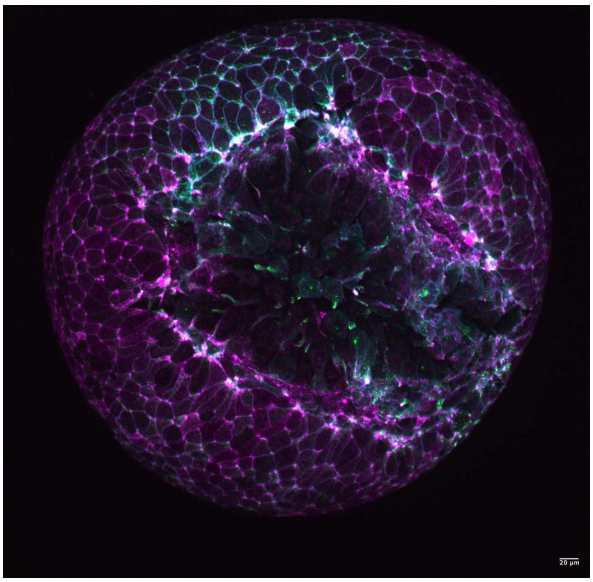
Københavns Universitet - Bachelorstudiet i fysik

Førsteårsprojekt 2025

Projekt nummer: 2025-15

Wound Healing



Forfattere:

Emil Pind	KU- ID: hzk585
Noah Merrild	KU- ID: cls766
Lau Holck	KU- ID: bzl287
Luca Koushede	KU- ID: vqk505

Vejledere:

Ala Trusina	Email: trusina@nbi.ku.dk
Jakob Schauser	Email: jakob.schauser@bi.ku.dk

Abstract

When an organism is wounded, the cells surrounding the injury move in order to close the wound. This project intends to determine the presence of wounds by analyzing the location of proteins within cells. Two analytical methods are developed to explore this: Using Principal Component Analysis and analyzing the angular placement of proteins within cells. Principal Component Analysis yields statistically significant results, proving the connection between wounds and the location of proteins within nearby cells. Angular analysis fails to statistically prove this connection, but the method still proves interesting, as it could provide a simpler representation of the data. Both methods could benefit from larger sets of data. This could make the angular analysis more valuable. It could also improve Principal Component Analysis, as the methods way of classifying cells could be improved.

Contents

Abstract **ii**

Introduction **1**

Method **2**

Principle Component Analysis 2

Angle to center 6

Results **8**

Principal Component Analysis 8

Angle to center 11

Discussion **12**

Common cause of error 12

Discussion of results from Principal Component Analysis 13

Direction of the wound 13

Deciding on a magnitude of polarization cut-off value 13

Discussion of results from angle to center 14

Comparison of the two methods and results 14

Conclusion **15**

Taking the project further 15

Appendix **16**

References **17**

Introduction

When an organism is cut, the defence system of the organism initiates a multitude of processes to heal the injury. One of these being that the nearest layers of cells move towards the wound, closing it. (Staddon u. a., 2018).

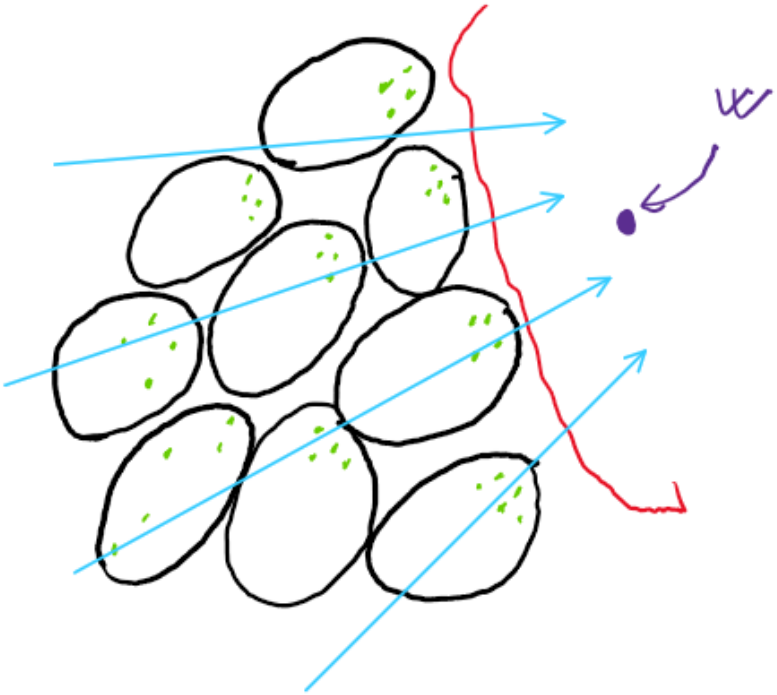


Figure 1: A general example of the placement of proteins polarizing a cell. The blue arrows demonstrate the overall movement of the cells. The purple dot is a point representing the location of the wound.

Along the direction of cell movement, a polarization of the proteins within the cells has been observed. This polarization is characterized by the protein $Arp\frac{2}{3}$ collecting at poles along an axis through the cell towards the wound (Staddon u. a., 2018). The correlation between the polarization of proteins and cell movement allows for quantifying the movement of cells, thereby also allowing the quantification of wound healing, by analyzing the location of proteins within cells. By analyzing images of cells, where it is known whether or not a wound is present, we have developed analytical methods in Python, which are capable of analyzing a given image of cells and concluding whether or not cells within the image are polarized, thereby concluding if a wound is present nearby. Two methods have been developed throughout this project: Finding the protein's position in terms of angle from the cell center, and using a modified version of Principal Component Analysis to find the axis of highest variance of proteins within cells relative to the cell center.

