BACHELOR'S THESIS

Modeling Calcium Dynamics in T Cells

submitted to the

Institute of
Analysis and Scientific Computing
TU Wien

under the supervision of

Assistant Prof. Dr. Andreas Körner

and

Dipl.-Ing. Alexander Edthofer

by

Ida Hönigmann
Matriculation number: 12002348

Vienna, December 6, 2024

Acknowledgement

I would like to thank Andreas Körner and Alexander Edthofer for proofreading and support throughout this bachelor thesis.

Thank you to Caroline Kopittke for the insight into T cells without this work would not have been possible, as well providing the Ca^{2+} concentration data.

I am also thankful to my friends and family for helping me during research and pointing me towards vastly improving the parameter choices in the approximation.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Bachelorarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 6. Dezember 2024

Ida Hönigmann

Contents

1. Introduction

As part of the bodies' defence against viruses T cells can undergo activation in their lifetime. They are named after the thymus, where they are differentiated into T cells. Whether and when a T cell activates is an interesting topic when studying immunology. However, measuring activation is often done indirectly by using the correlation between calcium concentration within the cell and activation. Measuring the calcium concentration leads to a time series that can be analysed for activation by experts.

This work aims to model the activation behaviour of T cells and provide an algorithm for automatic detection of activation. By using approximation algorithms the time series is fitted with sigmoid functions. This reduces the data from hundreds of values in a time series to few parameters of the approximation function. Additionally, the parameters are chosen to be valuable for interpretation by experts. This parameter representation of the data is then filtered and used for clustering the data into activated and unactivated cells.

The given algorithm can be used on any data set, provided a positive control and negative control is supplied. The most simple use case of finding the number of activated cells in the data set is described in more detail.

The proposed algorithm is tested with two of the most common types of T cells, human Jurkat cells and mouse 5c.c7 cells. Further details on the data the algorithm is applied to can be found in chapter 4.

The main question this work aims to answer is which criteria can distinguish between unactivated and activated cells. Additionally, a criterion for detecting cells which activated before the experiment began will be investigated.

When only looking at activated cells some interesting questions are whether there are different types of activated cells and how they are different. A typical pattern observed in activated cells is that they show oscillations of the calcium concentration. Analysing the frequencies of these oscillations might be interesting.

Lastly, differences between mouse and human cells show whether the proposed algorithm is applicable to the two most common types of T cells studied.

To summarize, the main research questions are:

- Which criteria can distinguish between unactivated, activated and pre-activated cells?
- Do different types of activated cells exists? How are they different?
- With which frequencies does the Calcium concentration oscillate after activation?
- Is there a difference between mouse and human cells?

This work starts with a chapter on optimization algorithms. In chapter 2 the relevant algorithm, trust region reflective algorithm, is attained from other algorithms, which are described as well.

Following is chapter 3 focusing on the biology of T cells. All relevant components of T cells for changes in calcium concentration are depicted. Their interconnections are outlined as well.

Next, chapter 4 describes the structure and experimental setup for retrieving the data. Some processing steps performed on the data are outlined.

The main focus of this work, approximating the calcium concentration, are portrayed in chapter 5. Here the approximation function is inferred and characterized. Parameter descriptions are provided. Pseudocode for the approximation algorithms are given and explained. The parameters found from the approximation of the data sets used in this work are analysed. Oscillations are approximated in this chapter as well.

In chapter 6 the clustering algorithms gaussian mixture model and k-means are characterized and applied to the output of the approximation. Visual representation of the clustering is shown.

Chapter 7 aims to answer the main research questions posed above by using the approximation and clustering described in the other chapters.

A final discussion of the results provided in this work is provided in chapter 8. The outlook is featured here as well.

The appendix features Python code, providing an implementation of the algorithms discussed throughout this work. It was tested on a Lenovo T470 ThinkPad.

2. Optimization Algorithms

An optimization problem is any problem where a function $f\colon X \to Y$ is given, and we search for the point $x \in X$ such that $f(x)$ is minimal or maximal. We define the minimum of f as a point $x \in X$, such that for all $\tilde{x} \in X : f(x) \leq f(\tilde{x})$. Sometimes this is called global minimum to differentiate it from a local minimum, which is defined as there existing a point $x \in X$ along with a neighbourhood $U_x \subseteq X$ of x such that for all $\tilde{x} \in U_x : f(x) \leq f(\tilde{x})$. The maximum is defined analogously. Obviously the minimum or maximum must not exist, as the example $f\colon (0,1) \to \mathbb{R}, x \mapsto x$ demonstrates by not having either. Investigating conditions on X, Y and f such that a minimum or maximum exists is mathematically interesting. However, when implementing an optimization algorithm the true minimum or maximum can sometimes not be found even if it exists and is instead replaced by a sufficiently good approximation.

2.1. Gradient Descent

An iterative algorithm for finding the minimum of a continuously differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is gradient descent. This method of steepest descent is described in detail by Chong and Żak[CŻ13]. As the name suggests, it uses information of the gradient ∇f. Locally, the negative gradient always points into the direction of greatest descent. The idea is to follow this direction for the next guess of the minimum, which can be expressed mathematically as

$$\hat{x} := x - \gamma \nabla f(x),$$

with γ being the step size. The pseudocode of this approach is given in algorithm 1.

Algorithm 1: Gradient Descent

input : $f\colon \mathbb{R}^n \to \mathbb{R}$ ... differentiable, $x_0 \in \mathbb{R}^n$, $max\_iterations \in \mathbb{N}$, $threshold \in \mathbb{R}^+$, $\gamma_n \in \mathbb{R}^+$
output: $x \in \mathbb{R}^n$

```
begin
    for n = 0 to max_iterations do
        if improvement is smaller than threshold then
            break
        end
        set or calculate step size γ_n
        x_{n+1} = x_n − γ_n ∇f(x_n)
    end
    x = x_n
end
```

If we consider a function with a local minimum, that is not a global minimum, gradient descent might not converge to the optimum. An example of such a function can be seen in figure 2.1 along with the first few points x_n of gradient descent. The starting value was chosen to not have convergence to the global minimum. For a different starting value the global minimum can be reached.

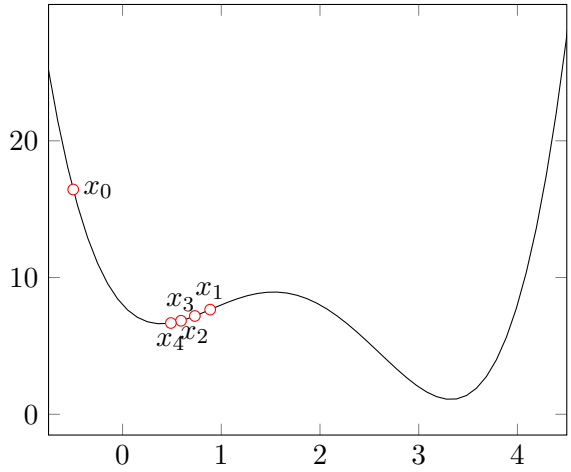

Figure 2.1.: The illustrated function has two local minimums. For starting value x_0 and particular step size, gradient descent approaches the local, but not global minimum.

Improvements can be made by choosing good step sizes, starting value or by starting with different values and comparing the results.

2.2. Least Square Problem Algorithms

We now focus on the least square problem and give an introduction into various algorithms solving this problem.

In the example dealt within this work we are given some data points $((x_k, y_k))_{k\in\{1,2,...,n\}}$ and want to find a close approximation in the form of a function $g(x, a_1, a_2, ..., a_m)$ where for every list of function parameters $a = (a_1, ..., a_m)$ we have the function

$$g_a(x)\colon \mathbb{R} \to \mathbb{R}, x \mapsto g(x, a_1, ..., a_m).$$

Searching for a good approximation can be reformulated as searching for the minimum of $r(a) := \sum_{k=1}^{n} |g_a(x_k) - y_k|^2$, or any other error function. This form of optimization problem is called the least square problem.

First, we want to discuss some variations of the problem. Easiest to solve are problems that can be formulated as the minimization of $||Ax - b||^2$. These are called linear problems and can be solved using calculus by $x = (A^T A)^{-1} A^T b$, provided the rank of A is full.

Often, we want to constrain the search for a minimum under some properties. For linear problems we can find a formulation as

$$\min_{x \in \mathbb{R}^n} ||Ax - b||^2,$$

subject to

$$Cx = d.$$

Finding a solution can be done by minimizing $||Ax - b||^2 + \lambda||Cx - d||^2$ for very large λ. Intuitively this works by simultaneously minimizing $||Ax - b||^2$ and $||Cx - d||^2$. By increasing λ we can determine the weighting or importance of getting $||Cx - d||^2$ as close to 0 as possible, which corresponds to getting solutions where $Cx \approx d$.

General least square problems are formally given as a residual function $r_f(x)$ which tells us whether a function f is a good approximation at the point x. We therefore want to find a way to minimize $||r_f(x)||^2$.

As $||r_f(x)||^2 \geq 0$ we can turn to the simpler problem of finding a root. However, a root must not exist, in which case we want to find the value closest to zero. This is then the minimum of the function. Some algorithm for minimization are discussed in the following subsections.

2.2.1. Gauß–Newton Algorithm

The Gauß-Newton algorithm is detailed by Björck in his book Numerical Methods for Least Squares Problems, [Bjö96]. The idea behind this algorithm is that it is easy to find the intersection with zero of a linear function. If we linearize $r\colon \mathbb{R}^n \to \mathbb{R}^m$ locally, we can approximate the root by finding it of the linear approximating function, which is described below. The idea is demonstrated in figure 2.2.

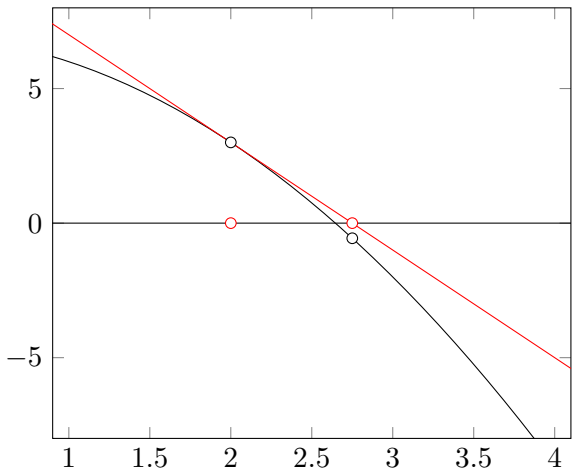

Figure 2.2.: By approximating the black function by a line an approximation of the root has been found.

