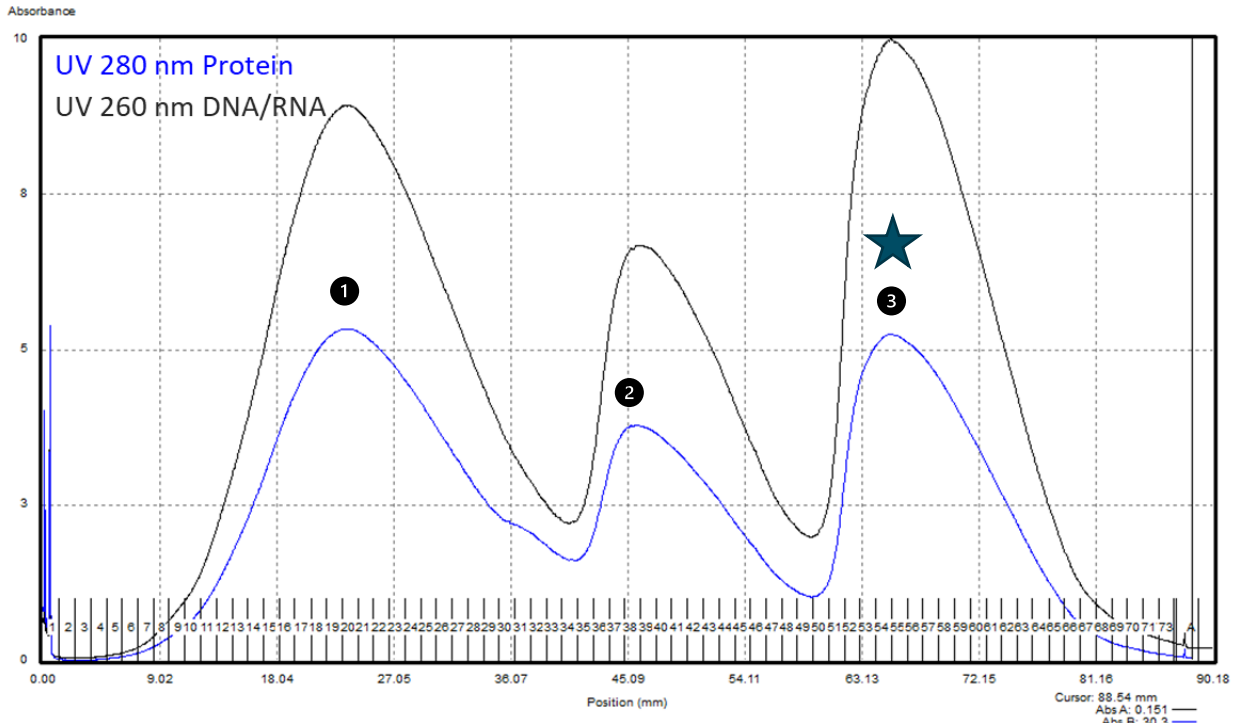


Figure S14. **The C6-D1 Sample was Applied to a 10%-60% Sucrose Gradient (w/v) resulting in Encapsulins present at approximately 50% Sucrose Concentration.** After applying a sucrose gradient to further separate the different proteins, the encapsulins are present in the third peak, wells 50-69, based on previous work done by W. Song.