Method

The first step in analyzing the data - for both methods - is to label the proteins by using a mask that has each pixel labeled with a whole number corresponding to a given cell. The proteins were labeled by converting the protein channel to a binary picture, where if pixels contained proteins they were given value 1, and otherwise 0. This binary picture is then multiplied with the aforementioned mask, giving each protein a label corresponding to its cell. The making of the binary picture introduces an important factor; a threshold, that, based on a pixels intensity, determines whether or not a given pixel should have value 1 or 0, and therefore be labeled as a protein. This reduced the noise in the images. The threshold was determined by looking at a histogram of the pixel intensity in the image, where a fitting threshold became apparent. The mask for each cell is also used to calculate the center of each cell, in x -and y-coordinates by:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

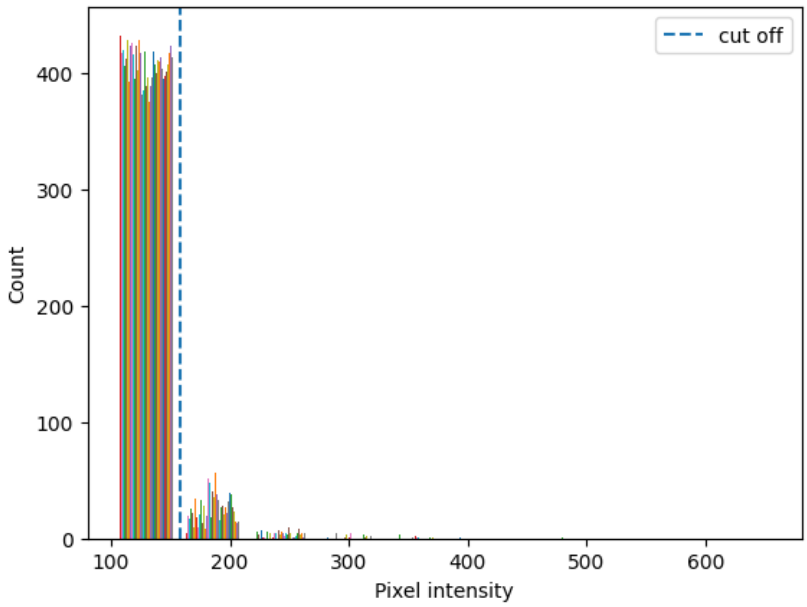


Figure 2: Example of histogram for a control picture

Principal Component Analysis

One of the methods for determining polarization we devoloped is through applying Principal Compononent Analysis. Principal Component Analysis, PCA for short, is a tool used mainly for representing large-dimensional data in fewer dimensions (Connelly, 2021). The application of PCA happens, however, to be incredibly useful

in the case of polarized cells, despite the data being 2-dimensional, consisting of x- and y-coordinates. PCA works by determining the axis along which the variance of data is greatest, and then placing a second axis orthogonally on the first. Our use of PCA results in two eigenvectors and two corresponding eigenvalues. The eigenvectors represent the two axis of PCA, the one of most variance and orthogonally on that axis. The eigenvalues represent the magnitude of variance in the direction of the corresponding eigenvectors.

Determining when a cell is polarized

In a non-polarized cell, the two eigenvalues from the PCA would be roughly equal, as there is no axis of significantly larger variance. In a polarized cell, the first eigenvalue is expected to be bigger than the second eigenvalue relative to the magnitude of polarization. Therefore, we have defined the magnitude of polarization as:

$$m_{polarization} = \frac{\lambda_1}{\lambda_2}$$

Where λ_1 is the first, larger eigenvalue, and λ_2 is the second, smaller eigenvalue. Deciding a cut-off value between polarization and non-polarization enables counting the number of polarized cells in a given image and then comparing this count to the total number of cells in the image, which is useful for determining the overall polarization within the image. Having a metric for the magnitude of polarization within each cell also allows us to quantify the magnitude of polarization in relation to a cell's distance from the wound.

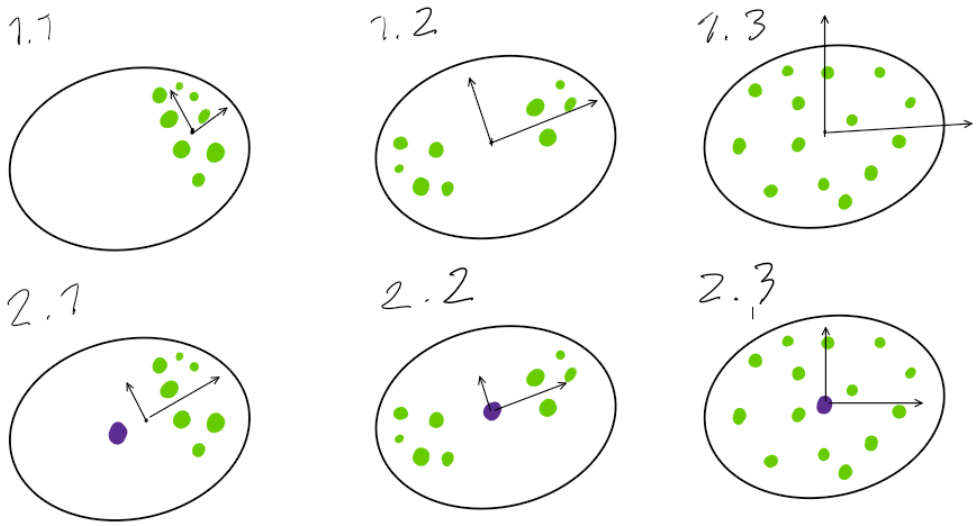


Figure 3: Illustration of 3 different protein distributions in cells and the eigenvectors from a Principal Component Analysis. Row 1 and 2 before and after the addition of data points in the center respectively. Cell 1 is polarized, but this is only seen through PCA after the addition of data points. In cell 2 and 3, PCA still results large $m_{polarization}$ for polarized cells, and small $m_{polarization}$ for non-polarized cells. This figure also shows how the center of the cell, purple dots, and center of proteins, where the eigenvectors start, overlap in cases that are not of the same nature as cell 1.