Iterating this step of linear approximating gives us the Gauß-Newton Method. In figure 2.3 we can see that indeed x_n seems to converge towards the root of the function.

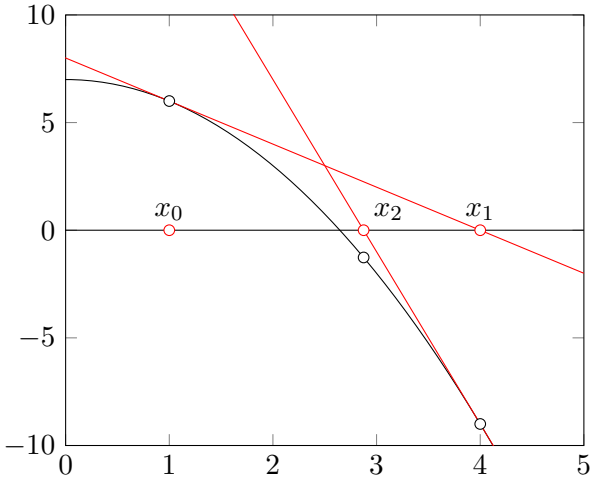

Figure 2.3.: Iteratively applying linear approximation gives the Gauß-Newton Method for approximating the root.

Defining the Jacobian matrix and using Taylor's theorem we get the linear approximation

$$Dr := \left(\frac{\partial r_i}{\partial x_j}\right)_{ij},$$

$$r(x) = r(a) + Dr(a)(x-a) + h(x)(x-a) \approx r(a) + Dr(a)(x-a) \text{ with } \lim_{x \to a} h(x) = 0.$$

Rewriting this as $r(x) \approx Ax - b$ where $A := Dr(a)$ and $b := Dr(a)a - r(a)$ gives us the algorithm for this method. As $Dr \in \mathbb{R}^{n \times m}$ we solve $Dr^T Drx = Dr^T b$ in order to get a system with square matrix. If $n = m$ we can skip this step and get the so-called Newton algorithm as a variant. It is well known that there exists a unique solution to this system if the rank of Dr is full.

Algorithm 2: Gauß-Newton

input : $r\colon \mathbb{R}^n \to \mathbb{R}^m$ ... differentiable, $x_0 \in \mathbb{R}^n$, $max\_iterations \in \mathbb{N}$
output: $x \in \mathbb{R}^n$

```
begin
    for n = 0 to max_iterations do
        if ||r(x_n)||^2 close enough to zero or ||x_n − x_{n−1}|| is too small then
            break
        end
        Calculate A_n := Dr(x_n) and b_n := A_n x_n − r(x_n)
        Solve A_n^T A_n x_{n+1} = A_n^T b_n
    end
    x := x_n
end
```

Gauß-Newton is guaranteed to find a local minimum x if r is twice continuously differentiable in an open convex set including x, Dr has a full rank and the initial value is close enough to x.

For the example demonstrated in figure 2.4 we can see that choosing a particular starting value leads to a loop in which only two points are explored as possible roots.

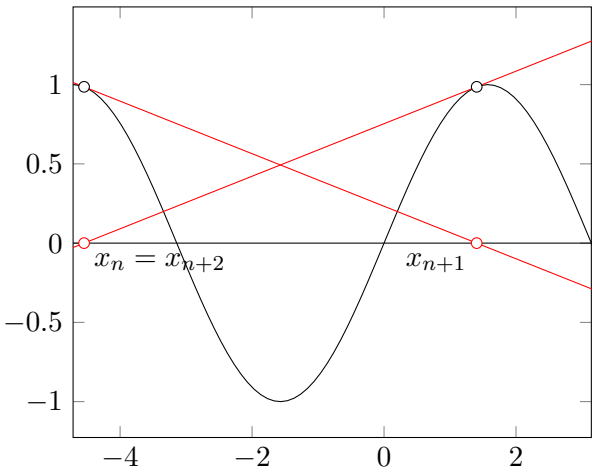

Figure 2.4.: For a poor choice of starting values Gauß-Newton can never find the root of the function $\sin(x)$.

Another even more extreme examples can be seen in figure 2.5, in which Gauß-Newton gets increasingly further away from the root, due to an increasingly flat incline the further we get from the root.

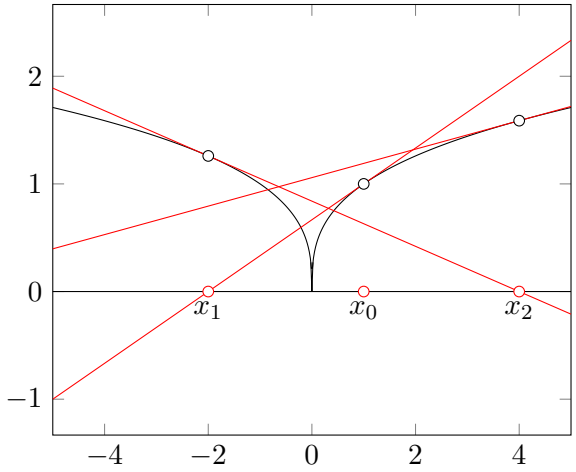

Figure 2.5.: Finding the root of the function $\sqrt[3]{|x|}$ using Gauß-Newton is only possible if the starting value x_0 is chosen as 0, which is the root. If we choose any other value the guess gets farther and farther away. Indeed, for any x_n we have $x_{n+1} = -2x_n$.

Gauß-Newton has two problems, the starting value being too far from the root and Dr not having full rank. This can be combated using the technique of dampening. Instead of moving the new guess all the way to the root of the linear approximation we only move part of the way. How far to move can be determined by a dampening factor λ_n or a constant λ. This results in the dampened variation of the Gauß-Newton algorithm, which can be described as

$$x_{n+1} := \lambda\hat{x} + (1-\lambda)x_n$$

where $\hat{x}$ is defined as the solution of

$$Dr(x_n)^T Dr(x_n)\hat{x} = Dr(x_n)^T Dr(x_n)x_n - r(x_n).$$

Algorithms using a similar approach are described in the following subsections.

2.2.2. Levenberg-Marquardt Algorithm

This section is following section 18.3 of the book Introduction to Applied Linear Algebra by Stephen Boyd and Lieven Vandenberghe[BV18].

As stated above, a shortcoming of Gauß-Newton is that for x far from x_n we must not have that

$$r(x) \approx r(x_n) + Dr(x_n)(x - x_n) =: \hat{r}(x, x_n).$$

Levenberg-Marquardt addresses this by minimizing

$$||\hat{r}(x, x_n)||^2 + \lambda_n ||x - x_n||^2.$$

The first part is the same as above, while the second objective expresses our desire to not stray away too much from the region where we trust the linear approximation. The parameter λ_n is a positive parameter specifying how far the trusted region extends.

Writing the above idea as a single squared norm to minimize gives us the problem

$$\min_{x\in\mathbb{R}^d} \left|\left| \begin{pmatrix} Dr(x_n) \\ \sqrt{\lambda_n} I \end{pmatrix} x - \begin{pmatrix} Dr(x_n)x_n - r(x_n) \\ \sqrt{\lambda_n} x_n \end{pmatrix} \right|\right|^2.$$

We observe that as λ_n is positive the left matrix has full rank. From this it follows that a unique solution exists.

The change of including λ_n translates into the algorithm as replacing solving

$$A_n^T A_n x_{n+1} = A_n^T b_n$$

in Gauß-Newton by solving

$$A_n^T A_n z + \lambda_n z = A_n^T b_n + \lambda_n x_n.$$

The question of how to choose λ_n arises. If it is too small, x_{n+1} can be too far from x_n to trust the approximation. If it is too big, the convergence will be slow. Whenever the

objective $||r(x_n)||^2$ decreased in the previous step, we decrease λ_{n+1} slightly. In the case that the last step was not successful λ_n was too small. Therefore, we increase λ_{n+1}.

Pseudocode of the resulting Levenberg-Marquardt algorithm is shown below. The stopping criteria of $||2Dr(x_n)^T r(x_n)||$ being too small is known as the optimality condition. It is derived from the fact that $2Dr(x)^T r(x) = \nabla||r(x)||^2 = 0$ holds for any x minimizing $||r(x)||^2$. Note that this condition can be met for points other that the minimum.

Algorithm 3: Levenberg-Marquardt

input : $r\colon \mathbb{R}^n \to \mathbb{R}^m$ ... differentiable, $x_0 \in \mathbb{R}^n$, $\lambda_0 > 0$, $max\_iterations \in \mathbb{N}$
output: $x \in \mathbb{R}^n$

```
begin
    for n = 0 to max_iterations do
        Calculate A_n := Dr(x_n)
        Calculate b_n := A_n x_n − r(x_n)
        if ||r(x_n)||^2 close enough to zero or ||2A^T r(x_n)|| is too small then
            break
        end
        Solve (A_n^T A_n + λ_n) z = A_n^T b_n + λ_n x_n
        if ||r(z)||^2 < ||r(x_n)||^2 then
            x_{n+1} := z
            λ_{n+1} := 0.8 λ_n
        else
            x_{n+1} := x_n
            λ_{n+1} := 2 λ_n
        end
    end
    x := x_n
end
```

Coming back to the example where Gauß-Newton failed, we once again consider the function from figure 2.5. In comparison, this time we are able to find a good approximation of the root using Levenberg-Marquardt. A few steps are demonstrated in figure 2.6.

Further improvements to this algorithm can be made using good starting values perhaps from the output of other algorithms or by letting the algorithm run multiple times with different starting values and comparing the results.

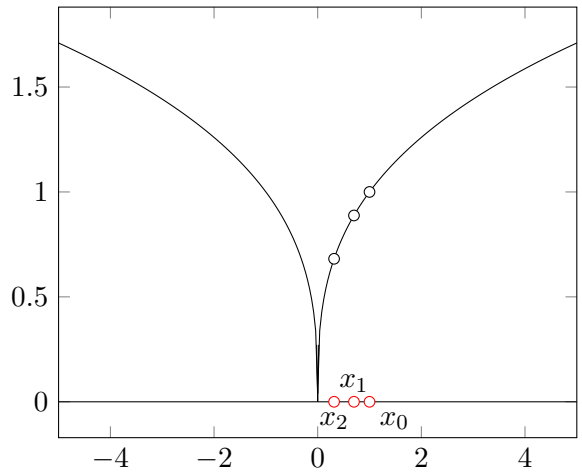

Figure 2.6.: In comparison to Gauß-Newton, Levenberg-Marquardt is able to find the root of the function $\sqrt[3]{|x|}$. From the starting value $x_0 := 1$ and using $\lambda_0 := 1$ the guesses x_n move towards the root $x = 0$.

