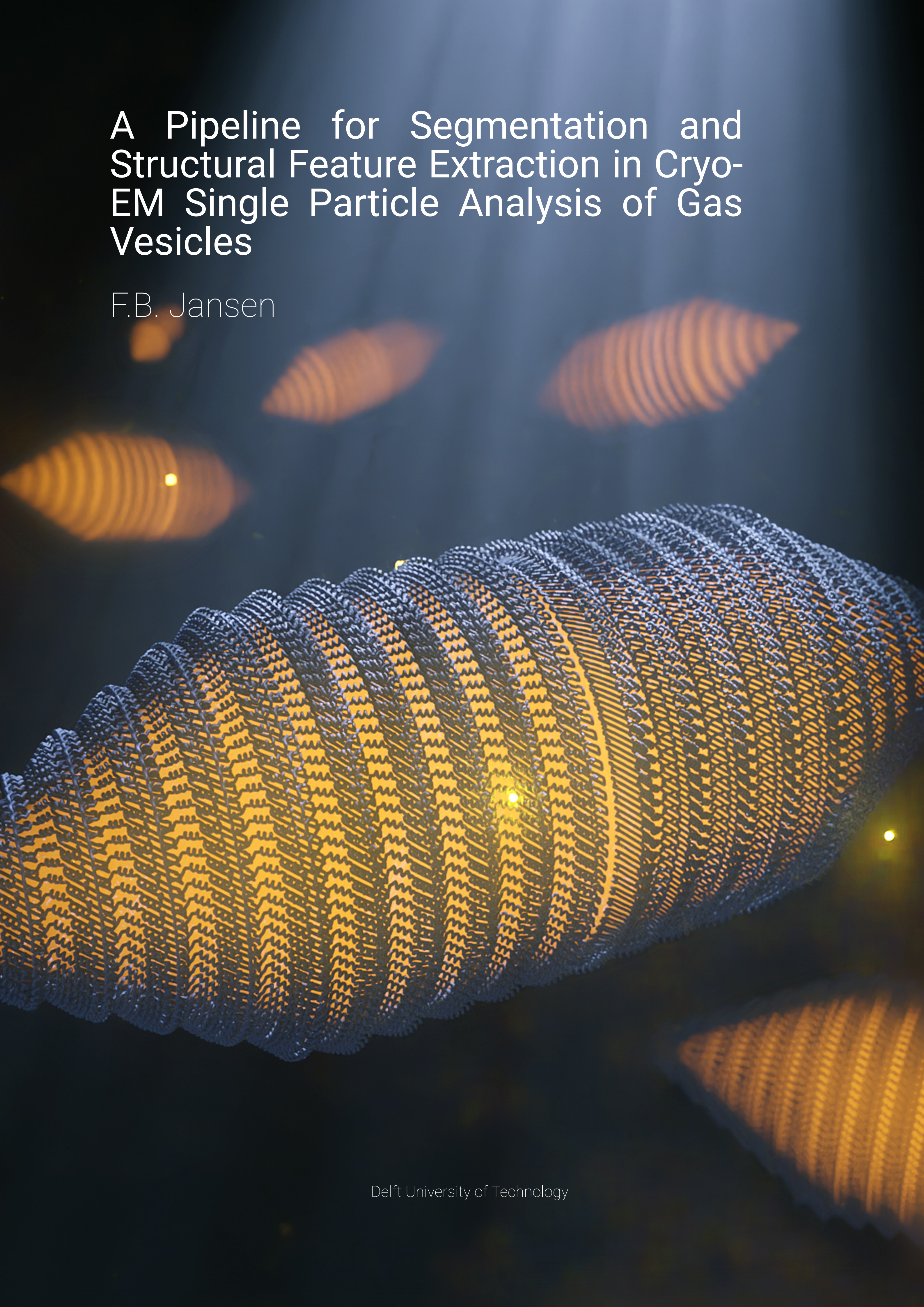
A Pipeline for Segmentation and Structural Feature Extraction in Cryo-EM Single Particle Analysis of Gas Vesicles

F.B. Jansen



Delft University of Technology

A Pipeline for Segmentation and Structural Feature Extraction in Cryo-EM Single Particle Analysis of Gas Vesicles

by

F.B. Jansen

to obtain the degree of Master of Science Nanobiology
at Delft University of Technology & Erasmus University Rotterdam,
to be defended publicly on Monday July 10, 2023 at 14:00 PM.

Student number:	4474740
Project duration:	September 19, 2022 - June 30, 2023

Thesis committee:

Assoc. Prof Arjen Jakobi,	Delft University of Technology, supervisor
Assoc. Prof. Greg Bokinsky,	Delft University of Technology
Assoc. Prof. Martin Depken,	Delft University of Technology
MSc. Stefan T. Huber,	Delft University of Technology, supervisor
MSc. Maarten Joosten,	Delft University of Technology, supervisor

Cover: Gas vesicle atomic model structure illustration by Stefan T. Huber
Style: TU Delft report style, with modifications by Daan Zwaneveld

Abstract

Gas vesicles, micrometer-scale protein structures that function as bacterial buoyancy providers, encapsulate gas in a highly optimized manner. While their atomic structure has been elucidated through single-particle analysis of cryo-EM images, certain structural and functional details remain uncertain. Its biogenesis - the formation and growth mechanisms - consequently remains elusive. Here we apply automated segmentation methods originating from cell imaging to cryo-EM images of gas vesicles to analyze the positions of gas vesicle features in a context-preserving matter. Subsequent whole gas vesicle processing is able to transform accurate gas vesicle segmentations into high-confidence structural feature location picks and statistics. This enables the formation of a sizeable data set containing 86k whole gas vesicles, improved resolution 2D class averages, and the potential for improved structural modeling. Combining high sample number contextual information enables inference on the dynamical properties of gas vesicle growth. Our findings validate recent atomic structure propositions and lend support to a stochastic monomer insertion growth model.

Contents

1

Introduction

1.1. Who runs the world? Proteins!

Proteins are a fundamental workhorse of all living things, without proteins no life. Proteins are large, complex molecules that play many critical roles in cellular processes. They do most of the work in cells and are required for the structure, function, and regulation of the body's tissues and organs.[1]

Studying proteins is therefore vital for understanding crucial processes in the fields of medicine, agriculture, and industry among many others. Additionally, a fundamental understanding of biology can have wide-reaching unforeseen benefits.[2] In order to maximize these potentials, atomic model structures of proteins are preferred since atomic resolution understanding is fundamental for understanding protein function. Given an atomic model, one is able to explain structural properties like hydrophobicity or reaction potential of a substance. More specifically, structural models are fundamental for drug-based design - knowing how a molecule interferes with a drug target is crucial. Such structural models are commonly obtained through analysis with cryo-electron microscopy (cryo-EM), x-ray crystallography, or nuclear magnetic resonance (NMR) spectroscopy.

Despite significant experimental effort, only a small fraction of the structures of known protein sequences have been determined.[3] This is due to the painstaking effort required to determine a single protein structure. Given a hard-to-crystallize protein, obtaining an atomic model can be prohibitively challenging. This can be due to many additional factors such as: hard-to-reproduce native protein physiology imaging conditions, small protein size, low protein yields, or heterogeneous protein shapes and structures among many other complicating factors.

Notably, there are also predictive models like Deepmind's AlphaFold that try to predict protein structure from its genetic sequence. However, the development of computational methods to predict protein structures has focused on physical interactions and evolutionary history. As a consequence, these methods have been limited in their utility for many biological applications due to their lack of accuracy in most cases where a close homolog has not been solved experimentally.[3]

1.1.1. Gas vesicles: bacterial ballast tanks

An example of such an elusive protein structure is that of the protein gas cylinder that many aquatic bacteria express for controlling buoyancy.[4] These structures, also referred to as gas vesicles (GVs), are self-assembling pure protein structures of only a single monomer thick wall, with sizes of around dozens of nanometers in width, and lengths ranging from tens of nanometers to the micrometer range (Figure 1.1).[4]

Their design is rigorously dictated by stringent specifications, owing to their primary function of offering buoyancy. First, the structure has to maximize the ratio of encompassed volume to structural volume, and it has to be able to be diffusive to gasses but keep water out.[6] Additionally, it has to be able to withstand pressure specific to the organisms they are expressed in. As a result, GVs are micrometer-scale protein structures that

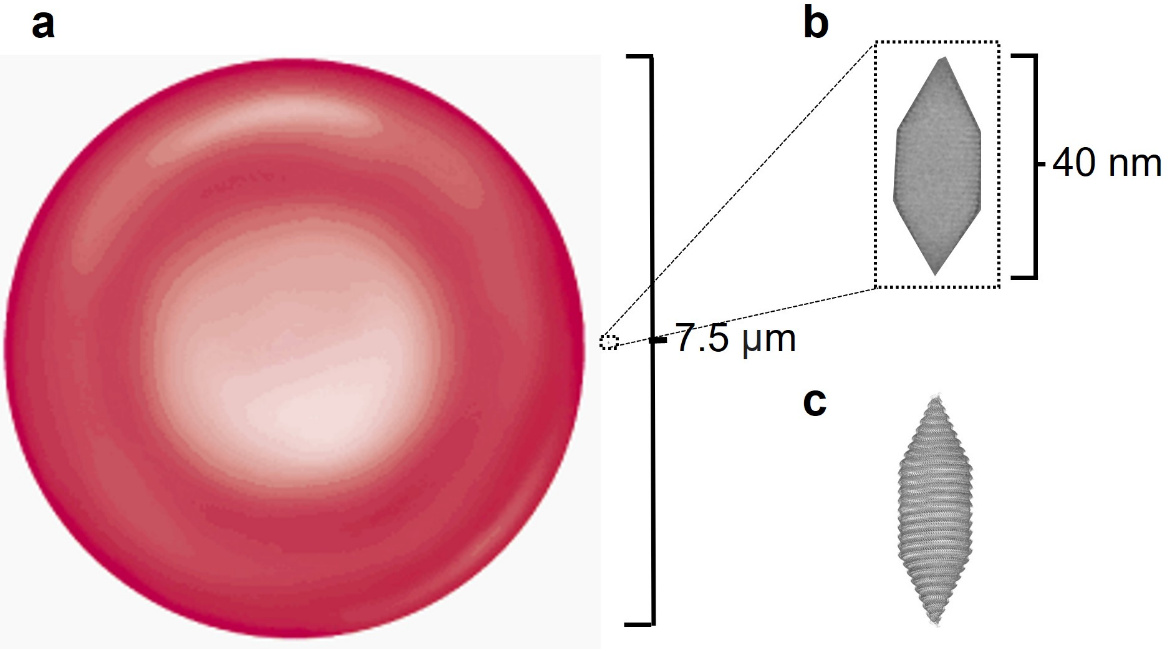


Figure 1.1: **GV dwarfed by RBC (a)** Diagram of human red blood cell (RBC) compared to **(b)** a 40 nm long gas vesicle (GV) cryo-EM image cutout. **(c)** Proposed atomic structure GV model of Huber et al.[5]

are only several nanometers thick - that despite having a structure surface thickness of one or two peptides GVs can still resist pressure up to several bar.[7]

Given that these buoyancy constraints are only varying in the amount of pressure the GVs have to sustain their structure is highly conserved across species. Generally, only the widths of the GVs tend to vary significantly between species since this has the most effect on the pressure resistance of the structure.[4] Their main protein subunits GvpA2 (which we from here on refer to by the name of its more often referenced homolog GvpA) and GvpC make up the shell and outer support structure, which form strong effective structures similar to supporting walls and corrugated ribs of canned goods containers respectively.[5] GVs are rigid structures comprised of two outward conal tips that transition into helical cylinders towards the center. The line in the center at which the two halves meet is called the seam. Along the seam Huber et al propose that a point exists at which the top helix transitions to the bottom helix through polarity reversal of the monomer insertions, this location is coined the polarity reversal point (PRP).[5] Besides being widely present protein structures with unique features, GVs are also interesting as potential acoustic reporters for ultrasound imaging.[8]

GVs are difficult to image using crystallization due to their odd physiology - heterogeneous in widths, lengths, tip shape, and PRP position. Consequently, conventional structure determination methods were for a long time not sufficient to resolve an atomic structure, despite intensive efforts.[4,9,10] Moreover, this heterogeneous appearance also gives issues for structural reconstruction in conventional single-particle-analysis (SPA) methods.[11,12]

1.2. Cryo-EM: the gold standard for structure determination of heterogeneous macromolecules

With the maturation of cryo-EM hardware and software, 3D volumes and corresponding de novo atomic resolution structures (<4Å) can now be reliably resolved for a large range of sizes and types of molecular biological structures (e.g. viruses, membrane proteins, RNA or DNA assemblies).[13] Significant improvements in imaging capabilities are in part possible due to the maturation of direct electron detector device (DDD) sensors, ever-increasing computational capacity and advancements in data processing software.[11] It is therefore that cryo-EM is the current gold standard for resolving atomic-resolution structures of heterogeneous difficult-

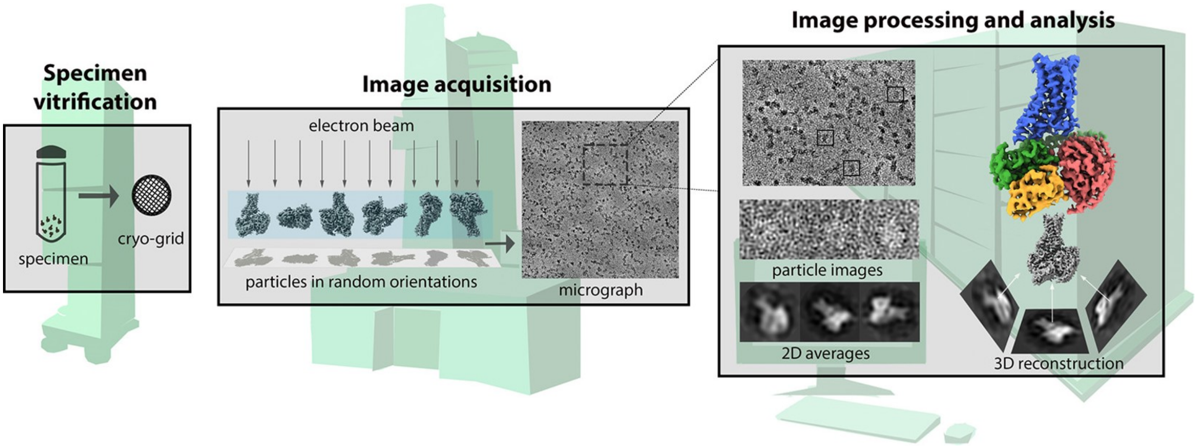


Figure 1.2: **Cryo-EM SPA Workflow** Diagram showing main steps in cryo-EM Single Particle Analysis (SPA) workflow. Sample preparation; isolation of particles and subsequent cryogenic plunge freezing of imaging grid. Image acquisition; electron beams generate 2D projections of particles, seen as contrast on the image. Image processing; individual particles segmented and extracted from images, aligned to orientation of prior 3D particle shape, generate 2D class average for back projections improvement of 3D model. Progressive iteration of particle alignment and 3D model refinement towards the final atomic model. Adapted from Akbar et al.[18]

to-crystalize macromolecules.[14] Given that many functional proteins (see GVs) fall in this category, cryo-EM presents a reliable and effective way of resolving this important set of proteins.[15] This efficacy is reflected in the exponentially increasing number of cryo-EM-resolved protein maps.[16,17]

Most maps are obtained through SPA, a method consisting of several key stages. First, the sample preparation phase involves protein purification. This often involves expressing the protein of interest through the insertion of a DNA vector in a suitable host organism like $E.Coli$. The cells are then lysed and often subjected to a pipeline of filtering through means of size exclusion, elution, or other separation strategies. The resulting isolated protein sample is then blotted onto an imaging grid. This imaging grid most often consists of non-magnetic materials like copper, gold, or carbon that form a support structure that does not deflect the electrons. The sample is then subjected to cryogenic plunge freezing. A process that cools so rapidly that no crystallization of water molecules can occur, preserving the sample in a near-native state.

The subsequent stage is image acquisition, where electron radiation interacts with the sample to generate two-dimensional (2D) projections of the particles onto the microscope detectors. Combinations of interacting and non-interacting electrons produce mostly phase contrast on the image plane that represents the effective electron-interacting density of the sample. Downstream this information is translated into a density map and through subsequent fitting an atomic resolution structure of the particle of interest.

Following the acquisition, image processing is undertaken. During this phase, the images are corrected for any drifting of the stage as a consequence of external vibrations or electron interactions among other factors. This process is called motion correction and is possible due to the generation of so-called movies by the microscope. This involves taking images at a high refresh rate. These movies can then be used to estimate the stage drift in between images and subsequently perform motion correction.

Additionally, the images are corrected for the influence of the contrast transfer function (CTF) of the electron signal. The CTF represents the degree of modulation and attenuation of the image signal as a function of spatial frequency. Changes in the CTF are a result of variations in the effective defocus, aberration of the microscope lens, and differences in ice or particle thickness. These values tend to drift throughout acquisition and within images. Additionally, the defocus is varied on purpose to cover separate parts of the Fourier space that the CTF would otherwise nullify. Covering the full spatial frequency range can be used for improved structural modeling. Overall, compensation for the CTF is crucial for resolving atomic resolution structures.

CTF correction involves estimating the parameters of the CTF, such as the defocus value and astigmatism, for each image (subsection). These parameters describe how the electron waves are phase-shifted and attenuated as they pass through the sample and reach the detector. The estimation is typically done by analyzing the power spectrum of the image, which is obtained by converting the image from the spatial domain to the frequency domain using a Fourier transformation.

Once the parameters are estimated, they are used to computationally reverse the effects of the CTF on the image. This involves adjusting the amplitudes of the Fourier components in the image based on the CTF parameters to correct for the attenuation.

Individual particles are picked and extracted from the motion and CTF-corrected images. The extracted particles are then clustered for image similarity. The resulting classes are averaged to create 2D class averages. Given that the image noise is stochastic, averaging over many images will increase the signal-to-noise ratio (SNR).

These higher SNR 2D class averages are then used to filter the particle set for erroneous outliers and subsequently use this set to align the individual particles to the orientation of an initial three-dimensional (3D) particle shape (volume). This can be understood as a Bayesian optimization problem where we incorporate prior information on the particle shape and smoothness to solve the orientation parameters of the image that best matches a projection of the 3D particle shape. To then subsequently, obtain an improved estimate of the volume with these orientation parameters using backprojection.[19] In the case of an established particle volume reconstruction program called CryoSPARC, this involves performing expectation-maximization (EM), where the expectation step applies a branch and bound algorithm to find orientation parameters.[20] The maximization step is done using stochastic gradient descent (SGD) towards the volume that best explains these parameters. The improved volume can then be used to once again obtain improved orientation parameters, this is done until an estimate of the volume converges (Figure 1.2).

1.3. What do we know?

The GV shells atomic structure has only recently been resolved using cryo-EM SPA (3.2 Å resolution), resolving a model that consists of GvpA subunits (~7 kDa) that make up the core.[5] However, some crucial aspects remain uncertainly mapped. More specifically, the precise structure of the GV tips and proposed PRP point are unresolved in 3D and only inferred indirectly through model building and inference from 2D views. This is a consequence of their difficult-to-model structure (heterogeneous in location for the PRP and heterogeneous in appearance for the tip). Additionally, the tip is a rather thin part of the GV compared to the main structure which means it's covered by a thicker layer of ice, which in turn decreases the signal and visibility of this point.

Huber et al propose a GV growth model that explains the found structure and fits with the main subunit's (GvpA) properties. As is shown in Figure 1.3, the images of GVs can be explained by the proposed model consisting of GvpA subunits. Huber et al propose that GVs form from two cone-like nucleations that meet at the seam. The GV is then proposed to grow from individual GvpA insertions at the point of the seam where the two cone helixes meet and the polarity of the helix GvpA structure reverses - the PRP. The structure and the proposed PRP location additionally offer an explanation for the transition from conical growth of the nucleation cones to cylindrical growth along the center. This growth transition is proposed to happen through an interplay of GvpA structure and resulting forces when growing at the PRP. This leads to variability in the cone angle at which the growth transitions due to cylindrical growth offering a lower subsequent energy state than conical growth.[5]

The precise mechanism of GV growth is still unclear. Current projections point to growing mechanisms adding monomers at the PRP, where either symmetric or stochastic expansion is possible. Symmetric when two opposing monomers are added simultaneously or stochastic when single monomers are inserted randomly. If the symmetric theory were to hold we would expect the seam and PRP always to be located at the geometric center of the GV.

Both the PRP and the tips of the GV cones have not yet been atomically resolved. This leads to uncertainty in regard to how GVs nucleate. That is in part hard to predict due to the highly hydrophobic components of the GvpA monomer, which leads to denaturation when exposed to the environment (cytosol). Additionally, uncertainty regarding the PRP leads to uncertainty on the precise growth mechanism of GVs. Therefore, we can affirm that the uncertainties surrounding the biogenesis of GVs, specifically the unclear aspects of nucleation and growth, stem from the complications that arise when modeling the inherent heterogeneous structure of GVs using conventional single particle analysis (SPA).

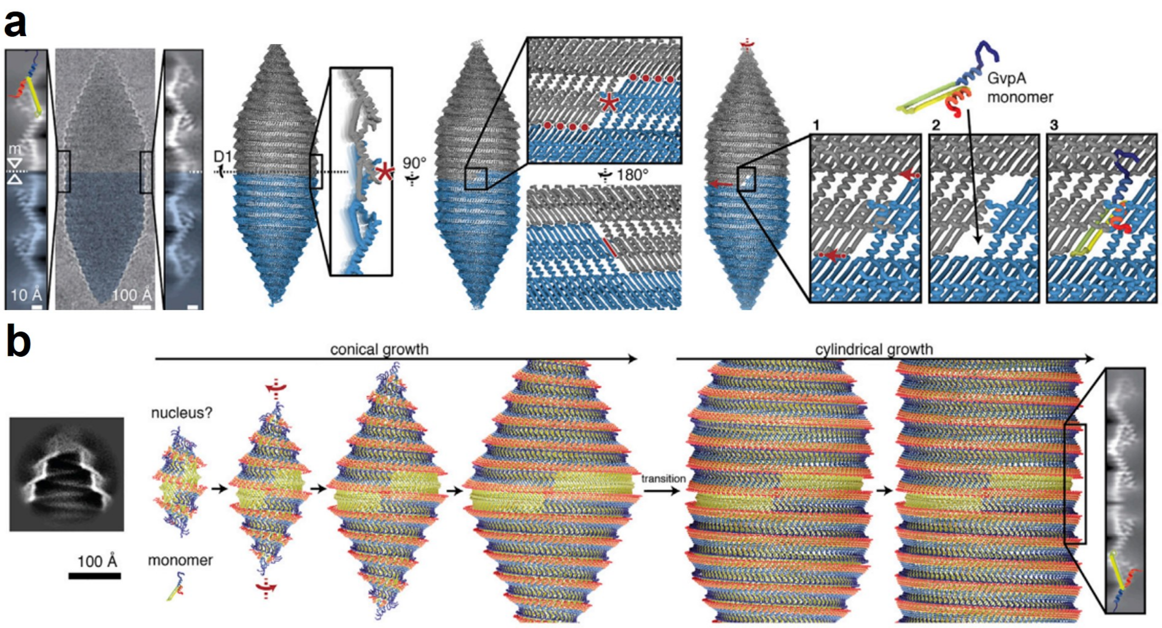


Figure 1.3: **GVs, what do we know?** **(a)** From left to right: A single GvpA monomer (3.2 Å resolution) fitted to the 2D class average of the GV seam. GV particle from cryo-EM image. Diagrams showing the proposed pseudo atomic GV model and monomer insertion mechanism at the seam and polarity reversal point (PRP). **(b)** 2D class average of GV tip, followed by the proposed growing GV mechanism. Starting with two nucleation cones, and transitioning from conal to cylindrical growth. Adapted from Huber et al.[5]

1.4. Issues stemming from GV structure (heterogeneity)

For resolving the tips and PRP or to perform inference on biogenesis conventional single-particle cryo-EM workflows are not applicable because these methods are not able to identify these regions while preserving contextual information such as the relative position of other points of interest of the same GV. This is a consequence of using conventional direct particle picking, such as the use of cross-correlation with a template. We can localize these points but contextual information on the GV the point belongs to is lost. Without this information, it is not possible to infer the location of the point of interest relative to other points of interest or measures such as the specific GV length or width. Without this information inference on the distribution and proposed underlying dynamics of PRP positions becomes infeasible.

Additionally, 3D reconstruction of GVs runs into issues using context-free particle picks when performing structure reconstruction algorithms such as homogeneous or helical reconstruction (these methods assume a homogeneous or helical structure respectively). This is a consequence of the non-homogeneous GVs structure and non-perfect helical structure respectively. Particle picks that do have contextual information can potentially be filtered for having a specific measure (length, width, or cone shape). Additionally, this context information can be used to validate the accuracy of the found points of interest. Obtaining GV feature-filtered higher confidence particle picks could result in improved structural modeling. To achieve context-preserving particle picks we have the need for extraction of individual whole GVs from images, to this end, we require the segmentation of images into distinct cell areas (instance segmentation).

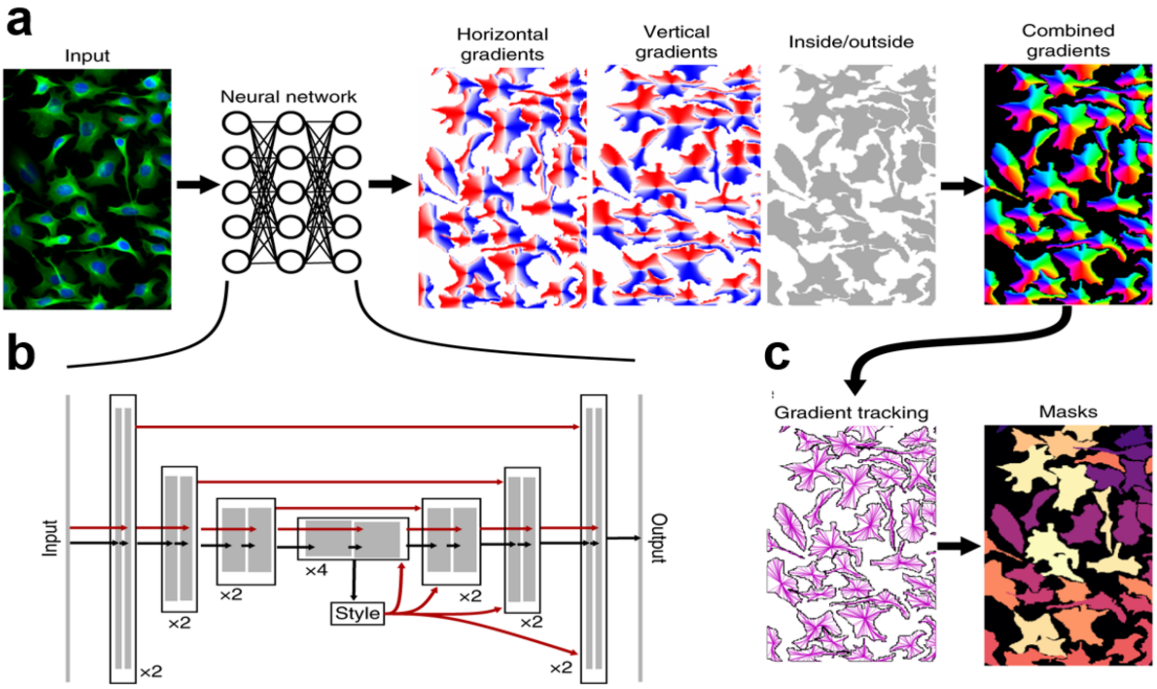


Figure 1.4: **Cellpose model workflow (a)** Input image to be segmented, fed through the Cellpose U-net model to obtain gradient predictions and a cell probability output. These outputs are combined to produce a gradient map of the segmented cells. **(b)** U-net model structure from Cellpose showing four downsample residual layers to extract feature information (style) and four upsample residual layers to identify features in a location-independent manner. **(c)** Combined gradient map is used to identify instance segmentation of cells that result in the Cellpose output masks. Adapted from Stringer et al.[21]

1.5. Machine learning for robust cell instance segmentation

In the last decade, Machine Learning (ML), more specifically Deep Learning (DL), has found its way across science and has also provided a solution for the problem of accurate, general-to-specific, reproducible cell segmentation. A similar issue to cell segmentation is being solved in the field of computer vision, namely, that of object instance identification.[22]

Imitating computer vision, the field of biology has found that the use of similar specific DL networks offers a robust solution to instance cell segmentation. DL methods, specifically those using convolutional neural networks (CNN) with a U-net architecture, have significantly outperformed other methods in image recognition tasks.[23]

Cellpose is a cellular segmentation network that incorporates the established DL methods.[24] Cellpose uses a Residual block U-net architecture to perform instance segmentation of images (Figure 1.4 b). Additionally, Cellpose incorporates transfer learning with human-in-the-loop capabilities to tailor a network to segment specific samples.[24] Transfer learning involves applying a pre-trained model as the starting point for training a sample-specific model. Applying a human-in-the-loop approach to transfer learning involves manually reannoting outputs generated by our transfer learned model and retraining - reannotating additional images and retraining until segmentation performance converges. The result is that Cellpose can use a relatively small number of images per type of cell to train a network.

The U-net model architecture sequentially downsamples and then upsamples the image to create feature maps (styles), integrating skip connections (Residual blocks) between layers of equal size, and global skip connections from the image styles. These styles are computed at the lowest resolution and applied to all subsequent computations.

As illustrated in Figure 1.4a, Cellpose takes a 2D image input and outputs three mappings of the input (horizontal/vertical gradient and inside-outside cell predictions). It does this by transformation trough using a simulated diffusion process that starts from the predicted center of the cell, generating spatial gradients that point toward the center. This provides a single normalized mapped flow direction from 0° to 360°.