Shortcomings of PCA and work-arounds

When using PCA the first eigenvector is placed along the direction of most variance. There are cases, however, where the protein distribution is concentrated at only one end of the cell. See cell 1.1 on figure 3. In this case, PCA is unable to determine the direction of most variance relative to the cell, as the variance in the distribution of proteins is not guaranteed to describe the direction of polarization; see again cell 1.1 figure 3. The resulting eigenvector will instead be pointing in the direction of most variance within the small clump of proteins, which is of little to no value when determining the polarization of the entire cell. Furthermore, the first eigenvalue will be small and is unlikely to be significantly larger than that of the second, orthogonal eigenvector. The resulting $m_{polarization}$ would then be similar to when the cell shows no polarization, and thus these cases would result in false negatives. To combat this shortcoming, we developed a method: Placing data points in the center of the cell equal to the total amount of proteins within the cell. This method ensures that, in cases where the proteins are located at only one end, applying PCA will still yield a larger eigenvalue along the direction of the cell's polarization, as there will now be a great variance from the center of the cell to the clump of proteins. See cell 2.1 figure 3. Adding data points equal to the number of proteins ensures that cells are not unequally affected by this process. In non-polarized cases after the added data

points, applying PCA will yield a similar, or slightly smaller, $m_{polarization}$, as both eigenvalues would be lowered by a similar scale. In non-polarized cells and cells, where the polarization is at both ends of the cell, the addition of data points in the middle of the cell reduces both eigenvalues, as the center of the cell and center of the PCA will be roughly the same. Having more data in the center of the PCA, reduces the variance. This change, however, we take into account when deciding on the cut-off value between polarization and non-polarization. In the final case, where the proteins are polarized at only one end of the cell, the addition of data points in the center of the cell will drastically increase $m_{polarization}$ thus changing these cases of would-be false negatives to positives, improving the overall separation of cells into polarized and non-polarized.

Quantifying polarization in relation to distance from wound

We additionally used PCA to analyze the correlation between polarization of cells and the distance from the wound. Here, a point is manually selected as the location of the wound. See figure 1 . The distance from the center of each cell to this point is calculated. We create an array containing the ratio $m_{\text{polarization}}$ for each cell. This array is sorted such that $m_{polarization}$ related to the closeness of the related cell in descending order. We then chose a cut-off for the ratio that we will characterize as polarized. The set of ratios is rebinned. The standard error on the mean is used to calculate the uncertainty on each rebinned data point (Barlow, 1989). This uncertainty is used in the linear fit. The covariance matrix of the fit is used to calculate the standard error for the parameters of the fit. The uncertainty from rebinning the data is the only uncertainty used as we have no reasonable measurement of the uncertainty on $m_{polarization}$ for each cell.

Statistics and Principal Component Analysis

To determine whether the results produced by PCA are statistically significant, we used two methods: t-test (of Standards und Technology, n.d.) and Bootstrapping (VanderPlas, 2016).

A t-test can be used on data that resembles a Gaussian-distribution. It uses a t-score, a measure of the difference of sample means in terms of standard error, and states whether this difference is statistically significant. We used Welch's t-test because our sample sizes and variances differ between datasets. The t-score, t, is given:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

In Welch's t-test, where N is sample size, $\bar{x}$ is mean and s is standard deviation.

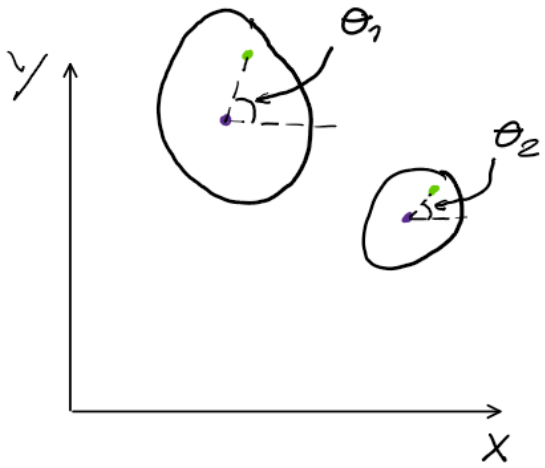


Figure 4: The measurement of the angles to the proteins in 2 cells.

For the t-test, we also calculate degrees of freedom, ν , which is done by:

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2\nu_1} + \frac{s_2^4}{N_2^2\nu_2}}$$

In this case, we use bootstrapped data for the t-test. Bootstrapping is a process where random numbers are taken from a sample. The mean value of the random numbers is calculated. This process is repeated, and by the CLT, it results in a normal distribution with a mean and variance related to the sample (VanderPlas, 2016). This is why we use bootstrapping as the data we have been given does not resemble a Gaussian enough. The mean and variance is then used to calculate the t-score and give a p-value. For a p-value below 0.05 the difference between the two sample means are judged to be statistically significant (DATAtab, n.d.).

Angle to center

For describing the polarization of cells, we have also developed a different method. This method defines not when a cell itself is polarized, but rather collects each protein's position related to the center of the respective cell, to determine whether or not the image as a whole indicates general polarization.

This method relies on simply analyzing the image as an xy-grid. For each cell the associated protein coordinates is used to calculate an angle θ. See figure 4 . The angle is calculated using the numpy function arctan2. This function takes in a y - and x-value and returns a value between $-\pi$ and π radians. To get the correct x and y values you take the difference in each axis from the cell's center coordinates.

$$\Delta x = x_{protein} - x_{center} \quad \text{and} \quad \Delta y = y_{protein} - y_{center}$$

When a wound is not present the surrounding cells are expected to show little to no signs of polarization, and the angles are hereby assumed to be distributed equally, thus resembling an uniform distribution. This can be represented visually in a polar histogram collecting the angles from all the cells in an experiment.