2.2.3. Algorithms for Bounded Least Square Problems

The algorithms described above do not consider bounds. For bounded problems two algorithms known as trust region reflective algorithm and dogleg algorithm with rectangular trust regions can be used.

Trust region reflective gets its name from the use of trust regions as in Levenberg-Marquardt as well as reflecting along the bounds. A more detailed explanation is provided by Branch et al. [BCL99]. If an iterative x_n lands outside the bounds set, it is replaced by a reflected value within the bounds. This ensures each iterative is feasible as a solution.

As the name suggests the Dogleg Algorithm with Rectangular Trust Regions uses rectangular trust regions as opposed to ellipsoids. Voglis and Lagaris describe this algorithm in more detail [VL04]. As the bounds are specified as a rectangle to stay within, this results in the intersection of trust region and bounds to be rectangular. The resulting minimizing problem is solved with an adequate algorithm, as one described by Nocedal and Wrigh [NW99].

In Python the library SciPy provides the three methods Levenberg-Marquardt, Trust Region Reflective and Dogleg with Rectangular Trust Regions for solving Least Square Problems.

3. T Cells and Calcium Concentration

Lymphocytes form a key component of the immune system. T cells are a type of lymphocyte and are responsible for responding to viruses, fungi, allergens and tumours. Different subtypes of T cells exist, that perform various responsibilities. Kumar et al. describe that they are transported throughout the body via the lymphatic system and blood see [KCF18].

Precursor cells are formed in the bone marrow. Once they are transported to the thymus they undergo maturation and selection to become T cells. Each cell forms receptors, called T cell receptors (TCR), that respond to one particular out of many (10^6–10^9) possible short pieces of proteins, called peptides. These peptides are attached to the major histocompatibility complex (MHC) present on antigens and antigen presenting cells (APC). Important aspects of the selection are ensuring that the T cells react to foreign peptides, but not to those present on the body's own cells. This is investigated by Ashby and Hogquist in their work [AH24].

In positive selection, cells in the thymus present peptides on their MHC. If a T cell is unable to bind, it will undergo apoptosis, a type of cell death. T cells, which were able to bind, receive survival signals. Negative selection verifies that T cells will not attack the body's own cells. This is done by only selecting T cells which only bind moderately to the peptides presented, as a strong bond suggests that these T cells would have a high likelihood of being reactive to own cells. A detailed analysis is provided by Hagel [Hag18]. If a T cell passed both the positive and negative selection it is transported to the periphery.

According to Ganong there are multiple types of peripheral T cells [Gan97]. Native T cells respond to new antigens. Cytotoxic T cells kill cells which present peptides on their MHC compatible with the TCR. Helper T cells activate other parts of the immune response. Memory T cells shorten the reaction time when the same antigen is encountered again at a later point in time. Suppressor T cells moderate the immune response.

The following sections describe the relevant components of T cells and activation.

3.1. Components of a T Cell

T cell components relevant in activation and subsequent changes in intracellular Ca^{2+} are listed below.

- **T cell receptor:** Receptor on the cell surface that can recognize peptides. By the simultaneous triggering of the TCR and co-stimulator, signalling is induced that leads to activation.
- **Co-stimulator:** A stimulation of co-stimulatory molecules is necessary in order for signalling to occur as part of activation.

- **Endoplasmic reticulum (ER):** A series of connected sacs in the cytoplasm that is attached to the nucleus. As described by Rogers, important functions are folding, modification and transportation of proteins [Rog24].
- **Ca^{2+} permeable ion channel on the ER:** There are several Ca^{2+} channels present on the ER. Some receptors are responsible for releasing Ca^{2+} into the cytoplasm, when the intracellular Ca^{2+} concentration is low. This is described by Schwarz and Blower [SB16].
- **Ca^{2+} storage in the ER:** Ca^{2+} is stored in the ER and can be released by Ca^{2+} permeable ion channels on the ER.
- **Cytoplasm:** The semi-fluid substance enclosed in the plasm membrane. It contains organelles, ions, proteins and molecules.
- **Stromal interaction molecule (STIM):** If the Ca^{2+} storage in the ER is depleted STIM proteins cluster where the ER is in the vicinity of the plasm membrane and assembles Ca^{2+} release activated Ca^{2+} channels, which then leads to uptake in extracellular Ca^{2+}. Further details can be found in Schwarz and Blower's work [SB16].
- **Plasm membrane:** A semipermeable structure forming the wall of the cell made up of lipids and proteins. Ganong describes the ion channels and transport proteins, that allow certain substances to move through [Gan12].
- **Ca^{2+} release activated Ca^{2+} channel (CRAC):** According to Stathopulos and Ikura these channels open after a decrease in ER-stored Ca^{2+} is sensed by STIM, these channels intake Ca^{2+} from outside the cell [SI13].
- **Cytoskeleton:** Ganong describes the cytoskeleton as a system of fibres within the cell, that allows it to change shape and move [Gan12].
- **Nucleus:** According to Cooper and Adams the nucleus is an organelle that stores most of the DNA, controls cell growth and cell division [CA22]. A double membrane separates it from the cytoplasm.

The two relevant components of APC, both present on the surface of the APC, are the

- **Major histocompatibility complex (MHC)**, which can present peptides, and the
- **Co-stimulator**, which can form a bond with the co-stimulator on a T cell.

All components are schematically shown in figure 3.1.

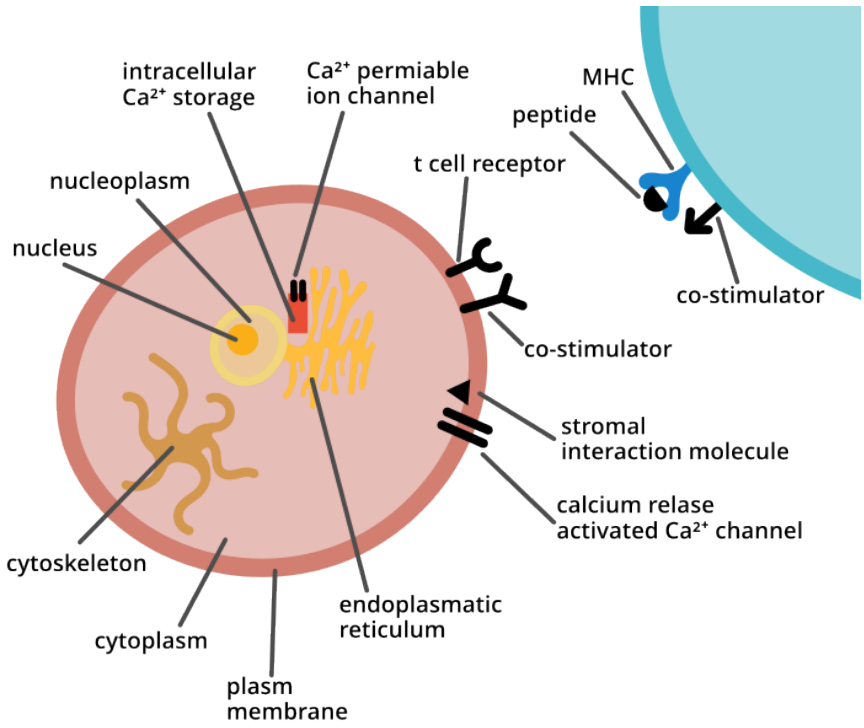

Figure 3.1.: Schematic view of a T cell and antigen presenting cell, with all relevant components.

3.2. Activation of T Cells

As Ganong describes, activation is necessary for T cells to divide and perform their functions [Gan97].

When a native T cell encounters a peptide on an APC that is compatible, a bond is formed between the TCR on the T cell and the peptide-MHC complex on the APC. This recognition can be triggered by less than ten molecules of foreign substance and is therefore described as near perfect. Sufficiently long contact is necessary between the APC and the T cell in order for the T cell to activate. The role of contact time in T cell activation is modelled by Morgan et al [ML23].

The presence of co-stimulatory molecules is needed for activation. The bond between the co-stimulatory molecules on the T cell and the one on the APC plays a role in signalling. Especially Ca^{2+} signals play a vital part in T cell activation.

According to Smith-Garvin, an increase of Ca^{2+} in T cells during activation is caused by the stimulation of Ca^{2+} permeable ion channel receptors on the ER membrane [SKJ09]. Ca^{2+} is released from the ER into the cytoplasm. Additionally, this decrease in Ca^{2+} is sensed by STIM, which leads to an influx of Ca^{2+} through plasma membrane CRAC channels.

As the intracellular Ca^{2+} concentration is dependent on the interaction between Ca^{2+} sources and sinks, a variety of different forms in Ca^{2+} concentration have been observed. Lewis lists examples, such as infrequent spikes, sustained oscillations and plateaus [Lew01].

Intercellular Ca^{2+} increase together with other signals lead to a redistribution of receptors, signalling molecules and organelles. This is described by Joseph et al [JRB14].

As Ca^{2+} is linked to activation and relatively easy to measure it is used in this work as data to analyse T cell activation.

4. Generating Data of the Calcium Concentration

From section 3.2, we gather that analysing the intracellular Ca^{2+} concentration gives us good insight in whether and when a cell activates. Additionally, it can be measured relatively easily by the method presented in this chapter.

4.1. Structure of the Data

First, we describe the structure of the data this work uses. The data matrix has one row for each combination of tracked particle and frame number. In this context cells are called particles as the recording might feature non-cells that are detected as a cell and recorded in the data set. The information stored for each particle and frame combination is described in detail in table 4.1.

Name	Data Type	Description
x	float64	Position of particle in pixels along the horizontal axis
y	float64	Position of particle in pixels along the vertical axis
frame	int32	Number of frame, with frame rate of 1 frame per second
mass short	float64	Brightness of cell in 340nm channel
mass long	float64	Brightness of cell in 380nm channel
ratio	float64	Calculated as mass short divided by mass long
particle	int32	Identification for each particle

Table 4.1.: Description and data type of all columns present in the data matrix.

One recording can have between 500 and 10000 particles. The length is between 700 and 1000 frames, which corresponds to approximately 11 to 17 minutes. The ratio recorded is typically between 0 and 5.

