Investigating Carboxysome Morphology Dynamics with a Rotationally Invariant Variational Autoencoder

Miguel Fuentes-Cabrera,[*,†] Jonathan K. Sakkos,[‡] Daniel C. Ducat,[‡,¶] and Maxim Ziatdinov[†]

†*Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA*

‡*Plant Research Laboratory, Michigan State University, East Lansing, MI, USA*

¶*Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA*

E-mail: fuentescabma@ornl.gov

Abstract

Carboxysomes are a class of bacterial microcompartments that form proteinaceous organelles within the cytoplasm of cyanobacteria and play a central role in photosynthetic metabolism by defining a cellular microenvironment permissive to CO_2 fixation. Critical aspects of the assembly of the carboxysomes remain relatively unknown, especially with regard to the dynamics of this microcompartment. Progress in understanding of carboxysome dynamics is impeded in part because analysis of the subtle changes in carboxysome morphology with microscopy remains a low-throughput and subjective process. Here we use deep learning techniques, specifically a Rotationally Invariant Variational Autoencoder (rVAE), to analyze fluorescence microscopy images of cyanobacteria bearing a carboxysome reporter and quantitatively evaluate how carboxysome shell remodelling impacts subtle trends in the morphology of the microcompartment over time. Towards this goal, we use a recently developed tool to control endogenous protein levels, including carboxysomal components, in the model cyanobacterium *Synechococcous elongatus* PCC 7942. By utilizing this system, proteins that compose the carboxysome can be tuned in real-time as a method to examine carboxysome dynamics. We find that rVAEs are able to assist in the quantitative evaluation of changes in carboxysome numbers, shape, and size over time. We propose that rVAEs may be a useful tool to accelerate the analysis of carboxysome assembly and dynamics in response to genetic or environmental perturbation, and may be more generally useful to probe regulatory processes involving a broader array of bacterial microcompartments.

Introduction

Cyanobacteria are prokaryotic autotrophs that are under investigation as an alternative chassis for the solar-driven conversion of CO_2 into useful bioproducts.[1–4] Like other members of the green photosynthetic lineage, carbon fixation is accomplished in cyanobacteria through the enzymatic activity of ribulose-1,5-bisphosphate carboxylase/oxygenase (Ru-

bisco).[5] Among the model cyanobacterial species, *Synechococcus elongatus* PCC 7942 (hereafter *S. elongatus*) has a well-developed genetic toolkit and has been the subject of extensive research on circadian rhythms, carbon metabolism, metabolic engineering, and carboxysome biogenesis.[6–9]

The carboxysome is a proteinaceous bacterial microcompartment that exists within the cytosol of cyanobacteria, encapsulates a phase-separated pool of Rubisco, and creates a micro-environment favorable to the carboxylation reaction.[10,11] Despite the carboxysome's central role in cyanobacterial metabolism, a complete picture of its biogenesis and remodeling remains elusive, though over the years several key studies have provided insights.[6,12,13] One open question within the field regards the degree to which the carboxysome is in dynamic exchange with cytosolic components, and if this microcompartment can be reconfigured once assembled. While it is well-documented that carboxysome size, number, distribution, and shell composition are modulated under different environments,[14–18] it is unclear if the observed restructuring is only true for newly-synthesized carboxysomes, or if pre-existing carboxyomes are sufficiently dynamic to be remodeled in response to changing conditions. For example, some evidence suggests that once carboxysomes are formed, they are static until they are ultimately degraded as a unit.[13] Several tools and reporter constructs have been developed in order to track the dynamics of carboxysome positioning and morphological features in living cells,[6,12,13,19,20] yet the variability of carboxysome features in natural populations and resolution limits of fluorescence microscopy complicate quantitative evaluation of changes in carboxysome size, composition, and intracellular positioning.

Towards a better understanding of carboxysome biogenesis, we sought to develop a quantitative method for investigating the temporal dynamics of carboxysome remodelling. In recent work,[20] we used a method for protein down regulation based on the Lon protease from *Mesoplasma florum* (*mf-lon*), allowing rapid inducible degradation of proteins of interest. This approach offers the advantage that carboxysome components can be specifically targeted in a manner that causes dynamic rearrangement of carboxysome morphology begin-

ning at an experimentally defined time point. We previously targeted the carboxysome shell protein CcmO: carboxysomes deficient in CcmO form a distinct phenotype related to their inability to close completely.[6] Using fluorescence microscopy, we previously showed that *mf-lon* degraded CcmO and led to carboxysome remodelling within 24-72 hours of activation. Here we describe a workflow that leverages the use of deep learning techniques to segment and analyze the fluorescence microscopy images and quantitatively evaluate trends in carboxysome degradation. The workflow is schematically represented in Fig. 1): it segments the images using Cellpose[21] and analyzes the resultant objects (individual carboxysomes, the set of carboxysomes per cell, and the cell themselves) using a Rotationally Invariant Variational Autoencoder, rVAE. We find that application of these deep learning techniques allows for high-throughput analysis of cyanobacterial microscopy data and that significant changes in carboxysome morphology can be confidently detected within hours of the initiation of carboxysome remodelling.

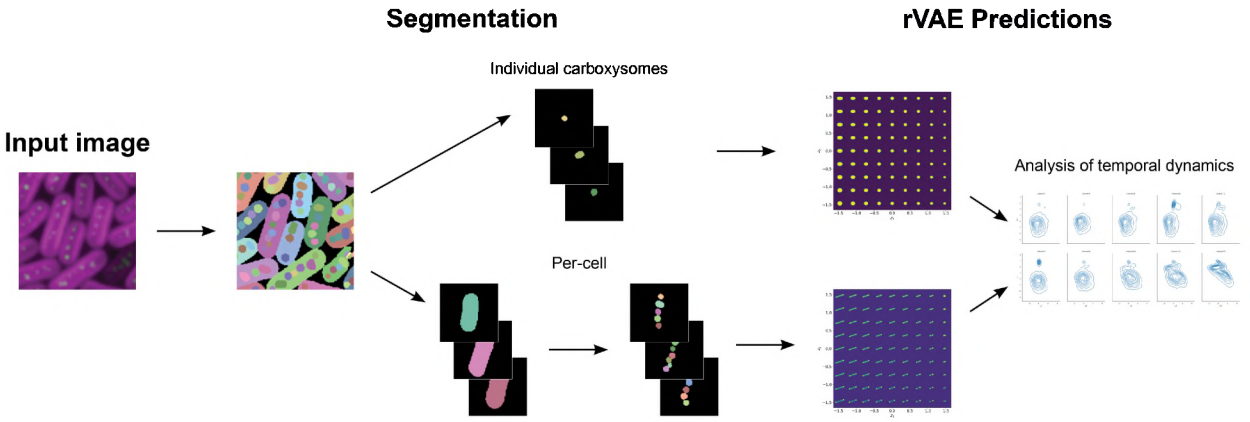


Figure 1: Overview diagram depicting the deep learning workflow in this work. Images are first segmented using Cellpose.[21] The resultant objects are individual carboxysomes, set of carboxysomes per cells, and cells. Changes with time in the shape and size of individual carbyxsomes, and size and position of the carboxysomes in a set, are then analyzed using a rVAE.

Methods

Microbial culturing conditions

S. elongatus cultures were grown in baffled flasks (Corning) with BG-11 medium (Sigma) supplemented with 1 g/L HEPES, pH 8.3, in a Multitron II shaking incubator (Infors HT). Cultures were grown under continuous light with GroLux bulbs (Sylvania) at 125 μ mol photons $m^{-2}s^{-1}$, 2 % CO_2, 32 °C, and 130 rpm shaking. Carboxysomes were degraded by inserting a protein degradation tag (PDT) at the C-terminus of the shell protein CcmO, which is essential for carboxysome closure, and expression of a non-native Lon protease from *Mesoplasma florum.*[20] For visualization of changes in carboxysome morphology, a second copy of the small subunit of Rubisco, rbcS, was tagged with a C-terminal fusion of mNeonGreen (mNG). For the induced samples, cultures were induced with 30 μM theophylline. More detailed culturing, genetic assembly, and transformation information was previously described.[20]

Microscopy

All experiments were performed on live cells in exponential growth. Images were collected with a Zeiss Axio Observer D1 inverted microscope with a Zeiss Plan Apochromat 100× lens. Epifluorescence images were collected of both chlorophyll *a* autofluorescence ($\lambda_{ex} = 545, \lambda_{em} = 605$) and mNG ($\lambda_{ex} = 500, \lambda_{em} = 535$).

