Machine Learning Analysis of Biofilm Wrinkle Morphology

Submitted for the Summer Research Project 2024 as part of
The Laidlaw Leadership and Research Scholarship

by
Britt M. Dewing
Supervised by Moe Vali and Diana Fusco

Fusco Lab, Biological Soft Systems Group, Cavendish Laboratory,
University of Cambridge

Abstract

This project outlines a method for isolating biofilm wrinkles in brightfield microscopy images of radial *B. Subtilis* colonies through the production of binary masks, generated using a Convolutional Neural Network architecture, U-Net. The necessary pre-processing of the training data is explained as well as the production of the ground-truth masks from hand-drawn annotations. Further to this a system of different wrinkle types is introduced based on morphological differences and existing literature on their biological differences. An attempt at segmentation of brightfield images into these types using both image-to-image translation and multi-class classification is made. Finally, it is demonstrated how Sholl analysis can be applied to binary masks of biofilm wrinkles to quantify wrinkle density. A method to analyse the variation of phenotypes across space and time inside and outside of the imaged wrinkles is presented.

1. Introduction

Biofilms are bacterial communities exhibiting remarkable properties (Fusco, Pestana, Tropini, 2021, p.1). They are characterised by the production of a protective extracellular matrix - a sticky substance produced by some of the bacteria in the biofilm. This extracellular matrix is seen to form complex wrinkle morphologies which vary under different stress and growth conditions (Krishnan et al., 2024).

Biofilms are a significant area of research because they cause many chronic diseases in humans (Hall-Stoodley, Costerton, Stoodley, 2004), (Krishnan et al., 2024, p.1) are resistant to antimicrobial agents and also pose problems in industry (for example, they are known to form on water pipes) (Fusco, Pestana, Tropini, 2021, p.1). In order to develop reliable methods for removing and destroying biofilms, researchers seek to understand how they form and how they respond to stress.

Wrinkle formation is an active area of research for this reason. There have been a variety of approaches to modelling this process but none have successfully captured the full range of wrinkle morphologies, which vary across species, maturity of biofilm and growth conditions (Krishnan et al., 2024, pp. 10-14). It is believed that there is a strong connection between phenotypic expression and wrinkle formation which adds another layer of complexity to this problem (Krishnan et al, 2024).

Bacterial cells in Bacillus Subtilis biofilms are one of two phenotypes: matrix or motile. It has been shown that these phenotypes are mutually exclusive (Krishnan et al, 2024, p.7). Cells of the matrix type are adapted to producing the extracellular matrix fluid, while cells of the motility type are adapted for movement, for example possessing flagella (Fusco, Pestana, Tropini, pp. 1-2). The distribution of phenotypes within the biofilm can be visualised using specific reporters. An example is shown in Fig 1.

In observing the distribution of the matrix and motility phenotypes over time inside and outside of the wrinkles, we can see that different regions of wrinkles seem to have different behaviour. This implies that different wrinkle types have different functionality, which is explored in Section 2.2.1.

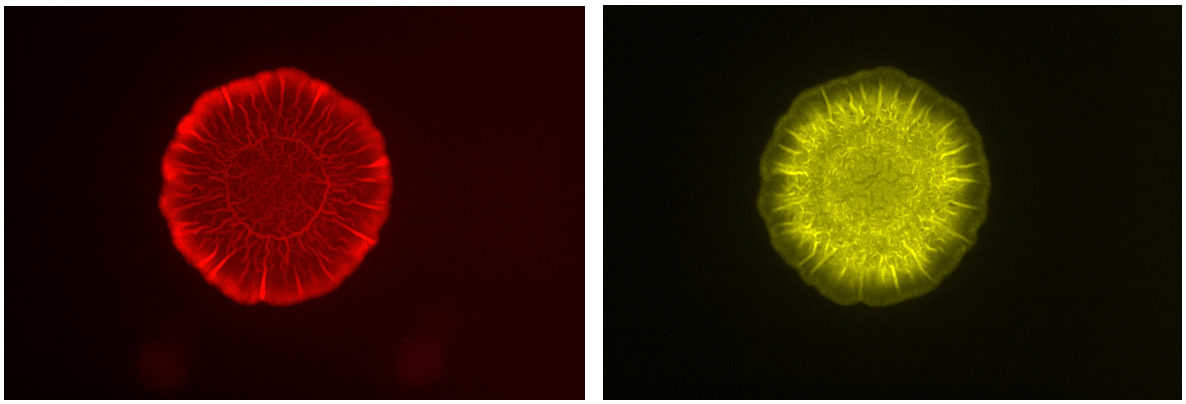


Fig 1: The distribution of matrix and motility phenotypes in the same B. Subtilis biofilm (Krishnan, 2022)

This research project had two aims:

1) Develop a time-efficient and accurate method to identify wrinkles in biofilms, to allow for quantitative analysis of wrinkle length, surface area, density and pattern
2) Develop a time-efficient and accurate method to segment the biofilm wrinkles into different types, so that such quantitative analysis can be performed on each type separately.

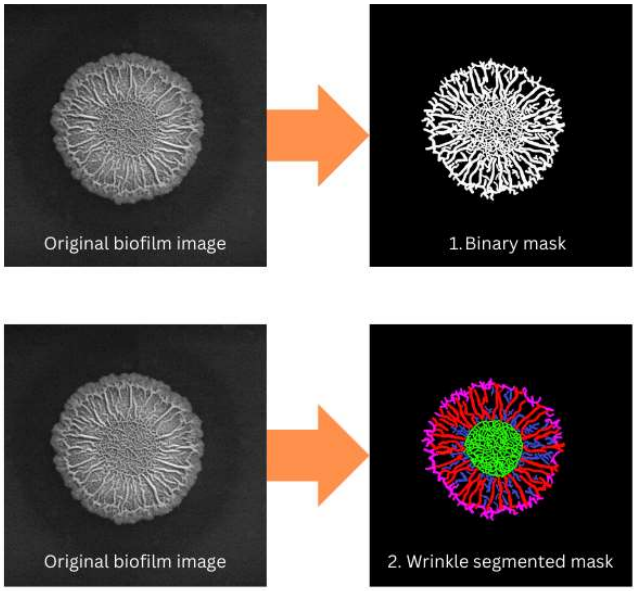


Fig 2. The two aims of this project. Original biofilm image by Krishnan (2022), masks generated in Python from hand annotations by Obaid (2024). Diagram created using Canva (2024).

In order to accomplish this, a Machine Learning (ML) algorithm was used to automatically generate wrinkle masks. To teach the algorithm how to recognise the biofilm wrinkles, it was trained on a labelled dataset containing pairs of original images and masks created from hand-drawn annotations (Obaid, 2024). The ML algorithm used is a convolutional neural network architecture called U-Net

(Ronneberger, Fischer and Brox, 2015), which has become a revolutionary tool in medical image segmentation tasks (Wang et al., 2024, pp.1-2) .

This research builds on very recent existing work by Cao et al. (2024) who successfully used U-Net to produce binary masks of biofilm wrinkles. This report furthers their work by demonstrating use of the method on a different dataset, exploring segmentation into different wrinkle types as well as more comprehensive analysis of the implications from the results.

In the following section I will outline my method for building the data analysis pipeline necessary to produce the wrinkle masks. The third section will present the results and their analysis. The findings of this report are then summarised in the conclusion.