Four recordings were generated, with two each from human and mouse cells. For each cell type a positive and negative control was measured. In a positive control the conditions are such, that in theory every cell should activate, while in negative control the conditions are such, that none should activate. Due to stress on the cells a few cells will activate before the recording starts, during the recording in the negative control or not activate at all in the positive control, regardless of the conditions. Stress can be caused by movement, changes in temperature and other factors.

4.2. Measuring the Calcium Concentration of T Cells

Jurkat cells and 5c.c7 primary mouse T cells are two T cell lines widely used. In order to be able to measure the intracellular Ca^{2+} concentration of cells they can be labelled with Fura-2. Martinez et al. depict how this method provides a way to record the Ca^{2+} concentration of multiple cells over a time period [MMS17]. Challenges encountered when using Fura-2 on certain cell types are described by Roe, Lemasters and Herman along with their respective solutions [RLH90].

After the cells have been labelled with Fura-2, a recording of up to 15 to 20 minutes can be generated. To achieve this the cells and stimulant are photographed once per second at both 340nm and 380nm wavelength. The resolution of the images are 1.6μm per pixel. By calculating the ratio of the two images at each pixel the Ca^{2+} concentration can be observed. An exemplary resulting image showing the ratio is depicted in figure 4.1. The T cells appear a lighter shade than the background when activated and darker when not activated.

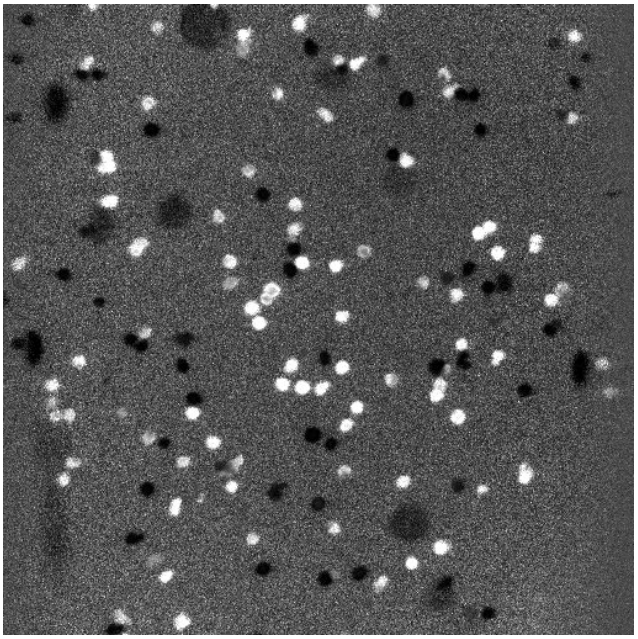

Figure 4.1.: Single frame showing the ratio of the 340nm and 380nm images from a recording of human Jurkat cells. Activated cells appear lighter, unactivated cells darker than the background. Big dark circles are out of focus cells that have not yet settled on to the plate.

To activate the cells in the duration of the recording they are transferred to a plate covered with replicas of the MHC-peptide complex normally present on APCs. This plate is then recorded as described above. For a negative control the plate is not covered with peptides, while for the positive control the peptide covering on the plate is very dense. Recordings of different densities in peptides lead to activation of a percentage of T cells.

4.3. Processing the Data

To track single T cells moving around during the video the sum of the 340nm and 380nm image of each second is calculated. This image provides the basis for separating T cells from the background. On this image all T cells will appear similarly light in colour. Therefore, it is used to track the movement of cells. Each cell is numbered, such that the same cell will have the same number during the video. For some cells the trajectory tracking is not perfect, resulting in a split of the numbering into multiple numbers for the same cell. The position and shade during both 340nm and 380nm as well as the ratio of each particle and each frame is then recorded into the data structure used in this work. The first roughly 50 frames at the start of the recording are discarded due to the video being out of focus. Additionally, cells only appearing in fewer than 300 frames are discarded as they most likely represent trajectories incorrectly tracked or split. The resulting data is then stored in a matrix structured as described in table 4.1.

5. Modelling the Approximation of Calcium Concentration

If we have a look at the typical trajectory of the calcium concentration in activated and unactivated cells, shown in figure 5.1, we can see differences emerging. For one, the maximum concentration value reached by most activated cells is higher. Another distinguishing feature is the presence of a steep incline at the moment of activation.

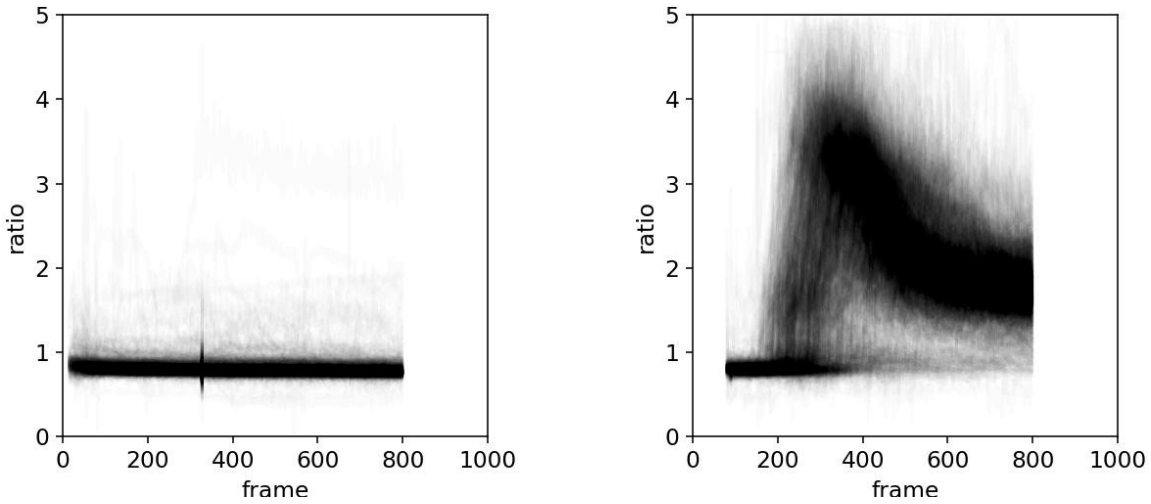

Figure 5.1.: Two plots of the overlapping calcium concentration time series of cells, on the left a negative control and on the right a positive control of mouse cells. It is assumed that most cells from the negative control did not activate, while most of the cells from the positive control did.

By modelling the time series with a function incorporating features such as the increase, maximum value and oscillations present in the decrease afterwards, we can extract these features more easily. By doing this, using approximation methods from chapter 2, we hope to have an easier method to answer the questions from the introduction.

5.1. Approximation Function

From studying the data in the two control groups we find to expect a function close to

$$f_{\text{unac}}(x) := u \in \mathbb{R} \tag{5.1}$$

for unactivated cells and

$$f_{\text{ac}}(x) := \begin{cases} \frac{a-u}{1+\mathrm{e}^{-k_1(x-w_1)}} + u & \textbf{if } x \leq t \\ \frac{a-d}{1+\mathrm{e}^{-k_2(x-w_2)}} + d & \textbf{else} \end{cases} \tag{5.2}$$

for activated cells. The parameters are described in table 5.1.

Variable	Description
u	average value before activation
a	value reached at the peak of activation
d	average value after activation
k_1	steepness of increase
k_2	steepness of decrease
w_1	time point at which the increase happens
w_2	time point at which the decrease happens
t	time point at which the increase ends, and the decrease starts

Table 5.1.: List of parameters occurring in function f_{unac} from 5.1 and function f_{ac} from 5.2 and their interpretation.

Figure 5.2 depicts the above functions (5.1), (5.2), and the relations to the parameters in unactivated and activated cells. The similarity between these functions and figure 5.1 can be observed.

For our model to make sense, we have to impose some conditions on the parameters. We expect

$$0 \leq u \leq d \leq a, \qquad w_1 \leq t \leq w_2, \qquad k_1 > 0 \qquad \text{and} \qquad k_2 < 0. \tag{5.3}$$

There are multiple ways in which the parameters of f_{ac} can be chosen to get a function similar to f_{unac}. If w_1 is very large or $u \approx d \approx a$ then f_{ac} approaches a constant value of u, thus approximating f_{unac}. If the approximation of a cell has parameters with w_1 very large or $u \approx d \approx a$ we can therefore expect it to be of an unactivated cell. Otherwise, it is more probable to be activated.

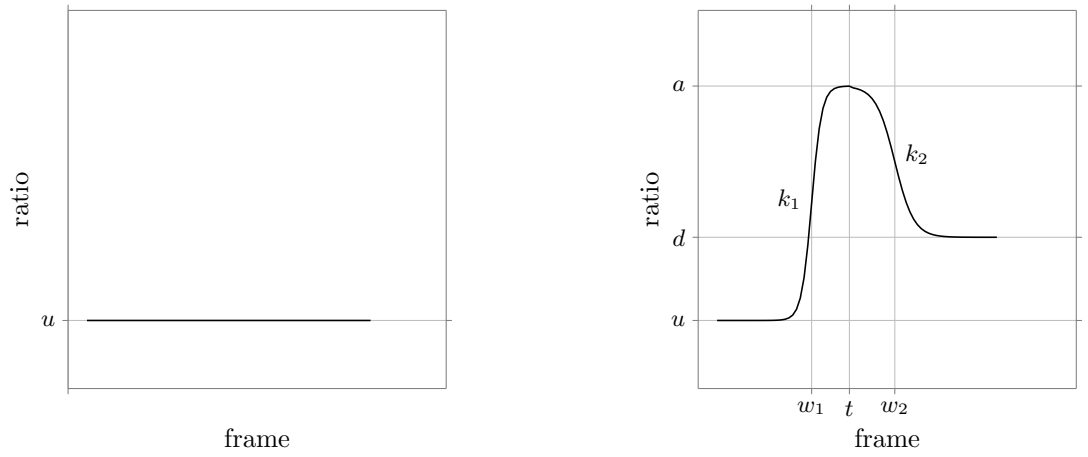

Figure 5.2.: Left shows the function f_{unac} defined in (5.1) with the parameter u. The right shows the function f_{ac} defined in (5.2) with the parameters u, d, a, w_1, t, w_2, k_1 and k_2.

5.2. Implementation of the Approximation Model

Now that we have defined our model functions we will implement a routine that fits such a f_{ac}-function through the data points of a particle recording.

First we give the pseudocode for approximating a single particles time series with the approximation function described above. It takes a (frame, ratio)-matrix of a single particle as input and returns the corresponding parameter list of the approximation.