The gradients are predictions of horizontal and vertical spatial flows for individual cells. Additionally, Cellpose predicts whether a pixel belongs to a cell. These predictions are combined to produce a flow field ("Combined gradients" in Figure 1.4).

The combined gradient is then used to create a watershed-like representation with fixed edges at the predicted not-cell pixel points. The basins of attraction of these points represent the predicted masks. Essentially, each pixel "follows the flow" along the predicted flow fields, converging toward its eventual fixed point. Lastly, the final masks are created by grouping pixels together based on their convergence to the same fixed point (Figure 1.4 c). This process ensures that all pixels assigned to a specific mask are related to the same cellular structure instance.

1.6. Can segmentation of cryo-EM images in a context-preserving manner inform us on GV biogenesis?

We propose to train a GV-tailored instance segmentation model (Cellpose 2.0) to accurately and robustly segment GVs from cryo-EM images. These segmented GVs are then extracted and subsequently processed to extract precise locations of interest (e.g. tips and seam) with corresponding contextual information (GV length, width, and distance to other points of interest) for answering questions on biogenesis.

Given that the growth mechanism is yet undetermined we want to deduce what kind of insertion growth model GVs follow. Previous studies support the case for the stochastic model through off-center seam observations. Therefore, we propose to test the stochastic model, by observing these instances of GVs with seam positions at distinct locations from the geometric center of the GV structure. Given a large number of these measurements, we can fit a distribution and subsequently validate if this matches with the expected distribution of this measure.

Additionally, we can numerically estimate the number of monomers per GV half given its width and length. We then argue that the proposed stochastic growth model can be understood as a binomial elementary random walk of a number of steps equal to the number of monomers of a whole GV (each insertion can be understood as a random walk to the left for the top GV halve and to the right for the bottom half). Therefore we expect the monomer count difference between monomer halves to follow the distributions associated with an elementary random walk. Therefore, we expect the seam-center offset (monomer difference) to converge to a normal distribution with a mean offset of 0 and a standard deviation of the square root of the number of monomers in the GV.

To obtain this information we will need the locations of tips and seams of whole GVs. These we can use to determine the ratio between GV halves to subsequently determine the drift away from the center. Then, we can perform the geometric fitting of the known monomer dimensions to these features to estimate the number of monomers in a GV. Given these objectives, we strive to answer the following research question

Can we apply context-preserving particle picking to elucidate the growth mechanism of GVs, and can a stochastic growth model accurately describe their biogenesis as evidenced by the distribution of seam-center offsets?

To reiterate, this involves using DL-based segmentation for context-preserved picked point-of-interest positions, possibly enabling the determination of the seam and tip structures. Additionally, precise GV location statistics on the seam and tip points could lead to inference on GV biogenesis by validating a stochastic monomer insertion growth model through analysis of seam-center offset statistics.
To achieve this in a feasible fashion automated instance segmentation is required. To this end, we will make use of a DL-based generalist cell segmentation package called Cellpose. Cellpose allows systematic identification of points of interest which can be subsequently used with template matching to obtain high-confidence positions of points like the tip and seams.

The remainder of this thesis discusses the implementation and training of a Cellpose GV model, the components of the GV processing pipeline, the theory for validating the proposed stochastic growth model, and the methods of 3D reconstruction in the Methods in Chapter 2. We present the resulting performance and output statistics from the GV pipeline with 3D model structure findings in the subsequent Results in Chapter 3. Then the inference on our results in Chapter 4, and lastly, future research recommendations in the Outlook in Chapter 5.

2
Methodology

2.1. Image processing

Our data set consists of 33k micrographs of GVs originating from *Bacillus megaterium* imaged with a 300kV Krios G4 E-CFEG Selectris Falcon 4i cryo-EM from the European Molecular Biology Laboratory (EMBL). The acquisition characteristics are identical to those of Huber et al.[5]

The motion-corrected movies are provided in .mrc format representing 4096 by 4096 pixels with a 1.518 Å/pixel resolution. For more efficient post-processing the corresponding images are downscaled and converted to a suitable image format. The image type is converted from the output format of the motion-correction algorithm (.mrc) to a format more suitable for subsequent image processing in Cellpose and downstream (.tif). The images are downscaled by binning, which is done why taking a square cutout of the images around the center of the power spectrum. The resulting images have reduced information in the high spatial-frequency domain, which should not affect the efficacy of segmentation or point-of-interest picking significantly. Cutting at half the original frequency range 'bins' the images 4 times, reducing the pixel size by 4 and effectively reducing the resolution and information by the same factor. This lowers storage space requirements significantly (16x) while increasing throughput through reduced computational processing costs. The unbinned images are used only when performing structure determination at the end of the workflow using CryoSPARC.

To resolve the structure of the GV seam and tips from a given cryo-EM dataset we will need to extract their exact positions. In a SPA cryo-EM workflow, these points are used to extract these regions of interest (particle picks). These picks are then often clustered and averaged to create 2D classes, which are then used to filter out outlier particles and to check the orientations and quality of the particle picks. 3D model building starts with a rough initial blob-like shape and then all the given input particles are aligned using the projection-slice theorem. These aligned particles are then used to refine the 3D density through backpropagation.[19] Then again, we realign our particles to this refined volume, which in turn can inform an improved 3D model. We reiterate this procedure until the 3D density reconstruction converges.

Getting the particle pick locations directly from unprocessed data could be done using a reference, however, conventional methods applying cross-correlation (template matching) do not maintain contextual information of the picked points of interest. Consequently, we would not be able to get corresponding features of interest - e.g. GV length, width, cone angles and distance to other types of points of interest. While it is precisely this contextual information that can facilitate the analysis of GV biogenesis.

Therefore we propose a GV processing pipeline that involves obtaining an instance segmentation of individual GVs using DL-based segmentation (Section 2), and subsequent GV processing, and only then we will apply template matching of our points of interest using a suitable reference. Performing the analysis in this way will maintain contextual information and allow for checking of the found location using this context of the isolated GV.

In order to obtain these high-confidence points of interest we will be formulating an image processing pipeline to utilize the high-accuracy instance segmentation. The following sections cover the components of this workflow in order of application.

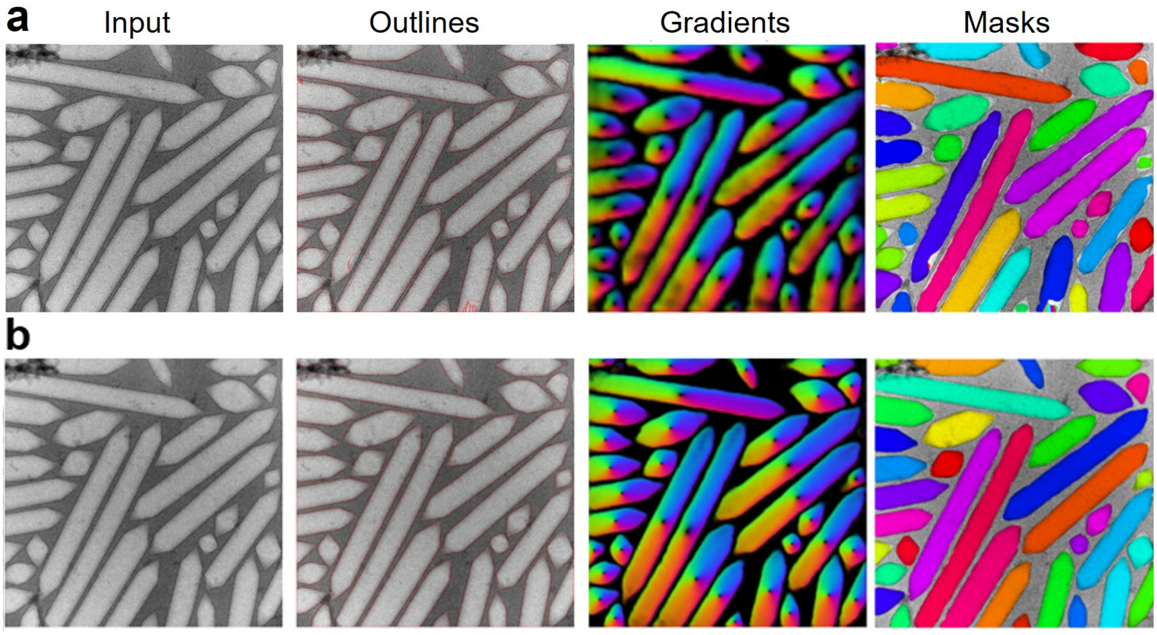


Figure 2.1: **Cellpose GV segmentation hyperparameter influence (a)** From left to right: input GV cryo-EM micrograph, Cellpose outputs - outline, combined gradient and masks for default hyperparameter settings (cell diameter: automatic, cell probability threshold: 0). **(b)** Same as in (a), for found optimal hyperparameters as shown in 2.3 (cell diameter: 150 pixels, cell probability threshold: 0).

2.1.1. Accurate GV segmentation

The first step in processing our input data is to obtain a rough idea of the locations of our points of interest from the accurate segmentation maps. These rough points can then be refined and improved upon to obtain high-confidence points suitable for inference on structure and biogenesis.

Instance segmentation - the separation of background and individual structure volume - is a suitable method for obtaining these initial rough estimates. Segmentation can, however, be sensitive to outlier shapes, such as GVs only partly in the image, or imaging impurities. Additionally, our method has to be robust to varying widths and lengths of GVs. We find that overlapping GVs and image impurities lead to improperly segmented GVs in build-in Cellpose models. It is for this reason that we chose to restrict our training set to only contain segments of non-overlapping and impurity-free GVs. Training with this type of data set significantly improved model performance.

To achieve the desired accurate robust instance segmentation of GVs we apply transfer learning and the human-in-the-loop approach of Cellpose 2.0. (Figure 2.2)[24]

We found that "cyto 2" from the Cellpose 1.0 model zoo offered the best initial GV segmentation performance. We then trained our custom GV Cellpose model using "cyto 2" by annotating only intact clearly visible GVs in 60 images. Ignoring difficult-to-annotate GVs (due to impurities or overlapping GVs) and removing bad segments in this category leads to training the best-performing segmentation model.

Additionally, our custom GV model requires a set of Cellpose hyperparameters ("Channels", "Resample", "Flow threshold", "Cell probability threshold", and "Cell diameter"). "Cell diameter" and "Cell Probability Threshold" are found to be of significant influence on segmentation performance. As can be seen, by the

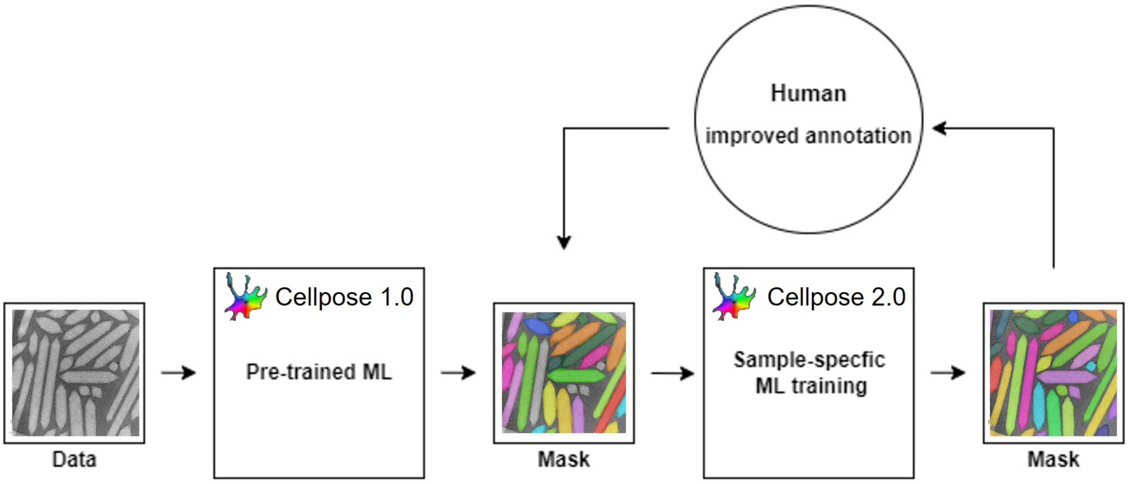


Figure 2.2: **Cellpose 2.0 transfer learning & human-in-the-loop** Diagram of model training workflow for training a GV instance segmentation model using the Cellpose 1.0 model zoo as the pre-trained starting model.[21] Subsequent retraining with Cellpose 2.0 using manually re-annotated outputs iteratively.[24]

wrongly shaped masks using the default model parameters in Figure 2.1a.

A critical parameter is the "Cell diameter". The models in Cellpose have been trained on images that were rescaled to all have a uniform diameter. Therefore, Cellpose needs a user-defined cell diameter (measured in pixels) as input, or it can also estimate the object size on an image-by-image basis. The automated estimate uses a two-step process that involves generating a "Style" vector from the network. Changing the diameter can impact the results, potentially causing cells to be either over-split or over-merged. We find that setting a set value of 150 performs best (Figure 2.3).

Cellpose also takes into account a parameter known as "Resample". This involves running the algorithm on your rescaled image, where the rescaling factor is determined by the diameter you input or determined automatically. The dynamics can either be run on the rescaled image size (Resample=False) or on the resampled, interpolated flows at the original image size (Resample=True). The choice between these two options affects the smoothness of the regions of interest (ROIs) and the speed of the operation. We find the best results by running with "Resample" set to "True".

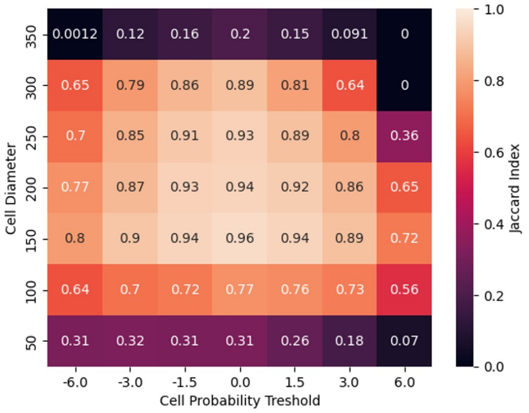


Figure 2.3: **Cellpose hyperparameter optimisation** Heatmap showing mean Jaccard index from running Cellpose on a set of ten GV micrographs for varying hyperparameters of interest. The Jaccard index is defined as a pixel-wise IoU(A, B) = $\frac{|A \cap B|}{|A \cup B|}$) - A Cellpose output, B manually annotated ground truth.

Flow threshold is another key parameter. This threshold restricts the error of the flows for each mask to ensure that the shapes recovered after the flow dynamics step are consistent with real ROIs. By adjusting this threshold, you can influence the quantity and shape of the ROIs that are returned. The channel selection is not of importance since we will be using only greyscale images and not segmenting any substructures (nuclei), therefore "Channels" will be set at [0,0].

Finally, there is the "Cell probability threshold" parameter. The network predicts a probability for each cell and pixels that score greater than this parameter value are used to determine ROIs. Tweaking this threshold allows you to change the number of ROIs returned, particularly from dim areas. For this value,

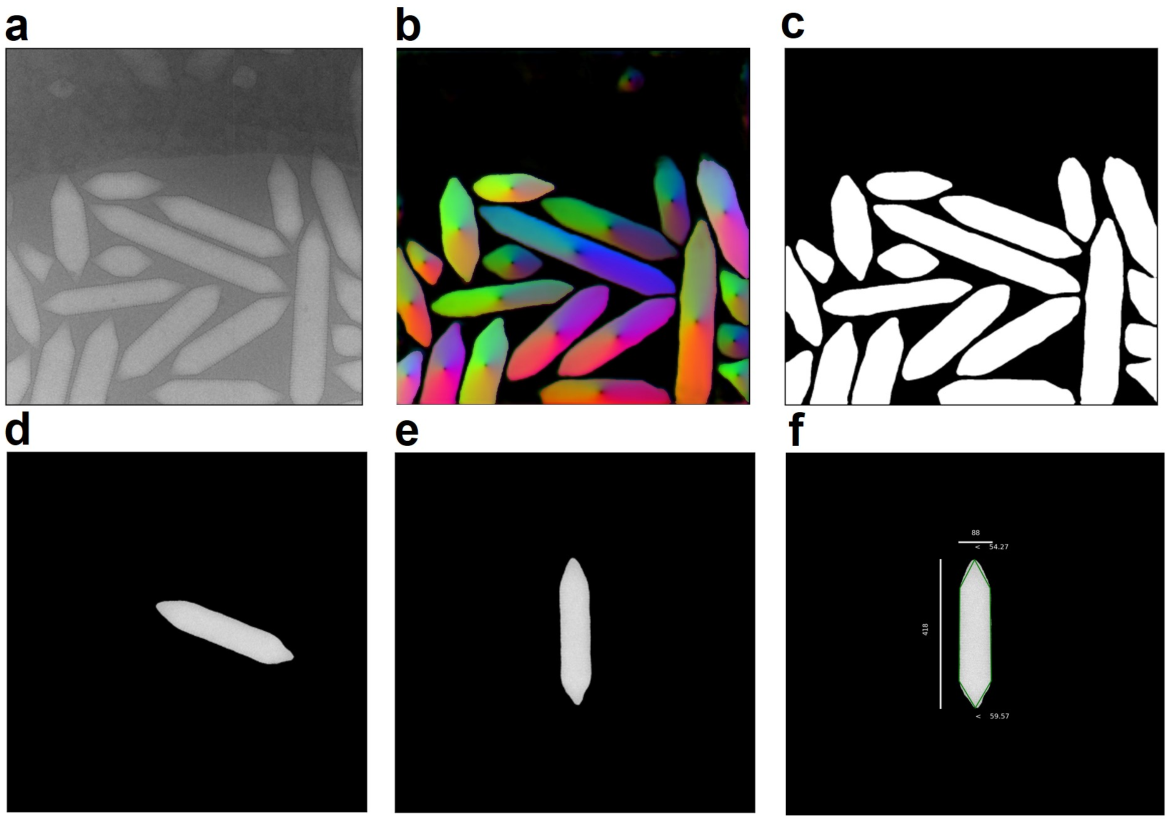


Figure 2.4: **GV processing pipeline example steps** **(a)** Input GV micrograph. **(b)** Cellpose combined gradient output. **(c)** Cellpose masks output. **(d)** Extracted GV. **(e)** Rotated extracted GV. **(f)** Preliminary points-of-interest and statistics from extracted GV.

we found the default value of 0 to work best (Figure 2.3). Once the hyperparameters are optimized we can segment our dataset. The next step is to use these segments to extract individual GVs for subsequent analysis. Figure 2.2 illustrates the difference in performance between the default and found optimized Cellpose hyperparameters.

2.1.2. Standardised formatting of extracted GVs through rotation and centering

In order to extract our GV feature of interest (GV length, width, and cone angles) we perform pre-processing on the GVs to obtain a standardized extracted GV (cutout) as is shown in Figure 2.4. This can then subsequently be used for the analysis of individual GVs. The pre-processing consists of particle extraction and then subsequent standardization. Extraction is done by multiplying the image with a mask corresponding to a GV instance segment. Standardization consists of rotating the GV to be upward and centering the GV in the center of the extracted image. After this pre-processing, we obtain upward-centered extracted GVs to subsequently extract our features from.

Rotation angle determination

After segmentation and extraction, the first standardization script that is applied to the extracted GVs is the rotation angle determination method. We find that using the Fourier transform of an extracted cutout allows computationally efficient, precise, and robust estimation of the rotation angle. This method relies on the distinct frequency patterns in the power spectrum originating from the repeated structure of the GV cell wall. For individually extracted GVs this wall will be homogeneously orientated within the image, producing a clear orientation signal in the power spectrum (Figure: 2.5). Additionally, we can see that the reference does have a constant inherent contrast signal that does not correlate to the reference input orientation. For this reason, the center part of the reference stack is blocked out to enhance correlation efficacy.

This method works by extracting a sub-cut around the center of the extracted GV, obtaining the power spectrum through the application of a Fourier transform, and then multiplying this spectrum with the spectrum of a reference image that is rotated by a rotation angle ($\alpha°$). We perform this multiplication for a range of 180 α ° s equally divided between 0° and 180° degrees of rotation. The Fourier spectra multiplication that results in the highest value (equivalent to real-space correlation) indicates that the input cutout was rotated by the corresponding angle $\alpha°$ (Figure 2.6).

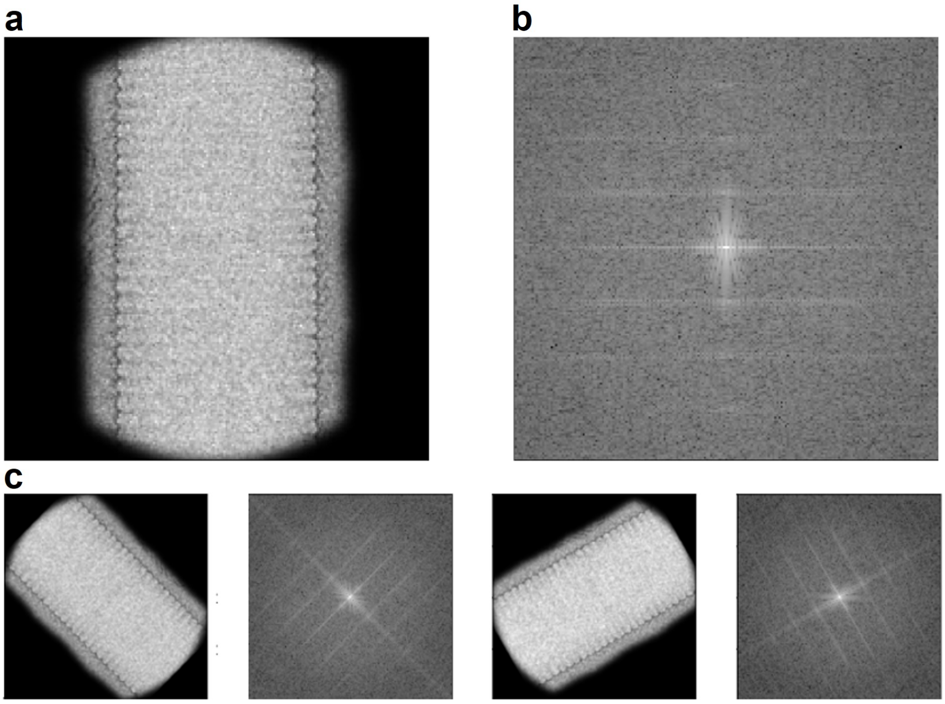


Figure 2.5: **Reference stack Fourier method GV rotation angle (a)** Micrograph cutout of chosen reference GV at 0° rotation. **(b)** Powerspectrum from applying an FFT to (a). Showing lay lines at angel parallel of GV shell direction. Sample specific contrast information at the center. **(c)** Micrograph cutouts and corresponding power spectra for 45° and 120° rotation.

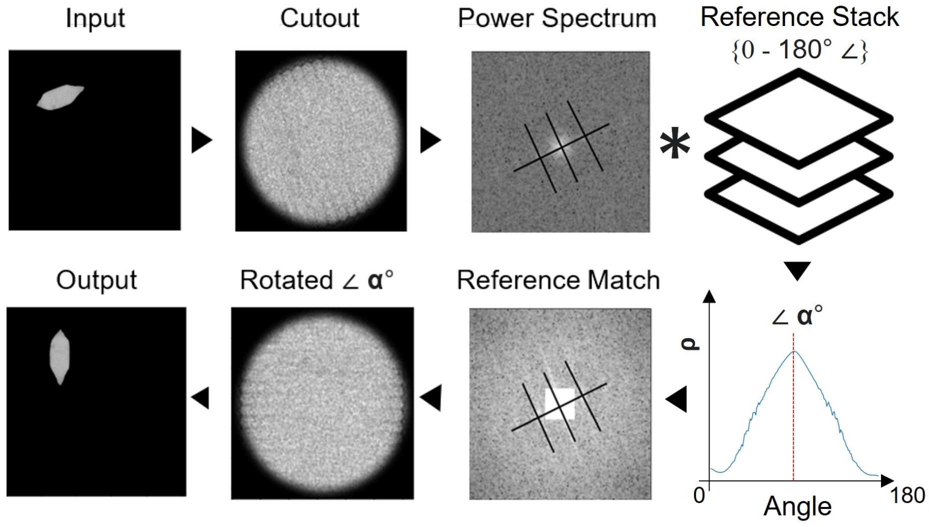


Figure 2.6: **Workflow Fourier method GV rotation angle** Starting from the input micrograph GV cutout; subsequent cutout around the center, FFT of the cutout for power spectrum cutout, multiplication of cutout power spectrum with reference stacks constituents, max sum reference constituent represents best correlation and corresponding index identifies found rotation angle $\alpha°$, reference match shows power spectrum from reference stack that best correlates with input - showing angle identifying ley lines in black. Rotated $\alpha°$ shows the rotated cutout. Output is GV rotated by $\alpha°$.

With an obtained rotation angle we can rotate and center the extracted GV to obtain a standardized GV to subsequently extract feature statistics from. In order to filter out potentially poorly orientated standardized cutouts we apply some preliminary filtering based on heuristically defined measures. To this end, we define the following test statistics

$$\text{Test 1: half\_distance} = \sqrt{\text{half\_dist\_x}^2 + \text{half\_dist\_y}^2}, \tag{2.1}$$

$$\text{Test 2: x\_y\_ratio} = \frac{\text{half\_dist\_y}^2}{\text{half\_dist\_x}^2}, \tag{2.2}$$

both functions find half_dist_(x or y) by taking the distance in the rotated outputs between the center and the tip or center-left for the y and x distance estimates respectively (in pixels). Test 1 tests that the half length of the GV is of a certain size, we filter for this since small GVs are harder to correctly estimate the rotation angle from using this method. Test 2 exploits the knowledge that a correctly orientated GV should have a reasonably large length-over-width ratio. We find that the combination of these two measures at heuristically defined cutoff levels produces the best results when testing a sample set of forty curated standardized extracted GVs. We take a rotated GV as correctly orientated when both these measure score above the heuristically defined limits (100 for Test 2.1, 2 for Test 2.2).

We find that for smaller GVs this method is less accurate and the resulting processed GVs tend to be rejected when tested by the previously proposed tests and corresponding cutoff values (Test 1 2.1 and Test 2 2.2) Wrong rotation angel estimation appears to happen because of the curvature of the cell walls around the center for smaller GVs. This results in a different Fourier spectrum pattern that no longer uniquely identifies the rotation angle of the GV. To address this problem we apply an additional Fourier method to these rejected (assumed-to-be-small) GVs. This additional is_small_gv method applies the same Fourier method but with a reference of a small GV. See Supplementary A for part of the is_small_gv method reference stack in Supplementary Figure A.1, and method in Supplementary Figure A.2.

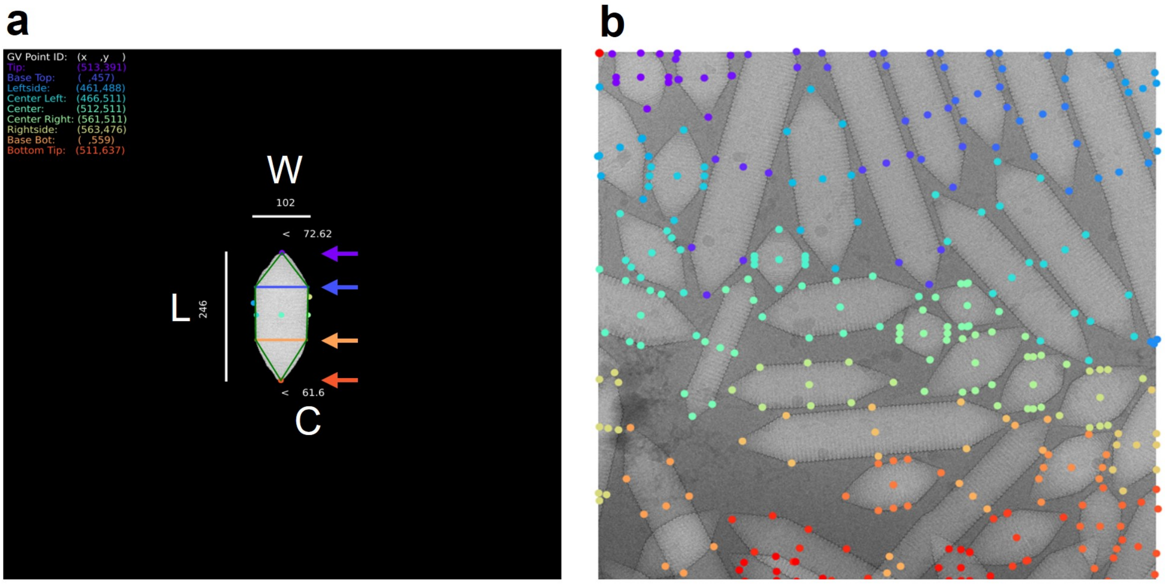


Figure 2.7: **Preliminary points-of-interest (a)** Proccesed GV (extracted, rotated, and centered) with corresponding preliminary statistics; width (W), length (L), and cone angel (C). Preliminary points of interest are indicated in the left top in colors [pixel values]. Example points of interest used for calculating the cone angles are the Tip (purple arrow), Base Top (blue arrow), Base Bot (orange arrow), and Bottom Tip (red arrow) **(b)** Preliminary points-of-interest from a single micrograph.

2.1.3. Preliminary points of interest

Within this work, we are interested in resolving and understanding current yet unresolved points of GVs. Our preliminary points of interest, therefore, include estimations of said points (see Figure 2.7). We focus on the seams and tip points (purple and red arrows). Additionally, we extract the points at which the GV structure transitions from a cone to a helical shape (blue and orange arrows). We do this to be able to obtain an estimate of the cone angles. The preliminary locations of these points are obtained through heuristic approaches on the standardized cutouts.