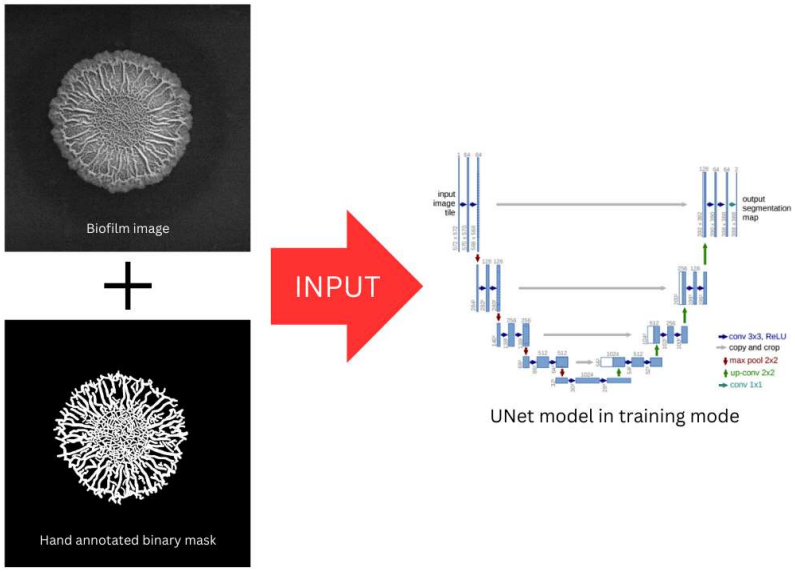


Fig 3.1. The training process. Pairs of biofilm images and corresponding hand-annotated masks form the training set, which is input into a U-Net in training mode. Biofilm image by Krishnan (2022), binary mask generated in Python from annotations by Obaid (2024), U-Net model diagram from Ronneberger, Fischer and Brox (2015, p.235, Fig 1). Diagram made in Canva.

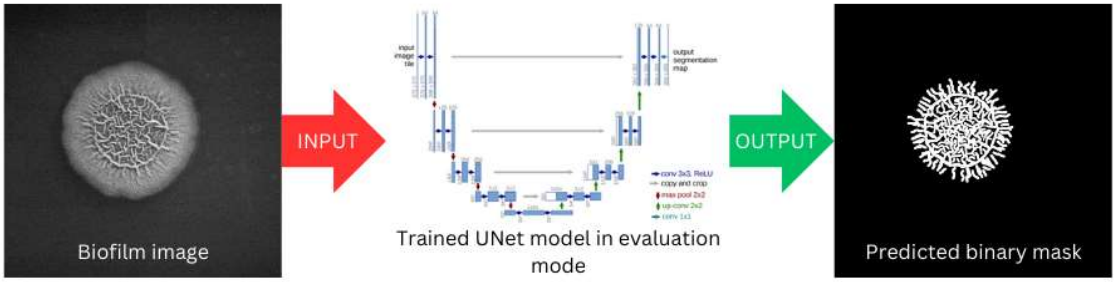


Fig 3.2. The testing process. Unseen biofilm images are input into the trained U-Net model, which is in evaluation mode. This produces a predicted binary mask. Biofilm image by Krishnan (2022), binary mask generated in Python from annotations by Obaid (2024), U-Net model diagram from Ronneberger, Fischer and Brox (2015, p.235, Fig 1.). Diagram made in Canva.

2. Method

2.1 Image preparation

Machine Learning tasks generally require very large datasets in order to sufficiently train the models. Depending on the task, this may be in the hundreds or hundreds of thousands. Images themselves contain thousands of pieces of information, therefore this task did not require such a massive dataset. Nevertheless, the size of the dataset used here - around 100 pairs of images - is too small, and thus limited the accuracy of the results. Given more time and access to a lab in which to grow more biofilm colonies, the dataset should have been expanded to at least 150 image pairs (Vali, 2024)

The images had to be selected to include a wide variety for generality, so that the resulting trained U-Net can be applied to most B. Subtilis images. However, it was determined that some images would confuse the model; for example images which were obscured by significant condensation on the petri dish, biofilms which were imaged upside-down or using other microscopy techniques. These images were removed from the dataset.

2.1.1 Background Removal

Furthermore, some images were deemed usable, but had variable, noisy and potentially distracting backgrounds. Two versions of the dataset were tested for the first task: the first contained unedited images, the second set had the image backgrounds removed.

The process of background removal was automated by a Python script using the computer vision library cv2. Contour detection was used to identify the edge of the biofilm. However, variable lighting conditions in the images meant that this alone frequently excluded part of the biofilm. Instead, the script attempts to correct for the variable lighting conditions first using adaptive histogram equalisation and various Gaussian blurs before applying the contour detection. From this, the smallest enclosing circle of that contour is found and a mask is generated. It is important to note that the mask is applied to the *original image* and not to the lighting corrected version, preserving the selected circle of the biofilm. The background is filled with a block grey colour. All the images were processed this way in one go, requiring only minutes on a standard personal computer for the full dataset

Even with this care taken to ensure accurate background removal, some “edge case” images had such poor or variable lighting conditions that the biofilm was not accurately detected. In this case the background had to be removed manually, using image editing software GIMP (2024).

2.1.2 Contrast Adjustment

Many of the biofilms used were imaged in varying sub-optimal lighting conditions: some images were over or under exposed, some were too shadowy. To counteract this some image editing was required.

However, inherent in many image editing techniques is a loss of the original features of the image. It was not known how this would affect the accuracy of the model.
It is possible to improve the contrast of an image in the following way without losing any features of the image:

1) Identify the maximum and minimum pixel values in the image. Hence find the range of the pixel values.
2) Assign the pixels with maximum value a new value of 255 and the pixels with minimum value a new value of 0. These are absolute white and black.
3) Correspondingly assign every intermediate value a new value between 255 and 0, stretching the range evenly.

For example if the maximum and minimum values were 0 and 51, a pixel with value 5 would be reassigned the value 25.

This technique worked well for some images. However, many images already contained values close to 0 and 255, so the range was not sufficiently stretched to make a difference. The method can to-that-end be altered:

1) Set upper and lower threshold percentiles
2) Pixels with any value in the upper threshold range of the image's pixel values are assigned 255. Pixels with values in the lower threshold range are assigned 0.
3) Assign every intermediate value a new value between 255 and 0, stretching only the range of these intermediate values.

This provides a greater visible increase in contrast. However, different pixel values falling in the upper or lower ranges are all set to the same values of 0 or 255. This means some features of the image may be lost, ironically in the process of trying to make the features clearer.

It was unknown how this would affect the process of training the U-Net. The aim was to find the optimal threshold percentiles that maximised the detectable features of the images. Therefore it was decided that the upper and lower percentiles would be set as *hyperparameters.* This is explained in Section 2.4.

Hence the contrast stretching process is treated as a transform applied to the images *just before* they are used in training. The contrast-stretched images are never saved, though once an optimum set of hyperparameters have been found it would be more efficient to save them with the optimised thresholds, rather than regenerating them.

2.2 Mask preparation

Having first prepared the images themselves, the corresponding masks needed to be created. For the second task, wrinkle type segmentation, these masks needed to distinguish between different wrinkle types.

2.2.1 Definition of different wrinkle types

In current literature there is no universal definition of different wrinkle types. Therefore, different wrinkle types needed to be identified. Four types were decided upon, which can be seen in Fig 4.