Our null-hypothesis is that all images, with or without wound, have a uniform spread of proteins. The intention being that only images of polarized cells would reject the null-hypothesis.

This is then tested for images with and without wounds by calculating the χ^2 value between the histograms and a uniform distribution. This is done by the formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where O_i is the frequency - counts in each bin. E_i is the expected frequency, which is N/k given a uniform distribution, where N - is the total number of observations, angles, and k is the number of bins. (Barlow, 1989)

To determine if the null-hypothesis holds true or not, the p-value is calculated using the χ^2. For calculating the p-value the function scipy.stats.chi2, using a χ^2 value and the degrees of freedom, is used. The degrees of freedom are given by $k - number\,of\,parameters$.

Like when applying PCA, we also used this method to analyse the correlation between the degree of polarization and the cell's distance from the wound. For this method, the standard deviation was used to define how polarized the cell was. This was done by taking the standard deviation of the angles in a single cell. To make sure that the proteins on opposite poles had the same value or angle, pi was added to all the negative angles in the respective cells. Thus, a low standard deviation means that the proteins have gathered along a single axis. The distance from the wound was calculated, and rebinning was done in the same way as when applying PCA. However, along the y-axis, is instead how polarized the cells within that bin are on average.

Results

The resulting methods both take a Tagged Image File, TIF for short, with 2 changes as input. One channel corresponding to the placement of proteins within the image, and one corresponding to cell walls.
The methods also use a specific mask for each image, that has each pixel labeled with a number corresponding to a given cell. Ultimately, after trying different approaches, we ended up working with masks given to us by our supervisor, Jakob Schauser. The making of these masks was not within the scope of the project and is not covered here.
The data analyzed in this project is two large images with wounds, ”Large wound 1” and ”Large wound 2”, and control. The control, ”No wound”, consist of three smaller images all from the same experiment.

Principal Component Analysis

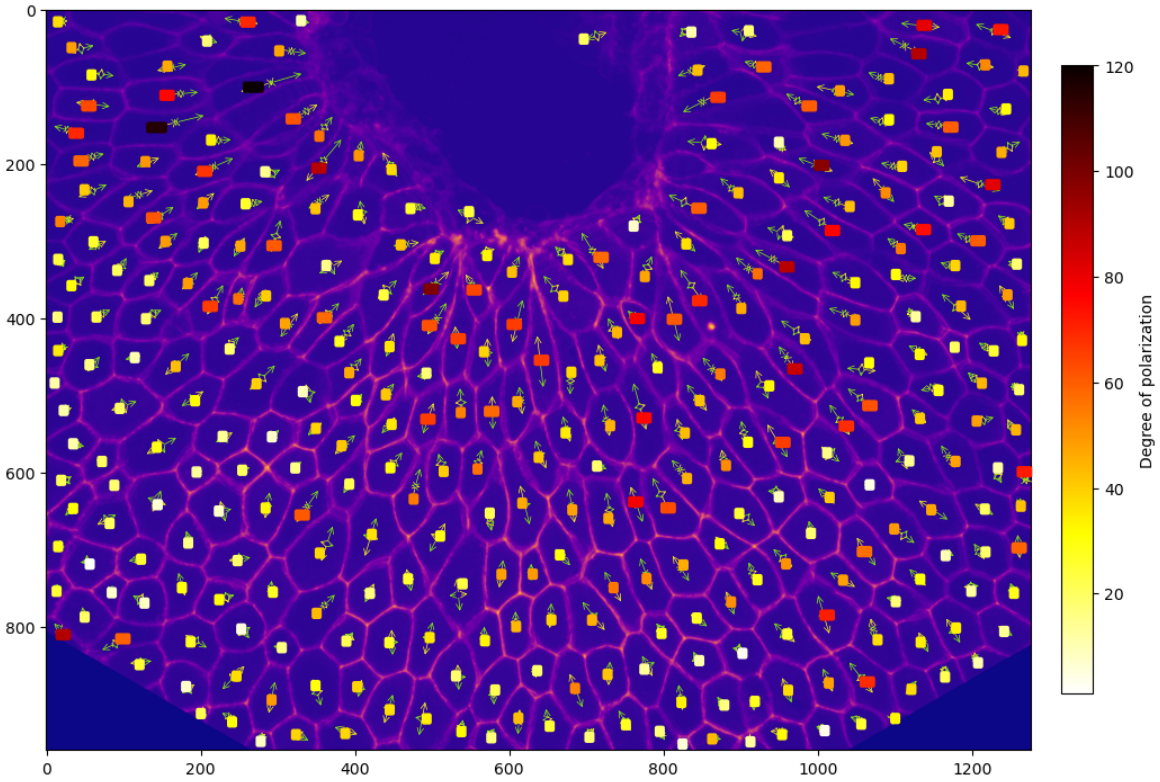


Figure 5: The result of a Principal Component Analysis on ”Large Wound 2”. The eigenvectors from the Principal Component Analysis are plotted as green arrows. Each cell is assigned a box that is colored based on the cell's $m_{\text{polarization}}$. The larger arrows, the first eigenvectors, point in the direction of the polarization and, therefore, the estimated direction of the wound.

For characterizing a cell as polarized or not, the threshold value of $m_{\text{polarization}}$ has been chosen to be 5. This cut-off is based on analyzing the results from applying PCA; this decision is further explained in the discussion.