The preliminary tips we find by taking the highest and lowest points of the standardized cutout. The preliminary seam point by taking the left-most and right-most points of the center. We find the end of the cone and the start of the helical base by tracing the pixel values starting from a tip and measuring when the width stops increasing for a certain number of steps.

We chose to extract a set of identifying points of interest from GVs to be used for subsequent analysis targeting to efficiently process GVs and keep output data manageable (Figure 2.7 b). These points can be combined to extract preliminary estimates of the GV length (L), width (W), and cone angles (C_{top} or C_{bot}) (Figure 2.7a).

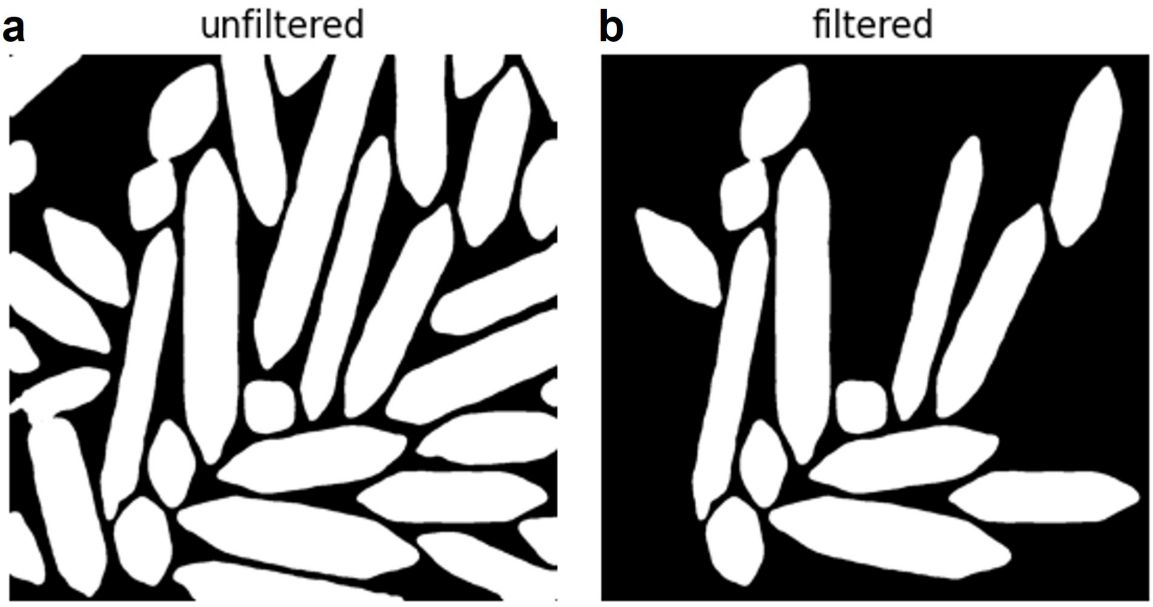


Figure 2.8: **Edge filtering example (a)** Cellpose mask output. **(b)** Result from applying edge filtering to masks from (a).

2.1.4. Filtering outlier segments

We argue that inaccurate segments will be separable from good segments through the application of systematic data analysis methods. For this, we apply positional and feature-based filtering.

Using positional information from masks it is possible to filter out GV masks that are on the boundary of the image and are therefore not showing a whole GV. Incomplete segments are more complicated to systematically analyze, therefore, filtering them out is preferred. We filter these cutoff GVs by checking the distance of their preliminary tip and center points to the edges of the image (Figure: 2.8).

Besides incomplete GV views due to image boundaries, inaccurate GV cutouts can also occur due to improper segmentation as a result of image irregularities. E.g. overlapping GVs, ice crystals obscuring view, and outlier GV shape that the GV segmentation network cannot accurately segment. Again, these segments are hard to systematically analyze, and filtering them out is therefore preferred.

To filter out inaccurate segments not at the edge of the image we apply k-means clustering to the preliminary features statistics obtained from the standardized cutouts. It can be argued that inaccurate segments will have distinct preliminary statistics and are therefore able to be grouped by k-means clustering. Resulting clusters are then selected for further processing based on cutout appearance and feature statistics.

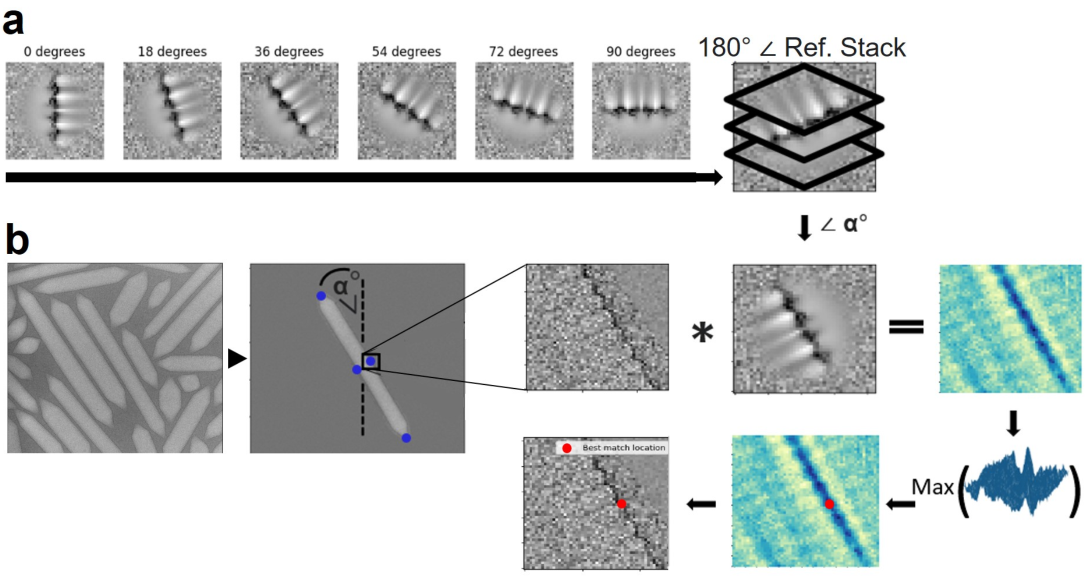


Figure 2.9: **Template match workflow (a)** Reference stack formation showing a subset of the 180 constituents representing rotations ranging from 0° to 180°. **(b)** Template match workflow from left to right; Input micrograph is extracted, an example target region (right seam) is extracted using a preliminary point-of-interest (blue dot), the target is cross-correlated with a constituent of reference stack (seam stack) corresponding to the rotation angle $\alpha°$- obtained from applying the Fourier angle method. Then the maximum value in the resulting cross-correlation map represents our found target (seam) match location (red dot).

2.1.5. Template matching

We require a high confidence level of accuracy if we are to reliably obtain structure statistics and points of interest. The heuristics methods used to obtain preliminary estimates of these points do not account for outliers and are not checked for true accuracy, therefore, a more verifiable method for obtaining these data points is required.

To increase the reliability of the picked points of interest - and subsequent statistics - we apply template matching through cross-correlation of the preliminary points of interest with a high-confidence structure reference. We center, scale, and contrast-invert these high-confidence references which are adapted from the work of Huber et al and constitute a 2D class average of respective points of interest (Figure 2.11).[5]

The template-matching workflow relies on cross-correlation verifying the location of points of interest to produce high-confidence estimates. To this end, cross-correlation is used by applying the equivalent of a real space convolution in the form of a Fast Normalized Cross-Correlation in the frequency domain, as defined by J.P. Lewis.[25] This procedure correlates the input target area (cutout around previously determined preliminary points of interest) and a real space reference stack constituent with a previously estimated rotation angel $\alpha°$. Notably, we find that adding Gaussian noise to the masking outside the reference and applying Gaussian filtering to the target increases correlation performance (Figure 2.9). We propose that Gaussian filtering decreases the image wide contrast gradient which increases the relative contrast between the areas of interest and their direct neighborhood - increasing correlation performance. Additionally, without the Gaussian noise added to the zero values at the outside of the masked reference, we find that the reference tends to align to the zero pixel values areas in the input area - decreasing correlation performance.

Due to the high computational cost involved in applying rotations, we avoid directly rotating the target area. Instead, we compare the target to an image in a rotated reference stack that has already been rotated by a specific angle. What angle ($\alpha°$) from the stack to choose from has been previously determined for the GV belonging to the target using the Fourier rotation method (Section 2.1.3).

Additionally, the references are also centered around the point of interest for more tractable result interpretation. For the seam and PRP reference, we find the x-axis centering of the references through projection

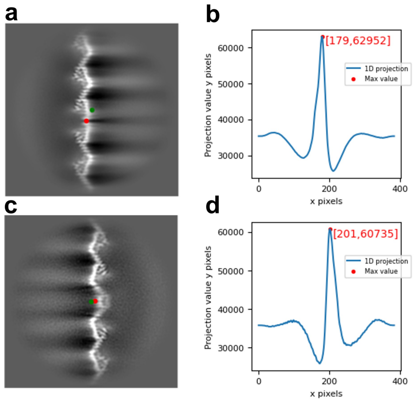


Figure 2.10: **Template match reference centering (a)** Seam reference from provided 2D class average of Huber et al[5] We indicated the centre of image (green dot) and found centre seam (red dot). **(b)** Summation output from projection of (a) onto the x-axis to find centre point (red brackets). **(c)** Same as (a) for a PRP reference. **(d)** Same as (b) for PRP reference

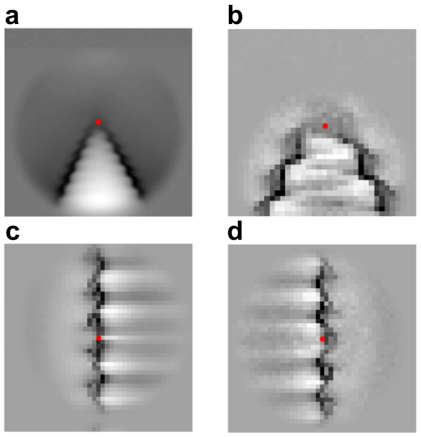


Figure 2.11: **References template matching method (a)** Wide view tip reference. **(b)** Narrow view tip reference. **(c)** Seam reference. **(d)** PRP reference. All references are previously obtained 2D class averages centred around template marking points (red dots), scaled to the pixel resolution of the input data (1.518 Å/px times a binning factor of 4), and have their contrast inverted to match that of the data.

onto the x-axis, see Figure 2.10. The other centers we pick manually. We also scale the reference to have the same pixel size as the target images. Finally, we apply contrast inversion to have matching contrast with the target (Figure 2.11). In order to automatically test the found template match points we propose the following (heuristically determined) output checks.

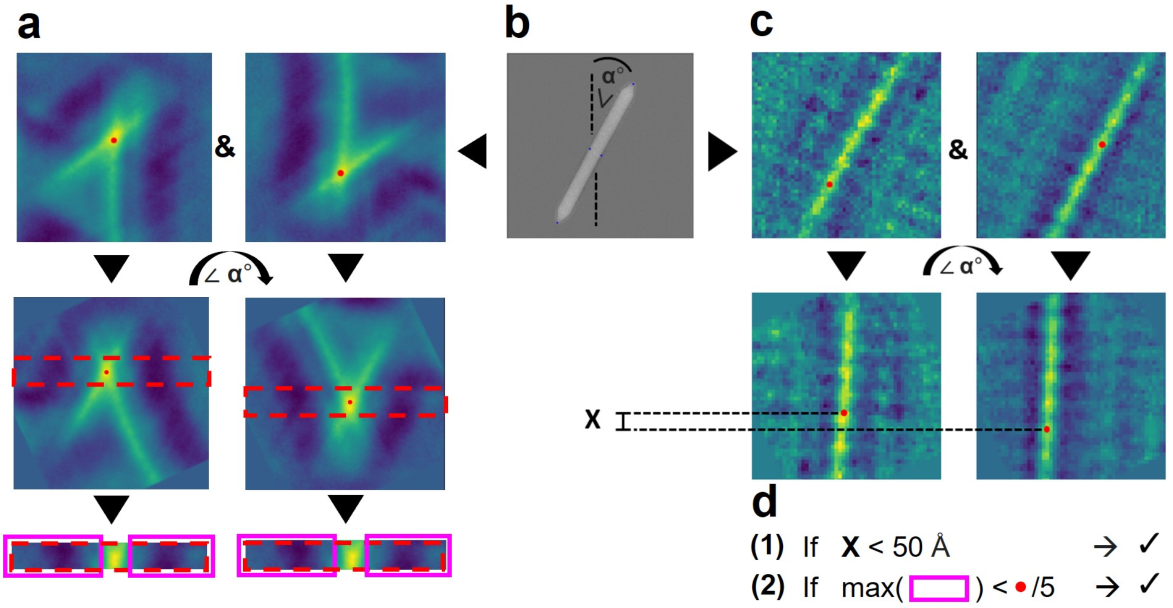


Figure 2.12: **Template match output testing (a)** Tip target match testing. Tip target correlation maps are shown with found tip target match (red dot). Correlation maps are rotated by rotation angle $\alpha°$ and a horizontal region is cut out (striped red box). These extracted regions are then subdivided into target match and residual locations (pink box). The highest correlation value in the residual locations is then compared to the tip target match correlation. **(b)** Input micrograph for template matching with previously found target locations and rotation angle $\alpha°$ **(c)** Seam target match testing. Seam target correlation maps are rotated by rotation angle $\alpha°$. **(d)** Test statistics seam (1) and tip (2). X is defined as the height difference of the seam target locations in the rotated cross-correlation maps. A pixel value scaled version of X is compared to the distance between GV helical rings (50 Å).

Testing seam and tip location matches

We test the confidence of the tip positions by applying a heuristically determined method of checking a subselection of the rotated cross-correlation maps for similarly high correlation coefficients (Figure 2.12a).

For testing the confidence of the seam positions we compare the height of the rotated cross-correlation maps for the seam targets of the two GV sides. This can be argued to be a robust method since within the GV structure the seam is expected to be at the same height on both sides of the GV (Figure 2.12 c). We take the margin of error to be the height of one protein helical ring (50 Å).[5]

Performing these operations in a scalable manner requires sound processing and data storage management. To this end, and to streamline and make it readily reproducible we use the workflow manager package SnakeMake.

2.2. Scaling the workflow

A challenging aspect of this method of structure determination is the amount of data that needs to be processed in a structured manner, both in terms of storage and computational requirements.

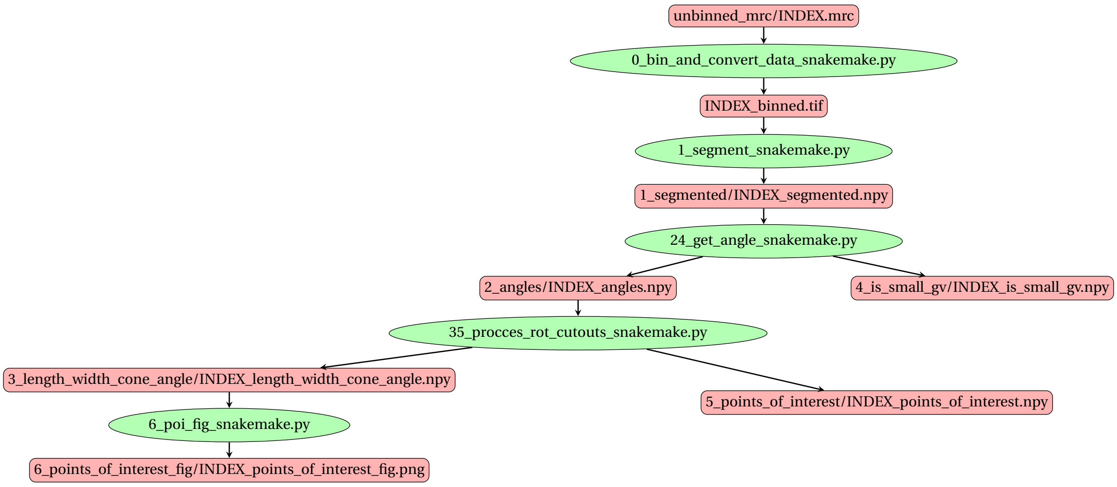

unbinned_mrc/INDEX.mrc
0_bin_and_convert_data_snakemake.py
INDEX_binned.tif
1_segment_snakemake.py
1_segmented/INDEX_segmented.npy
24_get_angle_snakemake.py
2_angles/INDEX_angles.npy
4_is_small_gv/INDEX_is_small_gv.npy
35_procces_rot_cutouts_snakemake.py
3_length_width_cone_angle/INDEX_length_width_cone_angle.npy
5_points_of_interest/INDEX_points_of_interest.npy
6_poi_fig_snakemake.py
6_points_of_interest_fig/INDEX_points_of_interest_fig.png

Figure 2.13: **Directed acyclic graph representation of the Snakemake GV processing pipeline** Red boxes represent data file paths with generalized naming conventions as used in Snakemake. Green boxes represent the functions (.py) that process and output data from and to their respective file paths. The naming convention follows that of the file-storing system as used in the GV Snakemake pipeline. E.g. storing Cellpose segmentation outputs from applying the "1-segment_snakemake.py" function to files that are stored as "Index_binned.tif" format in the "1_segmented ’folder, with the wildcard name "Index" and ending on _segmented.npy. We removed the visualization of non-direct input-output relations to enhance the interpretability of the DAG.

SnakeMake

Processing a single image takes in the order of minutes, it is, therefore, infeasible to conventionally process the 33k input images in a reasonable amount of time. We need a way to process images in parallel. To this end and to enhance the applicability of our GV pipeline we use the workflow manager package Snakemake.[26] Snakemake works on the principle of using directed acyclical graphs (DAGs) to represent the steps in a workflow to guide automated data processing. These DAGs can be followed to produce directions for processing an image, Snakemake is able to apply these processing steps in parallel to a given number of images using a given number of processing units (CPUs or GPUs).

Snakemake requires the user to define the steps of the workflow in a so-called "Snakefile". In this file, the steps are accompanied by their respective input and output file types. Snakemake is able to interpret the "Snakefile" into a DAG. Providing the Snakemake package with a Snakefile of your workflow and pointing it towards the outputs one wants to generate leads to Snakemake performing this analysis automatically.

It does this by generating a DAG from the Snakefile by going through the hierarchy of steps in the workflow until it comes to the point where the files needed to perform a step are present on the system. It will then perform all the subsequent steps and store the data as dictated by the Snakefile. The DAG corresponding to the GV processing workflow consists of four main scripts that perform the main tasks. An additional script is present to produce output figures to check the functioning of the pipeline, see Figure 2.13).

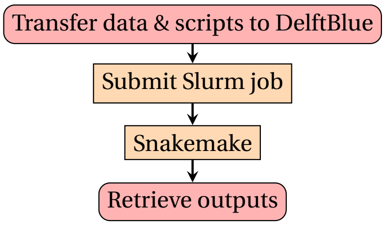


Figure 2.14: **Directed acyclic graph of Delftblue workflow** Red boxes indicate preparational steps for running scripts on Delftblue. Orange boxes present the running order on delft blue - submitting a slurm job (sbatch file type) that in turn runs the Snakemake workflow.

DelftBlue

Due to the size of our input data (33k images and 1Tb storage requirements), we require significant computational resources to perform our analysis. To this end, we make use of the Delft High-Performance Computing Centre (DHPC), DelftBlue Supercomputer.[27] DelftBlue enables the submission of sbatch scripts which can reserve a specified amount of computational resources for a specified time (e.g. fifty CPUs for ten hours).

Submitting an sbatch file that runs our Snakemake workflow enables us to perform our whole GV processing workflow on 33k micrographs in a day taking around 1900 CPU hours or around 3.5 minutes per image (Figure 2.14).

2.3. Statistical inference

Besides serving as a proof of concept for context-preserving Cryo-EM data processing capabilities for increasing imaging analysis efficacy, large-scale analysis also enables the collection of structure statistics on a previously unfeasible scale. We propose a set of measures to be collected and used for subsequent inference on GV biogenesis, more specifically to find support for the proposed stochastic monomer insertion growth model.

2.3.1. GV growth model

We seek to obtain improved structure information of the seam and tips to test the proposed stochastic monomer insertion model as proposed by Huber et al.[5] Additionally, statistics extracted from high-confidence seam positions can inform us on whether GVs appear to follow this proposed model. To this end, we will model the GV growing mechanism as a binomial distribution. We will first introduce the measure we define for performing inference on the GV growth model.

Monomer quantification

To estimate the extent to which the growth dynamics of a GV follow a stochastic insertion model we want to formulate a measure that is standardized for the width and length of the GV. To achieve this we will be using the estimated monomer counts per GV halve. The number of monomers offers a method of measuring the growth difference between the separate GV halves, which is normalized for differing lengths and widths of GV halves. Using the specific length, width, and cone angles we are able to estimate the monomer counts of the GV halves. With a significant number of GVs we can then in turn obtain information on the growth distribution of GVs. Given the central limit theorem, we can assume that a large sampling of these values will provide an accurate estimate of the true distribution of the differences between the monomer counts of GV halves. Inference on GV biogenesis can then be done by comparing the found distribution with the distribution as

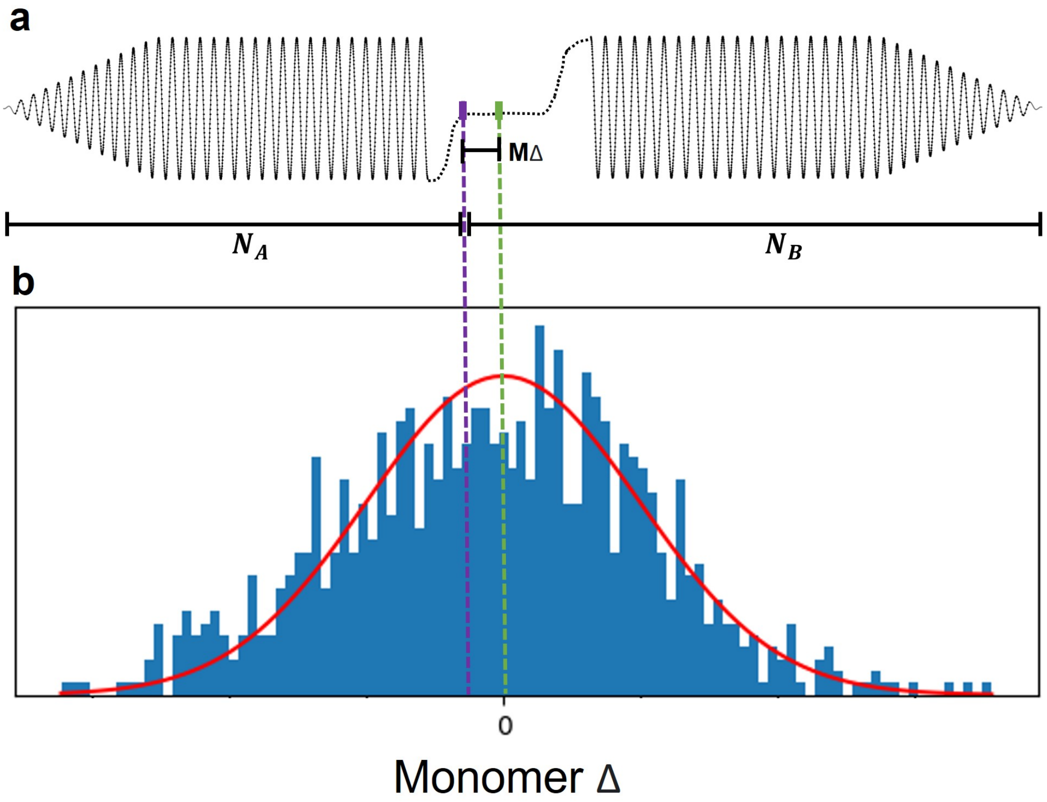


Figure 2.15: **Binomial growth model related to GV monomer counts (a)** GV monomer pseudo model of 4800 monomers (N) as proposed in Section 2.1.2 of GV halves with a cone angle of 60 °, length of 1400 (N_A) and 1600 Å (N_B), and width of 500 Å, leading to an off-centre seam location. The monomer delta measure ($M\Delta$) is defined as the difference between monomer counts of the halves of the GVs ($N_A - N_B$). This is equal to the number of monomers between the geometric centre (green dash) of the GV length and seam location (purple dash), which translates to an absolute $M\Delta$ larger than zero. The centre of the GV is drawn as an untangled line of monomers for illustrative purposes. **(b)** Proposed realisation of $M\Delta$ density histogram of GVs with a total monomer count between 4750 and 5000. The expected pdf from the proposed binomial growth model with $mean$ and σ of 0 and 70 is overlaid (red line).

expected in the proposed stochastic insertion growth model.

We use a numerical monomer fitting approach to estimate the number of monomers per GV halve based on the length, width, and cone angle of the GV halve. The approach assumes that the gas vesicle (GV) is a volume with a cylindrical center and a conical end (caps). Additionally, we assume that the monomers are evenly distributed along a helical path defined within this body, as done by Huber et al.[5]

We disregard GV halves that have a measure cone angle of fewer than 20 degrees since these are likely the result of errors in the upstream pipeline (this results in the discarding of 13 GVs). We use parameters of the previously resolved GV structure: a rotation (twist) of -3.874° and a rise of 0.525 Å per subunit. The number of monomers per helical rotation we find by dividing by the absolute of the rotation ($\frac{360}{3.874} \sim 93$). The pitch, in simple terms, is the vertical distance a helix ascends for one full 360° rotation. We calculate it by multiplying the number of monomers per rotation by the rise ($92.92 * 0.525 = 48.783$Å).

The length of the tip is computed using the trigonometric equation for a right-angled triangle, where the GV's radius is the adjacent side, the tip length is the opposite side, and the tip angle is the angle between them. Therefore, we calculate the GV tip length as

$$\text{tip length} = \frac{\text{radius}}{\tan(\text{tip angle})}, \tag{2.3}$$

the length of the cylindrical portion of the GV is then calculated as the total length minus the tip length. The helical path is defined using a parametric equation, where we iterate from 0 to the total number of turns in the GV. The function assumes that the radius of the helix remains constant within the cylindrical portion of the GV and decreases linearly to zero within the tip. The z-coordinate of the helix (along the axis of the GV) is calculated differently for the cylinder and the cap, taking into account the change in pitch due to the cone angle in the tip.

We then estimate the number of monomers by identifying the local minima in the path length modulo of the unit length of the helix (the length of one full turn), discarding the last few turns. The function assumes that each local minimum corresponds to the center of a monomer. For these calculations, we use the set of previously determined characteristic GV helical parameters (rise (0.525), run), pitch (48.8 Å), and twist (-3.87 °).[5] These we can keep constant since we are dealing with the same sample data and these parameters should not influence the average difference between halves - which is our measure of interest.

GV Center Point

An additional measure is a high-confidence GV geometric center point, which is found by taking the mean value of the high-confidence tip positions. This value is also used to calculate refined estimates of the distance-to-seam measure. Length and widths are obtained by taking Euclidean distances between the high confidence seam and tip points - that in turn inform the estimation of the monomer numbers per halve.

A method of quantifying the growth behavior of monomers is to model the insertion of monomers as following a binomial model of equal probabilities ($p = q = 0.5$). Given a binomial model, the distribution is defined by the number of picks N, and the number of times N_B an option corresponding to p is realized. Their GV counterparts are the number of monomers in a GV (N) and the number of monomers in one half (we pick the bottom half N_B) respectively. We then expect the standard deviation of N_B to converge to $E[N_B] = \frac{N}{2}$ and $\sigma[N_B] = \sqrt{N * p * q} = \frac{\sqrt{N}}{2}$. When we want to analyze the drift of the seam in regards to the geometric center we can measure $P = \frac{N}{2} - N_B$, where N_B represents the number of monomers on one side. Given that $\frac{N}{2}$ is taken as a constant the measure P also has an expected standard deviation of $\frac{\sqrt{N}}{2}$.

Monomer delta

Another way of describing the stochastic insertion is to model the growth as a random walk away from equal size GV halves. The deviation from a completely equal monomer growth distribution is then defined as the monomer difference between the top and bottom GV (Figure 2.15). We coin this measure the monomer delta ($M\Delta$), which is defined as

$$M\Delta = N_A - N_B, \tag{2.4}$$

With N_A representing the number of monomers in the top part of the standardized GV and N_B that off the bottom. The monomer walk (stochastic insertion) is a series of steps in which each step can be either towards the top or bottom half of the GV, with an equal probability of 0.5 for each direction. We can model this using a simple random walk with a step size of 1. In other words, we are assuming a stochastic single-monomer insertion growth model that follows a binomial distribution with an equal probability of insertion. We can therefore consider the positional drift of monomer insertion away from $M\Delta = 0$ as a representation of an elementary random walk.

$M\Delta$ can be understood in terms of the cumulative sum of these steps. Let's denote the cumulative sum at step N as S_N. In this case, S_N represents the $M\Delta$ at the total GV size of N monomers. The expected value of the $M\Delta$ ($E[S_N]$) is zero because, on average, an equal number of steps will be taken towards the top and bottom halves of the GV (equal amount of monomers added to N_A and N_B). This means that over time, $M\Delta$ is expected to not drift consistently in any particular direction.

The standard deviation of S_N ($\sqrt{E[S_N^2] - E[S_N]^2}$) is given by the square root of the number of monomers ($\sqrt{N}$). This reflects the fact that as time progresses and more steps are taken, the variability or spread of the $M\Delta$ increases. However, on average, the $M\Delta$ should not exhibit a systematic bias.

Therefore, the central limit theorem then implies that over a large number of GVs, the distribution of $M\Delta$values can be expected to approximate a normal distribution that corresponds to the distribution produced by an elementary random walk with a mean of 0 (no drift) and a standard deviation of $\sqrt{N}$. Given that we can rewrite $M\Delta$ to $N_A - N_B = N - 2N_B = 2(\frac{N}{2} - N_B)$ we have that the monomer delta is equal to twice our binomial measure P. Therefore we can expect the variance of $M\Delta$ to be equal to follow $VAR(2P) = 4VAR(P) = 4 * \frac{N}{4} = N$. In this way, we can see that measuring either the monomer difference or twice the normalized binomial distribution should lead to the same observed standard deviation.