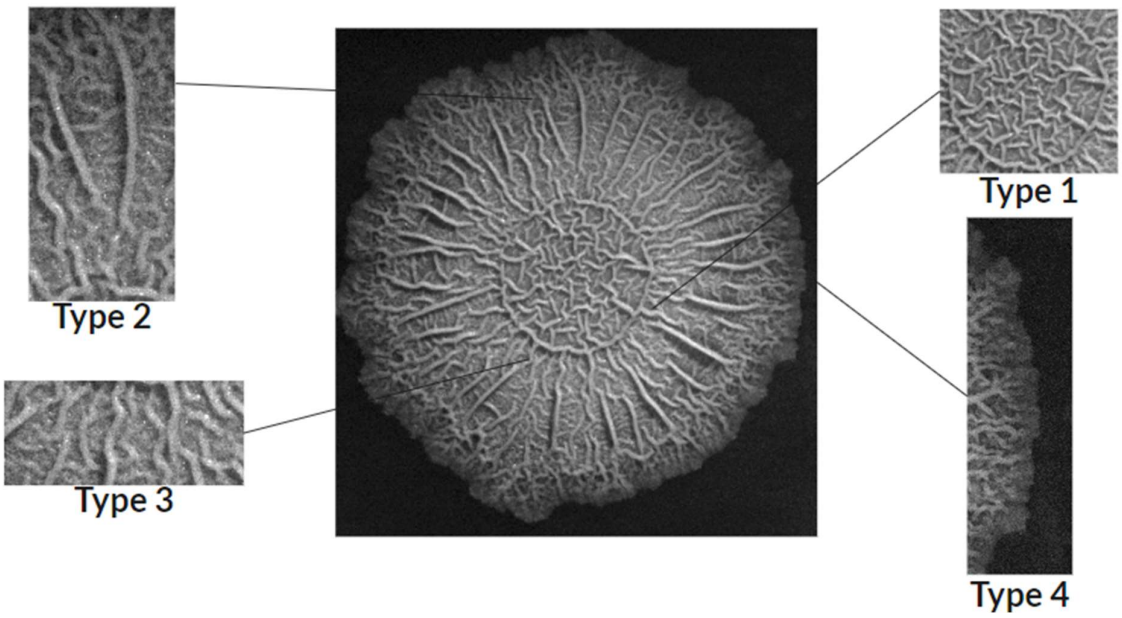


Fig 4.1. Four different regions of biofilm are highlighted, each predominantly containing a different kind of wrinkle. Image by Krishnan (2022). Diagram made in Google Slides.

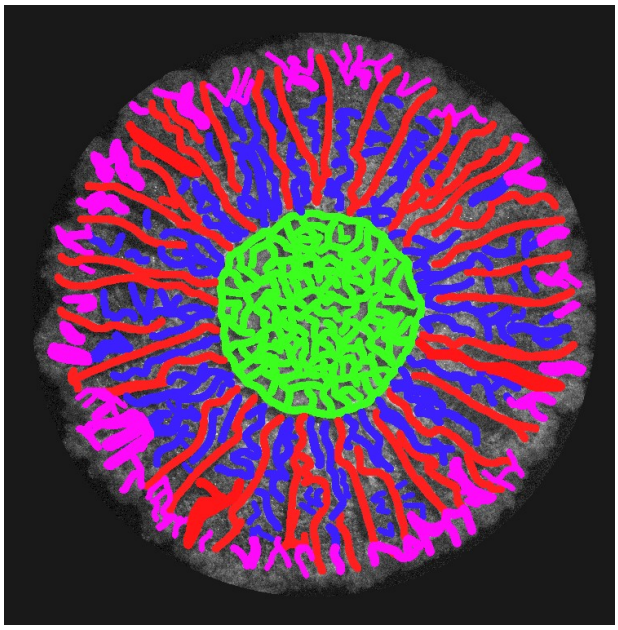


Fig 4.2 The different wrinkle types have been annotated with different colours. Image is a cropped version of a hand-annotated image mask by Obaid (2024).

Literature commonly refers to a central region (Krishnan et al., 2024, p.8), shown in Fig 4.2 in green. The growth of wrinkles in this region (Type 1) is in the early stages of biofilm development, when nutrients are abundant. This creates isotropic (non-directional) stress, leading to this ‘labyrinthine’ pattern. The growth of wrinkles outside of this region exhibits anisotropy due to the nutrient gradient between the nutrient depleted centre and the nutrient abundant edges.

The thick radial wrinkles shown in red (Type 2) have been observed to have a transportational role, distinguishing them from the blue and pink wrinkles. For example, these wrinkles may be used for antibiotic escape (Krishnan et al., 2024, p. 8).

The role of the blue wrinkles (Type 3) is not yet clear. However it is important to keep in mind that wrinkling is both a physiological stress response and a *mechanical* stress response (Krishnan et al., 2024). Therefore it is possible that these wrinkles do not have a purpose but are a byproduct of the growth rate.

Finally, the pink (Type 4) wrinkles observed at the edge of mature biofilms seem to have distinct properties. For example, some literature states they do not form under calcium deficient conditions (Krishnan et al., 2024, p.9) ; this could be due to the role of calcium in forming a rigid wrinkle structure (biomineralization). However, other works have also found similar results for Type 2 wrinkles.

2.3 U-Net

2.3.1 Model Architecture

The U-Net model was a revolutionary step forward in image segmentation tasks, and has been applied to many important challenges, especially in medical image segmentation (Wang et al, 2024, p. 2).

It is a convolutional neural network (CNN) involving several downsampling or encoding steps (Ronneberger, Fischer and Drax, 2015). These steps use convolutions with variable weightings to identify the features of an image. Each down-step identifies larger and more complex features. During the training process, the weights of these convolution matrices are adjusted. Then there are several upsampling or decoding steps, during which the information about the features gained from the downsampling is expanded to generate a new image - in this case the wrinkle mask. The upsampling steps are informed by the corresponding downsampling steps, which provides the U-Net with an improved understanding of the spatial organisation of the features. Fig 5 shows the U-Net architecture, with downsampling depicted on the left and upsampling on the right.

By training the U-Net for longer, the ‘loss’ - a measure of the difference between the output mask and the input ‘ground truth’ mask - is further minimised up to a certain point, beyond which the loss tends to plateau and may risk overfitting.

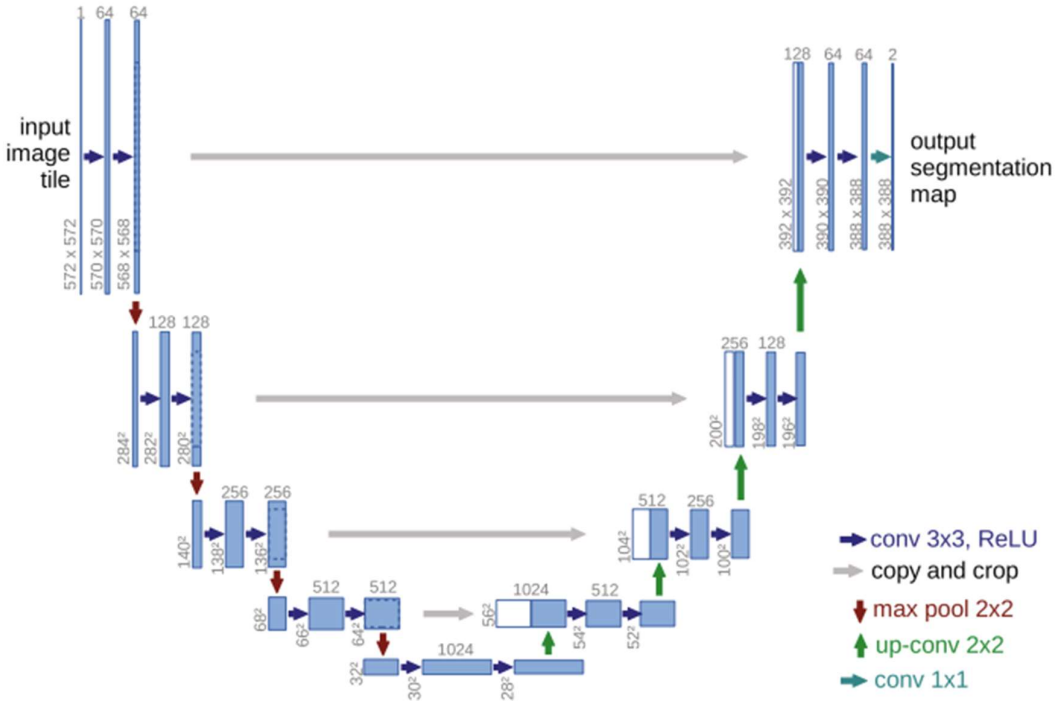


Fig 5. The U-Net architecture. Diagram from Ronneberger, Fischer and Drax (2015, p235, Fig 1.)