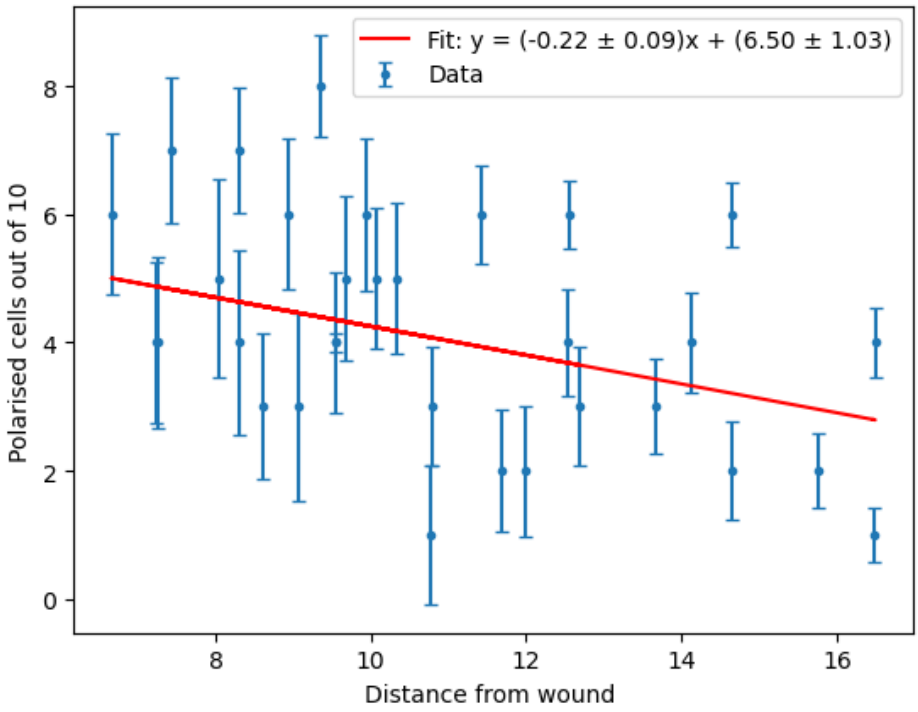


Figure 6: Negative correlation between distance from the wound and polarization of cells in figure 11. Data has been rebinned and the values on the y-axis is number of polarised cells out of 10. Binning 10 cells together was a qualitative decision that best illustrated the data.

The errorbars is the standard error on the mean, resulting from the rebinning of data. Notice that the slope of the fit shows that there is a statistically significant negative linear correlation between the distance from the wound and polarization of cells.

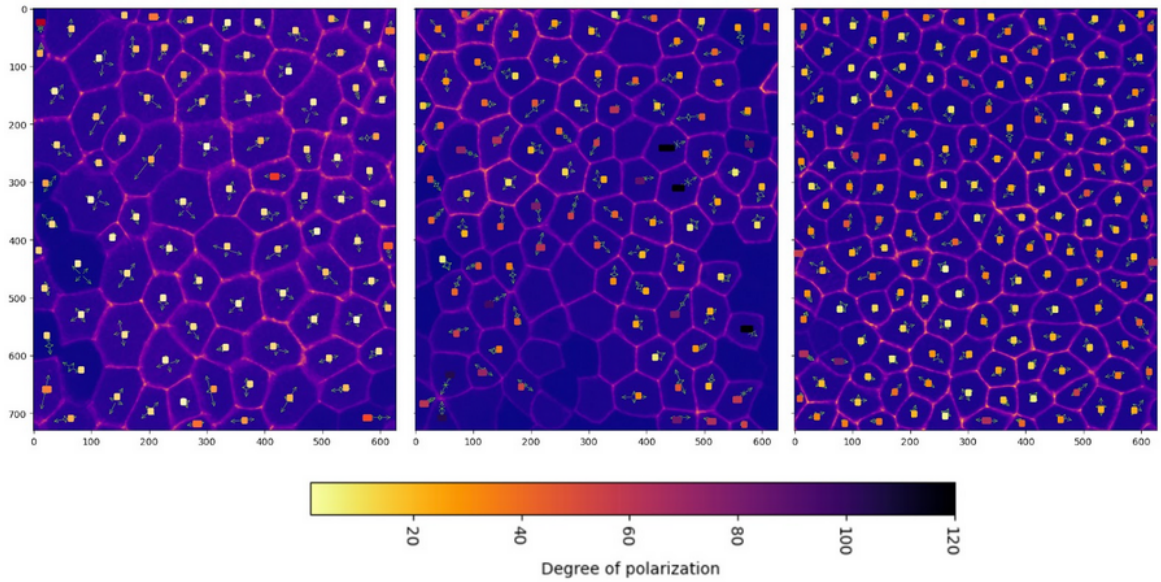


Figure 7: The result of Principal Component Analysis on the three control images without a wound. The same gradient was used for displaying $m_{polarization}$ as was used on figure 11