GV seam-center offset

Besides looking at the monomer difference we also wanted to look at the raw distance difference between the center of the GV and its seam position. To this end, we define the GV seam-center offset (D) as the Euclidean distance between the mean value of the found seam positions and the center of the mask or

$$D = \sqrt{\left(x_0 - \frac{x_1 + x_2}{2}\right)^2 + \left(y_0 - \frac{y_1 + y_2}{2}\right)^2}, \tag{2.5}$$

with (x1,y1) and (x2,y2) representing the left and right seam positions. Moreover, (x_0, y_0) represent the center GV coordinates.

2.4. Preliminary 3D reconstruction of GV tip and seam

We will be using our high-confidence location coordinates as particle picks for a conventional SPA workflow. Then we obtain 2D class averages and possible subsequent 3D volume estimates by using the cryo-EM SPA program CryoSPARC.[20]

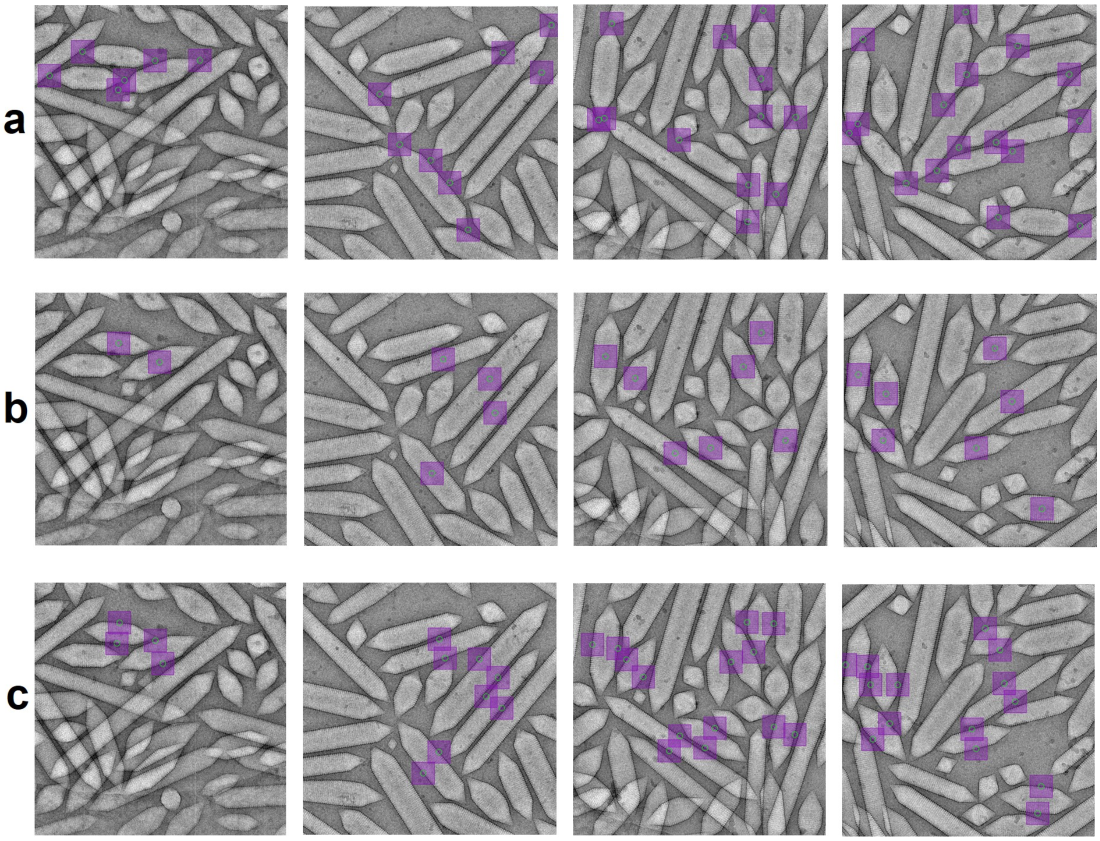


Figure 2.16: **Importing custom particle picks into CryoSPARC (a)** Tip particle picks. **(b)** Center seam particle picks. **(c)** Seam particle picks. Showing custom particle picks imported using CryoSPARC.tools into CryoSPARC for four micrographs. Images are captured from an "Inspect Particle Picks" job within CryoSPARC.

Custom Particle Picks
Integrating custom-picked particles into a CryoSPARC structure determination workflow is enabled through the use of CryoSPARC.tools package. This allows direct scripting interaction with CryoSPARC, enabling the loading of externally picked particle positions into CryoSPARC.

We will be importing three distinct high-confidence location sets. The first set contains the GV tip locations, the second the center coordinate at the seam, and the third will contain the individual seam picks (Figure 2.16 a, b, and c respectively).

For all sets, we will perform clustering and subsequent 2D class averaging to produce high-resolution 2D structure images.[12] When these 2D classes are sufficiently high-resolution we will then try and produce corresponding 3D model volumes through Ab into reconstruction for the tip. We use a "Helical Refinement" job to solve the structure of the GV tube around the seam.[28] Helical refinement using predefined helical symmetry parameters like the pitch and more than D1 symmetry are not possible since the seam and corresponding PRP disrupt the helical symmetry. Additionally, we tried using "Non-uniform refinement" for the tips and seams but this failed to improve upon the reached resolution.[29]

2.5. Code & model availability

Both the code for the pipeline ("*snakemake_data_pipline_package*") and the Cellpose GV model file ("20221010_*GV_segmenter*") are available by request through the Biomolecular Electron Nanoscopy lab repository (`https://gitlab.tudelft.nl/aj-lab`).

3

Results

3.1. Robust low error gas vesicle processing

After applying the GV processing workflow to our input set of 33k cryo-EM micrographs of GVs we obtain 1.15 million segments. These are filtered leading to the extraction of high-confidence features of 86K whole GVs. Here we present the outputs of the pipeline in order of application.

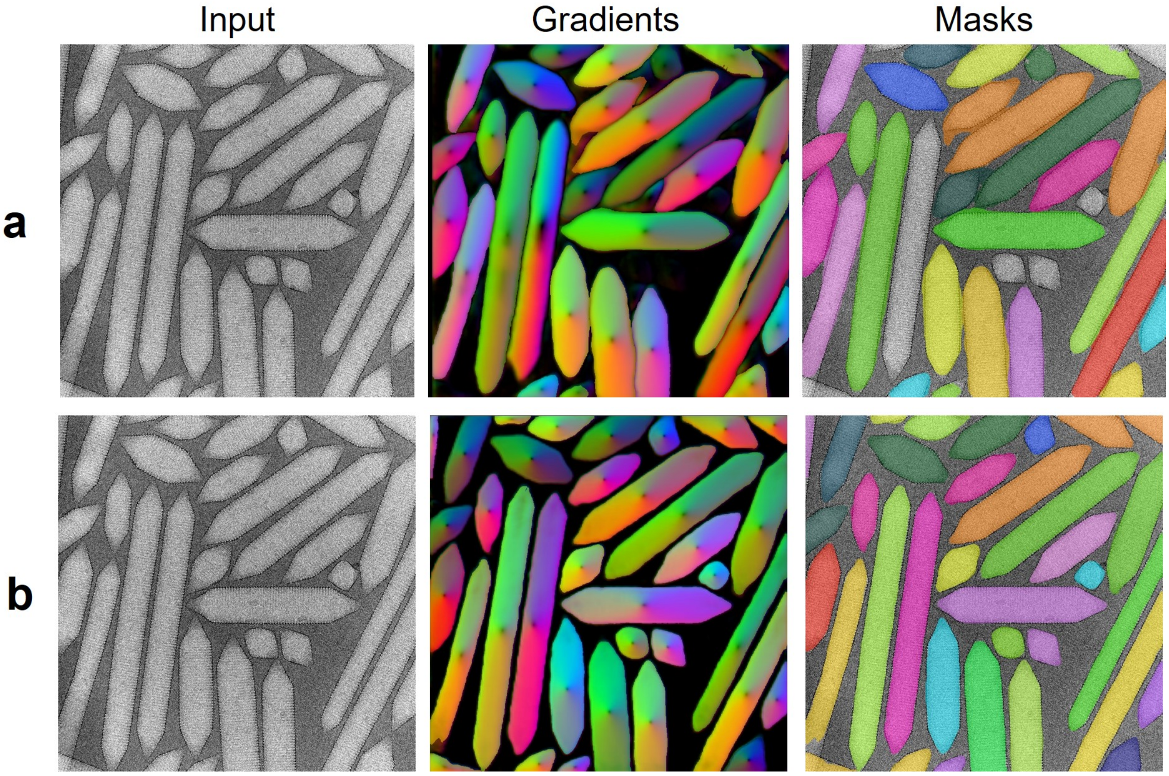


Figure 3.1: **Cellpose GV instance segmentation(a)** Result overview from applying the best performing Cellpose model zoo model (cyto2). Input colom shows a cryo-EM micrograph of purified *B. megaterium* GVs. Gradients show the resulting Cellpose combined gradients output. Masks show the result of combining the gradients with the independent cell pixel predictions. **(b)** Result overview from applying the custom-trained GV Cellpose model.

3.1.1. Standardisation of over a million GV segments

The custom GV model is able to identify a wide range of GV shapes and sizes accurately. The best performing Cellpose model zoo model *cyto2* tends to hallucinate or miss GVs and segment others as if they are more

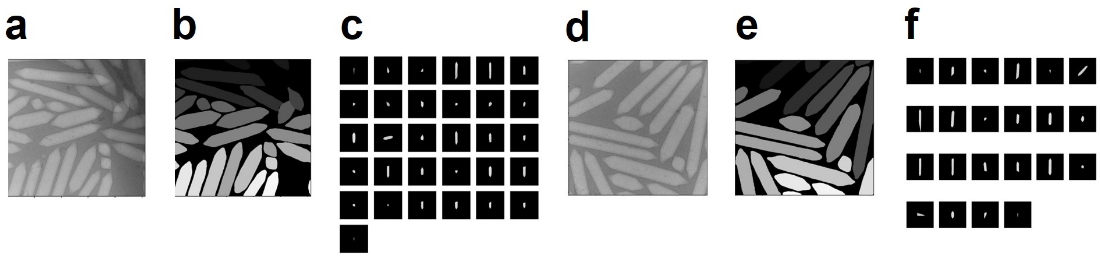


Figure 3.2: **Example GV processing pipeline outputs(a)** Binned input GV cryo-EM micrograph. **(b)** GV Cellpose model output masks. **(c)** GV processing pipeline output. **(d)** Same as (a). **(e)** Same as (b). **(f)** Same as (f).

cell-shaped (blob like, less rigid) (Figure 3.1). This leads to curved segmented GV walls and rounded-off tips. Given that a GV is comprised of rigid walls and sharp tips this is not beneficial to segmentation accuracy. The custom GV model does not suffer from these segmentation issues - segmenting straight lines for the GV cylinder and triangles for the tips. Additionally, we find that the custom GV model has no observed false positives. It does however tend to miss hard-to-segment GVs (overlapping or obscured due to imag impurities), but this we argue is an attractive feature since these kinds of GVs are unwanted for subsequent processing. The resulting 33k segmented GV images contain 1.15 million GV instance masks as per Section 2.1.2).

The processing pipeline is able to reliably segment, extract, rotate and centre the GVs (Figure 3.2). The network does tend to neglect to segment parts of GVs that are overlapping (Figure 3.2 b) and the processing wrongly estimates the rotation angle in a fraction of the GVs (Figure 3.2 c and f). Examples of the preliminary output points resulting from the processed GVs are shown in the Supplementary Figure B.1. However, these erroneous segments can be filtered out by the subsequent filtering steps. In the GV processing pipeline, the 1.15 million segmented GV instances are edge-filtered as described in Section 2.1.3. The resulting set contains 482k GVs.

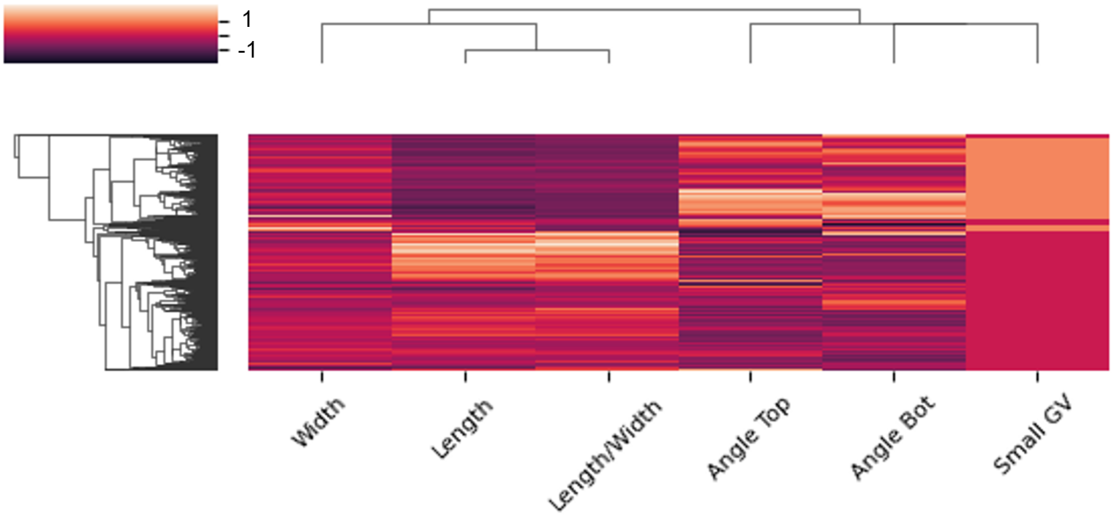


Figure 3.3: **Hierarchical clustering on GV features** Heatmap of hierarchical clustering output on a 100k random sample using a selected set of normalised GV features of the edge-filtered set.

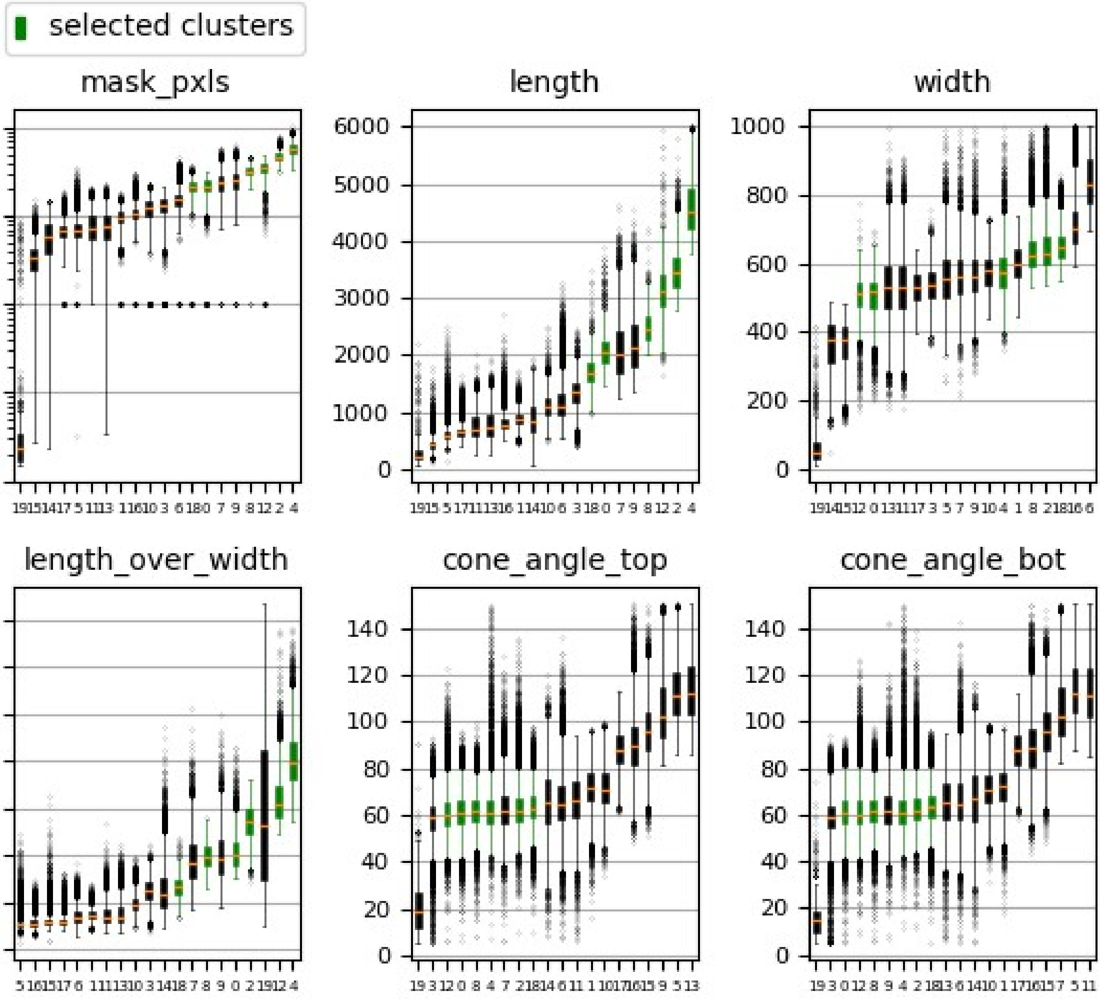


Figure 3.4: **Feature boxplots of (selected) clusters** Boxplot of selected features from preliminary points-of-interest for clusters resulting from k means clustering (with a k of 20). Input data set for clustering contained 482k GVs, selected clusters cover 174k GVs (14006 in Cluster 0, 9212 in 2, 5523 in 4, 21989 in 8, 8437 in 12, and 27310 in cluster 18).

3.1.2. GV segments are separable through clustering on structural features

Subsequently, we perform hierarchical clustering on a set of features of a 100k sub-sample of our edge-filtered set. This results in preliminary information on the distinct features of classes within the data set - see Figure 3.3. Performing PCA or t-sne does not lead to a clear distinction of erroneously segmented GV classes - both show many distinct clusters but not a clear bifurcation as desired. We, therefore, apply k-means clustering and determine the class of proper segments (rigid and accurate in shape) through inspection of feature statistics and standardised GV appearances. We observe that generally top and bottom cone angles are highly correlated, which is assuring assuming that a GV is composed of identical cone structures. Additionally, we observe that the GVs segmented with the is_small_gv method offer a distinct profile in terms of length and cone angles to those from the main method (Section 2.1.2). From this preliminary hierarchical clustering, we can conclude that the data appears separable based on the selected feature statistics.

We perform k-means clustering (k = 20) on these selected features as described in Section 2.1.4. We chose clusters (0,2,4,8,12,18) on the basis of their feature distributions - given that their mean cone angles (~ 60 °), mean widths (~600 Å) and mean lengths (1500-4500 Å) are consistent with previous research - as can be seen in Figure 3.4 (selected clusters in green).[4,5] In the figure we show box plots of the distribution of the preliminary features on a cluster basis. We select clusters mostly on the basis of having a desired width range of around 500 Å combined with consistent top and bottom cone angle estimations. Additionally, we inspect a random subset of processed GV samples from the selected clusters to check their general shape and rotation to establish the validity of the selected clusters, see Supplementary Figure B.2.

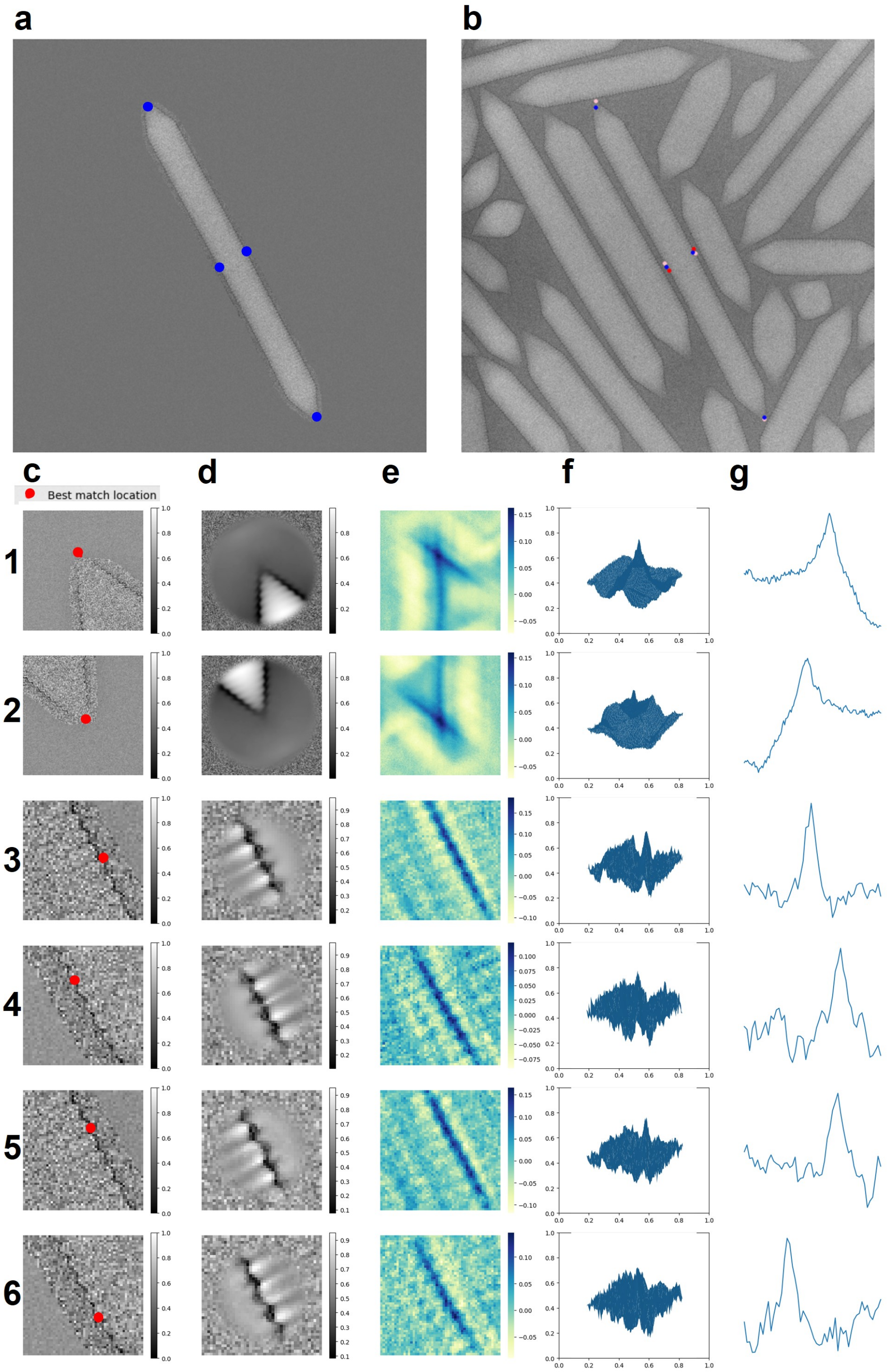


Figure 3.5: **GV template match output (a)** Extracted GV input with corresponding target locations (blue dots). **(b)** Target match location outputs of tips or seam (both pink dots), PRP (red dots) overlaid on input micrograph. **(c)** input target region with best match location (red dot). **(d)** Rotated target reference. **(e)** Cross-correlation map between target and reference. **(f)** 3D rendering of cross-correlation map (rotated at -45 °) **g** X-axis slice of cross-correlation map at y of best match. **(1-6)** represent template match flows of the top tip, bottom tip, right seam, left seam, right PRP, and left PRP targets respectively.

3.1.3. Template matching identifies exact tip and seam locations

In order to have higher confidence particle picks of our features of interest (seam and tip) we perform template matching (Fast Normalized Cross-Correlation) with a reference as described in Section 2.1.5. For the seam picks our estimate is bound to be more accurate since the preliminary seam pick was positioned at the geometric centre of the GV, while we expect and observe the seam to also be located off-center. The template matching location pick can be reasonably expected to correspond to the location of the seam of the GV as described in the testing of Section 2.1.5.

Filtering the GVs from the clustering step (174k GVs) results in the high-confidence tip and seam estimates of 106k top and 107k bottom tips and 94k seam location picks that pass the tests as described in Section 2.1.5. We use only GVs that have templated matched points for both tips and the seam, this results in a high confidence set of 86K whole GVs. An example output of an arbitrary GV from this finalised set is depicted in Figure 3.5 for target matches in Panel 1-4. Here we visually observe that the found tip and seam locations match the observed locations with high accuracy. The PRP location template match we find to generally be off (see Figure 3.5 Panel 5 and 6), this we propose is the result of the PRP being harder to template match due to the fact that the PRP will only be visible in a fraction of images. This is due to the requirement that the GV has to be orientated just right for the PRP to be visible in the cryo-EM projection image, which will not happen often given that the PRP is only a single point on the circumference of the seam. It is for this reason that we chose not to pursue additional analysis of our found PRP template match points. Further example outputs are shown in the Supplementary Figures B.4, B.5 and B.6.

Metrics	Values (±)
N	6702.66 (4844.34)
N top	3339.32 (2454.83)
N bot	3363.34 (2453.29)
M Δ	-8.01 (252.43)
Dist. center-seam [Å]	56.84 (78.32)
Seam diff. LR [Å]	-0.12 (12.26)
L/W	4.73 (3.6)
Width [Å]	572.56 (143.97)
Length [Å]	2628.7 (1655.34)
Top length [Å]	1310.4 (838.15)
Bot length [Å]	1318.3 (837.6)
Mask [px]	31028.68 (23446.38)
Angle top [°]	61.66 (12.96)
Angle bot [°]	61.77 (13.06)
Rot angle [°]	89.98 (103.35)

Figure 3.6: **Overview feature statistics** Mean and 95% confidence intervals of features of interest from Table 3.1. N represents the number of monomers in a GV, with *top* and *bot* referring to the GV halves. $M\Delta$ represents the monomer difference between top (N_A) and bottom halves (N_B) as defined in Equation (2.4). Dist. Center-Seam is our measure D as defined in Equation (2.5). Rot angle refers to the rotation angles found through the application of the Fourier angle method (Section 2.1.2)

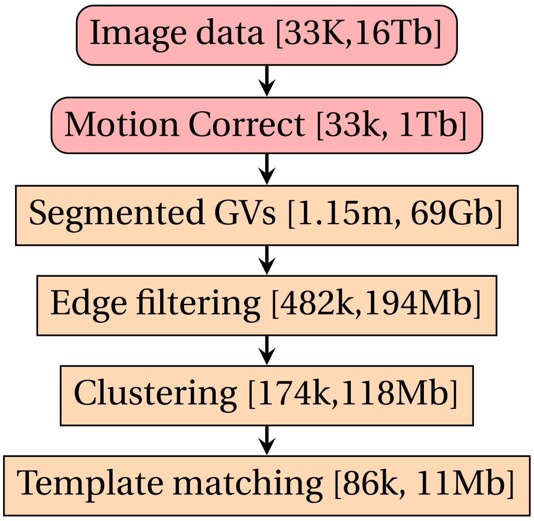


Image data [33K,16Tb]
Motion Correct [33k, 1Tb]
Segmented GVs [1.15m, 69Gb]
Edge filtering [482k,194Mb]
Clustering [174k,118Mb]
Template matching [86k, 11Mb]

Figure 3.7: **Data processing pipeline** Flowchart of the complete data processing pipeline. Boxes in red represent steps performed by collaborators, where the numbers between brackets represent [image number, data size]. Boxes in orange represent steps performed within this work, where the numbers between brackets represent [GV number, data size].

3.2. Output statistics show a varied GV set

Our ML-based segmentation and subsequent processing workflow as defined in Section 2 processed 33k cryo-EM images with a size of 1 Terabyte (Tb) to a segmented set of 1.15 million GV instances and then down to a data frame of 86 thousand high confidence GV points (Figure 3.7).

We subsequently use these high-confidence tip and seam locations to refine our estimates of the GV (halves) length, and widths. We calculate an improved length estimate from the distance between the tip and seam positions, and the widths from the distance between the left and right seam position. Additionally, the values are multiplied by the realised pixel size of the analysed images. This we calculate by multiplying the pixel size of the microscope by the binning factor (binning factor: 4, microscope pixel size: 1.518 Å/px).

These refined estimates are then used to estimate the number of monomers (N) at the top and bottom of the GV as described in Section 2.3.1. We chose to use the whole set mean cone angle of 61.72 ° as a constant when calculating the GV halve monomer counts because errors in our method of determining these angels could result in erroneous differences in cone angels between GV halves - this could potentially obscure our estimate of the monomer count differences. The proposed cone angle estimation errors are however of such small magnitude that on average the estimated cone angle should be representative of the true value. It will therefore not influence the validity of the mean cone angle estimation. Additionally, we argue that using a constant tip angle should not significantly affect the distribution of $M\Delta$ given that only a fraction of monomers are contained in the tips. These statistics are then used to calculate the monomer delta ($M\Delta$) and distance to the centre-seam (D) as defined in Section 2.3.

The resulting refined set of 86K GVs measures forms the basis for the rest of our analysis (Table 3.1). The summary statistics are in line with values found previously (Figure 3.6).[4,5] We find that the mean $M\Delta$ is significantly different from 0 (mean = -8.01; σ = 128.79, t (86k-1) = 18.24; $p \sim 0$). This is not what we expect to find given the proposed growth model but we contribute this to some systematic error within our pipeline and not to any real-world physiology since the orientations of the GVs are assumed to be uniformly distributed within the images. Therefore, any physiological reasons for seam-centre offsets should be smoothed out over the random sampling of orientations from GVs picked in the pipeline. Potentially, the found rotation angle (α°) has a systematic bias that results in shorter GV halves having a slightly higher chance of being rotated to the top of the GV. This could then result in a slight negative bias in our estimate of $M\Delta$.