2.3.2 U-Net for binary classification

The first task was to create a wrinkle mask which simply showed where there were wrinkles in the biofilm. This is known as binary classification, as the U-Net has to predict whether each pixel is or isn't part of a wrinkle.

Since the masks are coloured and contain their original background, binary masks needed to be prepared to train the model. Any colourless pixels - those in which each of the RGB values are equal - were reassigned a value of 0 and all other pixels were assigned a value of 1. The images which were output by the model were of the same format.

Previous work has shown that U-Net works optimally on images which have been patched into smaller images of size 256x256 pixels (Vali, 2024). Experimentation with using the whole images vs the patched images showed that this does indeed lead to a significant improvement.

This also greatly increases the size of the dataset, which is a positive thing. However, many of the image and mask patches did not contain any biofilm (being from the background). Removing these manually would have been very time consuming so a Python script was used to automate their removal, using edge detection with cv2. In total 1317 blank patches were removed and 993 non-blank patches were kept, halving the training time needed.

Two metrics were used to evaluate the performance of the model: the accuracy and the dice score. The accuracy is found by comparing each predicted pixel to the ground truth mask, summing the correct pixels and dividing by the total number of pixels. This measure of accuracy tends to be very high due, since the majority of both the predictions and the ground truths is black background. The dice score

metric is more sensitive and therefore provides a clearer sense of the performance of the model (Persson, 2021).

The code for the U-Net (including the model architecture, training module, data retrieval and associated utilities) in both the binary classification task and the wrinkle segmentation task is adapted from Persson (2021).

2.3.3 U-Net for Image Segmentation

There are two approaches taken for the second task of generating masks of the different wrinkle types. These are image-to-image translation and multi-class classification.

The former takes the RGB image and mask as input and outputs an RGB prediction. The model is provided with the full information from the annotation, which is unaltered, and the output is not restricted to certain values: every pixel in the output prediction could be any colour.

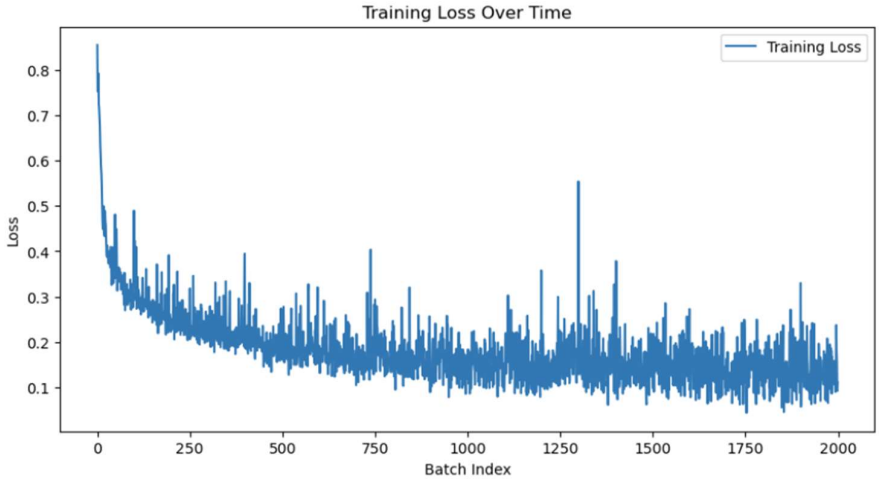
By contrast, multi-class classification requires that values 0-5 be assigned to the pixels in the mask (0 for background, 1-4 for the relevant wrinkle types). The prediction is therefore limited to these values as well, which are converted back to represent just 5 colours.
It was not known which of these approaches would work best. Image-to-image translation provides the model with more information, while multi-class-classification restricts the space of possible outputs. Each of these could provide an advantage. Both attempts were tried and the results are presented in Section 3.

The evaluation metrics for these tasks also differ, making quantitative comparison impossible, though their performances can broadly be compared by looking at the outcomes of each method. In the case of image-to-image translation the structural similarity index measure (SSIM) was used, while the multi-class classification was evaluated using ‘mean Intersection over Union’.

2.4 Optimisation of hyperparameters

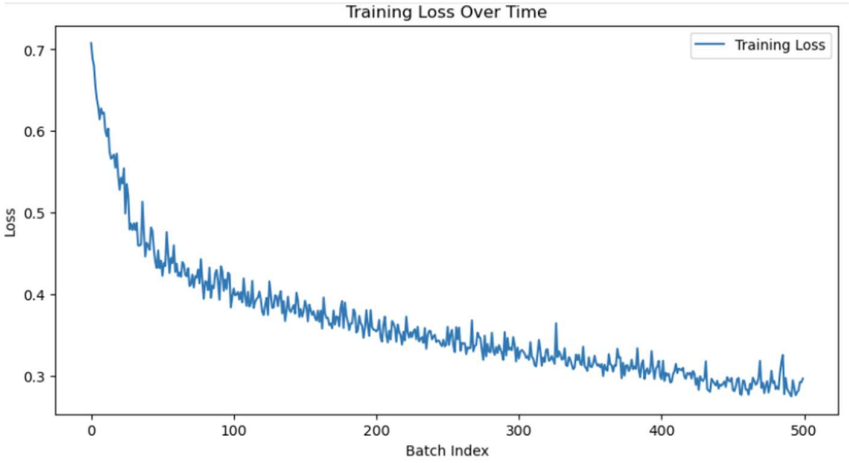
Hyperparameters are parameters used in ML algorithms which are not set by the algorithms themselves. That is they have to be set by the programmer. Hyperparameters are then varied manually by the user, or using an optimisation algorithm, to find the optimum values. In this context, this means that the upper and lower percentiles were varied, along with several other parameters and the effect on the accuracy of the trained model was observed.