Segmentation

A total of 90 images, each with dimensions (1460,1936) pixels, were segmented with Cellpose[21] and analyzed with an rVAE. The set of 90 images is divided into control and induced sample groups, which contain 42 and 39 images, respectively. Both groups contain a time series with images collected at 4, 8, 24, 48, and 72 hours post induction.

Segmentation was performed with Cellpose,[21] and for each image, masks for the carboxysomes and the cells were created. We used a diameter setting of 7 and 20 for generating the masks for carboxysomes and the cells, respectively.

From these masks, two types of sub-image stacks were obtained. One stack contains only individual carboxysomes, whereas the other contains the set of carboxysomes per cell. The latter was generated by using the cell channel mask, within which we selected the set of carboxysomes in each cell. The resultant images were padded to have a size of 115 x 115 pixels. The dimension of each stack for the control and induced groups are given in Table 1.

As it can be seen in Table 1, there are significantly less sub-images in the stacks for the set of carboxysomes than in the stack for individual ones. This is due to the position of the focal plane, which is inclined in some of the samples and causes some group of cells to be visualized different than others. This affects the segmentation, and the end result is that there are fewer cell masks than expected. Because these masks are used as the molds within which we selected the set of carboxysomes sub-images, the stacks for the set of carboxsyomes have less sub-images than the stacks for the individual ones.

rVAE analysis

The rVAE was implemented within the AtomAI Package[22] and, for stacks of images with 1 (3) channels, trained for 1000 (400) epochs using 3 fully connected layers for both the encoder and decoder. Each layer had 128 neurons and was activated by the $tanh()$ function, whose weights were optimized using the Adam optimizer with a learning rate of 0.0001.

The computational cost of training a rVAE was low. Training was performed in Google Collaboratory (Colab) using 1 GPU, and the computational time depended on size of the stack and the number of iterations chosen. For example, training the rVAE for 100 epochs on 327800 images of 40x40 pixels takes about 2 hours and 30 minutes in Colab. Most of the computational time, however, was spent on preparing the images for training. This included preparing the stacks for segmentation, performing segmentation, saving the stacks in Google

Table 1: Dimensions of the sub-images stacks for the control and induced samples. When the stack of images contain both the Chla and mNG channels, the tensor that characterizes the stack has one more dimension. For example, for a stack that contains one channel only, either Chla or mNG, the tensor might have the following dimension (4388, 115, 115, 1). The first number represents the number of images in the stack; the second and the third represent the width and height of each image; the fourth number can be seen as representing the 'color' of each image. This color can be Chla or mNG , however when the stack contains both channels, it has the following dimension (4388, 115, 115, 2).

	Time (hrs)	control	induced
Individual carboxysomes	4	(40619, 40, 40)	(42101, 40, 40)
	8	(53053, 40, 40)	(2756, 40, 40)
	24	(49713, 40, 40)	(41867, 40, 40)
	48	(23255, 40, 40)	(20408, 40, 40)
	72	(40378, 40, 40)	(13650, 40, 40)
Set of carboxysomes	4	(4388, 115, 115)	(4159, 115, 115)
	8	(7492, 115, 115)	(528, 115, 115)
	24	(7665, 115, 115)	(10325, 115, 115)
	48	(2977, 115, 115)	(4875, 115, 115)
	72	(6594, 115, 115)	(3981, 115, 115)

Drive, and reloading them. We have included in the Supporting the notebooks used to train the rVAE. They also contain links to the data used for training, which is directly downloaded from Google Drive.

Results and discussion

Elucidating complex biological interactions via microscopy, particularly those which change over time, is a challenging process. Subtle differences in protein localization, phenotypic variation within control samples, and observational subjectivity often preclude strong conclusions from being made based solely on observation.

In order to visualize dynamic changes in carboxysome morphology, we tagged the small subunit of Rubisco (RbcS) with a fluorescent protein, mNeonGreen (mNG, see Methods, Section). Rubisco is strongly concentrated to the carboxysome lumen in cyanobacteria,

therefore reporter fusions will localize as puncta organized in a characteristic pattern down the midzone of *S. elongatus* cells. We genomically integrated a second copy of the *rbcS* gene fused to mNG and expressed under the native Rubisco promoter. In order to have experimental control over a well-defined feature related to carboxysome morphology, we utilized the recently-described system for inducible down-regulation of the trimeric carboxysome shell protein, CcmO.[20] Briefly, this approach relies upon tagging the endogenous *ccmO* gene with a C-terminal protein translation quality control sequence that is orthogonally recognized by the *Mesoplasma florum* protease, *mf-lon*. By placing *mf-lon* expression under a riboswitch control element that is responsive to theophylline,[23] the targeted protein can be rapidly degraded in *S. elongatus* within minutes to hours of experiment-controlled expression of the exogenous protease. Because CcmO is an important shell protein required for the formation of a completely enclosed microcompartment surface,[24] loss of this protein leads to a well-established phenotype of enlarged, incompletely enclosed, and polar aggregates of Rubisco.[6]

We acquired epifluorescence images of both the cell chlorophyll *a* autofluorescence (Chla) and the carboxysomes (mNG) over the course of 72 hours with an uninduced control set (0 mM theophylline) and cells induced to express *mf-lon* with 30 μM theophylline. Prior to image analysis, segmentation was performed for the cell and carboxysome channels using Cellpose,[21] a recently developed algorithm specifically trained for cellular segmentation. Figures 2a,b show examples of segmentation for the channels containing cells and carboxysomes, respectively. A detailed explanation of the segmentation procedure and the image dataset is given in section *Methods*. From the segmented images, we cropped sub-images containing either individual carboxysomes or all the carboxysomes in a cyanobacterium. Examples of cropped sub-images are shown in Figs. 2c,d. We repeated this procedure for each segmented image, generating stacks of sub-images for the control and induced samples. Table 1 summarizes the sub-images stacks and their dimensions.

To improve analysis of microscopy data, we utilized a rotationally invariant Variational

Autoencoder (rVAE), which is comprised of an encoder and a decoder network. The encoder compresses images into a low-dimensional representation, known as the latent space. The vectors in this space contain the most relevant information of each image. The decoder reconstructs images starting from the latent space. During encoding, it's often desirable to disentangle relevant information from mere rotations and translations. Because we wish to understand how degradation affects the structure of the carboxysome, rotations and translations are irrelevant. A rVAE, unlike a "vanilla" VAE,[25] can disentangle rotations and translation during encoding,[26] and this is the reason why we used an rVAE in our analysis. Specifically, we used the rVAE implementation in Kalinin *et al.*[27] and AtomAI.[22] This rVAE encodes the images into unstructured latent variables (of which we used only two, denoted as $L1$ and $L2$) and the latent variables that encoded the rotational angle (L_θ) and the x-/y-translations ($L_{\Delta x}$ and $L_{\Delta y}$). In the sections below, we present results for $L1$ and $L2$ only. The results for L_θ, $L_{\Delta x}$ and $L_{\Delta y}$ are included in both the Jupyter notebooks that were used to run the analysis, as well as in the Supporting Information.