Algorithm 4: Approximation of the Calcium Concentration

input : particle data as (frame, ratio)-matrix
output: parameters describing the approximation

```
1 begin
2     set boundaries for parameters
3     set start values for parameters
4     use Trust Region Reflective Algorithm with boundaries and start values to get
        parameters
5     calculate corresponding approximation and add as fit_sigmoid columns to data
        matrix
6     return parameters
7 end
```

The parameters of f_{ac} used in the approximation are not independent of each other as we want to choose t to be the point at which the increasing part of the function, $(a-u)/(1+\mathrm{e}^{-k_1(x-w_1)})$, almost reaches the value a. We choose

$$t := w_1 - \frac{\log_e\left(\frac{1}{0.99} - 1\right)}{k_1} = w_1 - \frac{1}{k_1}\log_e\left(\frac{1}{99}\right),$$

as the function has had 99% of the increase of the sigmoid curve up to this point.

Setting the boundaries in line 2 is non-trivial. We have noted that the conditions from equation 5.3 are expected. We want to impose them using boundaries in which the parameters must lie. However, boundaries for each parameter must not depend on other parameters. We can circumvent this by changing the parameters to be relative to each other. As $u \leq d \leq a$ we choose to use the three parameters $u, d-u$ and $a-d$. Then, we can set the lower boundary to be 0 which ensures

$$0 \leq u \quad \wedge \quad 0 \leq d-u \implies d \geq u \quad \wedge \quad 0 \leq a-d \implies a \geq d$$
$$\implies 0 \leq u \leq d \leq a.$$

Using the same method, we choose the parameters $w_1 - \mathit{start}$ and $w_2 - w_1$, where *start* is the first frame in which the particle was tracked. The resulting boundaries are described in table 5.2, where we set min val, max val and median val as the minimum, maximum and median of the particles' ratio data respectively while start and end is the first and last frame where data was recorded for this particle.

The condition $t \leq w_2$ can be violated, but it is ensured that at least $w_1 \leq w_2$. The other conditions are met as $k_1 \in [0.05, 10] \implies k_1 > 0$ while $k_2 \in [-1, -0.01] \implies k_2 < 0$ and

$$t = w_1 - \underbrace{\frac{1}{k_1} \log(\frac{1}{99})}_{<0} \geq w_1.$$

Along with the boundaries we specify so-called starting values. These define to which values the parameters are set at the start of the approximation algorithm.

parameter	**lower bound**	**upper bound**	**starting value**
u	min val	max val	min val
$d-u$	0	max val	median val - min val
$a-d$	0	max val	max val - median val
$w_1 - \mathit{start}$	0	end - start	0
$w_2 - w_1$	0	end - start	(end - start) / 2
k_1	0.05	10	0.1
k_2	-1	-0.01	-0.03
d	min val	2 max val	median val
a	min val	3 max val	max val
w_1	start	end	start
w_2	start	2 end - 2 start	(start + end)/2

Table 5.2.: Upper and lower bounds as well as starting value for each of the parameters. The boundaries and starting values of d, a, w_1 and w_2 are derived from the parameters used in the implementation of the approximation, shown above the double line.

Starting values can have a big impact on the approximation reached by the algorithm. We want to choose starting values close to the expected resulting parameters. By choosing the starting value of $w_1 - \mathit{start}$ as 0, which corresponds to choosing $w_1 = \mathit{start}$, we favour

the first increase in the data to be the point of activation. Otherwise, we are more likely to mistake an oscillation later in the data as the activation point. As we do not know when the activation happens when setting the boundaries we guess that w_2 will lie somewhere in the middle. Therefore we choose $(end - start)/2$ as the starting value for $w_2 - w_1$. The other starting values are chosen as we expect u to be low, a to be high, d to lie somewhere in the middle. Experimenting showed that k_1 often has a value around 0.1 while k_2 lies around -0.03.

Using algorithm 4 we now describe a routine which handles reading the data, some necessary preprocessing steps and saving of the resulting parameter lists.

Algorithm 5: Approximation Loop

input : file containing data matrix as described in section 4.1
output: parameters of the approximation of all particles as a matrix

```
1  begin
2      read data
3      filter data
4      for each single particle do
5          particle data := (frame, ratio) columns of this particle
6          if length of particle data is too short then
7              skip
8          end
9          parameters := approximate(particle data)
10         optionally show ratio data and approximation
11         save parameters
12     end
13     return matrix of all parameters
14 end
```

Filtering the data is necessary as the ratio can be very large if the denominator is small. Values are therefore bounded to lie within the interval $[0, 5]$. Any values higher than 5 are almost certainly caused by measurement errors. Values below 0 are definitely incorrect, as both the denominator and enumerator are measured as the brightness of a pixel, which can not be negative.

The visualization in line 10 of algorithm 5 generates images such as figure 5.3. It depicts the approximation of two T cells, one of which is activated during the recording, while the other is not. The ratio data recorded is shown in black, while the approximation is shown in orange. It is important to note that the approximation performed on unactivated T cells still uses the same function f_{ac}.

This work uses the python scipy function `scipy.optimize.curve_fit(function, xdata, ydata, p0=starting_values, method='trf', bounds=(lower_bounds, upper_bounds))` as it provides all the necessary functionality. The method parameter `trf` stands for Trust Region Reflective, as described in section 2.2.3.

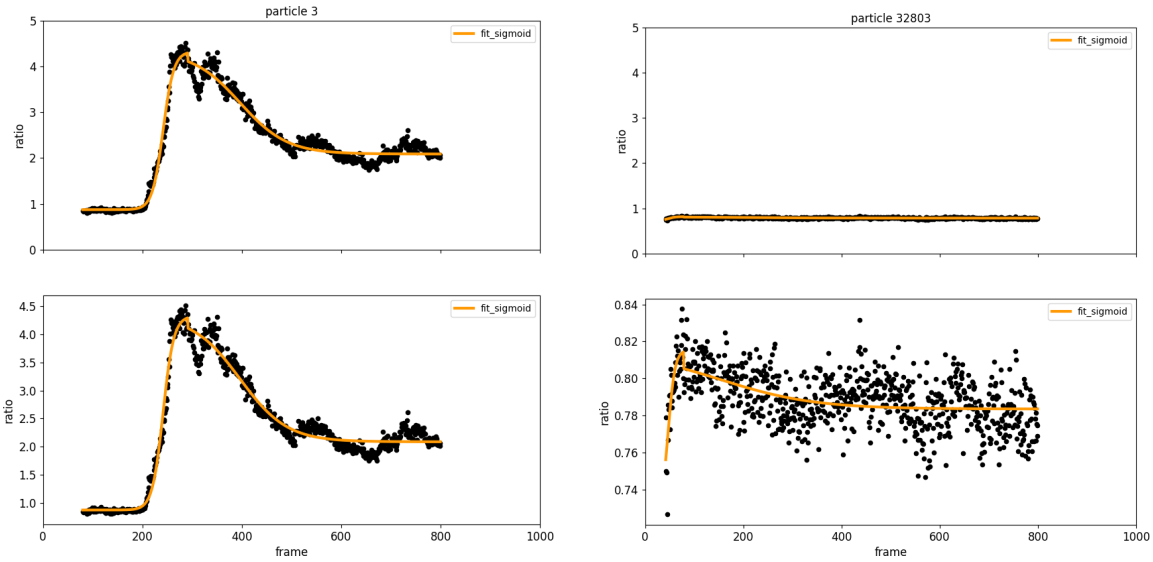

Figure 5.3.: The left plot shows the data in black and approximation in orange of an activated cell. The upper plot is scaled from 0 to 5, the lower one is scaled to fit the data. The right plot shows the same of an unactivated cell.

5.3. Analysis of the Approximation

We now give detailed information on the parameters found from the above approximation. Some statistics are found in table 5.3.

		Positive Control		**Negative Control**		
	Parameter	μ	σ	μ	σ	**Difference**
human cells	a	2.808	0.461	0.923	0.669	1.885
	u	0.663	0.521	0.613	0.311	0.05
	d	1.937	0.491	0.685	0.412	1.252
	k_1	0.263	0.428	0.524	0.963	-0.261
	k_2	-0.059	0.164	-0.163	0.292	0.104
	w_1	142.228	124.012	171.062	131.563	-28.834
	w_2	445.386	185.971	478.843	190.792	-33.457
mouse cells	a	2.9	0.907	0.876	0.186	2.024
	u	0.889	0.27	0.79	0.093	0.099
	d	1.749	0.407	0.804	0.129	0.945
	k_1	0.15	0.409	1.161	1.235	-1.011
	k_2	-0.1	0.195	-0.133	0.267	0.033
	w_1	295.809	77.207	100.712	112.352	195.097
	w_2	469.952	105.375	304.283	179.834	165.669

Table 5.3.: Average μ and standard deviation σ of the parameters retrieved from approximating the human cell data.

Figure 5.4 shows the distribution of the resulting parameters of the approximation. From the figure it seems the differences between activated and unactivated cells is biggest in the parameters activated value a, decreased value d as well as the steepness of increase k_1.

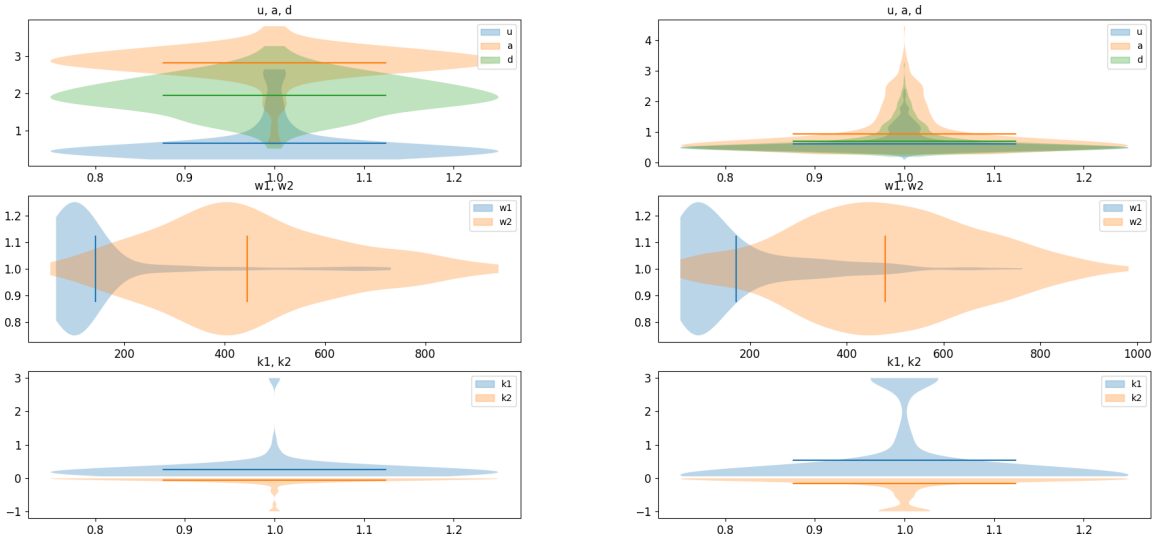

Figure 5.4.: Violin plots of parameters u, a, d, w_1, w_2, k_1 and k_2 from the approximations. The mean is shown by the bar. The width of the individual plots corresponds to the distribution density. The parameters of the positive control are on the left and those of the negative control are on the right. Both are of human cells.