A Python module - Optuna - was employed to automate the trial and error process of loss reduction. Fig 6 shows loss curves from two different sets of hyperparameters which have been trialled using Optuna on the binary classification task. It is clear that the set assigned a higher value (Fig 6.1) by Optuna minimises loss more effectively than the second set of hyperparameters shown in Fig 6.2.




```
[I 2024-07-29 21:56:35,014] Trial 9 finished with value: 20.256471633911133 and
parameters: {'batch_size': 8, 'num_epochs': 20, 'num_workers': 5, 'lower_percent
ile': 5.199122691610939, 'higher_percentile': 98.57804272626836, 'feature_1': 12
8}. Best is trial 3 with value: 20.390377044677734.
{'batch_size': 8, 'num_epochs': 20, 'num_workers': 8, 'lower_percentile': 8.8808
50013812923, 'higher_percentile': 98.57980601412677, 'feature_1': 64}
```

Fig 6.1. Plot of loss of the U-Net model on the binary classification task against the total number of batches trained for. The graph depicts the results from the 9th trial in the Optuna optimisation process, with a value of over 20. The plot is produced in Python using the package Pyplot from matplotlib.




```
[I 2024-07-29 21:48:49,887] Trial 5 finished with value: 10.387598037719727 and
parameters: {'batch_size': 16, 'num_epochs': 10, 'num_workers': 6, 'lower_percen
tile': 5.824905917691078, 'higher_percentile': 93.67315273214236, 'feature_1': 3
2}. Best is trial 3 with value: 20.390377044677734.
```

Fig 6.2. Plot of loss of the U-Net model on the binary classification task against the total number of batches trained for. The graph depicts the results from the 5th trial in the Optuna optimisation process, with a value of just over 10. The plot is produced in Python using the package Pyplot from matplotlib.

Given more time, the optimisation function could be run over the course of several days, running hundreds of trials, and therefore potentially improve on the existing optimum hyperparameters.

3. Results

3.1 Binary Classification

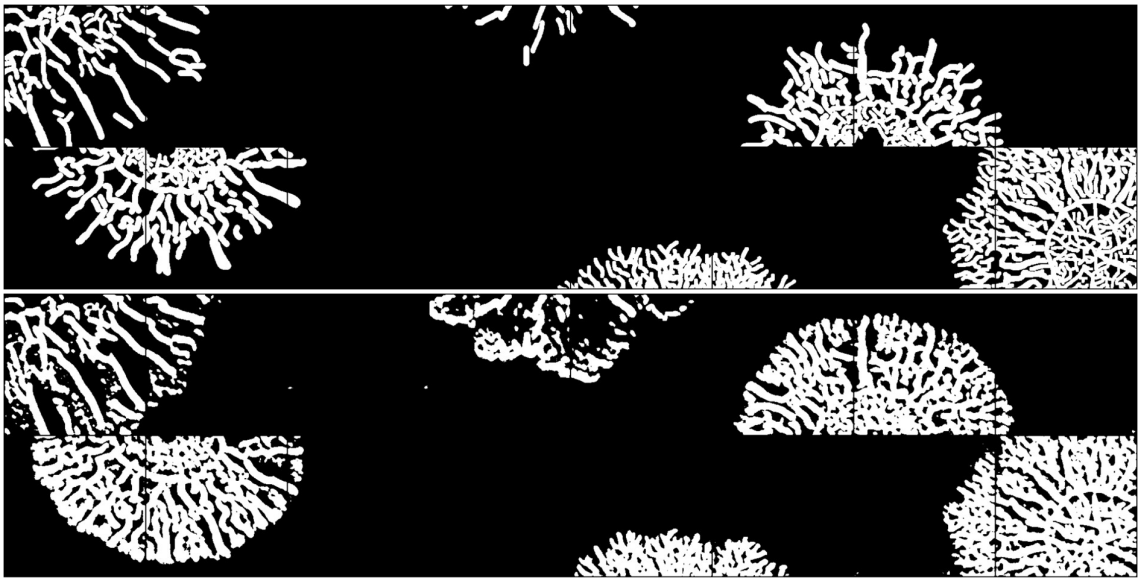


Fig 7. Predictions from the binary classification task (bottom) and the corresponding input ground truths (top)

With only limited training and optimisation, the U-Net model successfully predicted wrinkle masks for a variety of patches with good accuracy. It is observed that the model had a tendency to predict wrinkles where there are none in the hand-annotated masks. This is likely due to the contrast treatment of the original images; where large thresholds for the contrast stretching are used, it is likely that some details of the images are lost, causing this overfilling effect. This is particularly noticeable in the patches in the top-middle where the predicted mask contains many more wrinkles than the ground truth.
It is also observed that the predictions contain more small 'wrinkles' at the edge of the biofilms than the ground truths. It is possible that these are actually not wrinkles but the edge of the biofilm, which is being mis-detected by the model.

3.2 Image-to-image translation

The image-to-image translation approach did not produce results of any use. U-Net has been shown to work optimally on images of size 256x256 (Vali, 2024), so the initial approach of patching was employed, as with the binary classification. When this failed to produce useful results, the images were instead processed unpatched. This approach had the potential to lead to better results as the distinction between different wrinkle types comes from both their spatial organisation and their form. However, it significantly reduced the size of the dataset.

The SSIM for this prediction finished at 0.8063749670982361, where similar images would have a score closer to 1. Given that the U-Net was attempting to learn more complex patterns than in the binary classification task, it is not surprising that this led to poor performance. With a larger dataset, the image-to-image translation method may well be viable.

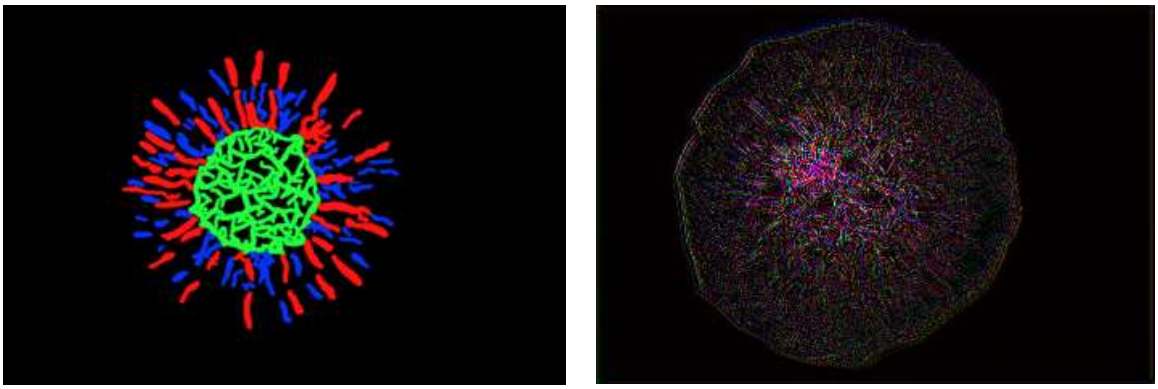


Fig 7. An example ground truth image (left) with corresponding image-to-image translation prediction (right). Ground truth image produced from a hand-annotated mask by Obaid (2024).

3.3 Multi-class classification

The multi-class classification method was adopted in response to the results of the image-to-image translation method, which frequently predicted colours - such as purple - which did not appear in the ground truths. In this case, the model had three output channels - RGB - which could be combined in any ratio. Therefore, one possible explanation for the presence of purple is that, in areas in which the model could not distinguish between red and blue wrinkles, a high intensity for both red and blue channels was predicted for the same pixels, combining to give purple.

Multi-class classification restricts the solution to include only five possible colours - black for the background and the four wrinkle colours. The images used were not patched, as before, in order to preserve information about spatial organisation.