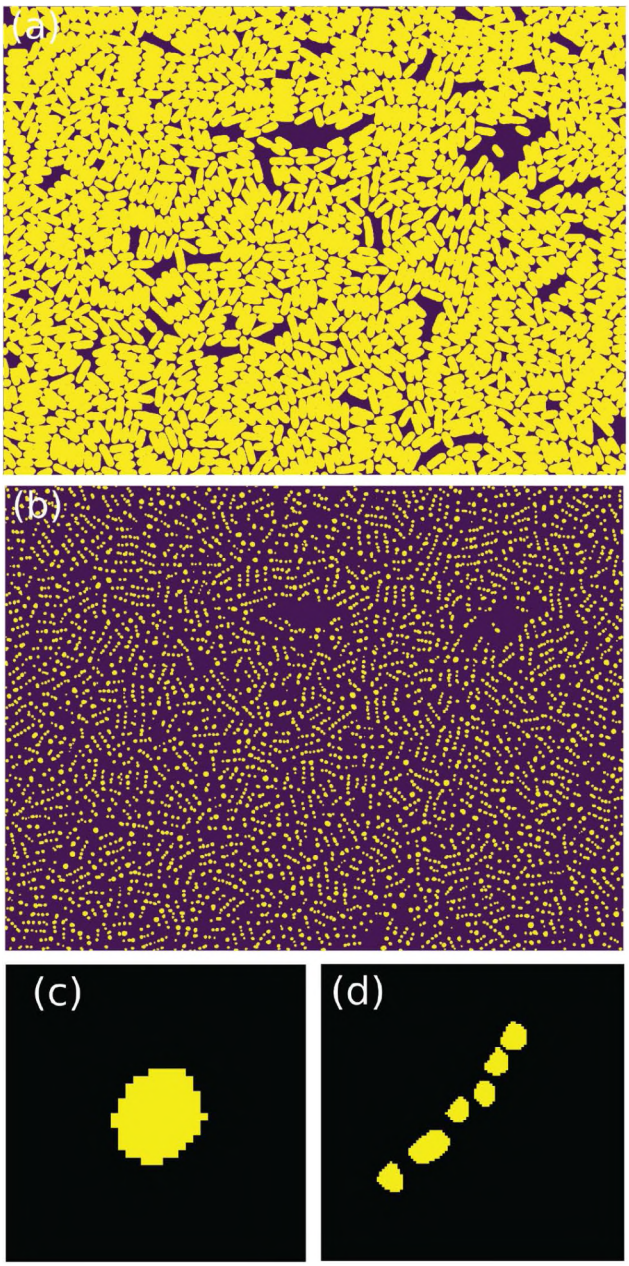


Figure 2: Segmentation of a fluorescence microscopy image shows: (a) the cyanobacteria cells; (b) the carboxysomes within all the cells; (c) an individual carboxysome; (d) the set of carboxysomes within a cyanobacterium.

Individual Carboxysomes

Images of individual carboxysomes decoded from the latent space are shown in Fig. 3a, where $L1$ and $L2$ vary along the x and y axis, respectively. For a more intuitive understanding of

what $L1$ and $L2$ represent, we decoded the images by varying $L1$ and fixing $L2$, and vice versa. We found that $L2$ was inversely correlated with carboxysome size, (Fig. 3b) whereas $L1$ represents the shape of the carboxysome, with high circularity in the $L1 > 0$ range and an elongated phenotype when $L1 < 0$ (Fig. 3c). The carboxysome shapes and sizes shown here correspond with a distillation of the myriad of carboxysome morphologies observed in both wildtype cells and those cells in which the CcmO has been degraded.[20]

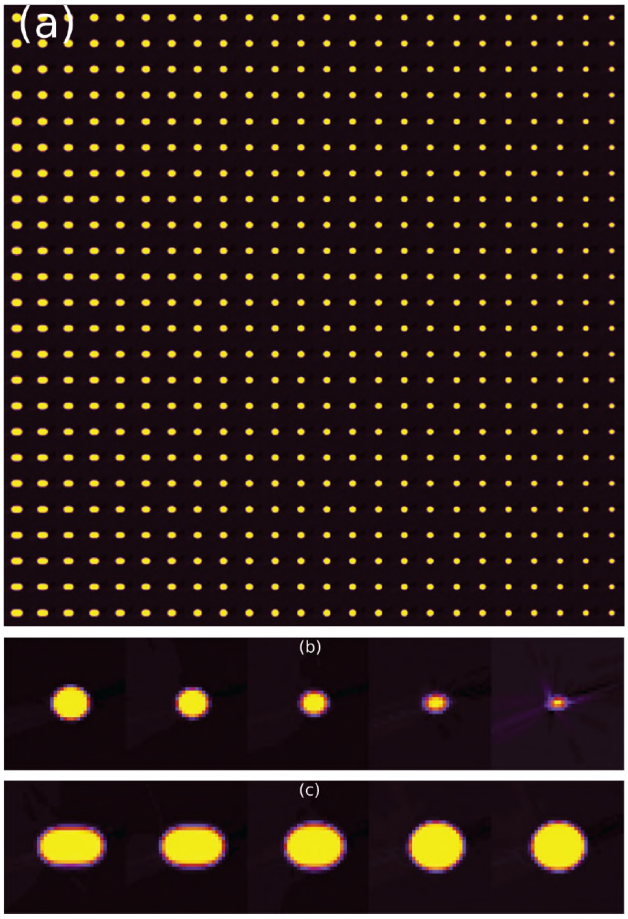


Figure 3: Decoded images of individual carboxysomes obtained by sampling the latent space for the control and induced samples. (a) Evolution of decoded images as a function of $L1$ (x-axis) and $L2$ (y-axis). $L1$ and $L2$ vary between $[-1.5, 1.5]$ in increments of 0.5; Decoded images at specific values of the latent variables: (b) $L1 = 0$ and $L2 = 0, 1, 2, 3, 4$; (c) $L1 = -1.5, -1, 0, 1, 2$ and $L2 = 2$. In (b) and (c) the axis denote the width and height of each image in pixels.

The $L1$ histogram for the control and induced samples are shown overlaid in Fig. 4a. For the control samples, a small change with time is observed, whereas for the induced samples

the change is significant and characterized by the appearance of a "shoulder" at $L1 > 1.0$ which grows steadily with time (see red arrow from 8 hours on). The $L2$ histograms for the control and induced samples are shown in Fig. 4b. For the control samples, variation is again small with time, whereas for the induced samples a decreasing with time of the large peak at $L2 = 0$ and the appearance of a new peak in the region $L2 = [-2, -1]$ are observed (see red arrows at 48 and 72 hours of induction).

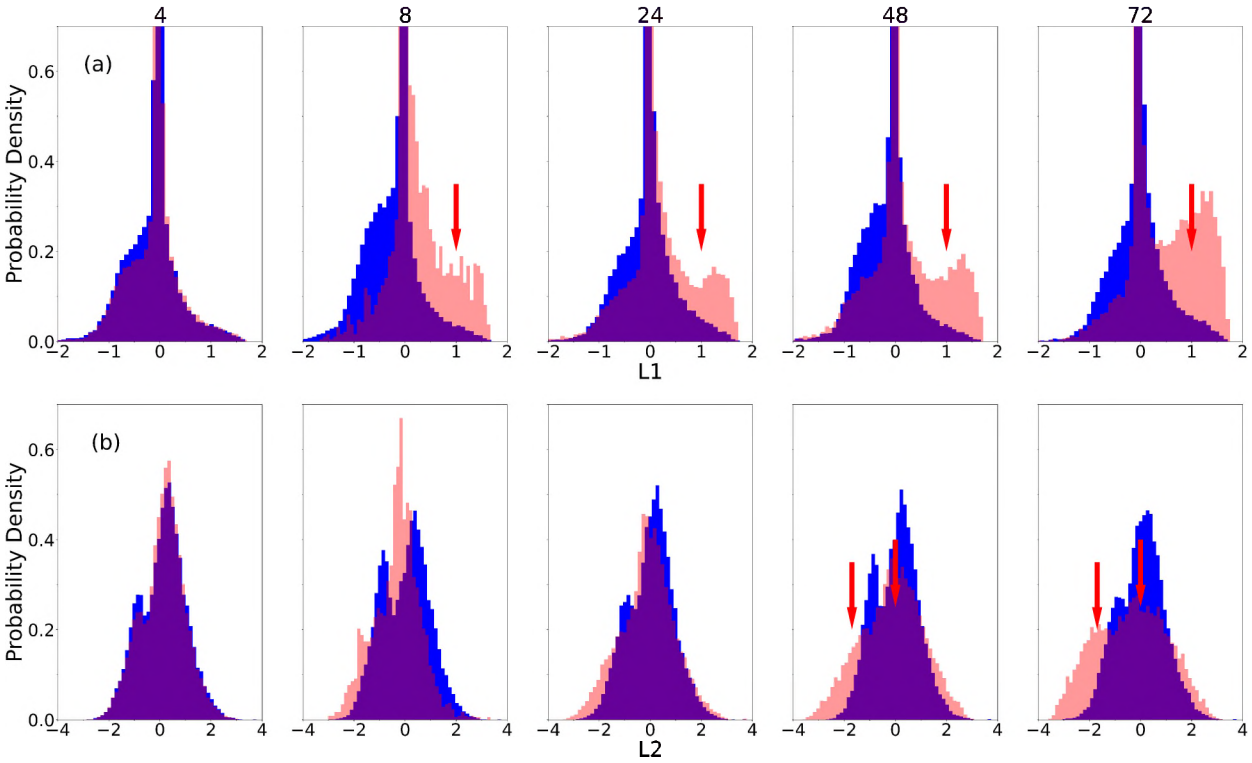


Figure 4: Overlaid histograms for the stacks of individual carboxysomes sub-images, Table 1, for both the control (blue) and induced samples (red). (a) histogram for $L1$; ((b) histogram for $L2$. The arrows indicate changes in the histogram that appear due to induction (see text for a discussion). The numbers $4, 8, 24, 48, 72$ above each figure denote the time in hours.