Compared to previous work, we find a slightly higher cone angle of 61.72 °, but this can be attributed to an erroneous estimation of the base height of the GV. Previous work found GVs ranging from 0.1 - 1 μm in length.[5] Our filtered set ranges from 0.1 - 0.6 μm. We propose that this is a result of filtering or potentially sample prep differences. The former is because longer GVs are more often cut off from the image field, and it could be possible that we used a smaller image view than previous studies. The latter since purifying the GVs involves centrifugal separation of the GVs from the bacterial host proteins and broken GVs. Within these iterative cycles of spinning and subsequent supernatant separation, the shear forces in both the spinning and subsequent pipetting could introduce a bias towards shorter and wider GVs due to their higher shear-force resistance leading to filtering for smaller GVs.[5]

Index	Image ID	N	N Top	N Bot	$M\Delta$	Dist. Center-Seam [Å]	Width [Å]	Length [Å]	Top Length [Å]	Bot Length [Å]
00000	00235_1-1	1972	1109	863	246	133.16	592.35	2604.83	1435.44	1169.38
00001	00235_1-1	1950	964	986	-22	13.57	606.78	2622.32	1300.58	1321.73
00002	00235_1-1	1168	726	442	284	141.75	692.23	1781.59	1030.11	751.48
00003	00485_1-3	2895	1484	1411	73	38.87	589.47	3713.11	1893.27	1819.84
...	...	...	...	...	...	...	...	...	...	...
86490	33085_1-3	1680	795	885	-90	79.16	622.31	2287.61	1065.80	1221.80

Table 3.1: **Refined GV statistics data frame** Refined statistics data frame of 86k GV resulting from applying the filtering pipeline showing selected GV identifiers and features. N represents the number of monomers in a GV, with top and bot referring to the GV halves. $M\Delta$ represents the monomer difference between top (N_A) and bottom halves (N_B) as defined in Equation (2.4). Dist. Center-Seam is our measure D as defined in Equation (2.5).

Cluster statistics

Additionally, we inspect the refined feature statistics on a cluster basis as can be seen in Figure 3.8. We find $M\Delta$ to be converging to a normal distribution around a mean of 0. Similarly, the distance to the centre-seam ("Dist. center-seam") also does not show any notable inter-cluster dissimilarities. The difference in the height of the estimated seam position of the left and right GV walls ("Seam diff LR") is interestingly bifurcated around 0. We expect this to be the consequence of the seam having a slight height gradient around its circumference due to the helical nature of both sides. A correctly estimated left seam position height will therefore always be off from the right counterpart.

The length-over-width measure ("L/W") illustrates the distinct type of GVs within the clusters. The corresponding width and length statistics show that this variation mostly stems from differences in GV lengths. This is in part a consequence of selecting for clusters with widths and cone angle profiles consistent with previous research. The top or bottom length and the mask pixel sizes ("Mask") also reflect this inter-cluster difference. Last, the estimated rotation angles of the GVs within the images are uniformly distributed within the possible range, as is expected given the arbitrary orientations and size of the set.

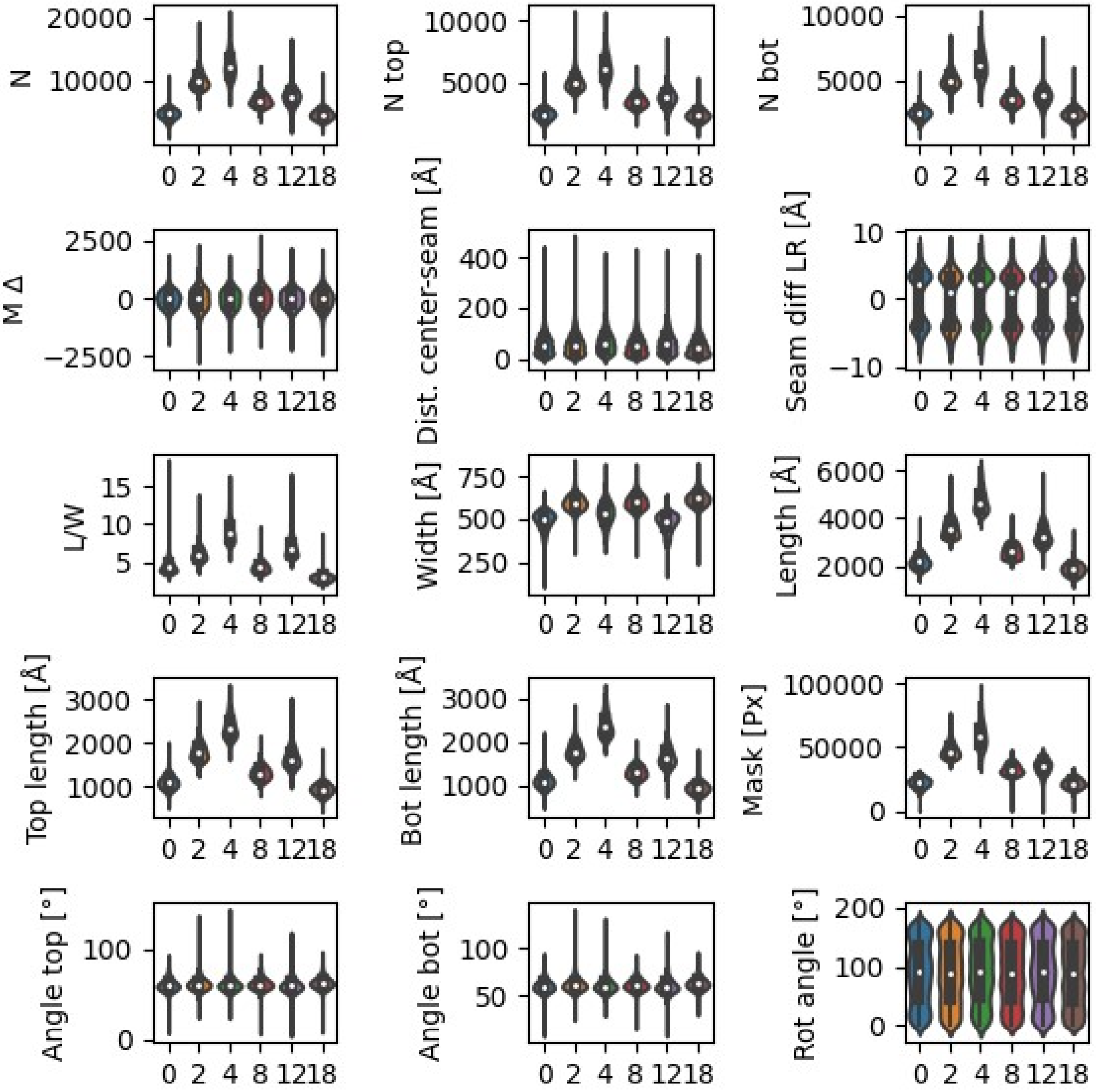


Figure 3.8: **Cluster sorted statistical overview refined GV data** Violin plot overview of refined measures sorted on selected clusters. N refers to the number of monomers. Top or bot refers to statistics concerning a GV half. $M\Delta$ represents the monomer difference as defined in Equation (2.4). The angle refers to the whole cone angle at respective GV halves. Rot angle refers to the rotation angles found through the application of the Fourier angle method (Section 2.1.2).

3.3. Growth dynamics appear stochastic

Using our refined statistics (Table 3.1) we are able to visualize and measure relations between variables to inform us about the proposed stochastic GV growth model. Within the following trend and feature relation analysis we chose to create bins along our data set along a feature of interest. We chose to bin to be able to infer marginalized relations that allow for clear visualization of results and allow estimation of the distributions of our relation of interest. Depending on the measure of interest we chose a bin size that is as small as possible to negate any underlying distribution gradient effects, but large enough to contain enough samples for sufficiently high confidence estimation of the distribution parameters. We subsequently perform analysis on a related feature of interest within these bins. Plotting the resulting outputs and performing regression analysis enables the identification of trends and possible relations between the features of interest. We chose to only take into account bins that contain a minimum of 100 particles to provide confidence in the stability of the estimates.

We chose to not include R^2 estimates of our regressions since most often this number does not add relevant information to what is already visually verifiable. Moreover, R^2 measures are only relevant when rigorous testing of the corresponding residuals has been performed, this we find to be outside the scope of the type of relations we propose to illustrate. Moreover, to account for the effects that binning might have on the distributions of the feature relations we chose to add parallel visualizations using so-called hexbin plots. These visualize the number of particles in a finely discretized feature relation space showing the underlying feature data. From these plots, we can infer densities and general shapes of the direct relations.

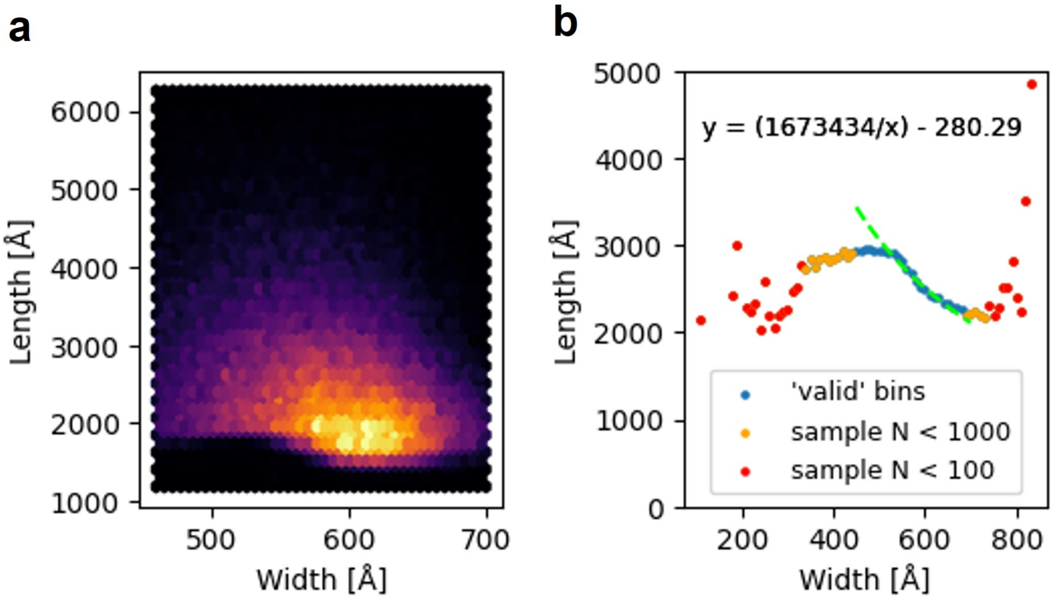


Figure 3.9: **Width vs. length (a)** Hexbin plot of the widths versus the lengths in our refined GV table 3.1. **(b)** Scatter plot showing the mean lengths of GVs within equally spaced bins (10Å) ranging from widths of 0 to 1000 Å. Red dots indicate bins with less than 100, Yellow with less than 1000, and blue with more than 1000 GVs.

Length, width, and monomer count relations

We first want to establish the relation between the length, width, and monomer numbers to help inference on their effect on ΔM. Knowing how these features relate to each other can help predict how they will be related to the number of monomers in a GV (N), which we expect to influence ΔM.

First, we explore the relationship between the GV width and length. Figure 3.9 a shows a hexbin plot, that relates the width to the length of GVs. We observe that most particles have a width of around 600 Å and a corresponding length of 1800 Å. Moreover, we find that for smaller GVs we have less density around the same length.

We perform binning along the width (bin size of 10 Å) and find a Sigmoid-looking relation for the domain of bins with more than 1000 GVs, see Figure 3.9. Given a constant growth rate and that the amount of monomers scales as $\pi * width$, we then expect that the length of a GV will be inversely related to its width. Since with an equal growth rate, a wider GV will grow less long than a more narrow one. Consequently, we expect the width and length to roughly be inversely related. Therefore we fit a power-law function following $y = \frac{a}{x} - c$. The resulting fit with an a of 1.6 million and c of 280 appears to not hold any physiological relevance. We propose that the distribution might be affected by the shear forces of GV purification leading to inverse selection for long and thin GVs as described in Section 3.2. We can however state that a general negative width-to-length relation appears valid.

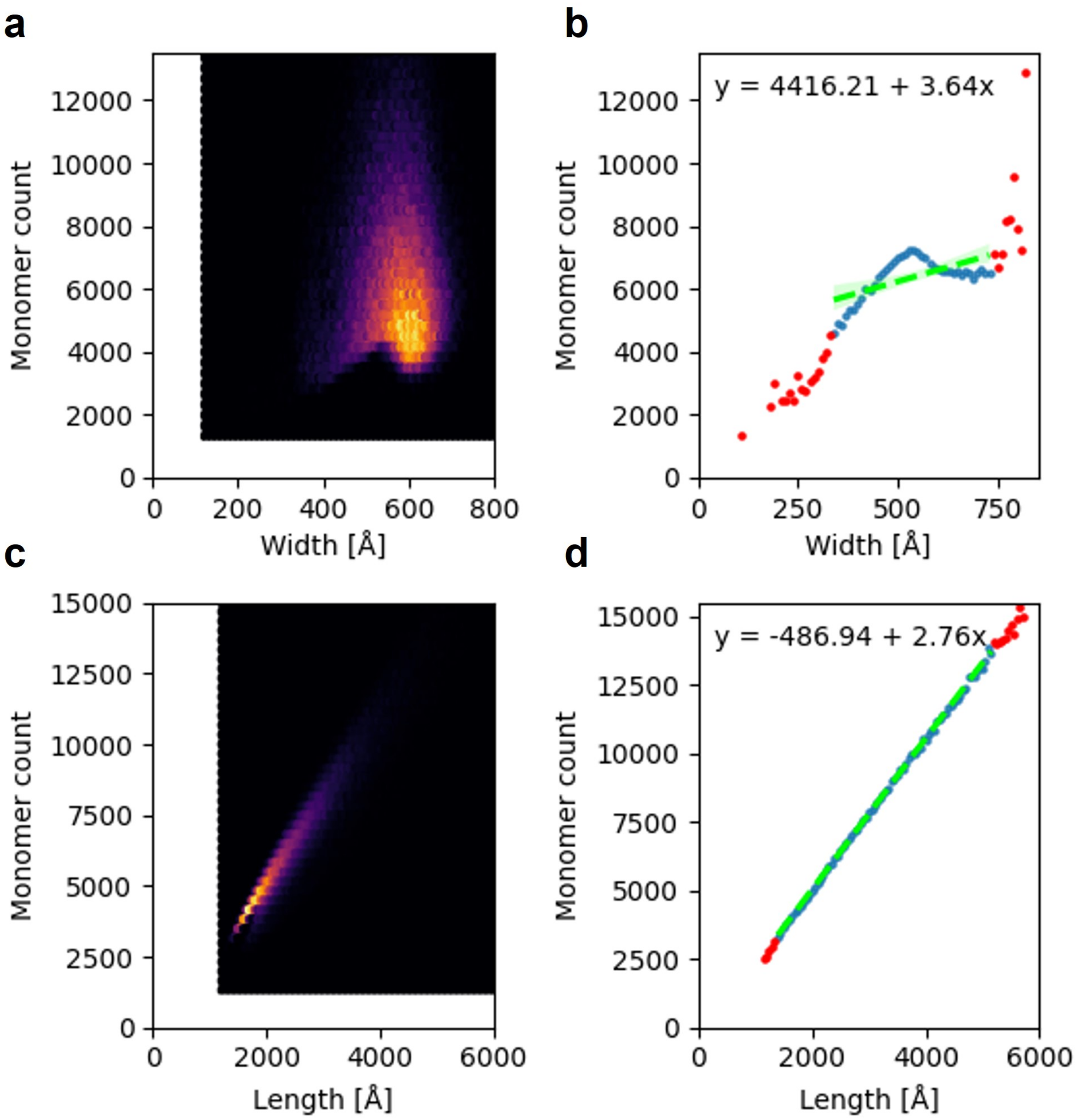


Figure 3.10: **Width and length vs. monomer count** (N) **(a)** Hexbin plot of the widths versus N in our refined GV table 3.1. **(b)** Scatter plot showing the mean widths of GVs within equally spaced bins (10Å) ranging from widths of 0 to 1000 Å. **(c)** Same as in (b) but of length versus N. **(d)** Scatter plot showing the mean lengths of GVs within equally spaced bins (50Å) ranging from lengths of 0 to 6500 Å. Red dots indicate bins with less than 100 GVs.

In Figure 3.10, we illustrate the estimated relationship between the width or length and the monomer count (N). The linear regression fits suggests a positive correlation between the length or the width and N. However, it is noteworthy that the width does not exhibit a similar linear relationship as observed with the length. This discrepancy can be attributed to our estimation of the monomer counts N_A and N_B using the corresponding GV half-lengths and widths, as detailed in Section 2.3.1. Furthermore, drawing upon the findings in Figure 3.9, which validates a negative width-to-length relationship. We hypothesize that this inverse width-length relation causes the monomer increase corresponding to a width increase to be negated by a non-linear decrease in length. This, in turn, results in the observed non-linear width-N relation.

Length vs. distance centre-seam

When we plot the lengths of GVs versus their center-seam distance (D) we find relations that appear linear (Figure 3.11 b). Figure 3.11 a presents a hexbin plot that shows the number of particles in the discretized surface of length versus D. From this, we observe that most particles are around a length of 2000 Å and 20 D and appear with a Gaussian-like spread around that point. We plot the mean D of GVs in equally spaced length bins (50 Å). When we fit a linear model through linear regression we find an intercept of 42.21 and a slope of 0.01. This intercept points to a 40 Å offset from zero, which could be caused by the allowed estimation error of the seam position of up to 50 Å (distance of one helical GV ring). Furthermore, the found slope points to a relation where a 1000Å increase in the length of a GV leads to an average added D of 10, which points to the expectation that D would increase with the length of the GV. This is in line with expectations of a stochastic growth model where a longer GV leads to an increase in the expected center-seam offset.

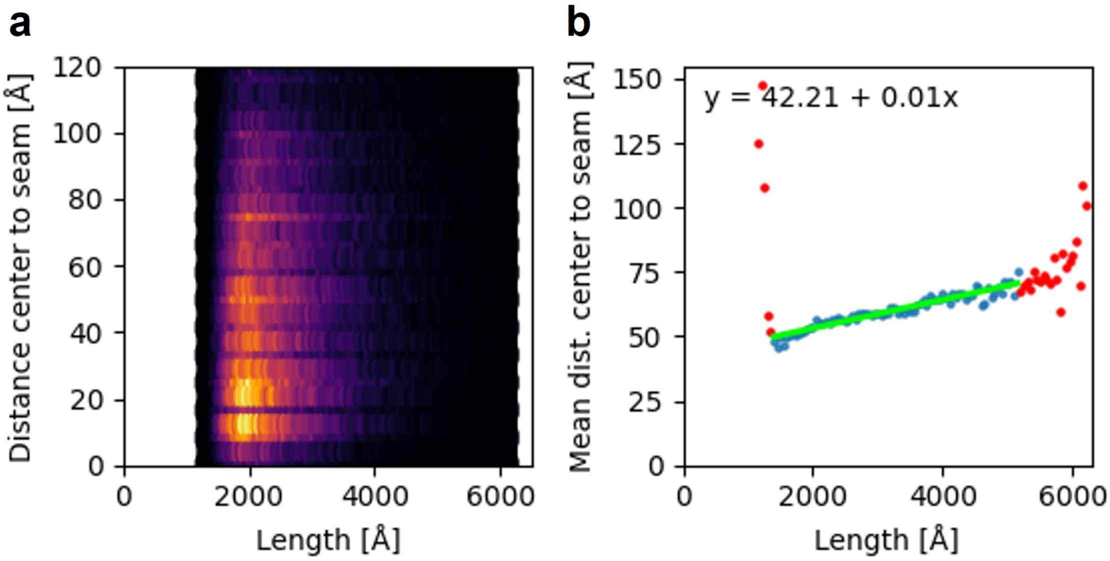


Figure 3.11: **Length vs. seam-center difference (a)** Hexbin plot of the lengths versus the distance from center to seam position (D) as defined in equation 2.5. **(b)** Scatter plot showing the mean D of GVs within equally spaced bins (50Å) ranging from a length of 0 to 6300 Å. Red dots indicate bins with less than 100 GVs.

Monomer difference relations

When performing a similar analysis of the relationship between length, width or monomer counts with the monomer difference between halves ($M\Delta$) we find additional relations supporting a stochastic model.

The first relation of interest is how the length influences $M\Delta$. We obtain an estimate by performing a binning of the lengths (bin size of 50 Å) and measuring the corresponding bin mean $M\Delta$. We find a linear fit with an intercept of -30.33 and a slope of 0 (Figure 3.12). Additionally, we observe an increasing variability in the $M\Delta$ as the length increases. This is in accordance with the expectation that with a stochastic growth model (converging to an elementary random walk in the limit) we expect the standard deviation of the mean distance traveled to increase in the order of the square root of the number of steps. Given the results in Figure 3.10 d, this is as expected. We do however expect the mean distance ($M\Delta$) to have an expectation of 0, which is not found to be the case. We expect this to be the result of some combination of improper estimations at points along the GV pipeline leading to a systematic error, following the proposed reason from Section 3.2.

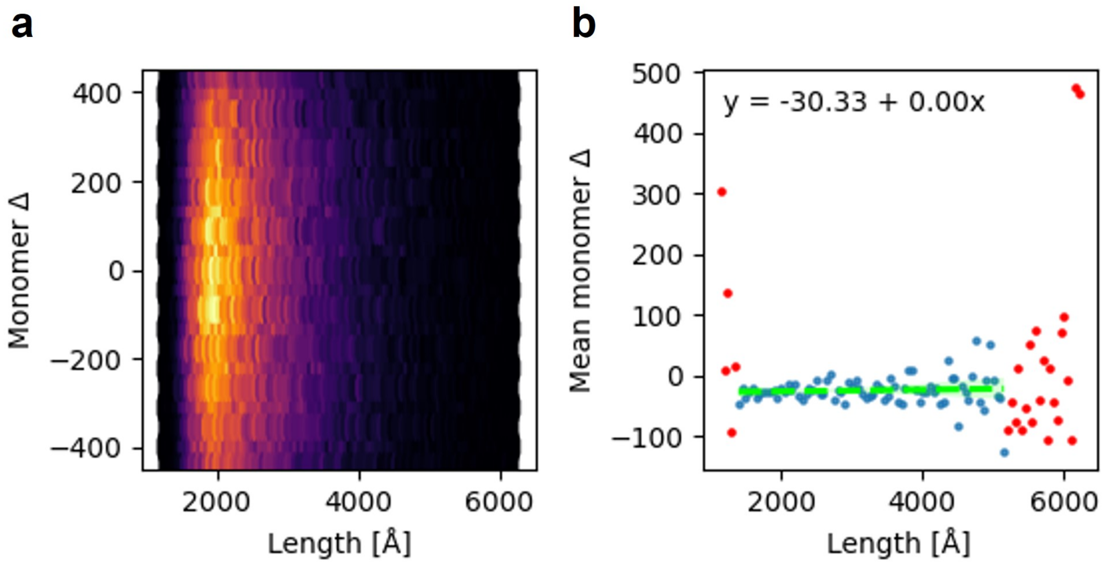


Figure 3.12: **Length vs.** $M\Delta$ **(a)** Hexbin plot of the lengths versus the monomer difference ($M\Delta$) as defined in equation 2.4. **(b)** Scatter plot showing the mean ($M\Delta$) of GVs within equally spaced bins (50Å) ranging from a length of 200 to 6300 Å. Red dots indicate bins with less than 100 GVs.

Besides the length, it's also interesting how the width relates to the monomer delta. We obtain an estimate by binning on width (bin size of 10 Å). The resulting linear regression intercept and slope are -15.40 and -0.02 respectively (Figure 3.13). We find the spread in mean $M\Delta$ of the widths to be a lot less variable than those of the length-$M\Delta$ relation. We propose that this is a consequence of the width having less of a straightforward relation to the number of monomers in the corresponding GV. An increase in width is proportional to the number of monomers of a single helical loop, but given an equal growth rate, this translates to a lower average GV length - as we have seen in Figure 3.9. Additionally, we expect that thinner GVs and therefore also longer GVs are more underrepresented in our data set due to a lower sheer stress resistance. It is these arguments that we propose to shape the relation between the width and N as shown in 3.10 b. Combining these effects we propose explains how the width seems to not significantly affect the variability in the spread of the mean $M\Delta$ estimates.

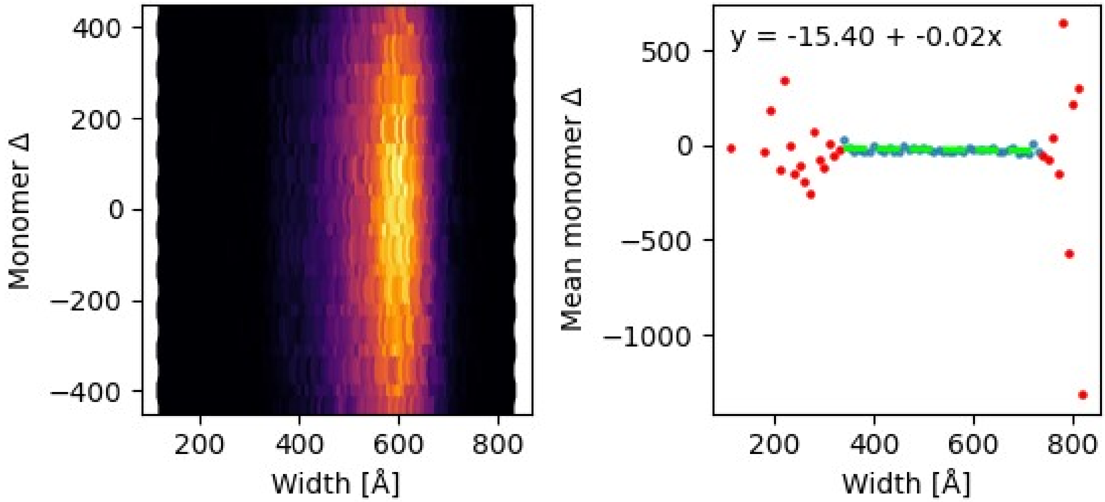


Figure 3.13: **Width vs.** $M\Delta$ **(a)** Hexbin plot of the widths versus the monomer difference ($M\Delta$) as defined in equation 2.4. **(b)** Scatter plot showing the mean ($M\Delta$) of GVs within equally spaced bins (10Å) ranging from a width of 100 to 850 Å. Red dots indicate bins with less than 100 GVs.

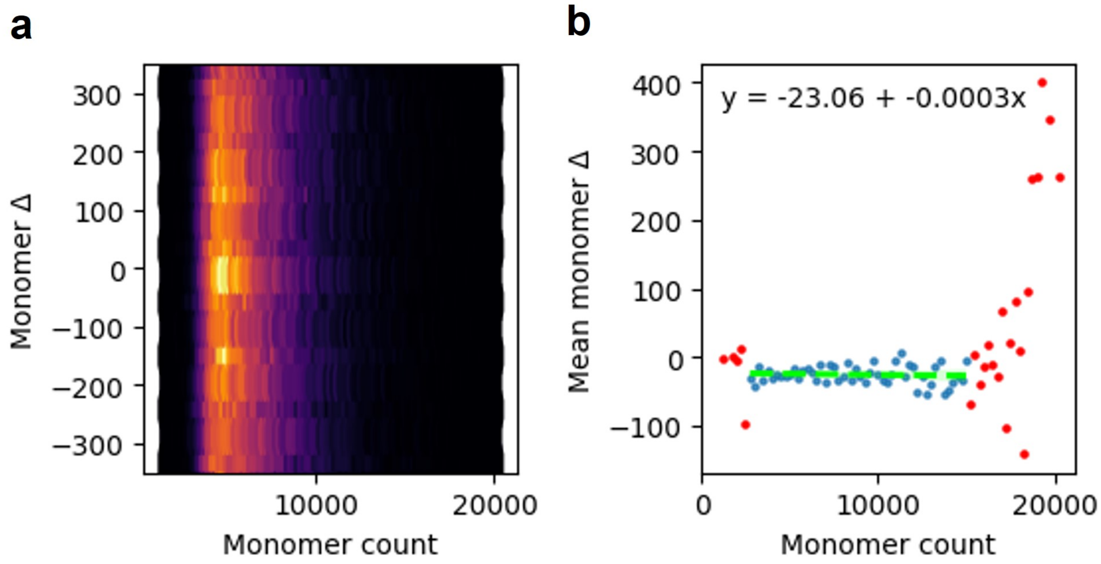


Figure 3.14: M **vs.** $M\Delta$ **(a)** Hexbinplot of the monomer counts M versus the monomere difference ($M\Delta$) as defined in equation 2.4. **(b)** Scatter plot showing the mean ($M\Delta$) of GVs within equally spaced bins (250 M) ranging from a length of 0 to 12500 M Å. Red dots indicate bins with less than 100 GVs.

When performing the same analysis but instead binning on the monomer count (N) (bin size is 250 N) we once again find a similar relation of increasing variability as the number of monomers increases but with a systematic downward error (Figure 3.14). A linear regression fit results in an intercept of -23.06 and a slope of -0.0003.

We want to test if the proposed elementary random walk distribution expectations of the standard deviation ($E[M\Delta^2] = N$ or $\sigma(M\Delta) = \sqrt{N}$) match those we can measure for the respective monomer count (N) bins. To this end, we once more bin over the number of monomers (bin size of 250 N) and calculate the standard deviation of $M\Delta$. We find that the expected standard deviation of $\sqrt{N}$ is less than the found standard deviation (Figure 3.15 a). We chose to only perform regression on bins with more than a thousand GVs (Figure 3.15 b) to increase the validity of the standard deviation estimate. As shown in Figure 3.15 c through linear regression we find an intercept of 287.02 and a slope of 1.40. This is quite far off from the expected standard deviation (yellow unit line). However, it does match the general trend that an increase in N translates to a higher standard deviation of $M\Delta$. Normalizing for the expected standard deviation we find a linear regression fit with an intercept of 9.09 and a slope of -0.05 (Figure 3.15 d). From the normalized relations, we find that the difference in the proposed model for the standard deviation and the found deviation appears to decrease with an increasing number of N.