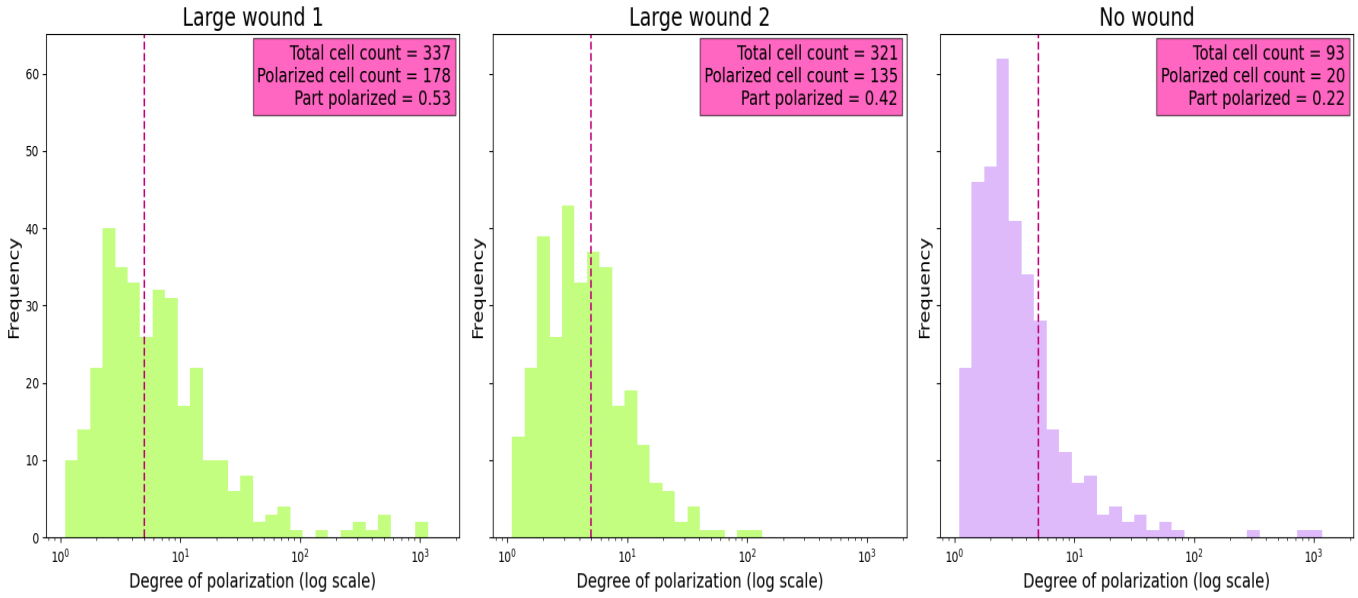


Figure 8: Histogram for each image of cells analyzed through Principle Component Analysis in the project. ”No wound” being the three control images. The histograms depict the $m_{polarization}$ for each cell of the images. The vertical line is the chosen minimum $m_{polarization}$ for characterizing a cell as polarized, 5.

By calculating the t-score and the p-value between each ”Large wound” distributions and the No wound distribution, it is clear that the data from the images with wound could not be drawn from the same distribution as the one without, as the p-values are below 0.05. As shown in the following table.

Results		
	Large wound 1 / No wound	Large wound 2 / No wound
t-score	215.99	111.57
p-value	0.0	3.7 * 10⁻14

Table 1: Comparison of the "No wound" data with Large wound 1 and 2 data, with t-score and p-value

Angle to center

The following histograms display the angles of proteins relative to the corresponding cell's center for each experiment. There is a clear distinction between the histograms that are associated with a wound, and thus is expected to show polarization, and the histogram that represents the test image (no wound).

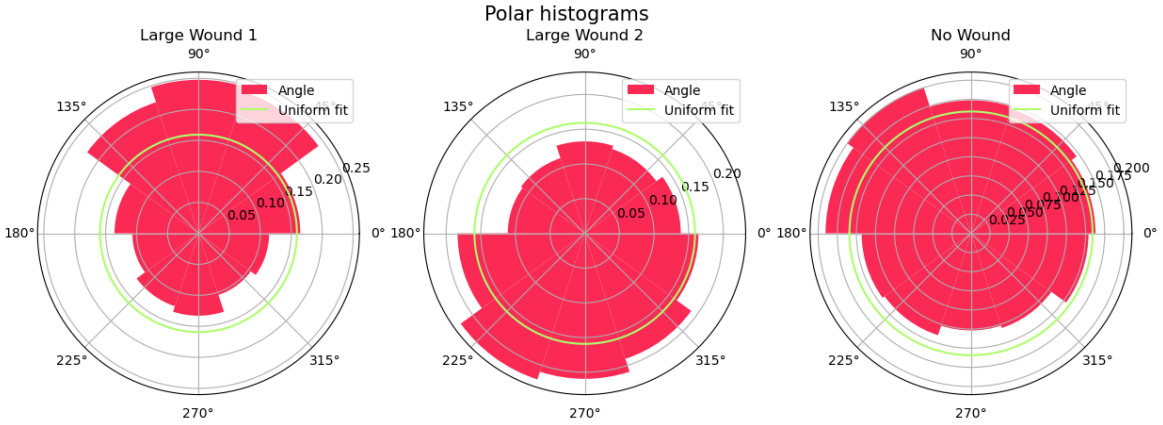


Figure 9: Normalized histograms of the angle to center method, with a line for the uniform distribution. For the histogram without wound, the three images from the same experiment were bundled together

The χ^2 value and the p-value are all represented in the following table:

Results				
	Large Wound 1	Large Wound 2	No wound	Synthetic
χ^2	8236	9159	2343	7.5
p - value	0	0	0	0.584

Table 2: Values for the different histograms and the corresponding tests.

The large values for χ^2, show that it does not fit well with a uniform distribution, so the null-hypothesis is rejected, as $p < 0.05$. While they unfortunantly reject the null-hypothesis, the no wound aswell, we can see a clear distinction between the

wound and no wound, with the χ^2 being 4 times larger suggest that the control still is better for a uniform fit.

To test the χ^2 value, synthetic data was drawn randomly from a uniform distribution between -pi and pi. The synthetic data shows that for a lower χ^2 value, the fit is better, thus resulting in a p-value greater than 0.05 which means the null-hypothesis cannot be rejected through synthetic data.