The results above indicated that induced degradation of CcmO changes $L1$ and $L2$ with time. To visualize what those changes meant in terms of carboxysomeal structural dynamics, we used the $L1$-$L2$ joint distribution. This distribution is shown in Fig. 5 for the control and induced samples. It is seen that at 72 hours the number of carboxysomes increases in the region defined by $L1 > 1$ and $L2 = [-2, -1]$. Therefore, decoding images from those

regions will visualize the structural changes caused by degradation.

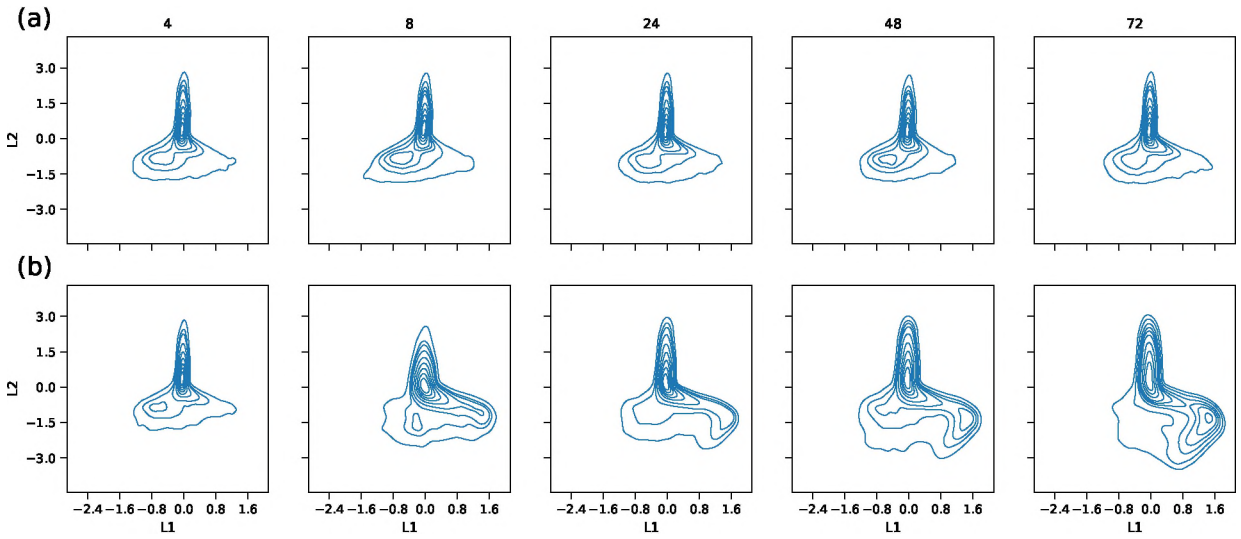


Figure 5: Contour plots showing the kernel density estimate (KDE) of the $L1$-$L2$ joint distribution for the (a) control and (b) induced samples of the individual carboxysomes set. The KDE is a smoothed out version of a histogram, and it enables computing a probability density function. The numbers $4, 8, 24, 48, 72$ above each figure denote the time in hours post induction.

The carboxysome images decoded in the regions $L1 > 1$ and $L2 = [-2, -1]$ are shown in Fig. 6, where $L1$ increases from bottom to top and $L2$ decreases from left to right (the specific values for $L1$ and $L2$ are given in the caption of Fig. 6). The carboxysome morphology becomes rounder when $L1$ increases and larger as $L2$ decreases. Taken together, these data highlight the change in carboxysome morphology as the result of proteolysis by *mf-lon*, with the carboxysomes getting larger and more round over time as Rubisco continues to aggregate in the absence of an intact shell. These observations are consistent with previous work.[20]

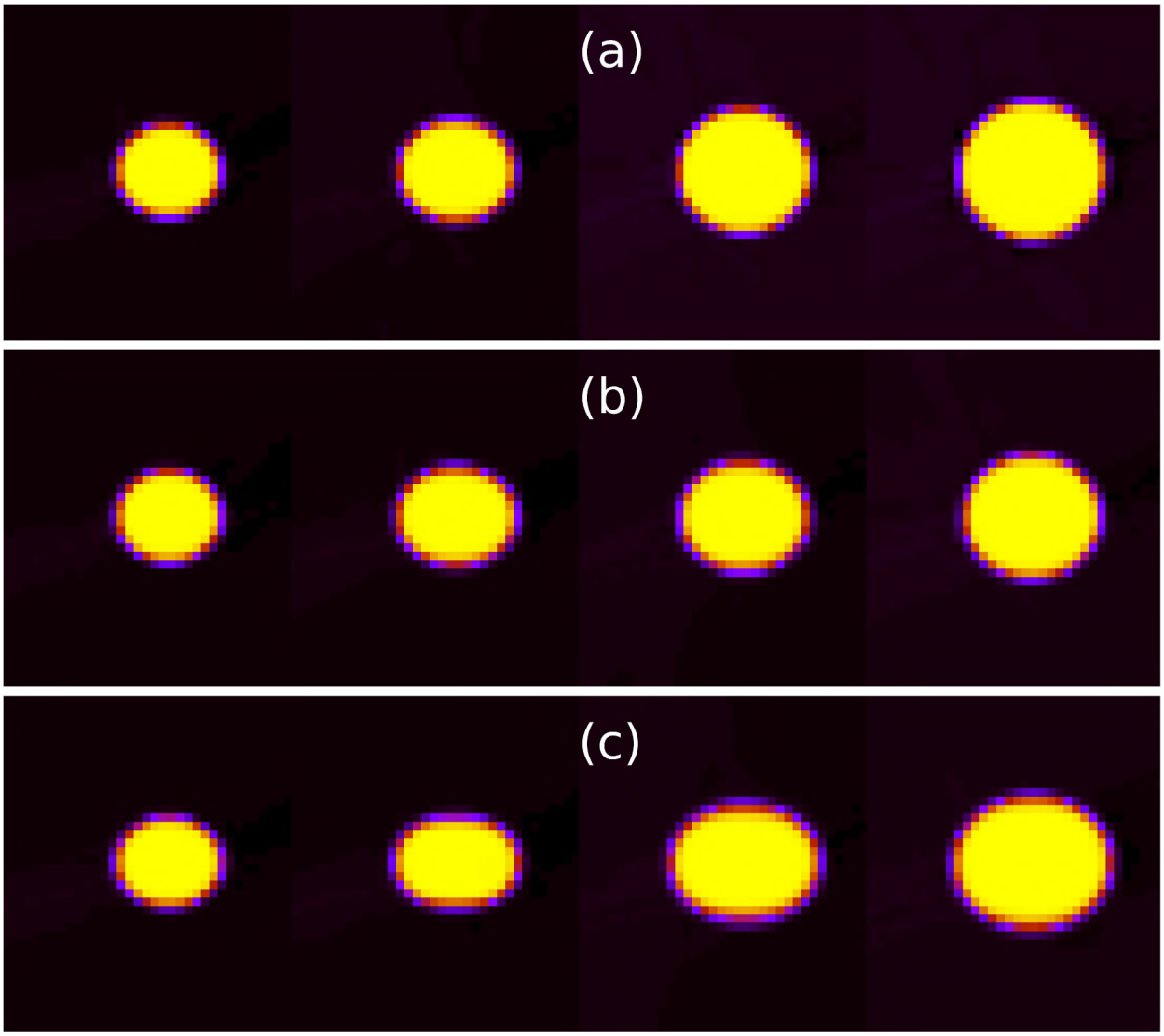


Figure 6: Images of individual carboxysomes decoded by sampling the latent space of the induced sample at specific values of $L1$ and $L2$: (a) $L1 = 1.6; L2 = -0.5, -1.0, -2.0, -2.5$ (b) $L1 = 1.0; L2 = -0.5, -1.0, -2.0, -2.5$ (c) $L1 = 0.0; L2 = -0.5, -1.0, -2.0, -2.5$. For illustration purposes, for each panel the figures were merged into a single one. Individual figures for each decoded image can be seen in the notebooks included in the Supporting Information. For these individual figures, the x- and y-axis represent the width and height in pixels.