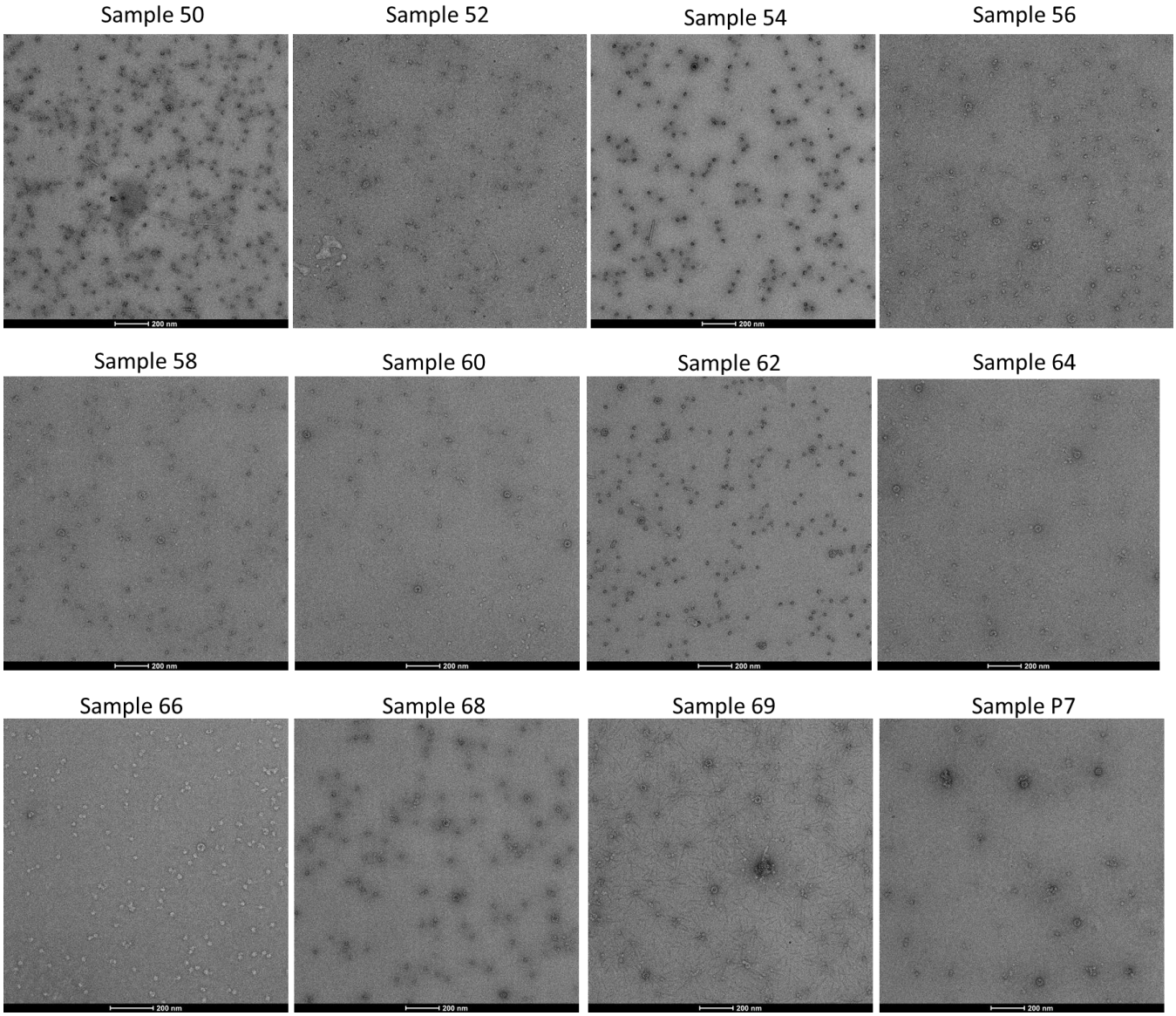


Figure S15**. VLP/Encapsulin Purification. NS Check After 1st Size Exclusion and 10%-50% Sucrose Gradient.** Each even number, plus samples 69 and P7 were screened by negative stain EM. The concentration of each sample can be found in table I.