The correlation between the degree of polarization and the cells distance from the wound, with the angle to center method is shown in the figure below:

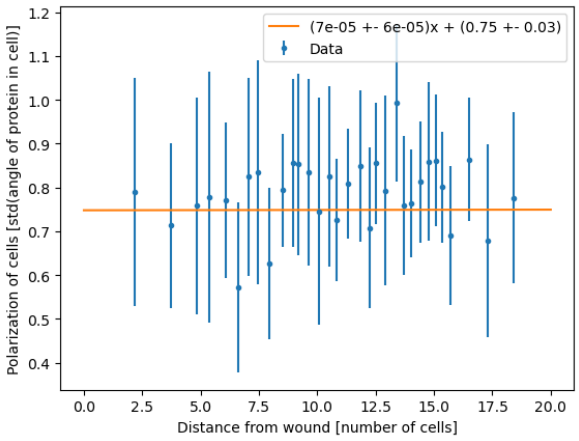


Figure 10: Correlation between distance from the wound and polarization of cells in figure 11 Data has been rebinned and the values on the y-axis is the average degree of polarization in the 10 cells binned together.

A decrease in polarization over distance and, thus, an increase in standard deviation is expected. The linear fit is increasing a tiny bit; however, it is nearly flat, and shows no real correlation when also considering the uncertainties. The expected positive correlation does, therefore, not show using this method of analysis.

Discussion

The main goal of the data analysis is to determine the presence of polarization. Both methods provide an answer, however they both have their own additional results and shortcomings.

Common causes of error

When labeling the proteins, a threshold was chosen to make the protein channel a binary image. The value of the threshold determines whether or not a pixel is

counted as a protein or not, which makes its impact on the results significant. Therefore, simply looking at histograms of pixel intensity to choose a threshold might not have been precise enough. To choose a more fitting threshold, a small algorithm determining a threshold, based on the histograms of pixel intensity, could be developed.
Another cause of error is that the images slice the cells around the image border. These "sliced" cells are often classified as polarized by the methods developed in this project even when they are not.

Discussion of the results from Principal Component Analysis

The statistical analysis on the histograms of polarization of cells from PCA conclude that there is a statistically significant difference in the average polarization of cells as a result of the presence of a wound. This method shows that there is a significant statistical difference between wound and no wound. However, from one single control, we cannot determine whether our control shows true non-polarization. This problem could be fixed by having multiple controls showing what we expect to be non-polarization. However, this is unlikely to have an effect because of the use of Bootstrapping, which somewhat works around this problem. Bootstrapping assumes data represents its own distribution; two controls should represent similar distributions, and thus, one control is sufficient.

Direction and location of the wound

As aforementioned, in addition to determining the presence of a wound through PCA, the eigenvectors from PCA also shows the direction of the wound. It is possible to calculate the position of the wound using the eigenvectors and cell centers, however, the clearest representation of the direction of the wound is the visual one already achieved, as seen in the PCA figures with wounds. These figures have easily readable arrows, which point in the direction to the wound. However, as the possibility of calculating the location of the wound may be of interest in further analytical work, it is still mentioned.

Deciding on a magnitude of polarization cut-off value

Applying PCA to a completely uniform distribution of proteins within a cell would be expected to result in two equal eigenvalues, as the variance along every axis would be expected to be equal. The resulting $m_{polarization}$ be roughly 1. This resulted

in the initial assumption that an $m_{polarization}$ of 2 or larger would mean a cell is certainly polarized. However, as is also seen in the PCA figures, polarized cells, in most cases, have an $m_{polarization}$ of around 10, while unpolarized cells, in most cases, have an $m_{polarization}$ of around 3. The cause of the high $m_{polarization}$ in non-polarized cells would be that there, in most cases, are not enough proteins in each cell for it to represent a completely uniform distribution, thus, the resulting $m_{polarization}$ is skewed by the lack of data points. However, as mentioned, the polarized cells often have a much higher $m_{polarization}$, and therefore, sufficient separation is still possible. Through trial and error, the final $m_{polarization}$ decided upon to separate polarized and non-polarized cells was settled at 5.

Discussion of results from angle to center

The Angle to center method faces a problem as it cannot, with the current statistical tools, cohesively show that an image is polarized. While it's quite clear from the plots, the χ^2 and p-value leave some to be desired. One reason is that at the moment, there is not enough data. More information would smooth out irregularities in the no wound, making it more closely represent a uniform distribution. To compensate for the lack in data, filtering out the images' edges, where only half cells are shown, which skew the data, could help.

Choosing bins for the histogram can have a large effect on how the data is presented. In this project, when choosing the bins, it was done by trial and error. This leads to a dilemma that must be faced: Not forcing the data to fit the desired result but still choosing appropriate bin sizes for the data. When a bin

When analyzing the degree of polarization in correlation to the cells distance from the wound, pi is added to the negative angels, so that a low std means the proteins have gathered along an axis. This, however, also affects the unpolarized cells, which results in a lower std, since the angles now only span from 0 to pi, which makes the spread of proteins relative to the center smaller. This makes the unpolarized cells harder to identify. Furthermore, cells on the image border might be sliced, which will make the cells furthest away from the wound seem very polarized. This could explain why the linear fit does not increase but instead goes flat.

Comparison of the two methods and results

Both methods seek to detect a difference in average polarization of cells as a result of a wound. This conclusion is reached with statistic significance through PCA but

not through analyzing angles to center. We also want to show a correlation between distance from the wound and polarization of the cells. This hypotheses is proven with PCA but not with analysis of angle to center. Furthermore PCA is capable of showing the direction from cell to wound; this is not possible when analyzing angles to center. Analyzing angles to center yield easily readable histograms and provide an overall simpler approach, but in this case fail to yield meaningful results.

Conclusion

By applying Principal Component Analysis to the sets of data, a statistically significant difference in the distribution of polarized and non-polarized cells in images with and without wound has been shown after using bootstrapping and a t-test. Through this difference, a correlation between protein-polarization and wound-presence can be concluded. Additionally, by deciding on a cut-off value for $m_{polarization}$ to classify a cell as polarized or non-polarized, a correlation between distance from wound and number of polarized cells can also be concluded. By calculating angles for each protein, histograms clearly showing polarization or not were created. However, through the χ-test, the hypothesis that an image without a wound should resemble a uniform distribution could not be proven statically. Using the standard deviation of the angle of proteins in a single cell as a measure of polarization did not show the expected correlation.