Set of Carboxysomes per cell

A similar study to the one performed above for the individual carboxysomes was performed here for the set of carboxysomes per cell. Images for the set of carboxysomes decoded from

the latent space are shown in Fig. 7a. Figure 7b shows that changing $L1$ while fixing $L2$, and vice versa, both reduce the number of carboxysomes per set, but depending on which variable is varied, the remaining carboxysomes can be separated into two or merged into a single one.

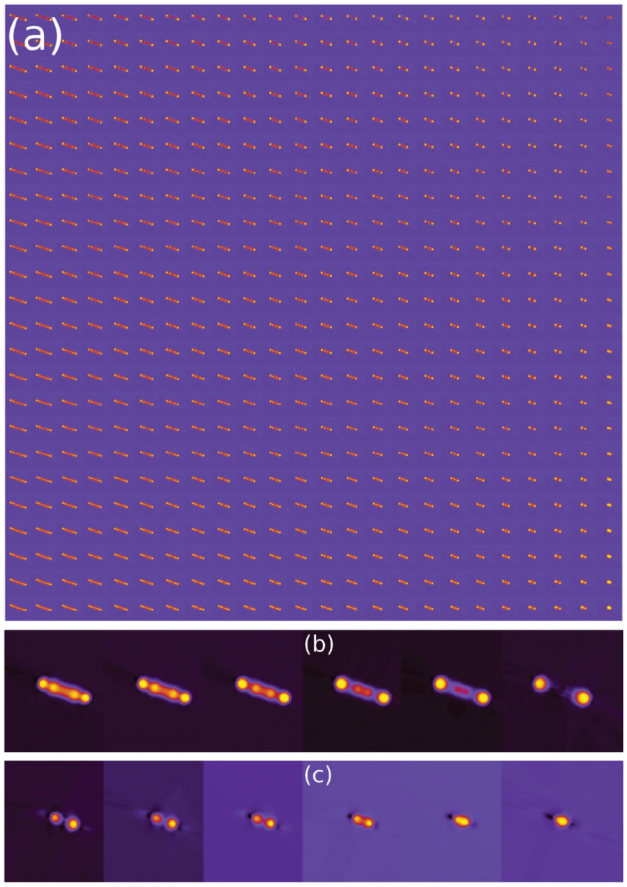


Figure 7: Decoded images of set of carboxysomes obtained by sampling the latent space for the control and induced samples. (a) Evolution of decoded images as a function of $L1$ (x-axis) and $L2$ (y-axis). $L1$ and $L2$ vary in $[-1.5, 1.5]$ in increments of 0.5; Decoded images at specific values of $L1$ and $L2$: (b) $L1 = -2, -1, 0, 1, 1.5$ and $L2 = 0$; (c) $L1 = 1.7$ and $L2 = 1.4, 1.5, 1.6, 1.7, 1.9, 2$. In (b) and (c) the x- and y-axis denote width and height in pixels

The $L1$ histogram for the control and induced samples is shown overlaid in Fig. 8a. For the control sample the $L1$ histogram is practically constant with time. However, for the induced sample the histogram significantly shifts beginning at 8 hours post-induction, and the distribution of carboxysome widens and moves to larger values of $L1$ (see red arrow at

72 hours). The $L2$ histograms for both the control and induced samples are shown in Fig. 8b. At 48 hours, the $L2$ histogram for the control sample shows new peaks, albeit small, for $L2 >= 2$. For the induced sample a large peak at $L2 >= 2$ appears as early as at the 4 hour (indicated also by a red arrow), and then additionally peaks located at $L2 >= 1$ increase with time; the latter are actually caused by a displacement of the whole histogram towards larger values of $L2$ (see red arrow at 72 hours).

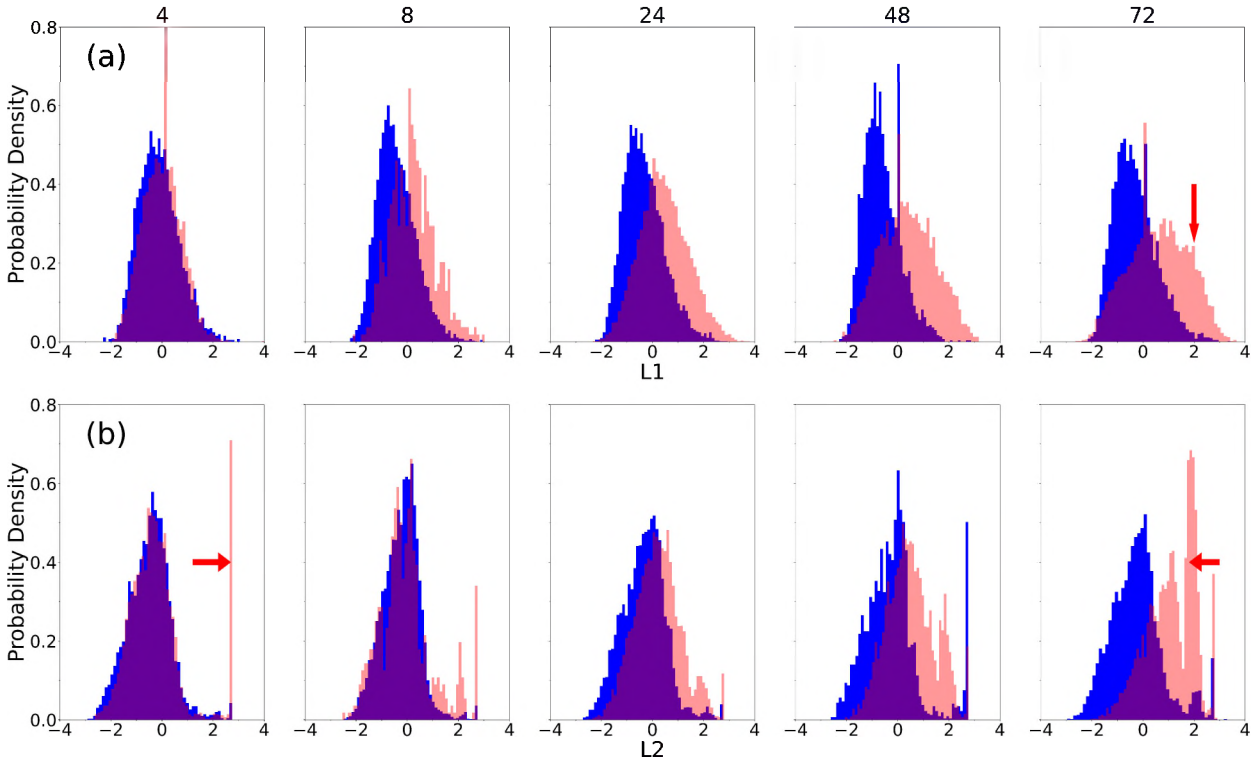


Figure 8: Overlaid histograms for the stacks of the set of carboxysomes sub-images, Table 1, for both the control (blue) and induced samples (red). (a) histogram for $L1$; ((b) histogram for $L2$. The arrows indicated the presence of peaks in the histogram (see text for a discussion). The numbers $4, 8, 24, 48, 72$ above each figure denote the time in hours.

Figure 9 shows the $L1 - L2$ joint distribution for the control and induced samples. For the control samples, a small incidence of events are categorized with $L2 >= 2$ values, which produced a ‘’hat" on the distribution that is consistent throughout all time points. For the induced samples, the “hat” also appears for values $L2 >= 1$ as early as at 8 hours post induction, and the proportion of events is dramatically increased at later time points. There’s

also an additional shifting of the distribution towards larger values of $L2$, consistent with the trend observed in the $L2$ histogram (shown in Fig.8b. Ultimately, after 24 hours post-induction, the joint distribution is split into two regions, the aforementioned "hat" region, where $L2 >= 1$, and a "skewed-like" region, where $L2 < 1$.