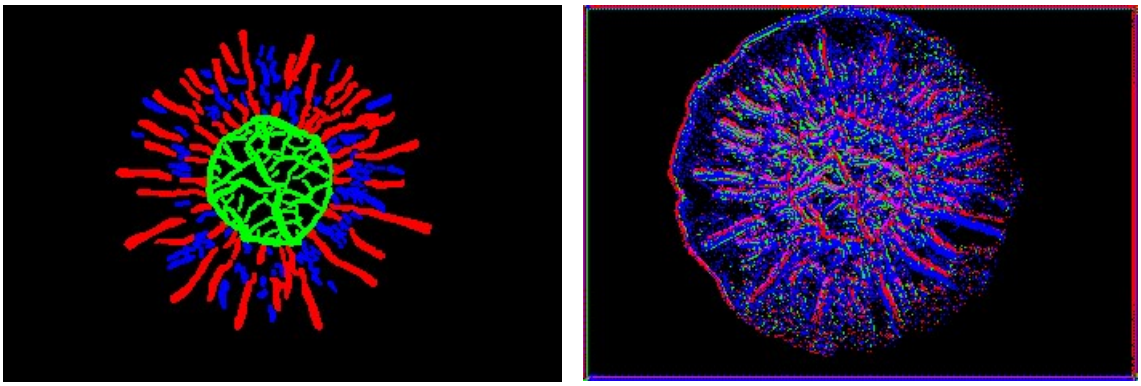


Fig 8. An example ground truth image (left) with corresponding multi-class classification prediction (right). Ground truth image produced from a hand-annotated mask by Obaid (2024).

The results were not useful, showing a seemingly random arrangement of colours across the biofilm, which varied considerably across different trials.

However, some promise was shown when the problem was reduced back to a binary classification problem, with the model instead recognising only Type 2 (red) wrinkles. This may indicate that with a

larger training set that U-Net is capable of distinguishing between different wrinkle types, one at a time. Nevertheless, I fear that the partial success of this may result from the distinct form of the Type 2 wrinkles - which are longer, thicker and straighter than other wrinkles; I do not believe that U-Net is the best choice for distinguishing between wrinkles based on spatial organisation. There are many more recently developed alternatives which fuse U-Net with alternative architectures to better enhance the model's ability to take into account local and global context of features (Wang et al., 2024, pp. 1-3).

There was also limited success in identifying Type 1 (green wrinkles), with predictions consistently identifying the central region as green but with very little precision in wrinkle identification. It appears that U-Net is not well suited to the joint demands of high- and low-level feature identification.

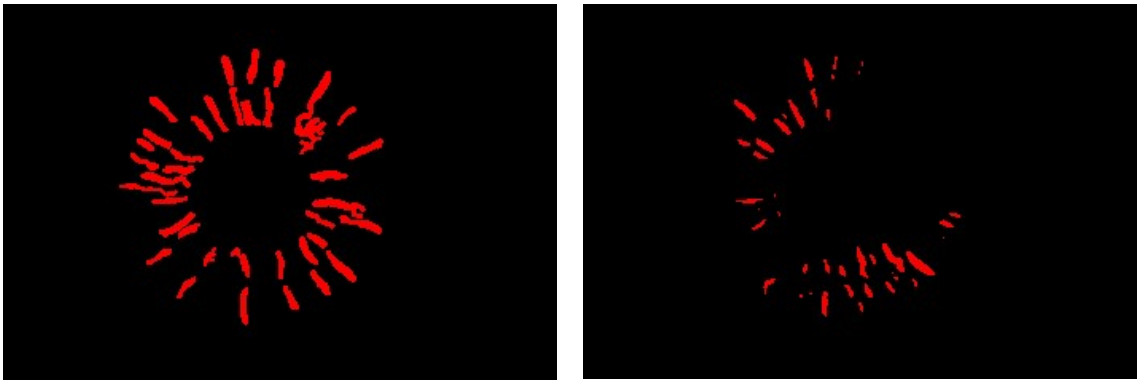


Fig 9. Type 2 wrinkle annotations ground truth (left) with corresponding predictions (right). Ground truth image produced from a hand-annotated mask by Obaid (2024).

Fig 10. Type 1 wrinkle annotations ground truth (left) with corresponding predictions (right). Ground truth image produced from a hand-annotated mask by Obaid (2024).

3.4 Methods of Analysis

It has been shown that binary masks of biofilm wrinkles can successfully be generated by a U-Net model, confirming the work of Cao et al. (2024) As demonstrated in their paper, Sholl analysis can be

utilised as an alternative method of distinguishing between different wrinkle ‘types’, though these are defined purely morphologically in Sholl analysis.
Sholl analysis is a method for determining wrinkle density (Binley, 2014). Concentric circles are drawn onto the wrinkle mask, and the number of intersections of the wrinkles with each circle are counted, summing to give a score for that circle and hence an indication of wrinkle density at that radius. The Cao team performed this Sholl analysis on mature biofilms (2014), much older than those in our dataset, which show different wrinkle morphology.

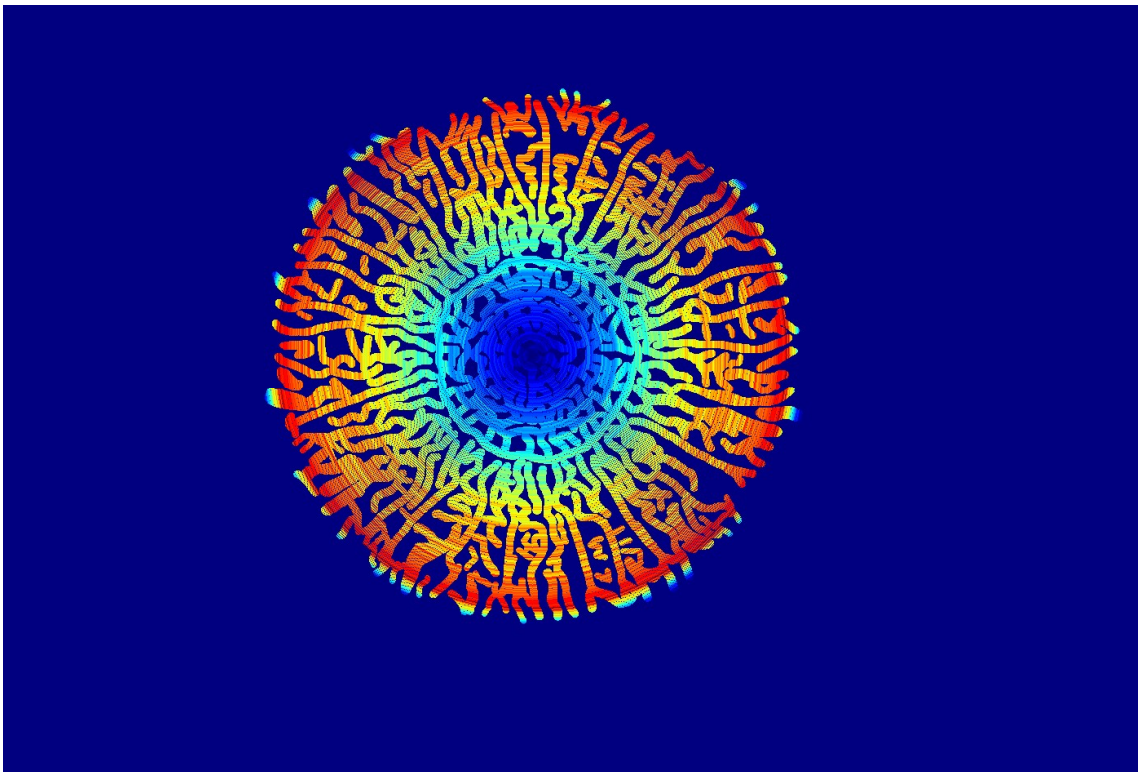


Fig 11. A binary mask of a biofilm is colourised using the values from the Sholl analysis. Cooler colours indicate lower wrinkle density while warmer colours indicate higher wrinkle density. Plot produced in Python using Google Colab’s patch for cv2.