Taking the project further

An interesting finding throughout the work is the possibility of combining the two methods used. With the current models, the PCA method is capable of analyzing any input image. With or without wound, and no matter where the wound is placed in the image. The angles from the center, while they yield different, useful results, are only capable of analyzing images wherein the wound, if present, is placed at the edge or just outside the image, as the same vertical and horizontal axis are placed on all cells. This presents a problem, as the ideal method of analyzing the images would not rely on making the qualitative decision as to wether or not the wound, if present, is sufficiently placed for the method to be able to analyze the image. If instead the axis found by applying PCA, the two eigenvectors, were used as the basis for the calculation of the angle, the position of a protein on a histogram would be relative to the axis orthogonal to the wound, found through PCA, and not the global horizontal and vertical axis. This would enable getting results from the application of angular mapping no matter the nature of the input image.

Appendix

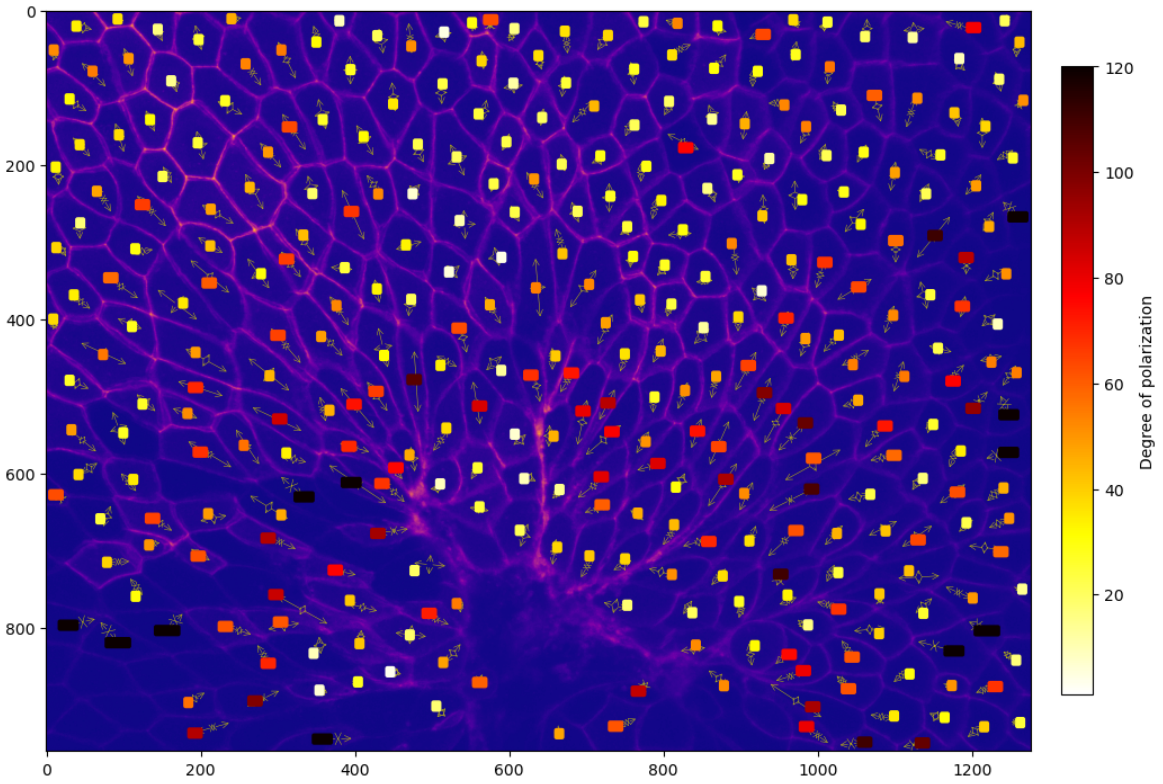


Figure 11: PCA analysis on ”Large wound 1” following the same model as Large Wound 2.

References

[Barlow 1989] BARLOW, R. E.: Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences. 2nd. New York : John Wiley & Sons, 1989

[Connelly 2021] CONNELLY, Bill: Principal Component Analysis (PCA) For Dummies. 2021. – URL https://www.billconnelly.net/?p=697. – Accessed: 2025-03-26

[DATAtab n.d.] DATATAB: p-value: A Beginner's Guide. n.d.. – URL https://datatab.net/tutorial/p-value. – Accessed: 2025-03-26

[Staddon u. a. 2018] STADDON, Michael F. ; BI, Dapeng ; TABATABAI, A. P. ; AJETI, Visar ; MURRELL, Michael P. ; BANERJEE, Shiladitya: Cooperation of dual modes of cell motility promotes epithelial stress relaxation to accelerate wound healing. In: PLOS Computational Biology 14 (2018), Nr. 6, S. e1006502. – URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006502

[of Standards und Technology n.d.] STANDARDS, National I. of ; TECHNOLOGY: Do two processes have the same mean? n.d.. – URL https://www.itl.nist.gov/div898/handbook/prc/section3/prc31.htm. – Accessed: 2025-03-26

[VanderPlas 2016] VANDERPLAS, Jake: Statistics for Hackers - PyCon 2016. 2016. – URL https://www.youtube.com/watch?v=Iq9DzN6mvYA. – Accessed: 2025-03-26

(Staddon u. a., 2018) (of Standards und Technology, n.d.) (Connelly, 2021) (DATAtab, n.d.) (VanderPlas, 2016) (Barlow, 1989)