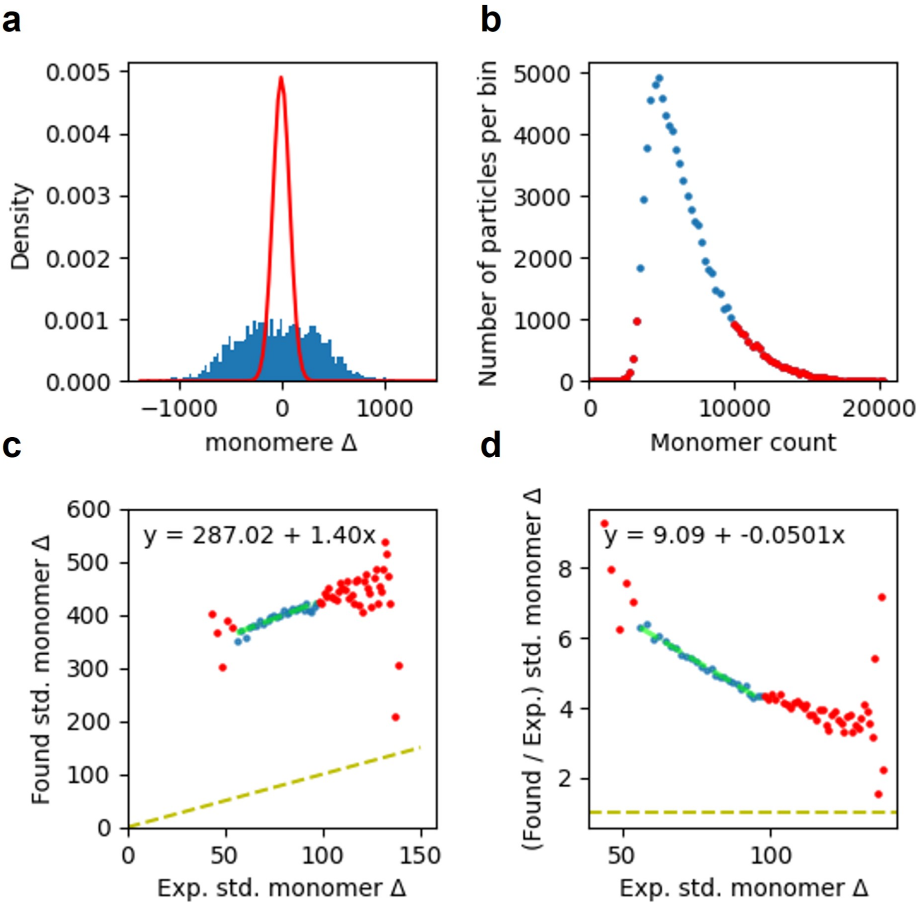


Figure 3.15: **Expected std.** $M\Delta$ **vs. found std.** $M\Delta$ **(a)** Histogram (blue) of the monomer difference ($M\Delta$) (Equation 2.4) for GVs with a monomer count (N) between 6000 and 6250. Overlaid is the pdf of a normal distribution with the proposed monomer difference in Section 2.3.1 ($N[0, \sqrt(N - error)]$ **(c)** Scatter plot showing the number of particles per bin for equally spaced bins ($250M$) ranging from 0 to 20000 M. **(c)** Scatter plot showing the expected std. $M\Delta$ versus the found std. $M\Delta$ of the bins of (b). textbf(d) Same measure as in (c) but normalized for the Expected std. $M\Delta$. Red dots indicate bins with less than 1000 GVs.

We suspect that an estimation error in the location of the PRP is at most the number of monomers in one ridge. The number of monomers in a ridge was found to be 92 for a width of 356 Å. Given that the amount of monomers in a ridge is related by a factor π to the width we can assume a linear relation. Therefore, we can vary the estimation error to the mean widths within a bin by scaling 92 N by the ratio of mean divided by the corresponding example width of 356 Å. This error we define as N_{error} We propose to take into account this standard deviation by modeling the error N_{error} by a standard uniform ranging from -2 N_{error} to 2 N_{error} (given that the error is doubled in the difference). We use that standard deviation should converge to $\sigma[U(a,b)] = \frac{(b-a)^2}{12}$ in the limit. Therefore, we can add the squared standard deviation to the expected variance to account for the estimation error (Figure B.7). This however does not seem to correlate to a more clear distribution. Additionally, we find that performing the same analysis on $\frac{M\Delta}{2}$ (Figure B.10) combined with the error estimation leads to the expected standard deviation to match the found standard deviation (Figure B.8). Interestingly, this we argue is interchangeable with comparing P to an expected standard deviation of $\sqrt{N}$, since $\frac{M\Delta}{2} = \frac{(N_A - N_B)}{2} = \frac{N}{2} - N_B = P$. We are, however, not able to formulate why half the monomer difference or P results in this correct fit with a twice as large as expected standard deviation of $\sqrt{N}$. It does however highly fit expectations as proposed for $M\Delta$ following a stochastic insertion model (Section 2.3.1).

To verify our model proposed relation between a binomial model as represented by measure P and $M\Delta$ we perform the same kind of analysis on P. P translates to comparing the standard deviation of N_A with the expectations from a binomial model of N picks as proposed in Section 2.3.1. The found relation using P (binomial model) closely resembles the one found by comparing $M\Delta$, which is as was proposed in Section 2.3.1. The results and corresponding analysis can be found in Supplementary Figure B.10 in Supplementary B.4.

3.4. Promising preliminary reconstructions

3.4.1. High resolution 2D class averages support the existence of the polarity-reversal-point and tip nucleations

Using CryoSPARC we obtained 2D class averages of our high-confidence points of interest set. We perform a "2D classification" job with close to default parameters, tweaking parameters iteratively in a heuristic-guided fashion. Importing our tip data set leads to the classification of sets of thousands of tip particles. This involves importing the points into an empty job using the CryoSPARC.tools package. These particles are then together with CTF and motion-corrected images combined to extract the points of interest. Then we perform a high pass filter of 300 Å for the seam center picks, 200 Å for the tips, and 100 Å for the seam picks. We chose and subsequently probe around these cutoffs based on the size of the particle extraction box size and the amount of contextual information that we think is required for proper centering. Applying a high pass filter reduces information not required for aligning the particles to distinct classes in 2D classification. We find that this increases the alignment and subsequently resolved resolution of the 2D class averages. Subsequently, we perform iterative rounds of 2D classification and particle extraction to create a subset of particles that are able to be used for high-resolution 2D class averages.

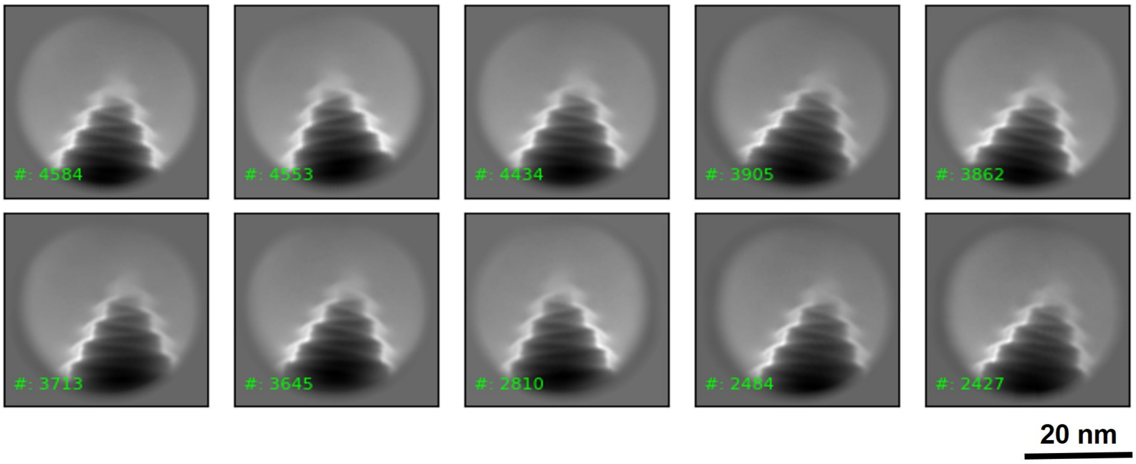


Figure 3.16: **2D class averages of tip particles** 2D class averages are displayed with the number of particles per class (mint #). Generated using CryoSPARC (tools).[20]

Our initial tip set extracted with a box size of 256 pixels (388.61 Å) results in high-resolution 2D class averages of Figure 3.16. In these classes, we observe the general cone helix with ridges visible with white lines. They distinctly show a conal monomer tip structure as proposed by Huber et al.[5] At this resolution the helical cone bands are clearly visible (Figure 3.16). The large box sizes does however lead to difficulties in alignment which results in a lower realised 2D class average resolution.

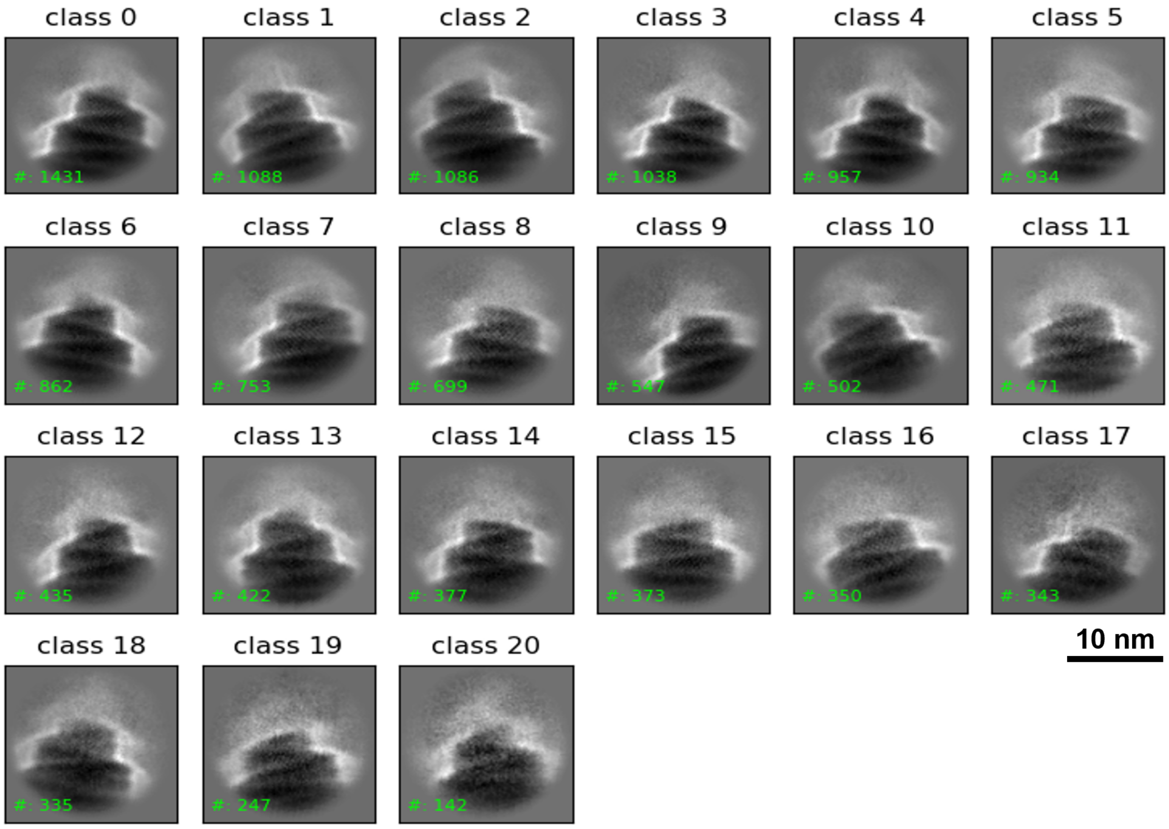


Figure 3.17: **Refined 2D class averages of tip particles** 2D class averages are obtained with a smaller box size than in figure 3.16. Displayed with the number of particles per class (mint #). Generated using CryoSPARC (tools).[20]

When we take a smaller cutout box (128 pixels, 194.30 Å) CryoSPARC is better able to center and form higher-resolution 2D class averages. We do however end up with significantly smaller classes after additional rounds of 2D classification and selection. While we might lose an overview of the whole tip structure we gain high-frequency information (Figure 3.17). Looking at class 7 of this set, we can clearly see the distinct traces of α_1 backbones of GvpA as proposed by Huber et al (Figure 3.18).[5]

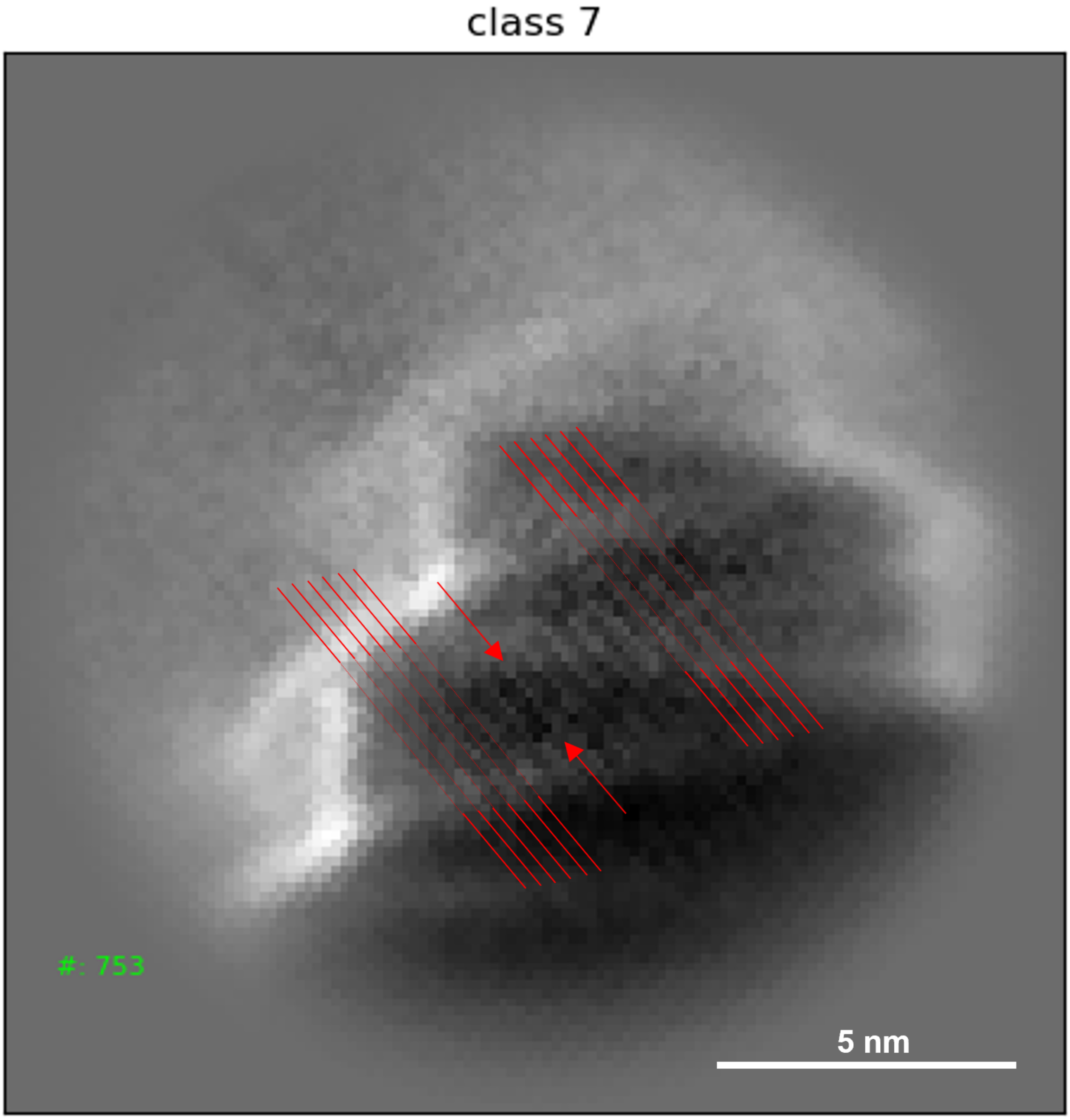


Figure 3.18: $\alpha - 1$ **linking visible in GV tips 2D class average** GvpA backbone $\alpha - 1$ chains are visible in class 7 of the 2D GV tip points averages. The red ley lines are drawn onto the $\alpha - 1$ chains on both the inner and outer spiral of the tip. A specific example of a distinct $\alpha - 1$ is indicated (red arrow). Displayed with the number of particles per class (mint #). Generated using CryoSPARC (tools).[20]

Using the center-seam particle set we are able to create high-resolution 2D class averages that show the symmetrical mirror axis that the seam creates within the GV wall structure where GvpA monomers reverse in structural direction (white arrow, Figure 3.19). Additionally, the seam is the point of the GV where the two halves meet, at the seam these halve are closer than the average distance between helical ribs. This causes a smooth appearance.

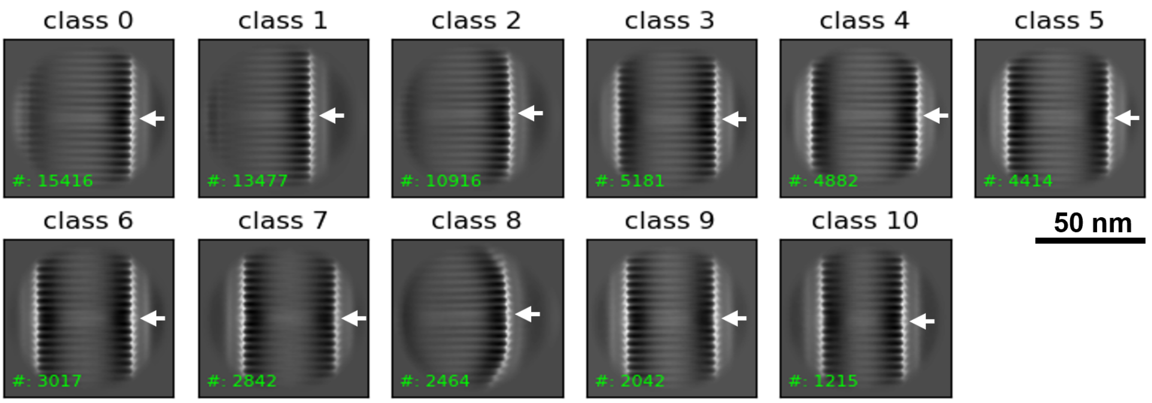


Figure 3.19: **2D class averages of seam center particle** GvpA monomers reverse in structural direction (white arrow). Displayed with the number of particles per class (mint #). Generated using CryoSPARC (tools).[20]

Importing our seam particles allows us to create high-resolution 2D seam class averages at a resolution of 4 Å reported by CryoSPARC (trough FSC comparison). These clearly show the features of the seam and PRP as proposed previously (Figure 3.20).[5] The location of the seam is clear by the symmetrical mirror point that the seam creates within the GV wall structure where GvpA monomers reverse in structural direction (white arrow). We find classes that show an asymmetric reversal, indicating that this class is close to showing a 2D projection of the PRP point (blue arrow). Additionally, we find a class that has a near-perfect overlap of the two opposing GvpA sidechains, indicating that this class is showing the PRP (blue arrow). These 2D classes support the existence of the PRP.

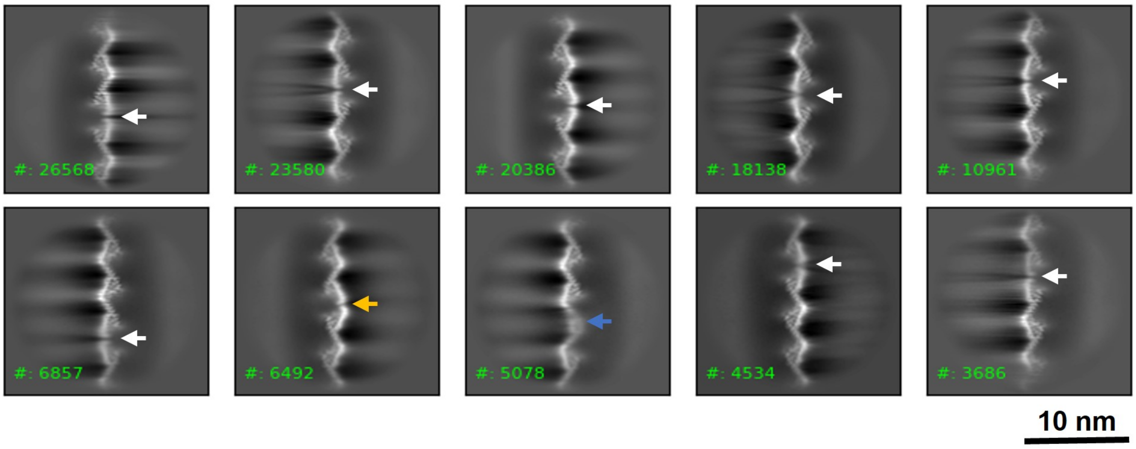


Figure 3.20: **2D class averages of seam particles** GvpA monomers reverse in structural direction (white arrow). Asymmetric GvpA structure reversal indicates that this class is close to showing a 2D projection of the PRP point (blue arrow). near-complete overlap of the two opposing GvpA sidechains, indicating that this class is showing the PRP (blue arrow). Displayed with the number of particles per class (mint #). Generated using CryoSPARC (tools).[20]

3.4.2. 3D volumes show the feasibility of seam and tip reconstruction

We perform 3D volume reconstruction as described in Section 2.4 on the seam-duo and tip particles independently. The resulting structures can be compared to previous densities for the seam and the proposed pseudo-atomic model for the tip. From these comparisons, we propose that our preliminary resolved volumes are supportive of previous findings and indicate the possibility of resolving the tip and seam at high resolution.

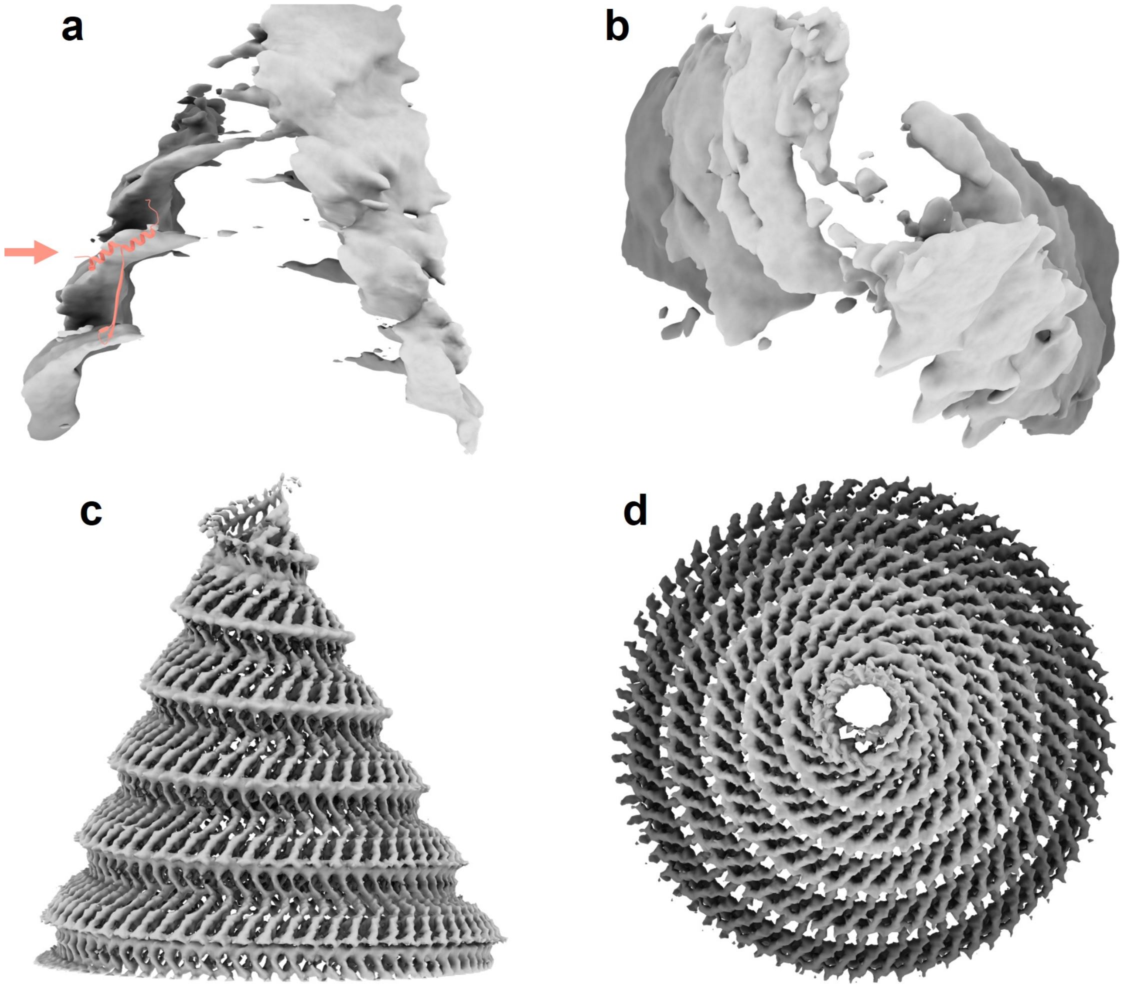


Figure 3.21: **Tip volume reconstruction (a)** Output volume of 3D reconstruction (Ab inito) from tip data set (Table 3.1) with a GvpA2 monomer fitted to the structure (pink arrow). **(b)** Volume of (a) rotated by 90 ° in x. **(c)** Simulated 3D density from the proposed pseudo-atomic model by Huber et al.[5] **(d)** Model of (a) rotated by 90 ° in x. Volumes and densities are rendered in ChimeraX 1.4.[30]

Tip volume reconstruction

We find that using the particle picks of the 2D class average from extracted particles with a cutout outbox size of 256 pixels (388.61 Å) results in a structure that most closely resembles the previously proposed models (Figure 3.16). The structure is constructed using an "Ab-initio" CryoSPARC job that tries to reconstruct a volume from scratch given C1 symmetry (no assumed symmetry). We are not able to resolve a structure that maintains the symmetry and resolution as the proposed pseudo-atomic model from Huber et al, see Figure 3.22.[5] The found model in Figure 3.22 a resembles the pseudo-atomic model in Figure 3.22 c from only one direction, we visually validate this aligning a GvpA monomer to the angled ridge of the volume. Looking at a 90 ° rotation of this volume in Figure 3.22 b we see that this does not match the proposed volume as can be seen in Figure 3.22 d. We propose that the CryoSPARC "Ab Inito" job lands in a local optimum that only conforms to the proposed structure in one dimension and that the shifted cone halves as visible in panel b represent an erroneous local optimum. We propose that a correct tip shape reconstruction is difficult as a consequence of the GVs never being imaged in an upward position in the ice, which is due to their length being significantly longer than the ice thickness leading to a preferred side ward orientation. It is therefore that the upward orientation is not well represented in our data set, which leads the SGD procedure of CryoSPARC to fail to correctly optimize for the volume from this direction.[12]

To circumvent this issue we performed volume refinement with a low-pass filtered (30 Å) version of panel d as a reference, which is from the pseudo-atomic model from Huber et al.[5] This led to a general structure that we would expect for 90 ° rotation view. It did not however result in an immediate improvement in resolution, which means that at that point the map does not provide meaningful insights. Given that the found preliminary volume does seem to contain structural information matching with previous expectations we propose that with additional refinement and heuristic tweaking obtaining an improved structural tip model appears feasible. However, due to time constraints, this was outside the scope of this work.