In the use of the younger biofilms from the Fusco lab dataset, the Sholl analysis plot is not so useful in distinguishing different regions of wrinkles, as the increase in the number of intersections is generally quite steady. Considering instead the colour map it may be possible to define the dark blue, green, orange and red regions as separate wrinkle types.

However, the sole use of Sholl analysis to distinguish between wrinkles is a flawed method as it makes no distinction between wrinkles that occupy several of these regions - such as Type 2 wrinkles. Type 2 wrinkles have been shown to have a different functional role, but Sholl analysis fails to distinguish them from Type 3 wrinkles which occupy the same region(s).

Nevertheless, Sholl analysis can be an interesting metric by which to compare biofilms. This application is not possible without the use of a binary wrinkle mask which can be hand-drawn or generated using a U-Net, as demonstrated in this report.

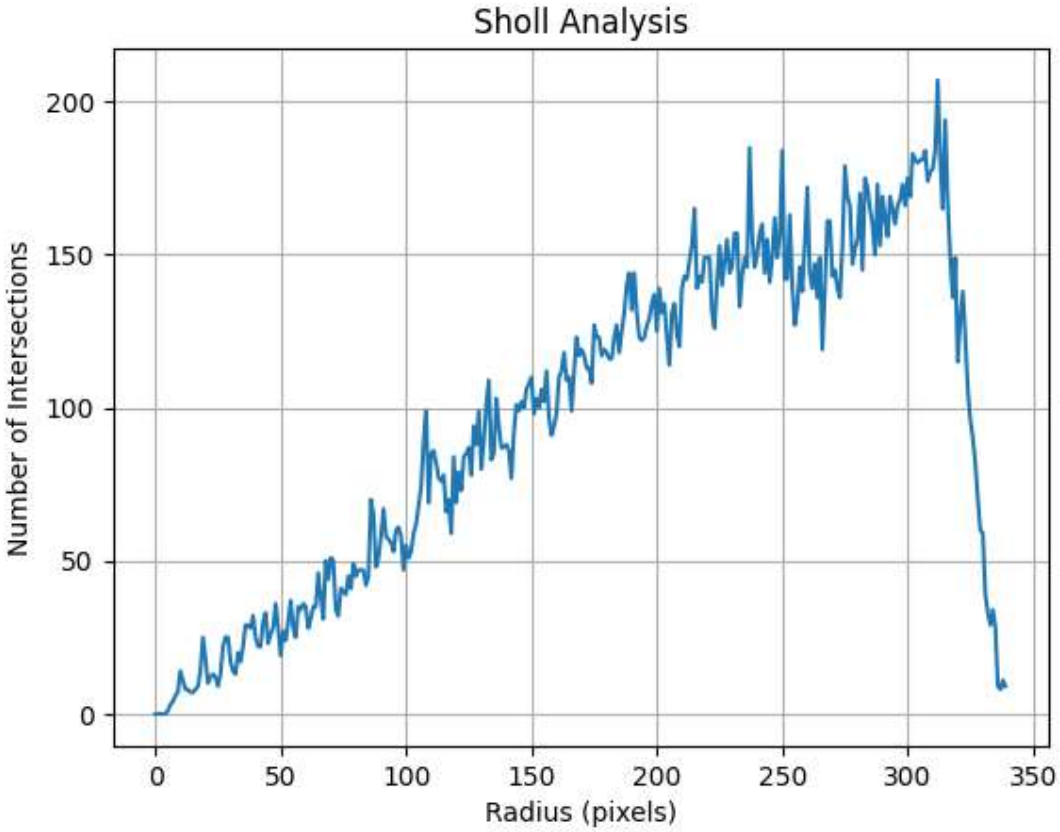


Fig 12. A plot of the Sholl analysis for the same biofilm mask as in Fig 11. The number of intersections of circles of different radii with the wrinkle mask are calculated and plotted against the corresponding radii. Plot is produced in Python using Pyplot from matplotlib.

A further application for these wrinkle masks is analysis of biofilm phenotype data inside and outside of the wrinkles. For example, consider a plot of average intensity of a fluorescent image at specific radii against the radii, as inspired by the work of Yulin Du (2024). These plots can be generated for every time-step in a timelapse to analyse the change in intensity across space and time. By applying a wrinkle mask to the image, this average intensity can be calculated inside or outside of the wrinkles and the phenotype behaviour inside vs outside wrinkles can hence be quantitatively compared. Note that the plot in Fig 13. is produced using a binary mask generated from a hand-annotated image (Obaid, 2024), in order to demonstrate the results possible with a high accuracy mask.

The plot reveals that motile phenotype expression varies with distance from the centre in broadly the same way both inside and outside the wrinkles. However, apparent differences in intensity in dual reporter images inside and outside wrinkles has been used to evidence the idea that phenotype distribution and wrinkle morphology are intrinsically linked (Fusco, Pestana, Tropini, 2021, p.1). Therefore this method of analysis could have promising results when applied to a timelapse, and to biofilm colonies grown in varying conditions.

If different wrinkle types can be successfully segmented then intensity analysis of these types individually could also provide some interesting insights.

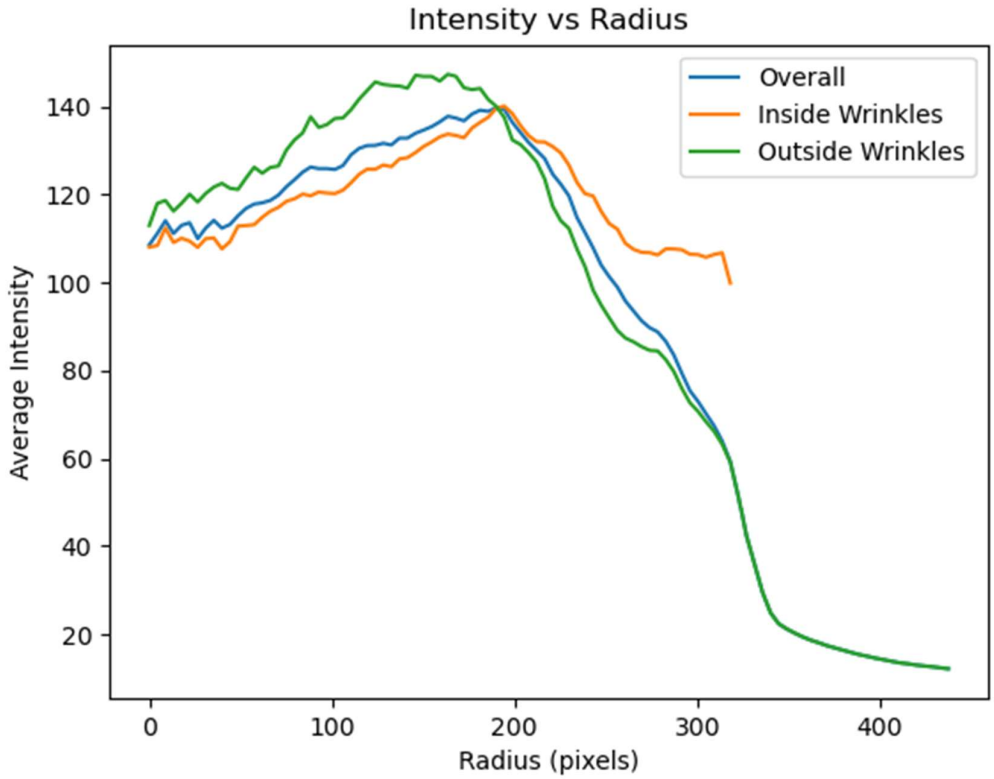


Fig 13. Intensity analysis on fluorescent images of biofilm, including inside and outside the wrinkles. Plot is produced in Python using Pyplot from matplotlib.