As the datasets are not perfectly labelled, meaning there are activated cells in the negative control and vice versa, we have relatively high standard deviation.

We can use the mean and standard deviation of each of the parameters to find data points that can be considered outliers. We expect wrongly-labelled data, e.g. activated cells in the negative control, to be an outlier in the parameters a and d. However, activated cells in the positive control might have a decreased value d that is very low, around u. This makes it difficult to distinguish activated from unactivated cells when looking at the parameter d. Therefore, we choose a as the only parameter when filtering for these kinds of outliers.

A particle from the positive control dataset that has a value in parameter a higher than the median should still be classified as activated. Only a value lower than some threshold indicates an unactivated cell. The same holds for values of a lower than the median in the negative control dataset. In short, we want to filter out particles with a high value of a in the negative control and those with a low value of a in the positive control dataset. Therefore, the threshold has to be specified as a lower and upper bound in multiples of the standard deviation.

We give pseudocode for the detection of outliers in algorithm 6.

Algorithm 6: Find Outliers

input : data as matrix of all particle parameters of the approximation, threshold as pair, parameters_used as list
output: set of indices of the outliers

```
1  begin
2      outliers := empty set
3      for parameter in parameteres_used do
4          mean := mean(data[parameter])
5          std := standard_deviation(data[parameter])
6          interval := [mean - std · threshold[0], mean - std · threshold[1]]
7          outliers = outliers ∪ {data[index] : data[parameter] ∉ interval}
8      end
9      return outliers
10 end
```

The question of how to choose the threshold will be discussed next. As we do not have information on what percentage of cells behaved correctly in the positive and negative control we do not have enough information to choose threshold values without guessing. Instead, we can manipulate the threshold as a multiple of the standard deviation until we filter out incorrectly labelled data, but would filter out correctly labelled data points if we increase the value. This trial and error approach led to different values for each of the four control datasets, which can be seen in table 5.4.

dataset	lower bound	upper bound
human positive	mean -3 std $= 1.582$	∞
human negative	$-\infty$	mean $+0.5$ std $= 1.306$
mouse positive	mean -2 std $= 1.41$	∞
mouse negative	$-\infty$	mean $+3$ std $= 1.445$

Table 5.4.: Thresholds in outlier detection in the different datasets.

Naturally we can use the same outlier detection with different parameters to find particles where the approximation failed to yield a good result.

These results will be used in section 7.1 to remove wrongly-labelled data from the datasets.

5.4. Adding Oscillation to the Approximation

In order to answer the questions from chapter 1 concerned with the oscillations happening in the decrease of the Ca^{2+} concentration we want to model them as well. We use a method often used when analysing oscillating data, called Fourier Transformation.

Fourier Transformation is used when an application is concerned with cyclic temporal

data. Examples are sound waves, seismic data or oscillations of a skyscraper in strong wind. This data can be represented as a function of amplitude over time. Most of the time we are not interested in the amplitude at a specific point in time, as a temporal shift would represent very similar information.

As the function of oscillations in the Ca^{2+} concentration in T cells is almost cyclic we might be interested in a decomposition into simple cyclic functions, such as sinoid functions. Then, we can analyse the most prominent frequencies and their respective amplitudes. This gives a representation of the data, that can be easier to interpret. Fast Fourier Transformation (FFT) is an algorithm that transforms temporal data into such a representation of a weighted sum of sines. Pseudocode and a detailed description are provided by Cormen et al[Cor+09].

As the oscillations happen in the decrease of the Ca^{2+} concentration we apply FFT to that part of the data. We can use the frequency with the highest amplitude and use them to further analyse the oscillations. After having found this estimation of the frequency we can use the method from before, but with a sine function instead of the sigmoid functions. In detail, we once again utilize `scipy.optimize.curve_fit`, this time using the function

$$f(x, A, \omega, p) := A \cdot \sin(\omega \cdot t + p),$$

with parameters amplitude A, angular frequency ω and phase p. The starting value for the approximation can be set using the results of the FFT. We denote the frequency with the highest amplitude returned from the FFT by $freq_{guess}$ and the standard deviation of the approximation residuum time series with $std(Y)$. Then, we can set the start value of A to be $\sqrt{2 \cdot std(Y)}$, the one of ω to be $2\pi|freq_{guess}|$ and the one of p to be 0.

The drawback of this method is that only oscillations that can be represented by a single sine function can be modelled. However, for our purpose we can assume that the oscillations are of this form. This gives an even better approximation of the data, which can be seen in figure 5.5. We store the data gathered from the approximation as a list of frequency, amplitude and phase.

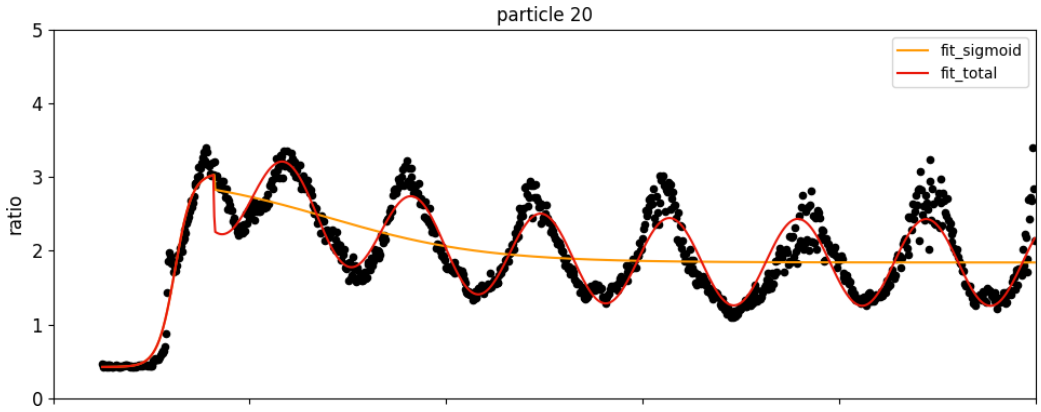

Figure 5.5.: The data of an activated cell with heavy oscillations is shown in black, simple approximation in orange and the approximation with FFT added in red. The first plot is scaled from 0 to 5, the second one is scaled to fit the data.

6. Clustering Algorithms and Application

The objective of classification is to find assignments between data points and categories. For some applications this can be done by taking correctly labelled data and comparing a new data point to the data points in different categories to see which category best fits. This is called supervised clustering. One such algorithm is k-nearest-neighbour. In our context the issue with this approach is that the data is only labelled as to which experiment it came from, e.g. positive control in human cells or negative control in mouse cells. However, as noted before not all cells from these experiments behaved as we expected them to, e.g. some cells activated in the negative control or did not activate in the positive control. Therefore, we choose to use a clustering algorithm which does not have the need for classified training data. In contrast, this is called unsupervised clustering.

6.1. Gaussian Mixture Model

This section follows the article Gaussian Mixture Model by Reynolds [Rey+09].

Often gathered observations are distributed as a normal distribution. These distributions have a density function

$$g(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

with parameters D, called the dimension, μ, called the mean vector, and Σ, called the covariance matrix.

As we are concerned with clustering data points we expect the observed data points from different clusters to have different parameters in the normal distribution they come from. Assuming we have n different data sources gives us normal distributions $g_i(x|\mu_i,\Sigma_i)$ where $i = 1,...,n$. Additionally, we might have more data points being generated from some normal distributions while less from others. We can express this using another parameter for the weights w_i with $i = 1,...,n$. To normalise the weights we set the constraint $\sum w_i = 1$.

The distribution describing the entire dataset can be described with the distribution

$$p(x) = \sum_{i=1}^{n} w_i g_i(x|\mu_i,\Sigma_i). \tag{6.1}$$

Gaussian mixture model is a method to retrieve these parameters w_i, μ_i and Σ_i for some D dimensional data points generated from n normal distributions. From these parameters it is easy to cluster the data, as we know where data points from the different clusters are expected to be positioned.

From equation 6.1 we expect every Σ_i to be independent of each other. In the context of gaussian mixture models this is called having a full covariance matrix. However, we can

eliminate some of the variables in the covariance matrix if we choose a diagonal covariance matrix. Additionally, we might specify to use the same covariance matrix for all i, which is called tied in this context.

Choosing a full covariance matrix is not necessary even if the data is expected to have statistically independent features, as the overall density is compromised from multiple normal distributions with diagonal Σ_i. This enables us to model correlations between features.

The question now is how we can derive the parameters w_i, μ_i and Σ_i. We choose the approach which chooses the parameters where the likelihood of data being generated by these parameters is maximal. The likelihood can be expressed as

$$L(w_i, \mu_i, \Sigma_i|X) = p(X|w_i, \mu_i, \Sigma_i) = \prod_{t=1}^{n} p(x_t|w_i, \mu_i, \Sigma_i)$$

with $X = (x_1, ..., x_n)$ being the recorded data. As $L(w_i, \mu_i, \Sigma_i|X)$ is non-linear in the parameters, deriving the maximum is not trivial. Instead, we use an iterative approach which approaches the solution. Define $\lambda = (w_i, \mu_i, \Sigma_i)$. Simplifying to a diagonal covariance matrix gives us the iterative algorithm where we define the successor values $\bar{\cdot}$ as

$$\begin{aligned}
Pr(i|x_t, \lambda) &:= \frac{w_i g(x_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{n} w_k g(x_t|\mu_k, \Sigma_k)}, \\
\bar{w_i} &:= \frac{1}{n} \sum_{t=1}^{n} Pr(i|x_t, \lambda), \\
\bar{\mu_i} &:= \frac{\sum_{t=1}^{n} Pr(i|x_t, \lambda) x_t}{\sum_{t=1}^{n} Pr(i|x_t, \lambda)}, \\
\bar{\sigma_i^2} &:= \frac{\sum_{t=1}^{n} Pr(i|x_t, \lambda) x_t^2}{\sum_{t=1}^{n} Pr(i|x_t, \lambda)} - \bar{\mu_i^2}.
\end{aligned}$$

for w_i, μ_i and σ_i^2 respectively. One can show that with this iteration rule we have $p(X|\bar{\lambda}) \geq p(X|\lambda)$. The value $Pr(i|x_t, \lambda)$ is known as the a posteriori probability for the i-th component.