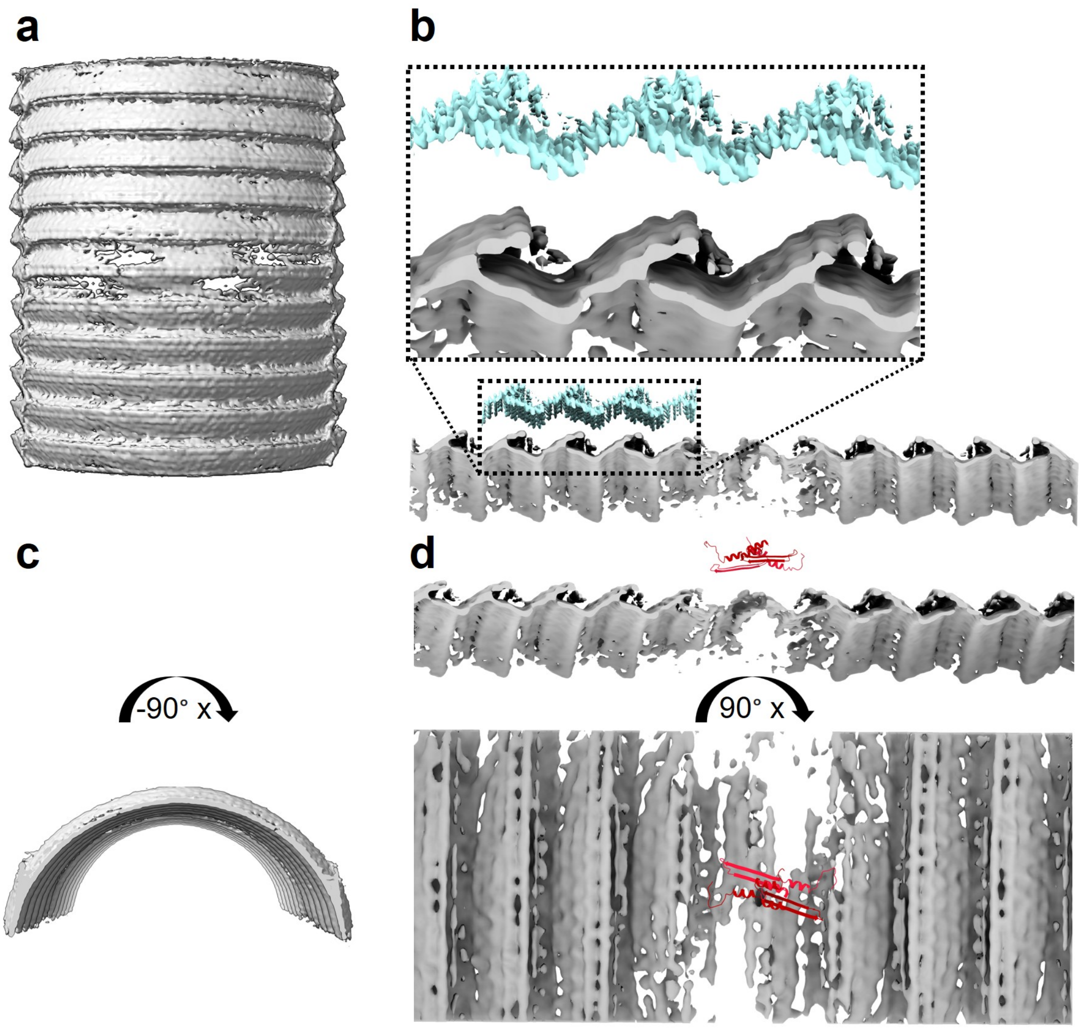


Figure 3.22: **Seam volume reconstruction (a)** GV tube reconstruction around seam points of Table 3.1 after applying a non-uniform refinement CryoSPARC job with an initial helical model of 500 Å (outer: 510 Å, inner: 490 Å). **(b)** Rotated (-120 ° x, 90 ° y) masked volume from (a) compared to seam volume of Huber et al.[5] **(c)** Volume from (a) rotated by -90 ° in x. **(d)** Model from (b) compared to PRP of the pseudo-atomic model from Huber et al.[5] Volumes and models are rendered in ChimeraX 1.4.
30

Seam volume reconstruction

We continue with the particle picks of the highest resolution 2D class average (class 3,4,5 and 6) from the 2D class steps with a cutout outbox size of 625 pixels (Figure 3.19) and apply a helical refinement job in CryoSPARC with an initial helical volume with a width of 474 Å (outer: 480 Å, inner: 468 Å), as described in Section 2.4. These initial volume values we find by measuring the width of the largest class using ImageJ.

Interestingly, the resulting mask consists of half a GV tube in which we can clearly resolve the rib structure (Figure 3.22 c). We propose that during helical refinement the algorithm is only able to clearly resolve to half of the GV tube for most of the particles. This is a consequence of the particles being from 4 classes with distinct GV widths, which means that a full helical model would never be able to fit all the particles. Half a helical tube however is able to accommodate a larger range of helix widths. In essence, this convergence of half a mask is similar as would be the next course of action when fitting on multiple width classes of particles, that is to use half a helix mapping as the initial model to fit to.

Looking at the seam containing GV helix reconstruction in Figure 3.22 b we compare it to the found densities from helical reconstruction off non-seam containing GV picks that were obtained by conventional helical reconstruction (blue densities).[5] We can confirm that our model converges to the correct global structure when comparing the angle and size of the repeating GV ridge pattern.

Additionally, when we observe the seam containing GV helix reconstruction (Figure 3.22 a and d) we find that density is missing around the proposed location of the seam. We compare the location of this missing density to an extracted PRP of two monomers from the pseudo-atomic model of Huber et al (red monomers).[5] We find the location of the densities approximately lines up with the proposed symmetrical axis that is the result of the two meeting GV halves, supporting the proposed model containing a PRP. We argue that the reason density is missing at this point is a consequence of the high variability of the location of the PRP for individual GV particles, which leads to the inability to assign clear densities to these locations. In this way, we argue that the densities also support a stochastic growth model given that when the PRP would always be at the same location we would not expect to observe missing densities to this extent.

In addition, we tried helical reconstruction on individual classes but were not able to reach a resolution where we were able to resolve the rib structure and seam location to a similar extent. We propose that difficulties in reconstruction are in part a consequence of the challenge of aligning 2D particle images in such a way that the PRP is at the correct orientation since it is a weak (heterogeneous) feature of the images. Moreover, mismatches in the alignment of particles could result in additional artifacts in the reconstructed densities.

4

Conclusion

We performed machine-learning-based instance segmentation of gas vesicles, subsequent processing, and filtering on a massive scale. Analyzing tens of thousands of images to target over a million individual gas vesicle (GV) structures. Performing this analysis in a scalable, tractable, and reproducible manner by leveraging workflow packages and high-performance cluster computing capacity. Our resulting high-confidence structural feature set describes 89k whole GVs.

Given the uniqueness of the GV structure and performing cryo-EM single-particle analysis (SPA) using segmentation, the workflow has been custom-built and formulated through a heuristic process of optimization. An improved theoretical foundation could lead to more carefully designed processing procedures to improve processing accuracy and confidence of inference on the obtained feature statistics. As an example currently, the pipeline applies a rudimentary form of k-means clustering to try and filter out erroneous segments. Additionally, the pipeline has issues standardizing small GVs, which leads to their under representation in our data set. This could introduce unforeseen biases that potentially convolute set statistics and lead to the perceived feature relation inconsistencies.

The separability of GV segments clustered on structural features could be probed by relaying the selected versus unselected clusters using statistical distance visualization methods like PCA or t-SNE. The processing of small GVs could be improved by performing Fourier angle analysis on half the GV and using a half-small GV reference. This would negate the problem of opposing GV wall directions making extraction of the rotation angle inconsistent.

Given that we find many instances of seam-center offsets and a significant non-zero mean seam-center offset, we propose that a simultaneous top and bottom insertion model appears highly unlikely. Additionally, given a simultaneous insertion model we would expect less to no variability in the center-seam offset. While in fact, we see more variability than a stochastic single-insertion model would predict.

More specifically, we propose that the higher-than-expected mean seam-center offset variability might either result from erroneous measurements or introduced biases during processing. Alternatively, it might hint towards a mechanism where insertion is not following equal probabilities or that the GV halves do not start growing together from equal sizes. Further even, subsequent insertion could have a certain degree of serial correlation. This we argue could be a consequence of subsequent insertion to a single GV halve being more energetically favorable. Another possibility is that differences in GV halve widths might lead to unequal insertion probabilities that result in our observed significant seam-center offset variability. Further research could probe this question by measuring width distances between halves and trying and relate those to differences.

Using a high-confidence structural feature set, we find evidence that supports previously posed structural properties and characteristics of GVs. We identify the presence of α_1 side chains in GV tips, which points to the validity of the proposed pseudo atomic model of Huber et al.[5] Additionally, we find strong evidence that the position of the so-called polarity reversal point's (PRP) is subject to drift from the geometric center of the GV. The found distributions of the seam-center offset and its corresponding monomer drift represent a higher variability than expected from an equal probability insertion model. However, given this high variability, we argue that this is still evidence supporting the hypothesis that GVs self-assemble by insertion of individual GvpA monomers at the PRP in a stochastic fashion as opposed to the alternative model of pairs of monomers being inserted synchronously on both halves.

5

Outlook

Context-preserving segmentation and processing of cryo-EM images in a single-particle-analysis (SPA) of GVs, as done in this work, can serve as a proof of concept of how improved computational capabilities (accurate machine-learning-based segmentation and parallel processing) enable targeted extraction of information from the high imaging throughput that cryo-EM methods can produce. All the while preserving contextual information important for sound inference. Substituting this kind of workflow in conventional SPA pipelines has the potential to open up dynamical analysis avenues of macromolecules.

Structural reconstruction
Increased accuracy and robustness of particle picking through training of customized machine-learning-based networks like Cellpose offers avenues to obtain new order of magnitudes of particles to be used in conventional cryo-EM SPA workflows. Additionally, more fine-tuned segmentation models might allow particle picking of previously unfeasible to segment structures. In the case of GVs, we are able to obtain a significant number of particles that form high-resolution 2D class averages of structure components of interest (seam and tips). Future research could use this expanded data set size to obtain higher-resolution 3D modeling structures of previously difficult-to-map areas.

Feature statisticss
Being able to flexibly segment and process many different types of GVs enables the collection of expansive feature sets spanning most of the realizations of GV types. This in turn facilitates the probing of feature relations that can offer inference on structural and dynamical properties.

Fast processing of homologs
Given the ease of reanalyzing new data sets once a custom model and processing pipeline have been built, analyzing adapted data sets and obtaining feature statistics could help with genetic-to-protein function or structural assays. For the case of GVs, creating slight GV gene homologs and testing for differences in feature statistics offers a way of better understanding how these genes influence biogenesis.

Method generalisation
Performing inference by obtaining statistics and providing custom particle pick positions for future downstream analysis is now enabled by readily implementable DL-based segmentation solutions and workflow management tool kits. This kind of workflow could be applied to other structure-resolving problems where a high throughput of data of highly heterogeneous structures is being produced.

We expect that with more accessible and readily implementable (parallel) workflow packages a wide scope of new research avenues will appear in regions of science where there is currently an under capacity to scalably and reproducibly evaluate the data being obtained.

6 Acknowledgements

For the duration of the last ten months I have been delighted to be part of the Electron Nanoscopy Lab, headed by Arjen Jakobi. I would like to thank Arjen for the good advice, encouraging guidance, and trust he allotted me to work independently. I greatly enjoyed my time at the lab and am very grateful for the welcoming and friendly atmosphere, inviting me to join in all regards of being part of a lab - from celebrating Sinterklaas, going out for dinner, or going on lab retreats.

In particular, I would like to thank Maarten and Stefan for supervising me in such an enthusiastic and engaging manner, helping me in many of the challenging stages of my thesis, and providing me with lots of interesting and fun discussions. I consider myself very lucky to have had such a nice duo of daily supervisors. I would also like to say thanks to the rest of the group: Alok, Cecilia, Clémence, Lennart, Natasha, Tanja, and Wiel for the engaging conversations over lunch.

F.B. Jansen
Delft, July 2023

Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular Biology of the Cell. 4th edition. Garland Science, 2002.

[2] Concepts of Biology – 1st Canadian Edition. BCcampus Open Publishing, 2023. Accessed: 2023-06-23.

[3] AlphaFold Team. Highly accurate protein structure prediction with alphafold. Nature, 2020. Accessed: 2023-06-23.

[4] A.E. Blaurock and A.E. Walsby. Crystalline structure of the gas vesicle wall from anabaena flos-aquae. Journal of Molecular Biology, 105(2):183–199, 1976.

[5] Stefan T. Huber, Dion Terwiel, Wiel H. Evers, David Maresca, and Arjen J. Jakobi. Cryo-em structure of gas vesicles for buoyancy-controlled motility. Cell, 186(5):975–986.e13, 2023.

[6] Felicitas Pfeifer. Distribution, formation and regulation of gas vesicles. Nature Reviews Microbiology, 10:705–715, 2012.

[7] P.K. Hayes and A.E. Walsby. The inverse correlation between width and strength of gas vesicles in cyanobacteria. British Phycological Journal, 21(2):191–197, 1986.

[8] M. Shapiro, P. Goodwill, and A. et al. Neogy. Biogenic gas nanostructures as ultrasonic molecular reporters. Nature Nanotech, 9(2):311–316, 2014.

[9] T.J. McMaster, M.J. Miles, and A.E. Walsby. Direct observation of protein secondary structure in gas vesicles by atomic force microscopy. Biophysical Journal, 70(5):2432–2436, 1996.

[10] Hussein M. Ezzeldin, Jeffery B. Klauda, and Santiago D. Solares. Modeling of the major gas vesicle protein, gvpa: From protein sequence to vesicle wall structure. Journal of Structural Biology, 179(1):18–28, 2012.

[11] Yifan Cheng, Nikolaus Grigorieff, Pawel Penczek, and Thomas Walz. A primer to single-particle cryo-electron microscopy. Cell, 161:438–449, 04 2015.

[12] Ali Punjani, John Rubinstein, David Fleet, and et al. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. Nature Methods, 14(3):290–296, 2017.

[13] Ewen Callaway. The revolution will not be crystallized: a new method sweeps through structural biology. Nature, September 2015.

[14] Werner Kühlbrandt. The resolution revolution. Science, 343(6178):1443–1444, 2014.

[15] Kazuyoshi Murata and Matthias Wolf. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. Biochimica et Biophysica Acta (BBA) - General Subjects, 1862(2):324–334, 2018. Biophysical Exploration of Dynamical Ordering of Biomolecular Systems.

[16] Ewen Callaway. Revolutionary cryo-em is taking over structural biology. Nature, February 2020. The number of protein structures being determined by cryo-electron microscopy is growing at an explosive rate.

[17] EMDB entries statistics. `https://www.ebi.ac.uk/emdb/statistics/emdb_entries_year`. Accessed: 2023-06-10.

[18] Shirin Akbar, Sukanya Mozumder, and Jayati Sengupta. Retrospect and prospect of single particle cryo-electron microscopy: The class of integral membrane proteins as an example. Journal of Chemical Information and Modeling, 60(5):2448–2457, 2020. PMID: 32163280.

[19] Sjors H.W. Scheres. A bayesian view on cryo-em structure determination. Journal of Molecular Biology, 415(2):406–418, 2012.

[20] A. Punjani, J.L. Rubinstein, D.J. Fleet, and M.A. Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. Nature Methods, 14:290–296, 2017.

[21] Carsen Stringer, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. Nat Methods, 18:100–106, 2021.

[22] J. Chai, H. Zeng, A. Li, and E. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. 6:100134, 2021.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90, 2017.

[24] Carsen Stringer and Marius Pachitariu. Cellpose 2.0: how to train your own model. Nat Methods, 19:1634–1641, 2022.

[25] John P Lewis. Fast template matching. In Vision interface, volume 95, pages 15–19. Quebec City, QC, Canada, 1995.

[26] F Mölder, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, V Sochat, J Forster, S Lee, SO Twardziok, A Kanitz, A Wilm, M Holtgrewe, S Rahmann, S Nahnsen, and J Köster. Sustainable data analysis with snakemake [version 1; peer review: 1 approved, 1 approved with reservations]. F1000Research, 10(33), 2021.

[27] Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 1). `https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1`, 2022.

[28] CryoSPARC Guide. Job: Helical Refinement. `https://guide.cryosparc.com/processing-data/all-job-types-in-cryosparc/helical-reconstruction-beta/job-helical-refinement-beta`, 2023. [Online; accessed 24-June-2023].

[29] Ali Punjani, Hao Zhang, and David J. Fleet. Non-uniform refinement: adaptive regularization improves single-particle cryo-em reconstruction. Nature Methods, 17(12):1214–1221, 2020.

[30] T.D. Goddard, C.C. Huang, E.C. Meng, E.F. Pettersen, G.S. Couch, J.H. Morris, and T.E. Ferrin. Ucsf chimerax: Meeting modern challenges in visualization and analysis. Protein Sci., 27(1):14–25, 2018.

A

Methodology Supplementary

A.1. Fourier angle method

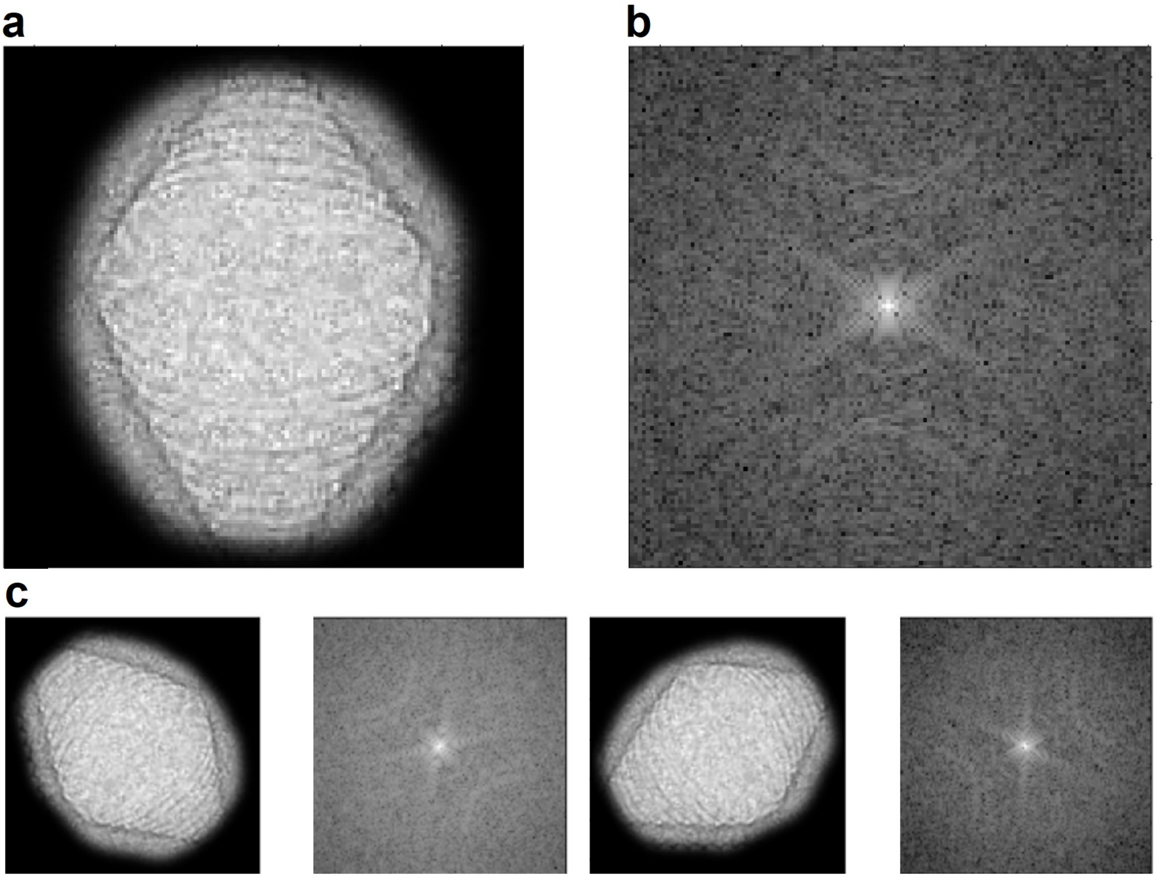


Figure A.1: **Reference stack Fourier method small GV rotation angle a** Micrograph cutout of chosen small reference GV at 0° rotation. **b** Powerspectrum from applying an FFT to (a). Showing laylines at angel parallel of GV shell direction. Sample specific contrast information at center. **c** Micrograph cutouts and corresponding power spectra for 45° and 120° rotation

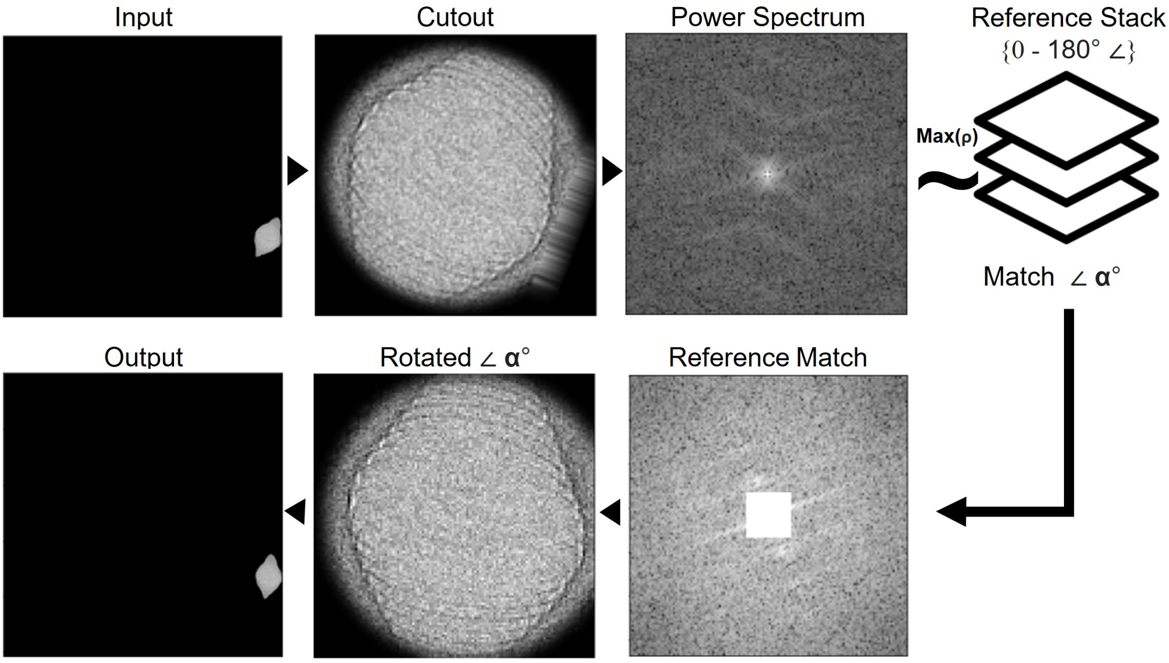


Figure A.2: **Workflow Fourier method small GV rotation angle** Starting from the input micrograph GV cutout; subsequent cutout around the centre, FFT of cutout for power spectrum cutout, multiplication of cutout power spectrum with small GV reference stacks constituents, max sum reference constituent represents best correlation and corresponding index identifies found rotation angle α°, reference match shows power spectrum from reference stack that best correlates with input. Rotated α° shows the rotated cutout. Output is GV rotated by α°.

B

Results Supplementary

B.1. Processing results

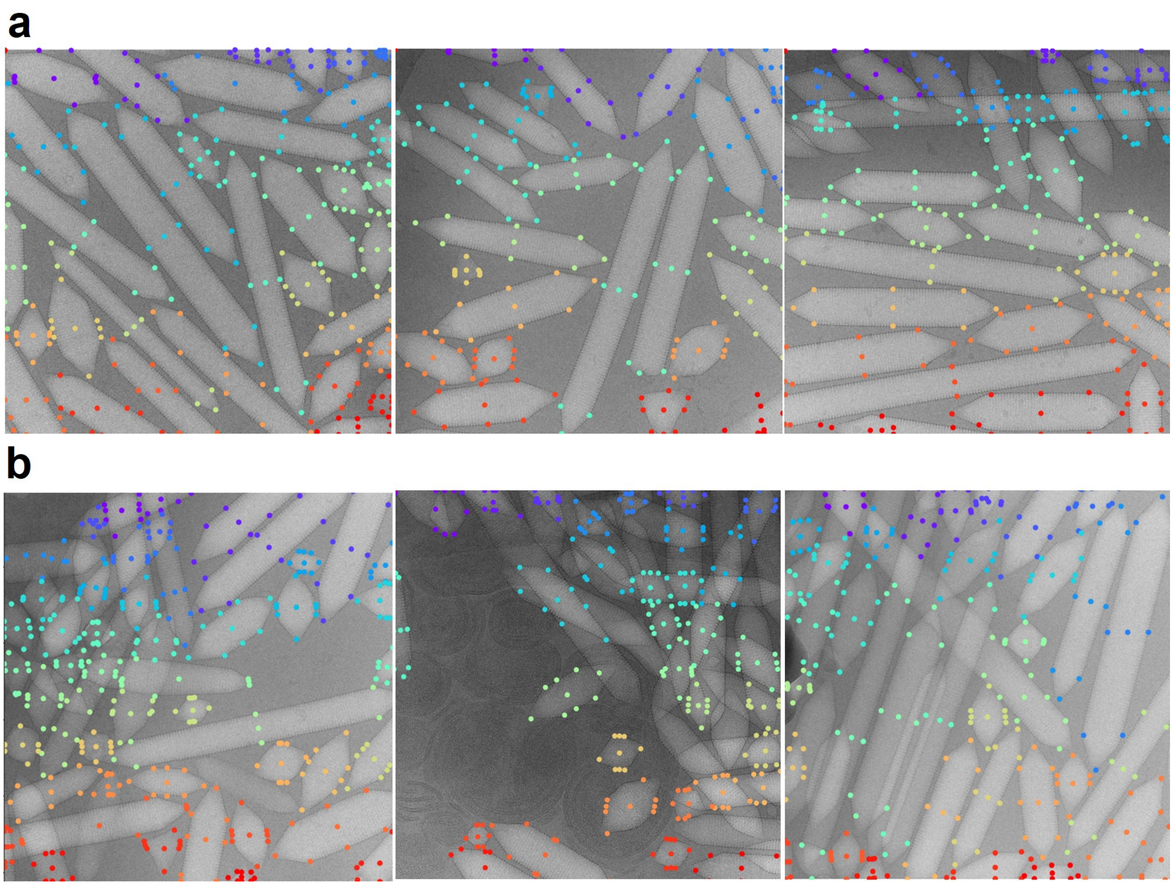


Figure B.1: **Example preliminary points on images (a)** Example set of preliminary points from well-segmented images **(b)** Example set of preliminary points from harder-to-segment images

B.2. Clustering results

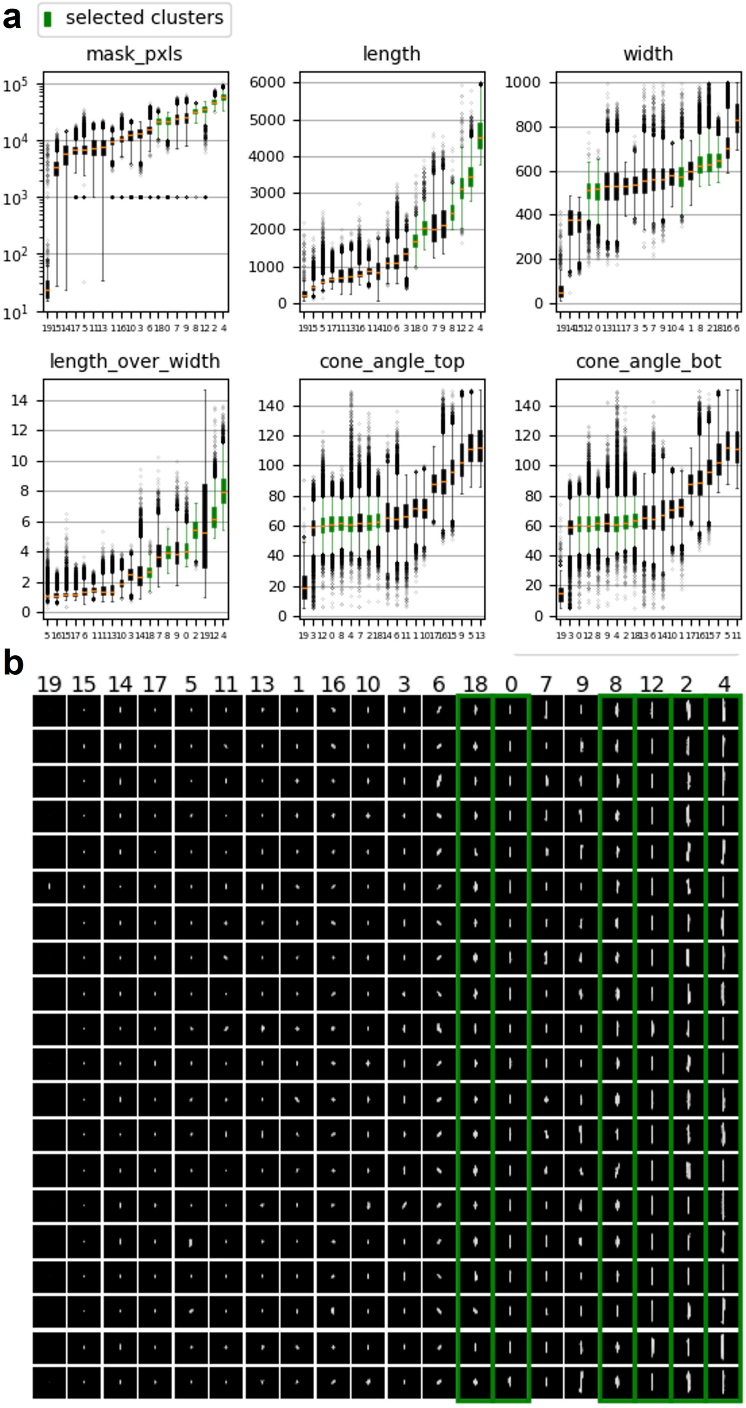


Figure B.2: **Selected k means GV clusters (a)** Boxplots of k-means clustering feature statistics showing selected (green) and unselected clusters (black). **(b)** Visualisation of a random sample of 20 standardised extracted GVs per cluster (Section 2.1.2). Showing selected (green) and unselected clusters. Clusters have been sorted from lowest to highest average mask size in pixels (left to right).