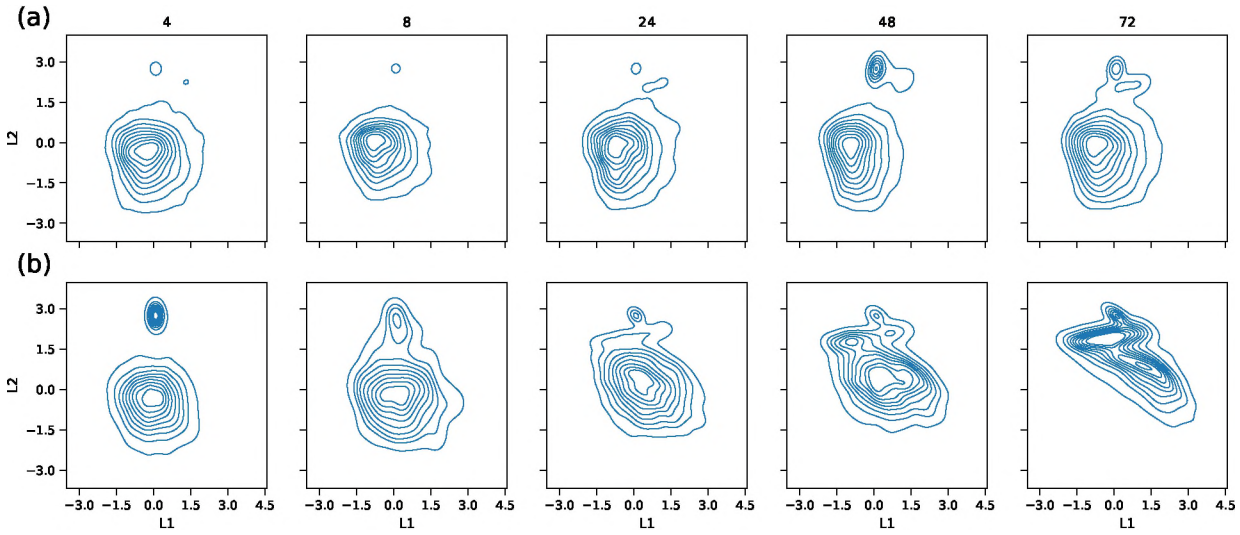


Figure 9: Contour plots showing the kernel density estimate (KDE) of the $L1$-$L2$ joint distribution for the (a) control and (b) induced samples of the set of carboxysomes data. The numbers $4, 8, 24, 48, 72$ above each figure denote the time in hours post induction.

The Fig. 10 shows the decoded images from the "hat" and "skewed-like" regions of the $L1 - L2$ joint distribution of the induced samples. Images were decoded by passing through these two regions either laterally, that is varying $L1$ while fixing $L2$, or vertically, varying $L2$ while fixing $L1$. Figure 10a,b show the lateral pass for two different values of $L2$, whereas Fig. 10c,d show the vertical pass for two different values of $L1$. For the set of carboxysomes per cell, it's somewhat more difficult to assign a physical explanation to the variables $L1$ and $L2$ than it was for the individual carboxysomes, where $L1$ was correlated to shape and $L2$ was inversely correlated to size. Nonetheless, in Fig. 10 it is seen that changes in $L1$ are correlated with the movement of carboxysomes towards the poles, whereas changes in $L2$ are correlated with carboxysome aggregation. In both cases, the number of carboxysomes per set decreases with time. While such trends are clear in Fig.10 (and by extension in

Fig.9, since the images in Fig.10 were decoded by selecting regions of Fig.9), it's significantly more difficult to understand how more aggregation or more movement towards the poles might affect the histograms and $L2$ and $L1$ shown in Fig.8. In any case, the observation that carboxysomes merge into one is significantly more prevalent in the induced sample than in the control. Indeed, merging is what causes the "skewed-like" region in the $L1 - L2$ distribution of the induced sample (see Fig.9b).

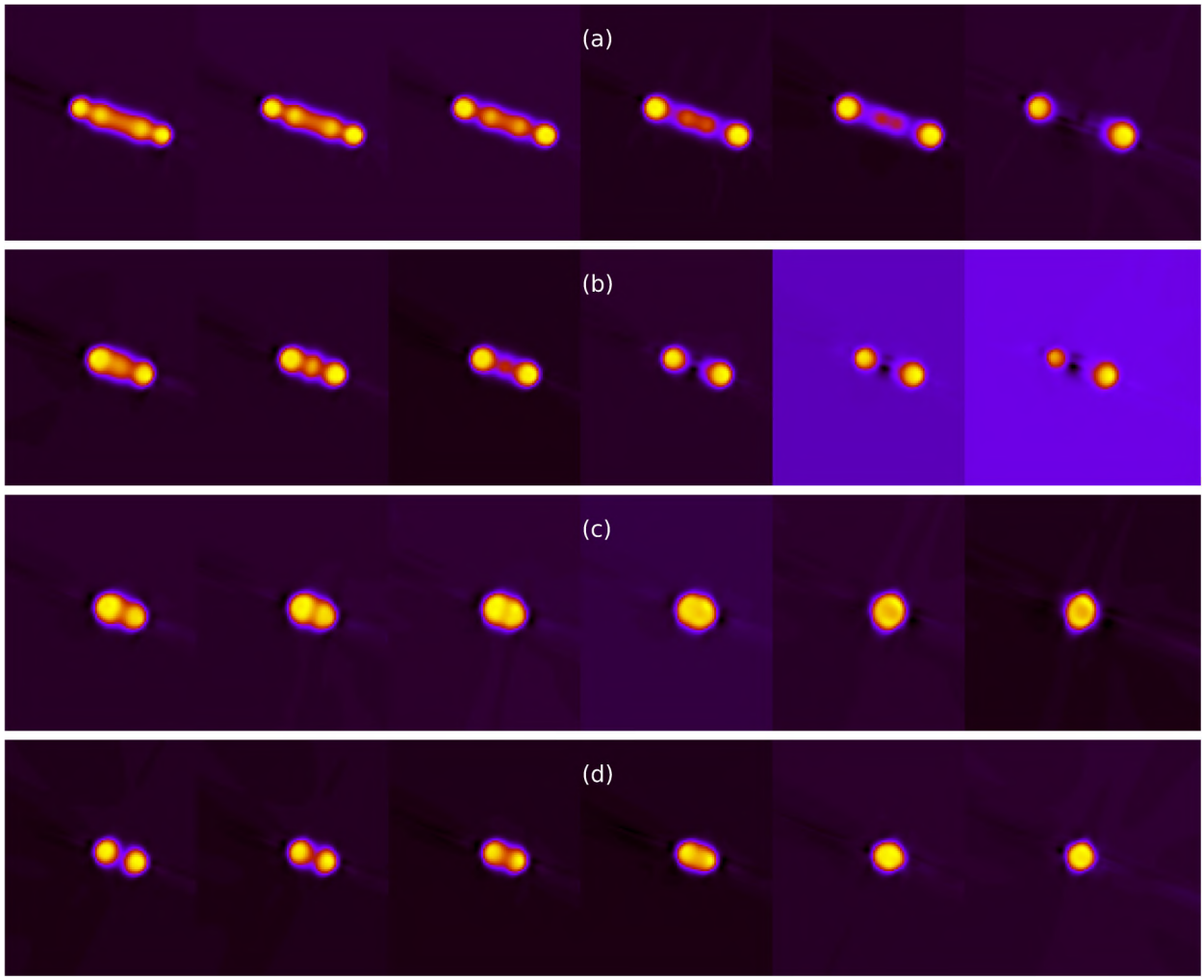


Figure 10: Decoded images from the latent space along the following selected regions in the $L1 - L2$ joint distributions (a) $L1 = -2, -1, 0, 1, 1, 5, 2.5$, $L2 = 0$; (b) $L1 = -2, -1, 0, 1, 1, 5, 2.5$, $L2 = 1$; (c)$L1 = 0$, $L2 = 1.4, 1.5, 1.6, 1.7, 1.9, 2.0$; (d) $L1 = -1.5$, $L2 = 1.4, 1.5, 1.6, 1.7, 1.9, 2.0$. For illustration purposes, for each panel, individual images were merged into a unique image. The individual images have width and height given in pixels.