From the parameters w_i, μ_i and Σ_i we can get the probability of a data point belonging to the i-th cluster. This data point can be either one of the ones used to get the parameters or a new one.

6.2. K-Means

We want to explore a second clustering algorithm as well. Like gaussian mixture model this method assigns data points to clusters after having been trained with unlabelled data.

Once again we assume to have k different data sources from where the data stems. As the data from every of these resulting clusters will be close to each other, the variance within the cluster is relatively small compared to the variance of data points from different clusters.

The variance of data points $(x_n)_{n=1,...,m}$ is calculated as

$$\mathrm{Var}((x_n)_{n=1,...,m}) = \frac{1}{m}\sum_{n=1}^{m} ||x_n - \mu||^2, \qquad \text{where} \qquad \mu = \frac{1}{m}\sum_{n=1}^{m} x_n.$$

We can therefore formulate the problem as a minimization problem. Using k clusters $S_1, ..., S_k$, with $|S_l|$ data points in each, we want to minimize the sum of variants

$$\sum_{l=1}^{k}\sum_{x\in S_l} ||x - \mu_l||^2 = \sum_{l=1}^{k} |S_l|\mathrm{Var}(S_l).$$

We want to show that this is equivalent to minimizing

$$\sum_{l=1}^{k}\frac{1}{|S_l|}\sum_{x,y\in S_l} ||x - y||^2.$$

Denoting the expectation with $\mathbb{E}$, this follows from the equality

$$\begin{aligned}
&\frac{1}{|S_l|}\sum_{x,y\in S_l} ||x - y||^2 \\
&= \frac{1}{|S_l|}\sum_{x,y\in S_l} ||x||^2 - 2||x||||y|| + ||y||^2 \\
&= \frac{1}{|S_l|}\left(|S_l|\sum_{x\in S_l} ||x||^2 - 2\sum_{x,y\in S_l} ||x||||y|| + |S_l|\sum_{y\in S_l} ||y||^2\right) \\
&= \frac{1}{|S_l|}(|S_l|^2\mathbb{E}(S_l^2) - 2\left(\sum_{x\in S_l} ||x||\right)\left(\sum_{y\in S_l} ||y||\right) + |S_l|^2\mathbb{E}(S_l^2)) \\
&= 2|S_l|\mathbb{E}(S_l^2) - 2|S_l|\mathbb{E}(S_l)^2 \\
&= 2|S_l|\mathrm{Var}(S_l).
\end{aligned}$$

Minimizing is hard, but is typically solved by the approximation using the following algorithm consisting of two steps.

- Assign all data points to clusters $S_1, ..., S_k$, by choosing the cluster S_l with the closest mean μ_l.
- Update the means μ_l by calculating them as the mean of the data points assigned to S_l.

We formulate this mathematically as the successor $\hat{S}_l$ and $\hat{\mu}_l$ of S_l and μ_l respectively being

$$\hat{S}_l := \{x : ||x - \mu_l|| \leq ||x - \mu_j||, \forall j \in \{1, ..., k\}\}, \qquad \hat{\mu}_l := \frac{1}{|\hat{S}_l|}\sum_{x\in\hat{S}_l} x.$$

From the definition of $\hat{S}_l$ and $\hat{\mu}_l$ it follows that

$$\sum_{l=1}^{k} \frac{1}{|\hat{S}_l|} \sum_{x,y \in \hat{S}_l} ||x-y||^2 \leq \sum_{l=1}^{k} \frac{1}{|S_l|} \sum_{x,y \in S_l} ||x-y||^2.$$

The convergence of the algorithm towards a minimum seems likely. The choice of initial values can have a big impact on the result.

6.3. Implementation of the Clustering Algorithm

Python offers an implementation of gaussian mixture model and k-means with the sklearn package. The function with parameters relevant to us is `sklearn.mixture.GaussianMixture(n_components, covariance_type, n_init)` and `sklearn.cluster.KMeans(n_clusters, n_init)`. The number of components `n_components` or `n_clusters` can be any positive integer. The `covariance_type` can be one of `'full'`, `'tied'`, `'diag'` or `'spherical'` and describes what type of covariance matrix is used. The parameter `n_init` lets the algorithm run multiple times and returns the parameters of the best clustering achieved.

As the input of the clustering we use data points of the form `[a, u, d, k1, k2, w1, w2]`, where `a, u, ..., w2` are the parameters of the approximation from chapter 5. As our goal is to separate data points from the four data sets mouse cells and human cells each with a negative and a positive control, we use all particles as input. The pseudo code in algorithm 7 describes the steps performed to reach a clustering of the data.

Algorithm 7: Cluster into Activated and Unactivated Cells

input : parameters of the approximations of all particles in all data sets
output: assignments to different clusters for each particle, parameters specifying each cluster

```
1 begin
2     initialize gaussian mixture or k-means by specifying the number of components
        and covariance_type for gaussian mixture
3     apply clustering method to matrix of all parameters of the approximations of
        all particle data sets
4     assign particles to clusters according to clustering results
5     compare assignments from clustering to those of the data set the data stems
        from
6     return assignments to clusters, parameters specifying every cluster
7 end
```

When comparing different covariance types in the gaussian mixture we see that using `'diag'` we have the lowest error rate. The details are shown in table 6.1. Why reducing the number of parameters in the covariance matrix can yield better results is described in section 6.1.

Using a diagonal covariance matrix we can now try to separate the data sets and visualize the results. As the data is 7 dimensional we show lower dimensional representations of the

Covariance Type	Percentage
`full`	13.23%
`tied`	12.7%
`diag`	7.17%
`spherical`	31.52%

Table 6.1.: Error as a percentage of particles being assigned to the wrong component.

data both as it is assigned according to the data set it stems from as well as the assignment from algorithm 7.

Choosing good axis for visualizing high dimensional data is tricky. One approach for minimizing the data lost by the lower dimensional representation is called principal component analysis. Axis are specifically chosen to maximize variance along those axes, which we assume corresponds to information displayed along the axis. The drawback is that the new axis can be more difficult to intuitively understand. The results can be seen in figure 6.1.

By comparing which data points are from which data set and where they were assigned by the clustering method we can find which cluster most likely belongs to the positive control, and which cluster corresponds to the negative control.

The relevant information derived from the gaussian mixture clustering is the means and covariances of the four components. For k-means clustering the means of the components are relevant. In both cases we can decide which cluster a new data point belongs to from this information. A use case might be to find percentages of activated cells in an experiment. Distinguishing between mouse and human cells does not have a clear use case. When specifying `n_components=2` we assume to have a cluster for activated and a second for unactivated cells.

In comparison to the approach focusing on outlier detection in section 5.3, we now have an approach that is not only independent on parameters specified by a user, but can also be applied to a greater set of problems. A proposed way of answering the research questions from chapter 1 using these methods is described in chapter 7.

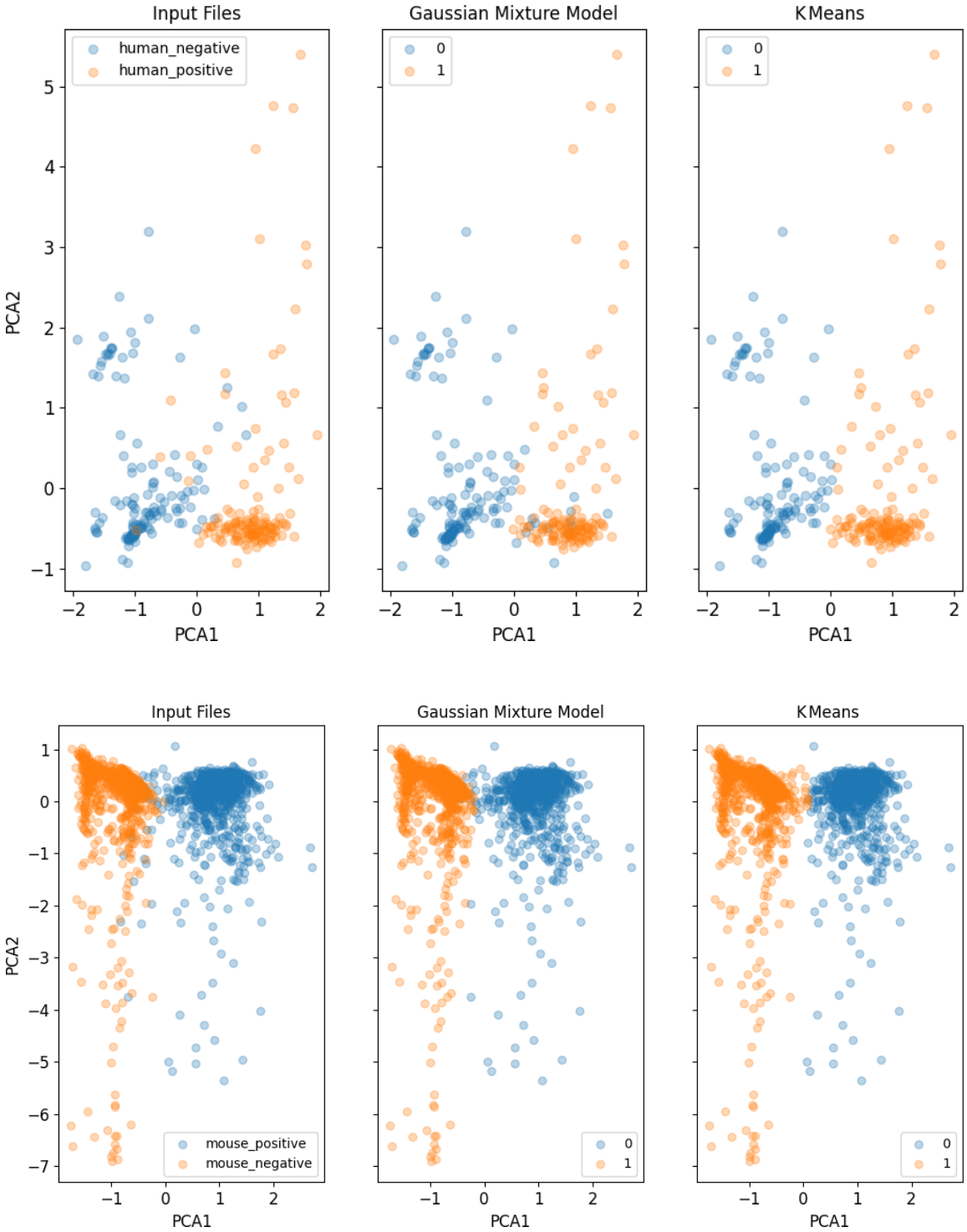

Figure 6.1.: Visualisation of the clustered data points in human T cells on the top and mouse T cells on the bottom row. The left image depicts the clustering according to positive and negative control. The middle image shows the clustering according to gaussian mixture and the right image depicts the clustering according to k-means.