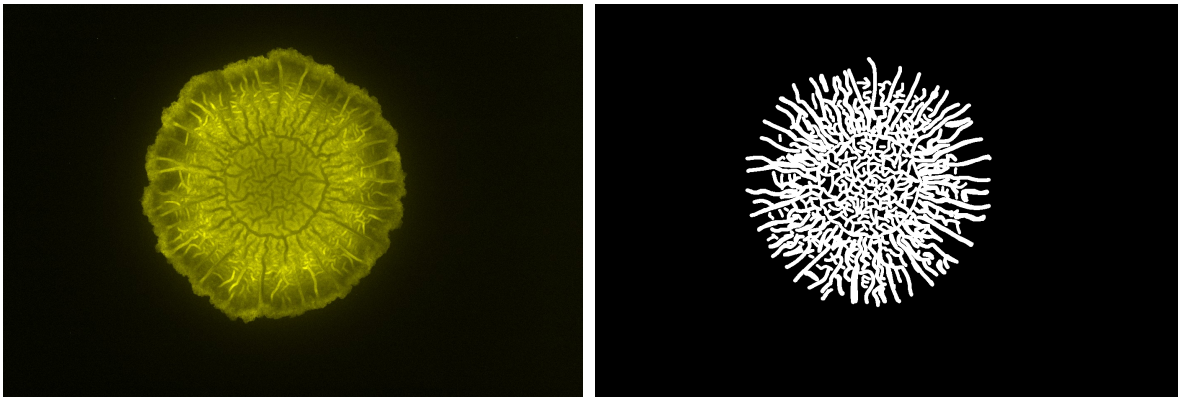


Fig 14. Fluorescence image (left) used for the intensity analysis in Fig 13 and the binary mask applied to analyse inside and outside the wrinkles. Biofilm image by Krishnan (2022) and binary mask produced from hand annotated image by Obaid (2024).

4. Conclusion

In summary, the first aim – to create a method of isolating wrinkles within a biofilm – has been accomplished to a good standard. With further training of the model and optimisation, this provides an efficient and accurate method of extracting the wrinkles for further analysis. The data preparation has

been automated as much as possible so that the dataset can be extended – or indeed so that the code may be reused on a different dataset.
A system of different wrinkle types based on their morphologies, functional and biological differences was introduced. U-Net was trained to segment the brightfield images into these types with little success, despite the use of two alternative methods: image-to-image translation and multi-class classification. It is suggested that use of an alternative model architecture – such as a Vision Transformer or hybrid model – may have better handling of longer-range dependencies, necessary to preserve the spatial context of the different wrinkle types. Two methods of analysis are performed on binary masks, including Sholl analysis. The second method introduces a way of analysing the variation of phenotype intensity across space and time inside and outside wrinkles separately.

5. Acknowledgements

The code used for generation of both binary and segmented masks (the implementation of the U-Net architecture and the associated dataset, utilities and training scripts) was adapted for use from the work of Persson (2021).

OpenAI's ChatGPT tool was used to generate parts of the code used for the above tasks, as well as used in the development of scripts used for data preparation and analysis. It was also used to identify where errors in the code may lie, as well as to suggest the most efficient or appropriate approaches to implementing the code. A history of all relevant parts of ChatGPT conversations will be provided upon request.

The images used in the training and validation datasets were produced by various members of the Fusco Lab in the Biological Soft Systems Group, Cavendish Laboratory, University of Cambridge. These members include: Nikhil Krishnan, Abhirup Mookherjee, Akhilesh Kumar Verma, Yulin Du, Anna Tarodi and Racha Majed.

Several development environments have been used when writing and editing the Python scripts used for this project. These include Visual Studio Code, Google Colab and Jupyter Lab. When developing locally in VS Code, Python version 3.9 was used.

Python packages used: cv2 for ComputerVision and image-handling (and Google Colab patches's version of cv2); numpy for numerical operations; torch, torch.nn, torch.optim, torch.utils.data and torchvision for machine learning; Pyplot from matplotlib for plotting and displaying graphs; aicspylibczi for czi file handling; PIL for image handling; scipy.ndimage for finding the centre of mass of images in the intensity analysis; imageio.v2 for image handling; time; random; os and shutil for file-handling; tqdm to track training progress; optuna for optimisation; json for reporting parameter values and skimage.metrics for SSIM.

Both Mamba and Conda were used as package and environment managers.

6. References

Binley, K. E. et al. (2014) 'Sholl analysis: A quantitative comparison of semi-automated methods', *Journal of Neuroscience Methods*, 225, pp. 65-70.

Cao, X., et al. (2024) 'A new method: Characterize and quantify biofilm wrinkles by U-Net and Sholl Analysis', *Science Direct*, 237, p105131. Available at: https://doi.org/10.1016/j.biosystems.2024.105131

Du, Y. (2024) *Group Meeting Presentation* [Presentation]. Cavendish Laboratory, University of Cambridge. July.

Fusco, D., Pestana, L. R., Tropini, C**.** (2021) *Early Career Application research proposal*. Unpublished research proposal. Cavendish Laboratory, University of Cambridge, College of Engineering, University of Miami, Department of Microbiology, University of British Columbia. Krishnan, N., Knight, J., Tropini, C., Ruiz Pestana, L. and Fusco, D**.** (2024) *The what, when, where, and why of wrinkly morphology in biofilms*. Unpublished research review. Cavendish Laboratory, Department of Physics, University of Cambridge, UK; Department of Microbiology and Immunology, University of British Columbia, Canada; School of Biomedical Engineering, University of British Columbia, Canada; Humans and the Microbiome Program, Canadian Institute for Advanced Research, Canada; Department of Civil and Architectural Engineering, University of Miami, USA.

GIMP for Windows (2024) (Version 2.10.38) [Computer program]. Available at: https://www.gimp.org/downloads/ (Accessed: 5th July 2024).

Gingichashvili, et al. (2022) 'An open-source computational tool for measuring bacterial biofilm morphology and growth kinetics upon one-sided exposure to an antimicrobial source.' *Scientific Reports*, 12(1) p16125. doi: https://doi.org/10.1038/s41598-022-20275-8.

Hall-Stoodley, L., Costerton, J. and Stoodley, P. (2004) 'Bacterial biofilms: from the natural environment to infectious diseases', *Nature Reviews Microbiology*, 2, pp. 95-108. Available at: https://doi.org/10.1038/nrmicro821

Krishnan, N. (2022) 'Untitled' *[Timelapse of brightfield and dual-reporter images of radial biofilm colonies].* University of Cambridge, Cavendish Laboratory, Biological Soft Systems Group. Unpublished photographs.

Obaid, M. (2024) 'Hand-annotated image masks'. Commissioned work, Fiverr.

Persson, A. (2021) *PyTorch Image Segmentation Tutorial with U-NET: everything from scratch baby*. 2 February. Available at: https://www.youtube.com/watch?v=IHq1t7NxS8k (Accessed: 9 July 2024).

Ronneberger, O., Fischer, P., Brox, T. (2015). 'U-Net: Convolutional Networks for Biomedical Image Segmentation'. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

Vali, M. (2024) Conversation with Britt Dewing, July.

Wang, Z., et al. (2024) 'Mamba-U-Net: U-Net-Like Pure Visual Mamba for Medical Image Segmentation '. arXiv preprint arXiv:2402.12345. Available at: https://doi.org/10.48550/arXiv.2402.05079.