In a previous work[20] it was found that induction produced one carboxysome aggregate that located at the pole of the cell. The results in the previous sections were obtained by training the rVAE to a stack of sub-images that had one channel only, *i.e.* a channel for a single carboxysome or a channel for the set of carboxysomes per cell. Although those analysis showed that the number of carboxysomes per cell decreases due to aggregation, it was not possible to determine the location of the carboxysome aggregate in the cell. For this purpose, the rVAE has to be trained on multi-channel stacks of images, where one channel contains the cell background (Chla) and the other channel contains the carboxysomes within the cell (mNG). In this manner, spatial correlations between the set of carboxysomes and the cell can be determined. Thus, we proceeded to create a stack of sub-images that had the same dimensions as those shown in Table 1 under the entry "Set of carboxysomes", except now each sub-image had the two channels aforementioned. The rVAE was trained to this multi-channel stack of sub-images and the $L1 - L2$ joint distribution that resulted is shown in Fig.11. It is seen that the distribution for the control samples changes only at the 48 and 72 hour time points, and even so not significantly. However, the distribution for induced samples at 4 hours already show two slightly separate classes (regions), and this distinction increases further as time progresses. We decoded the images across these two classes by fixing $L1 = 0$ and varying $L2$ from -3 to 2. The resultant images are shown in Fig.12: clearly, the number of carboxysomes per cell not only diminishes with time, but also the resultant aggregate locates to the pole of the cell, in agreement with the observations in Ref.[20]

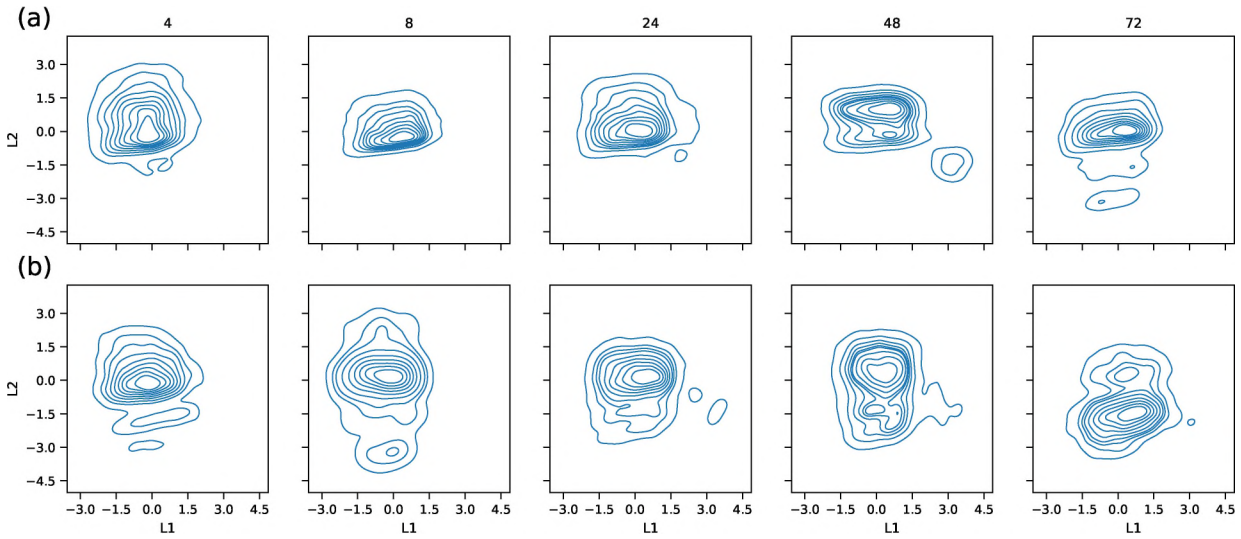


Figure 11: Contour plots showing the kernel density estimate (KDE) of the $L1$-$L2$ joint distribution for the (a) control and (b) induced samples of the for the multi-channel sub-images containing the masks for the set of carboxysomes and the host cell. (a) control and (b) induced samples. The numbers $4, 8, 24, 48, 72$ above each figure denote the time in hours.

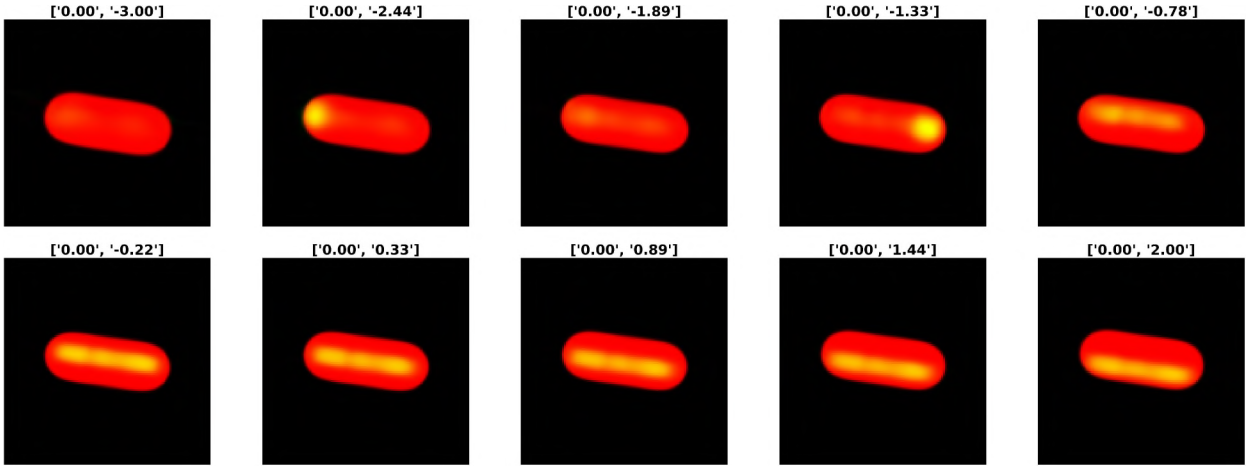


Figure 12: Decoded images from the latent space along the following selected regions in the $L1 - L2$ joint distributions, Fig.11: Decoded images from the latent space along specific regions in the $L1-L2$ joint distributions of Fig.11. The $L1$ and $L2$ values from each figure was decoded are given on top of each image as $[L1, L2]$. For $[L1, L2] = [0, -3]$ the carboxysome appears very light. Training the rVAE to more epochs did not improve this fact. The x- and y-axis of each image represent the width and height in pixels, respectively.

CONCLUSIONS

Herein, we demonstrate that rVAEs can be utilized for high-throughput analysis of carboxysome puncta to gain quantitative data regarding the remodeling of bacterial microcompartments. Prior observations of carboxysomes across multiple cyanobacteria species has suggested that these bacterial microcompartments may be adjusted in abundance, size, composition, and/or positioning to tune their properties for different environments. For instance, carboxysome β-carboxysome operons (main and satellite) display differential expression under different illumination intensities and $CO2$ availability.[14,15,18,28,29] Yet, it remains unclear if these changes in features are dynamically regulated in existing carboxysomes, or if they are only encoded in newly-assembled microcompartments. This uncertainty persists in part due to insufficient molecular tools and limitations in the inherent resolution limits of light microscopy, which confound analysis of carboxysome populations over time. In this context, fluorescence microscopy imaging of the alterations in carboxysome features in response to an environmental or artificial stimuli can provide valuable insights on structural dynamics. However, the amount of data produced by modern microscopy approaches is such that manual evaluation has insufficient throughput and sensitivity; only a qualitative analysis is feasible to evaluate subtle changes in population-level properties. Quantitative analysis can shed more light into the carboxysome biogenesis and dynamic processes, and for this purpose, deep learning techniques are especially suitable. Here we demonstrate that a type of deep learning technique known as a Rotationally Invariant Variational Autoencoder is capable of revealing structural changes from fluorescence microscopy datasets, including changes in the shape and size of carboxysomes and the number of carboxysomes inside of the cell. These high-throughput analyses are capable of detecting significant changes at time points as early as 4 or 8 hours post-induction of a carboxysome shell degrading circuit, whereas manual evaluation was only able to subjectively report changes 24 hours after the genetic intervention.[20] This work reveals that variational autoencoders can play a very important role in detailing the dynamic processes (e.g., remodelling, biogenesis) of carboxysomes and,

by extension, bacterial microcompartments in general.

Acknowledgement

This research was conducted at the Center for Nanophase Materials Sciences, which is a DOE Office of Science User Facility. This work was also funded by National Science Foundation Grant 1517241 (to D.C.D.), funding from Department of Energy Grant $DE - FG02 - 91ER20021$ (to D.C.D.) at the MSU DOE-PRL. Additional support for the research was provided by NSF Award 1845463 (to D.C.D.). Training the rVAE to a multi-channel set of images was performed at the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231. M.F.-C. also wishes to thank Sergei Kalinin and Kevin Roccapriore for insightful communications about rVAEs.

Supporting Information Available

The notebooks used in this paper were written in Jupyter and run in Google Colaboratory or NERSC. The notebooks for performing segmentation and for training the rVAEs can be found in these github repositories https://github.com/miguel-fc/carboxysomes and https://github.com/Jsakkos/mf-lon-cyanos-rvae.