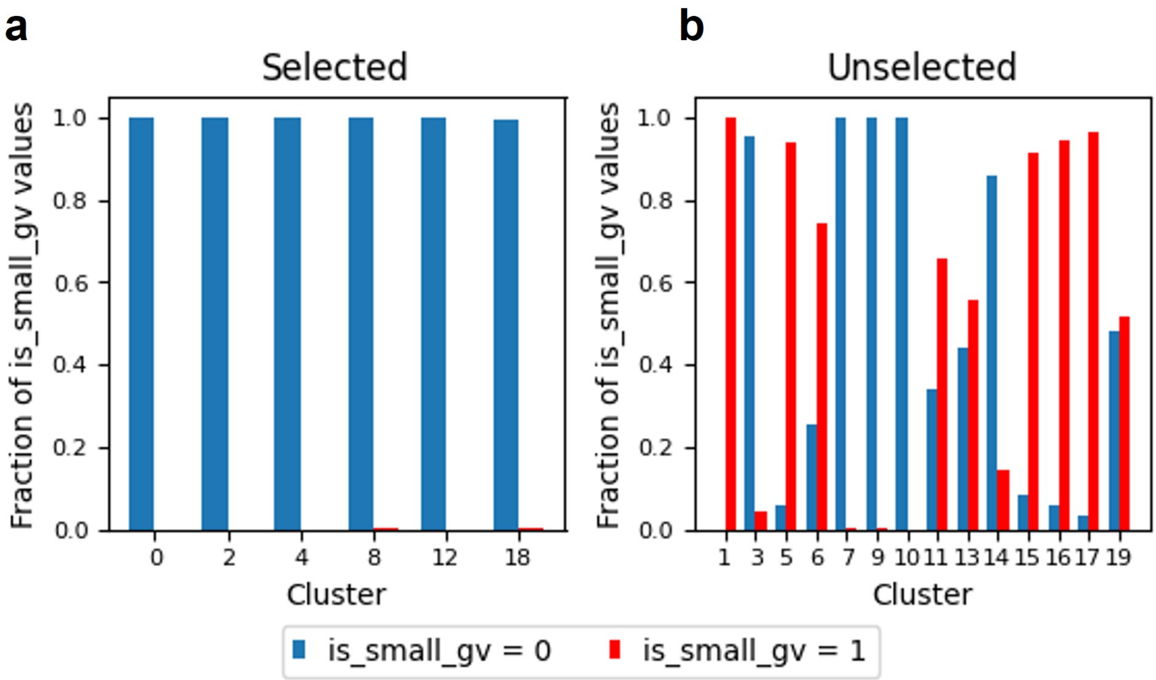


Figure B.3: **Small GV selection in clustering** Plotting the proportion of segmented GVs per cluster. **(a)** Proportion of is_small_gv segments in the selected clusters. **(b)** Same as in (a) but for unselected clusters.

B.3. Template matching results

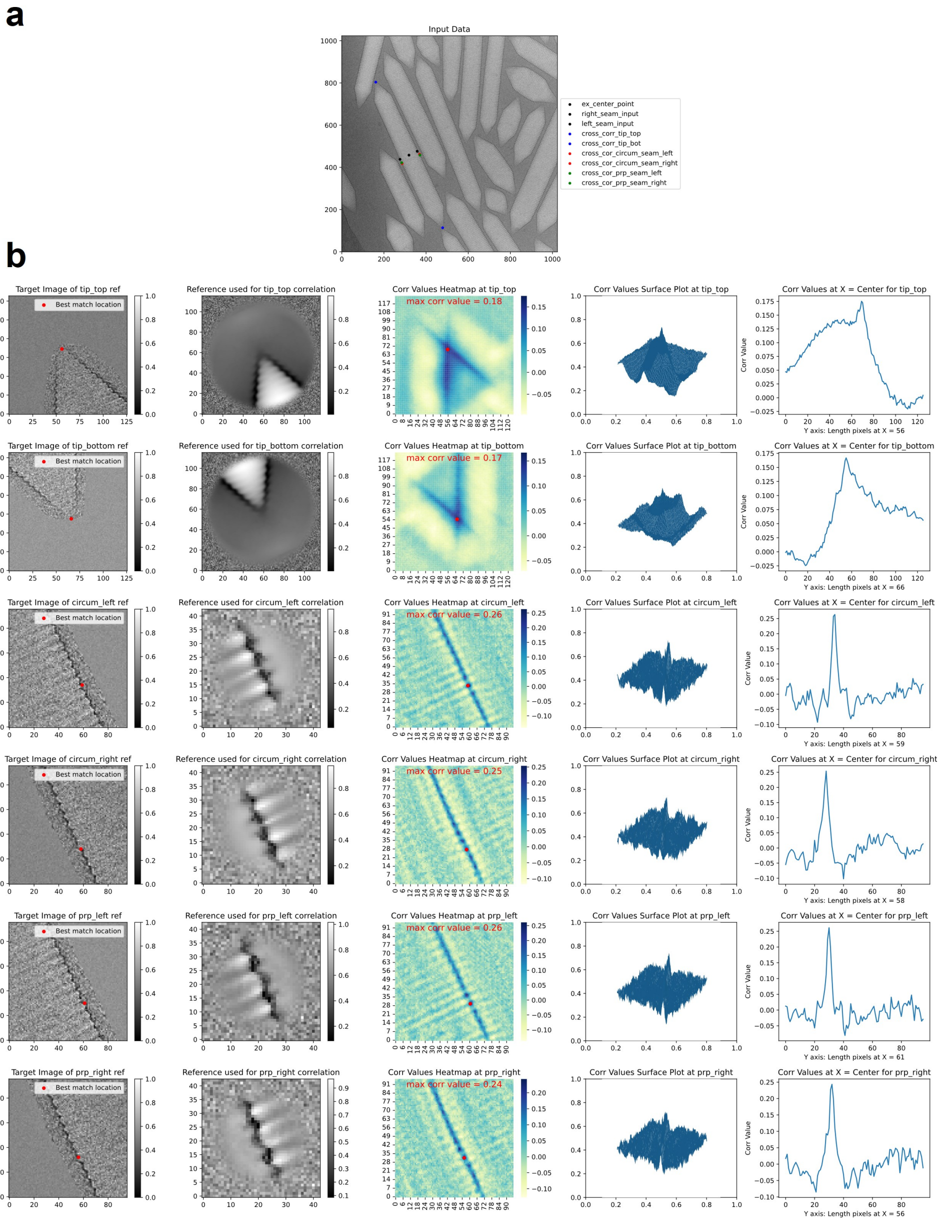


Figure B.4: **GV template match output (a)**Example output points from template match flows drawn on the input image. Shows the top tip (blue dot), bot tip (blue dot), right seam (red dot), left seam (red dot), right PRP (green dot), and left PRP (green dot) targets. **(b)** Example output from template match flows of the top tip, bot tip, right seam, left seam, right PRP, and left PRP targets (top to bottom).

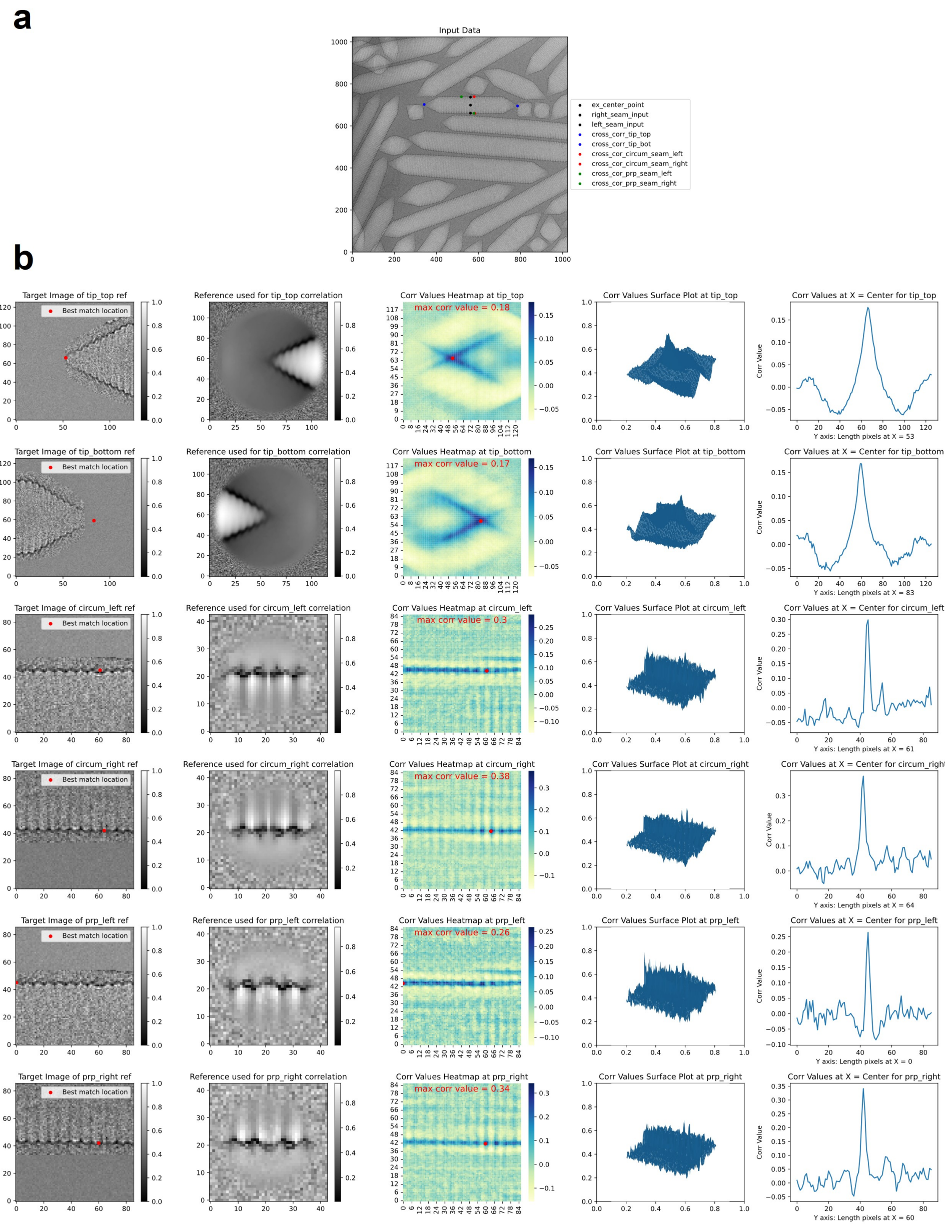


Figure B.5: **GV template match output (a)**Example output points from template match flows drawn on the input image. Shows the top tip (blue dot), bot tip (blue dot), right seam (red dot), left seam (red dot), right PRP (green dot), and left PRP (green dot) targets. **(b)** Example output from template match flows of the top tip, bot tip, right seam, left seam, right PRP, and left PRP targets (top to bottom).

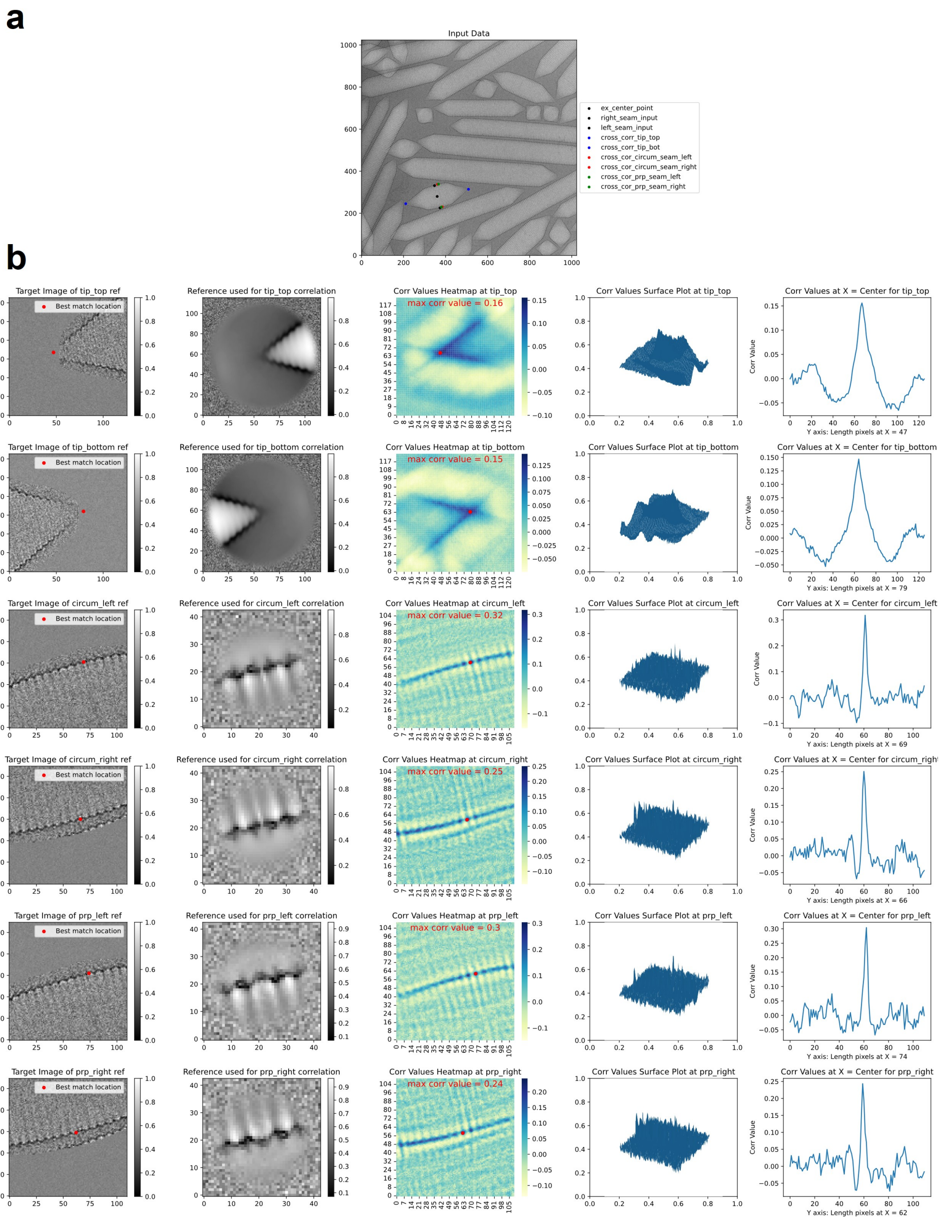


Figure B.6: **GV template match output (a)**Example output points from template match flows drawn on the input image. Shows the top tip (blue dot), bot tip (blue dot), right seam (red dot), left seam (red dot), right PRP (green dot), and left PRP (green dot) targets. **(b)** Example output from template match flows of the top tip, bot tip, right seam, left seam, right PRP, and left PRP targets (top to bottom).

B.4. Statistical results

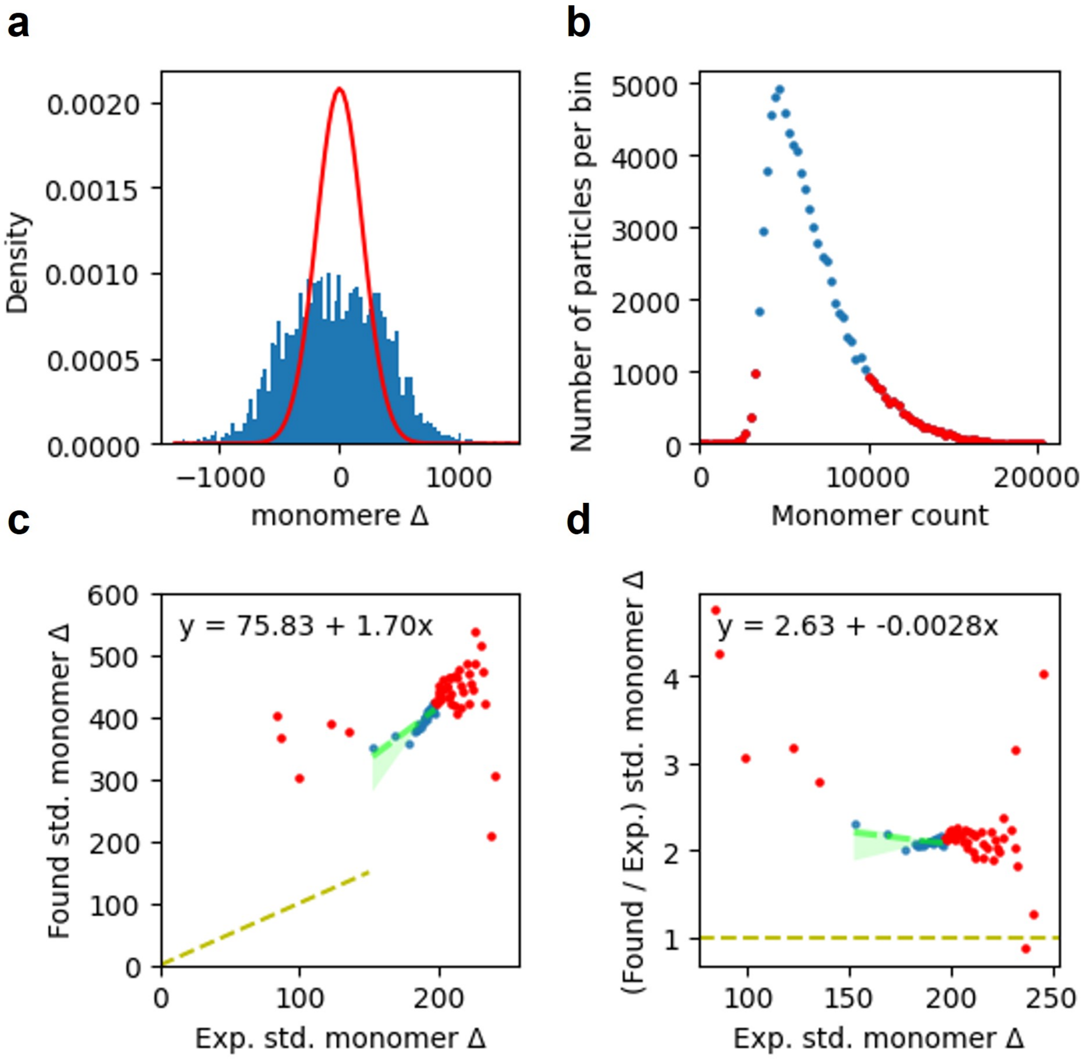


Figure B.7: **Estimation error adjusted Expected std. vs. found std.** $M\Delta$ **(a)** Histogram (blue) of the monomer difference ($M\Delta$) (Equation 2.4) for GVs with a monomer count (N) between 6000 and 6250. Overlaid is the pdf of a normal distribution with an additional width-varying estimation error for the proposed monomer difference in Section 2.3.1 ($N[0, \sqrt{N+N_{error}}]$ **(b)** Scatter plot showing the number of particles per bin for equally spaced bins ($250M$) ranging from 0 to 20000 M. **(c)** Scatter plot showing the estimation error adjusted expected std. versus the found std. of $M\Delta$ of the bins of (b). **(d)** Same measure as in (c) but normalised for the Expected std. $M\Delta$. Red dots indicate bins with less than 1000 GVs.

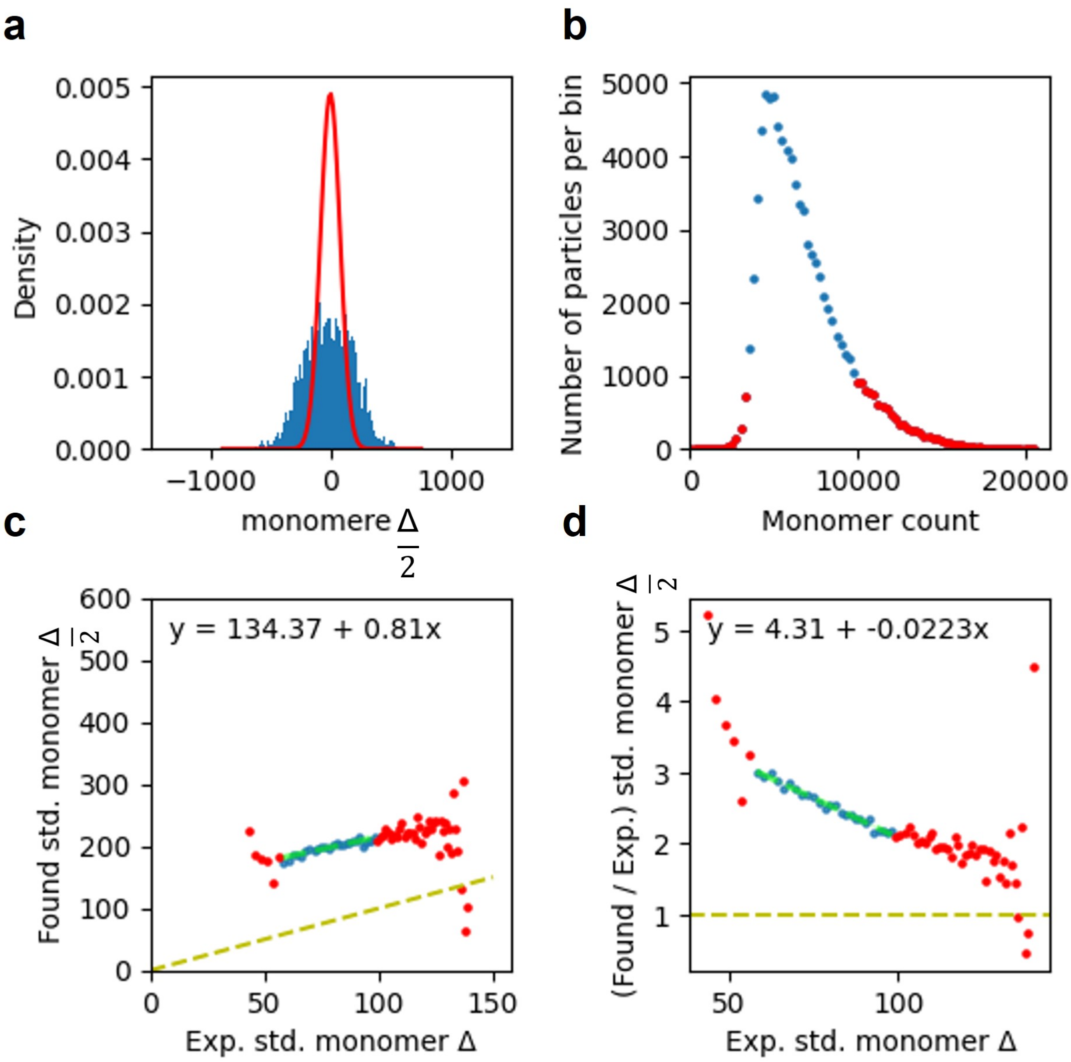


Figure B.8: **Expected std.** $M\Delta$ **vs. found std.** $\frac{M\Delta}{2}$ **(a)** Histogram (blue) of half the monomer difference ($\frac{M\Delta}{2}$) (Equation 2.4) for GVs with a monomer count (N) between 6000 and 6250. Overlaid is the pdf of a normal distribution with the proposed monomer difference in Section 2.3.1 ($N[0, \sqrt{N}]$ **(b)** Scatter plot showing the number of particles per bin for equally spaced bins (250M) ranging from 0 to 20000 M. **(c)** Scatter plot showing the expected std. $M\Delta$ versus the found std. $\frac{M\Delta}{2}$ of the bins of (b). **(d)** Same measure as in (c) but normalised for the Expected std. $\frac{M\Delta}{2}$. Red dots indicate bins with less than 1000 GVs.

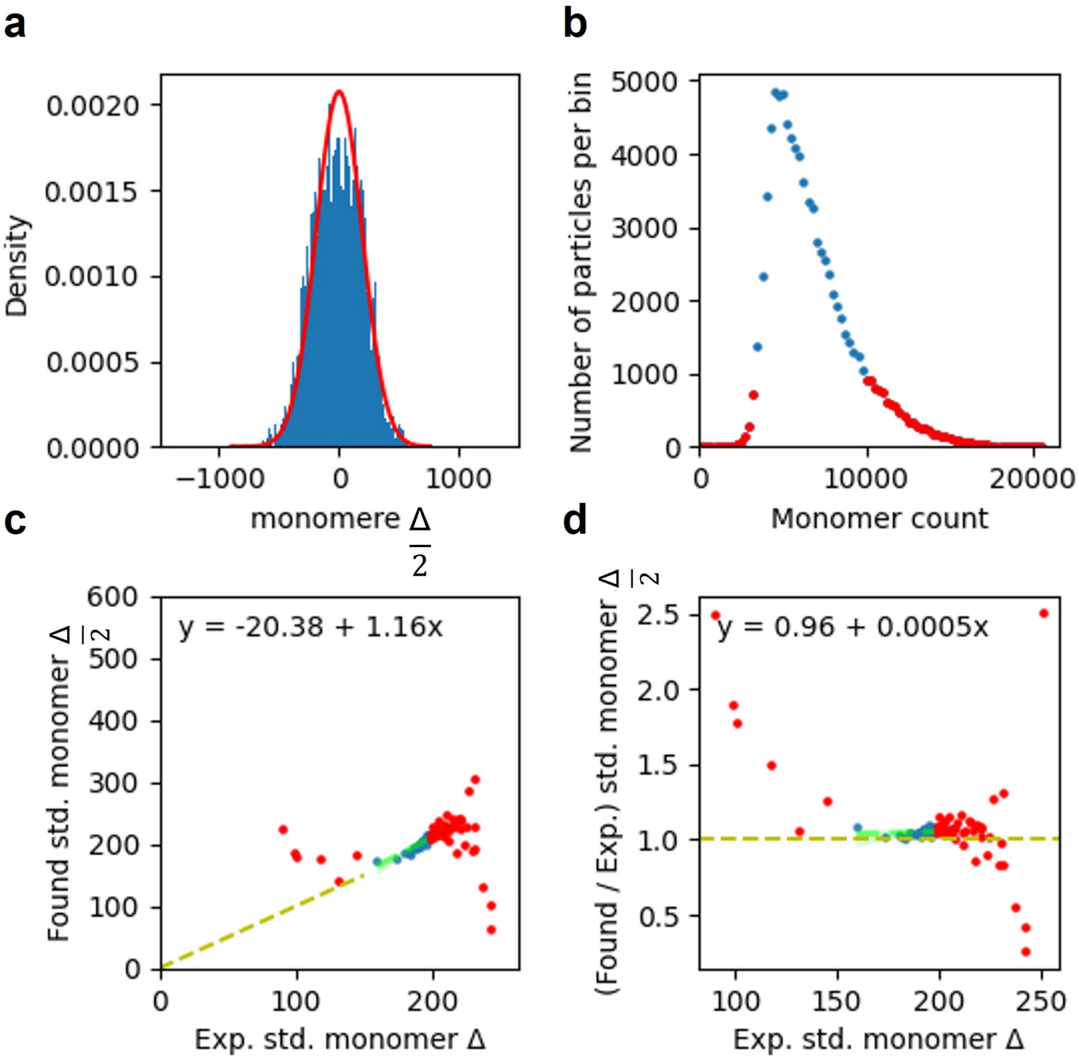


Figure B.9: **Estimation error adjusted expected std.** $M\Delta$ **vs. found std.** $\frac{M\Delta}{2}$ **(a)** Histogram (blue) of half the monomer difference ($\frac{M\Delta}{2}$) (Equation 2.4) for GVs with a monomer count (N) between 6000 and 6250. Overlaid is the pdf of a normal distribution with an additional width-varying estimation error for the proposed monomer difference in Section 2.3.1 ($N[0, \sqrt{N + N_{error}}]$ **(b)** Scatter plot showing the number of particles per bin for equally spaced bins (250M) ranging from 0 to 20000 M. **(c)** Scatter plot showing the estimation-error adjusted expected std. $M\Delta$ versus the found std. $\frac{M\Delta}{2}$ of the bins of (b) . **(d)** Same measure as in (c) but normalised for the Expected std. $M\Delta$. Red dots indicate bins with less than 1000 GVs.

We perform a linear regression on the found relation we find an intercept of 146.24 and a slope of 1.41, which also points to a positive linear relation of the standard deviations but does not fit expectations from the binomial model (Figure B.10 a). The normalized relation results in a linear regression fit with an intercept of 8.85 and a slope of -0.09. Like the fit to the normalized random walk model expectations, we find that the error with the expected standard deviation appears to decrease with an increasing number of N. The found relation using P (binomial model) closely resembles the one found by comparing $M\Delta$, which is as was proposed in Section 2.3.1.

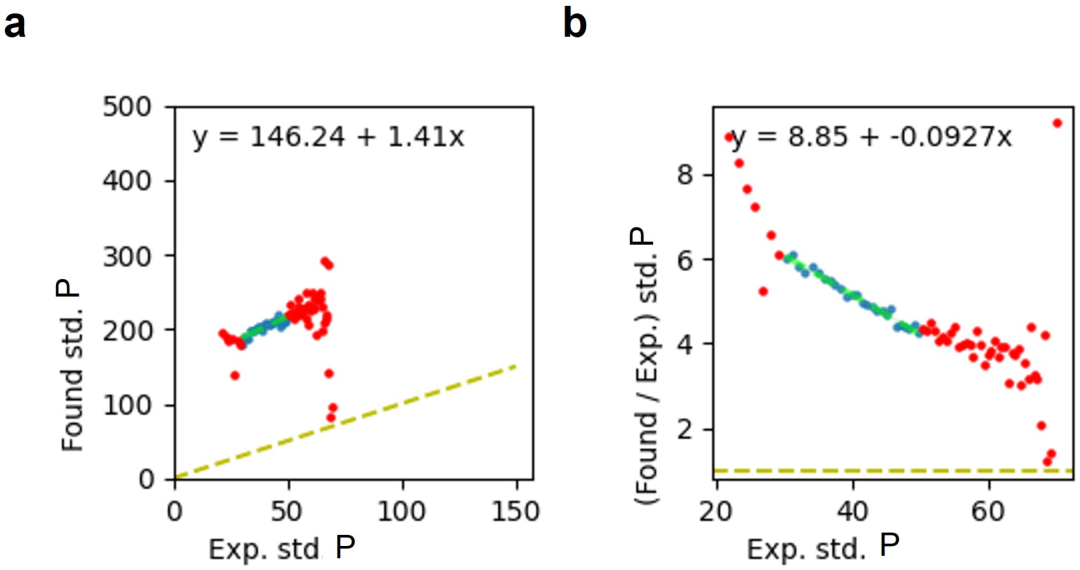


Figure B.10: **Binomial expected std. vs. found std.** P **(a)** Scatter plot showing the expected std. P under the binomial model as proposed in Section 2.3.1 versus the found std. P of the equally spaced bins (250N). **(b)** Same measure as in (a) but normalized for the expected std. P. Red dots indicate bins with less than 1000 GVs.