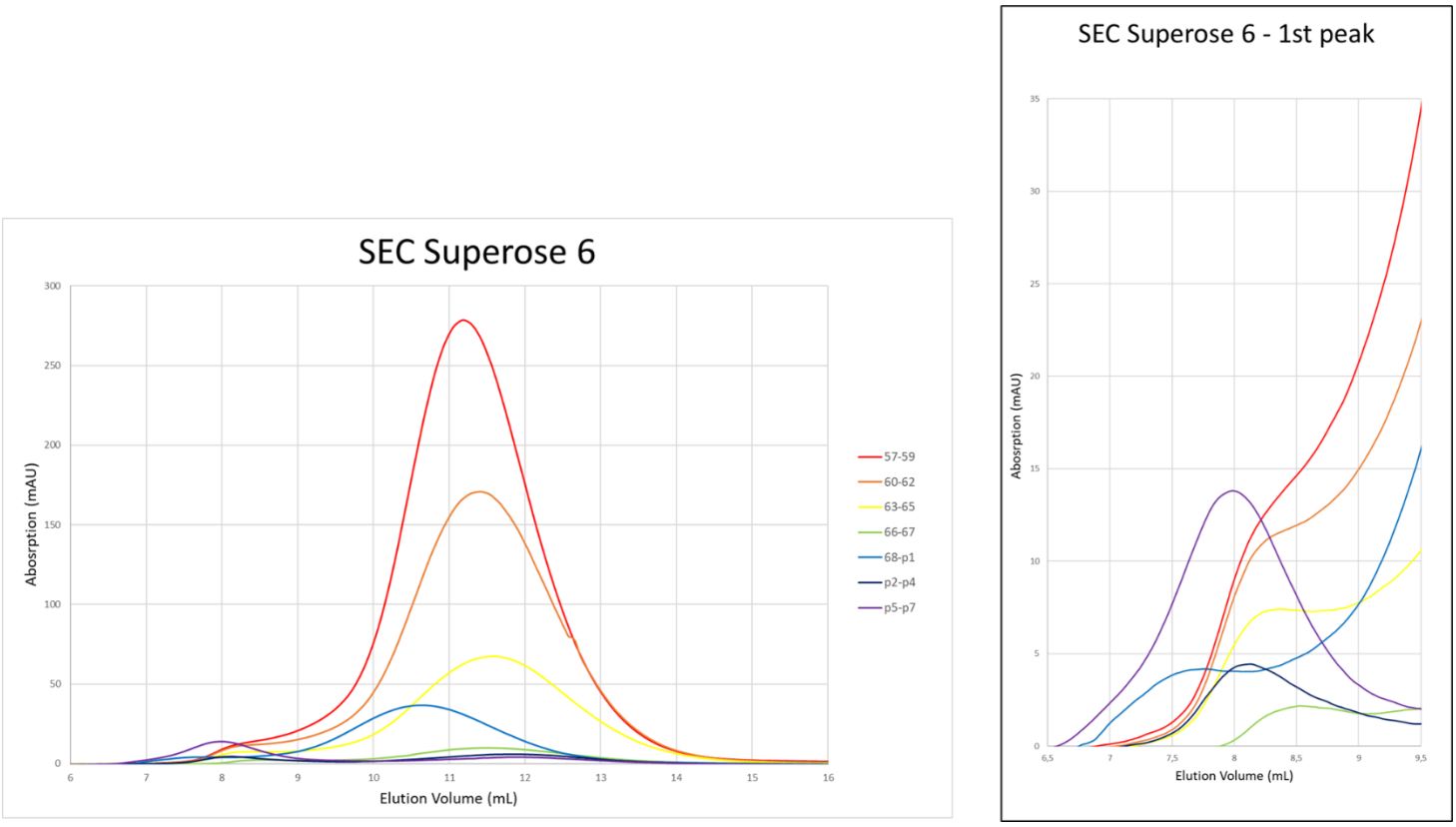


Figure S16**. Second Size Exclusion Superose© 6 whereby Encapsulins eluted at 8 mL.** The graph on the left shows the absorption peaks for the different sample mixtures, including a legend to indicate which graphs belongs to which sample. On the right is a zoom-in on the smaller protein fraction peak at approximately 8 m, where the encapsulins should elute. Especially the most bottom fractions, like P5-P7 and P2-P4 show a peak at the 8 mL.

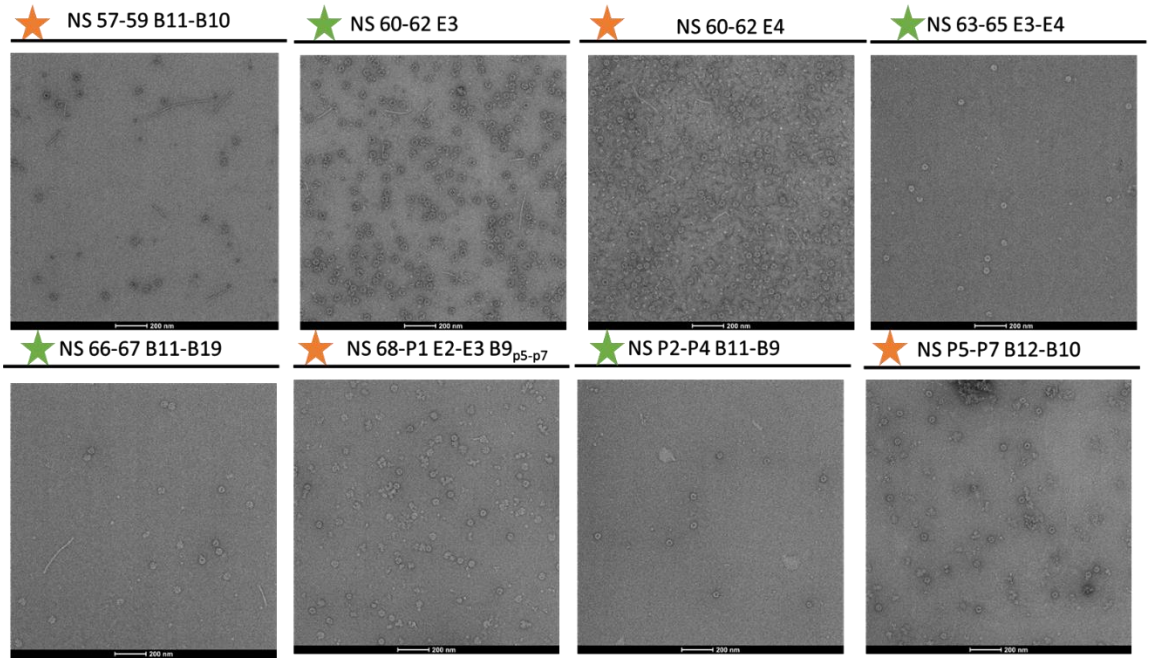


Figure S17**. VLP/Encapsulin Purification. NS Check After 2nd Size Exclusion.** Fractions containing Virus-Like Particles elute around the 8mL from the Superose6 column.The samples belonging to this peak were used for negative-staining screening. Based on their purity, the samples were combined into either a more pure or less pure sample, indicated with a green or orange star, respectively. The concentration of the samples can be found in table II.