References

(1) Ducat, D. C.; Way, J. C.; Silver, P. A. Engineering cyanobacteria to generate high-value products. *Trends in Biotechnology* **2011**, *29*, 95–103.

(2) Santos-Merino, M.; Singh, A. K.; Ducat, D. C. New applications of synthetic biology

tools for cyanobacterial metabolic engineering. *Frontiers in bioengineering and biotechnology* **2019**, *7*, 33.

(3) Angermayr, S. A.; Rovira, A. G.; Hellingwerf, K. J. Metabolic engineering of cyanobacteria for the synthesis of commodity products. *Trends in biotechnology* **2015**, *33*, 352–361.

(4) Knoot, C. J.; Ungerer, J.; Wangikar, P. P.; Pakrasi, H. B. Cyanobacteria: promising biocatalysts for sustainable chemical production. *Journal of Biological Chemistry* **2018**, *293*, 5044–5052.

(5) Tcherkez, G. G.; Farquhar, G. D.; Andrews, T. J. Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proceedings of the National Academy of Sciences* **2006**, *103*, 7246–7251.

(6) Cameron, J. C.; Wilson, S. C.; Bernstein, S. L.; Kerfeld, C. A. Biogenesis of a bacterial organelle: The carboxysome assembly pathway. *Cell* **2013**, *155*, 1131–1140.

(7) Ditty, J. L.; Canales, S. R.; Anderson, B. E.; Williams, S. B.; Golden, S. S. Stability of the Synechococcus Elongatus PCC 7942 Circadian Clock under Directed Anti-Phase Expression of the Kai Genes. *Microbiology* **2005**, *151*, 2605–2613.

(8) Sun, Y.; Huang, F.; Dykes, G. F.; Liu, L.-N. Diurnal Regulation of In Vivo Localization and CO2-Fixing Activity of Carboxysomes in Synechococcus Elongatus PCC 7942. *Life* **2020**, *10*, 169.

(9) Wishiwaki, T.; Satomi, Y.; Nakajima, M.; Lee, C.; Kiyohara, R.; Kageyama, H.; Kitayama, Y.; Temamoto, M.; Yamaguchi, A.; Hijikata, A. et al. Role of KaiC Phosphorylation in the Circadian Clock System of Synechococcus Elongatus PCC 7942. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 13927–13932.

(10) Turmo, A.; Gonzalez-Esquer, C. R.; Kerfeld, C. A. Carboxysomes: metabolic modules for CO2 fixation. *FEMS microbiology letters* **2017**, *364*.

(11) Wang, H.; Yan, X.; Aigner, H.; Bracher, A.; Nguyen, N. D.; Hee, W. Y.; Long, B.; Price, G. D.; Hartl, F.; Hayer-Hartl, M. Rubisco condensate formation by CcmM in β-carboxysome biogenesis. *Nature* **2019**, *566*, 131–135.

(12) Chen, A. H.; Robinson-Mosher, A.; Savage, D. F.; Silver, P. A.; Polka, J. K. The Bacterial Carbon-Fixing Organelle Is Formed by Shell Envelopment of Preassembled Cargo. *PLoS ONE* **2013**, *8*, 1–13.

(13) Hill, N. C.; Tay, J. W.; Altus, S.; Bortz, D. M.; Cameron, J. C. Life cycle of a cyanobacterial carboxysome. *Science Advances* **2020**, 1–9.

(14) Badger, M. R.; Price, G. D.; Long, B. M.; Woodger, F. J. The environmental plasticity and ecological genomics of the cyanobacterial CO2 concentrating mechanism. *Journal of experimental botany* **2006**, *57*, 249–265.

(15) McGinn, P. J.; Price, G. D.; Maleszka, R.; Badger, M. R. Inorganic carbon limitation and light control the expression of transcripts related to the CO2-concentrating mechanism in the cyanobacterium Synechocystis sp. strain PCC6803. *Plant Physiology* **2003**, *132*, 218–229.

(16) McGinn, P. J.; Price, G. D.; Badger, M. High light enhances the expression of low-CO2-inducible transcripts involved in the CO2-concentrating mechanism in Synechocystis sp. PCC6803. *Plant, Cell & Environment* **2004**, *27*, 615–626.

(17) Huang, L.; McCluskey, M. P.; Ni, H.; LaRossa, R. A. Global gene expression profiles of the cyanobacterium Synechocystis sp. strain PCC 6803 in response to irradiation with UV-B and white light. *Journal of bacteriology* **2002**, *184*, 6845–6858.

(18) Woodger, F. J.; Badger, M. R.; Price, G. D. Inorganic carbon limitation induces transcripts encoding components of the CO2-concentrating mechanism in Synechococcus sp. PCC7942 through a redox-independent pathway. *Plant Physiology* **2003**, *133*, 2069–2080.

(19) MacCready, J. S.; Hakim, P.; Young, E. J.; Hu, L.; Liu, J.; Osteryoung, K. W.; Vecchiarelli, A. G.; Ducat, D. C. Protein gradients on the nucleoid position the carbon-fixing organelles of cyanobacteria. *Elife* **2018**, *7*, e39723.

(20) Sakkos, J. K.; Hernandez-Ortiz, S.; Osteryoung, K. W.; Ducat, D. C. Orthogonal Degron System for Controlled Protein Degradation in Cyanobacteria. *ACS Synthetic Biology* **2021**, acssynbio.1c00035.

(21) Stringer, C.; Wang, T.; Michaelos, M.; Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *NATURE METHODS* **2021**, *18*, 100+.

(22) Ziatdinov, M.; Ghosh, A.; Wong, T.; Kalinin, S. V. AtomAI: A Deep Learning Framework for Analysis of Image and Spectroscopy Data in (Scanning) Transmission Electron Microscopy and Beyond. 2021.

(23) Nakahira, Y.; Ogawa, A.; Asano, H.; Oyama, T.; Tozawa, Y. Theophylline-dependent riboswitch as a novel genetic tool for strict regulation of protein expression in cyanobacterium Synechococcus elongatus PCC 7942. *Plant and Cell Physiology* **2013**, *54*, 1724–1735.

(24) Sutter, M.; Laughlin, T. G.; Sloan, N. B.; Serwas, D.; Davies, K. M.; Kerfeld, C. A. Structure of a synthetic β-carboxysome shell. *Plant physiology* **2019**, *181*, 1050–1058.

(25) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. 2014.

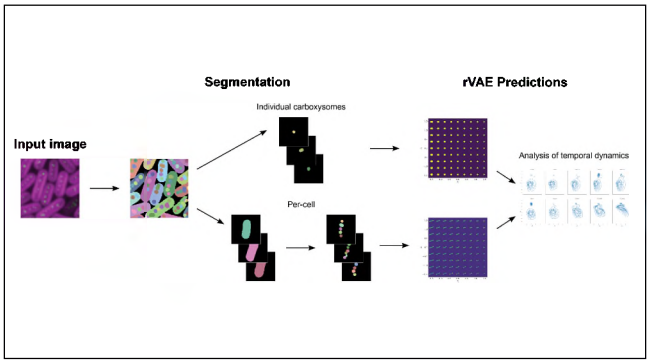
(26) Bepler, T.; Zhong, E. D.; Kelley, K.; Brignole, E.; Berger, B. Explicitly disentangling image content from translation and rotation with spatial-VAE. 2019.

(27) Kalinin, S., V; Zhang, S.; Valleti, M.; Pyles, H.; Baker, D.; De Yoreo, J. J.; Ziatdinov, M. Disentangling Rotational Dynamics and Ordering Transitions in a System of Self-Organizing Protein Nanorods via Rotationally Invariant Latent Representations. *ACS NANO* **2021**, *15*, 6471–6480.

(28) Sommer, M.; Cai, F.; Melnicki, M.; Kerfeld, C. A. β-Carboxysome bioinformatics: identification and evolution of new bacterial microcompartment protein gene classes and core locus constraints. *Journal of Experimental Botany* **2017**, *68*, 3841–3855.

(29) McKay, R. M. L.; Gibbs, S. P.; Espie, G. S. Effect of dissolved inorganic carbon on the expression of carboxysomes, localization of Rubisco and the mode of inorganic carbon transport in cells of the cyanobacterium Synechococcus UTEX 625. *Archives of Microbiology* **1993**, *159*, 21–29.

TOC Graphic


Segmentation
rVAE Predictions
Individual carboxysomes
Input image
Analysis of temporal dynamics
Per-cell