7. Results

Combining all the methods and algorithms from the previous chapters, such as approximation, clustering and outlier detection, gives us insight into the Ca^{2+} concentration data. We give an algorithm for detecting activated T cells, look at differences between mouse and human T cells, analyse the oscillations and investigate whether there are different types of activated T cells.

7.1. Proposed Algorithm for Detecting Activated T Cells

A relevant question this work aims to provide an answer to is, how pre-activated, unactivated and activated T cells can be distinguished.

First we give a method for filtering the pre-activated cells from a dataset. From their nature we expect a high value in Ca^{2+} concentration at the start of the recording. Using the approximation from chapter 5 it is easy to get the approximate Ca^{2+} concentration value at the start of the recording, as it is the parameter u. Using the algorithm 6 with parameters threshold as $[\infty, 0.5]$ and parameters_used as $[u]$ gives good results. It returns the indices of particles, which are pre-activated in the data sets of the positive controls.

After having filtered out pre-activated particles, we want to distinguish between unactivated and activated particles. For this we propose the following steps:

1. get positive control, negative control and experiment recordings
2. transform each particle time series of all three data sets to the parameter list by approximating it with a combination of sigmoid functions, according to chapter 5, using the algorithm 5
3. normalise each of the parameters of all data sets, such that the mean is 0 and the standard deviation is 1
4. use outlier detection, which is described in algorithm 6, to filter out non-conforming cells from both the positive and negative control, as well as pre-activated cells, and particles where the approximation yielded suboptimal results
5. sample particles from the filtered positive and negative control groups to match the number of particles in the both groups
6. use clustering method, as one of the two described in chapter 6, with parameters of selected negative and positive control as input to get the clustering parameters
7. predict the membership of the experiment particle parameters to the clusters to get a prediction of activation

The third and fifth step is beneficial when clustering, as many clustering methods expect equally sized clusters that are centred and have equal standard deviation.

This algorithm is implemented in Python and shown in appendix A. Applying this to a positive, negative control and an experiment data sets of mouse T cells gives the exemplary results shown in table 7.1. The experiment dataset contains T cells that came in contact with a medium density of MHC-peptide complex replicas. The number of activated cells detected in the positive and negative control are shown as a reference point. Comparing them in the case illustrated suggests that in this case gaussian mixture model is the better clustering algorithm.

	file	activated	out of	percentage
gaussian mixture	negative control	47	969	4.850%
	positive control	932	969	96.182%
	experiment	716	890	80.449%
k-means	negative control	53	969	5.470%
	positive control	926	969	95.562%
	experiment	687	890	77.191%

Table 7.1.: Output of the proposed algorithm applied to three files of mouse T cells.

The algorithm can be adapted by using different clustering methods, or specifying other methods of separating the particles based on the parameters derived.

7.2. Difference between Mouse and Human Cells

Now, that we have an algorithm for detecting activated cells, we might ask whether there are notable differences between mouse and human cells. Plotting all data points in a single plot, as seen in figure 7.1, shows whether differences are to be expected.

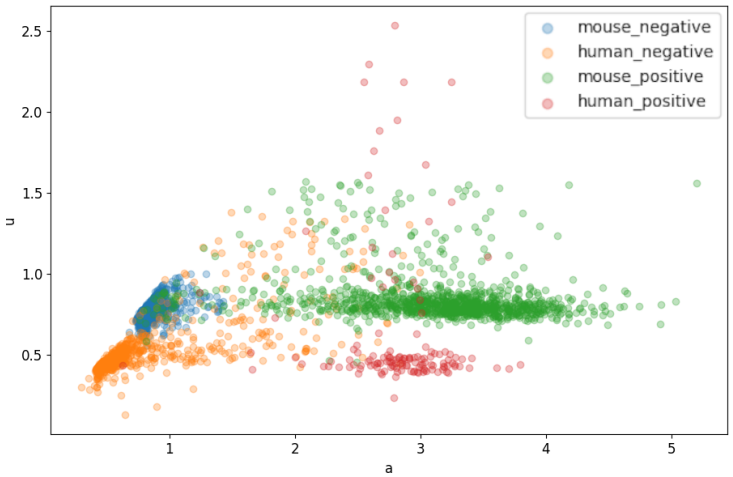

Figure 7.1.: All data points from the four control data sets plotted along the axes activated value a and unactivated values u.

Differences in the parameters a and u are present. On average the unactivated value u and activated value a are bigger in mouse T cells than the human counterparts. Looking at each of the parameters gathered from the approximation is possible by using the means and standard deviation of each parameter. These values are displayed in table 7.2. It shows the mean μ and standard deviation σ of the negative and positive control of mouse and human T cells.

			a	u	d	$k1$	$k2$	$w1$	$w2$
mouse	**pos**	μ	3.083	0.85	1.698	0.08	-0.052	300.77	491.275
		σ	0.72	0.159	0.33	0.085	0.07	82.451	104.304
	neg	μ	0.862	0.77	0.788	1.262	-0.075	88.644	304.531
		σ	0.106	0.057	0.053	1.219	0.148	99.833	184.08
human	**pos**	μ	2.854	0.553	1.916	0.23	-0.033	113.595	438.763
		σ	0.357	0.294	0.44	0.219	0.073	59.7	175.998
	neg	μ	0.864	0.543	0.604	0.58	-0.19	169.394	528.006
		σ	0.56	0.197	0.274	1.022	0.332	142.358	229.078

Table 7.2.: Mean μ and standard deviation σ of the parameters of the approximation for positive and negative controls in human and mouse cells.

By comparing the means from table 7.2 we see that indeed the mean of a and u are higher or about equal in mouse cells. Differences in d, $k1$ and $k2$ are less pronounced and differ between positive and negative control. Differences in $w1$ and $w2$ are probably caused by differences in timing between the different recordings of the control groups. Therefore, comparing the means does not give much valuable information. In conclusion the differences are biggest in the parameter u.

7.3. Oscillation in Decrease

Looking at the frequencies, amplitudes and phases returned from the approximation described in section 5.4 gives us an idea with which period the typical T cell oscillates in Ca^{2+} concentration during the decrease after activation. Figure 7.2 shows violin plots of all three parameters in the positive control on human T cells. The means in this data set are 0.005 for the frequency, 0.01 for the amplitude and -0.001 for the phase. The details of the density distribution are shown in the width of each of the violin plots.

As some T cells do not oscillate, the approximation sometimes returns a misleading frequency with a low amplitude. We want to filter these cells before looking at the typical period with which the Ca^{2+} concentration oscillates. We choose arbitrary threshold values, that seemed sensible to the author. Namely, all particles with an amplitude less than 0.1 or a frequency less than 0.005 are removed before once again looking at the average values. Now the average frequency is 0.008 with an average amplitude of 0.223. This corresponds to a period of oscillation being $1/freq = 125$ seconds. Indeed, looking at the example provided in figure 7.2, this seems like a good estimate.

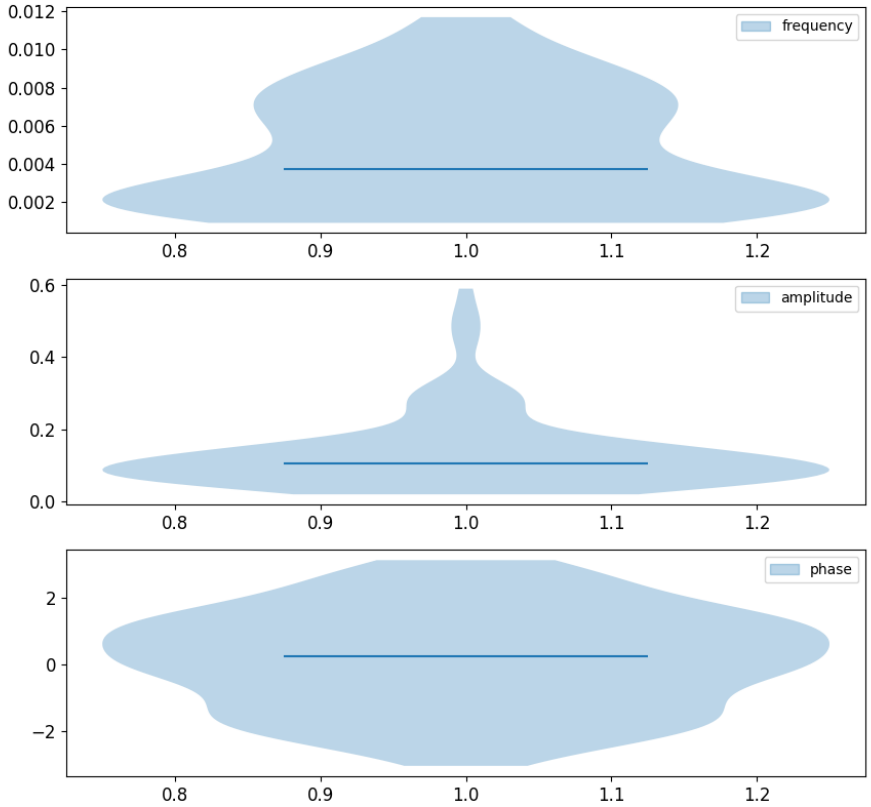

Figure 7.2.: Violin plots of the frequencies, amplitudes and phases of the oscillation approximation in the positive control dataset of human T cells.

Lastly we compare this to the positive control in mouse cells. Before filtering the averages are 0.007 for the frequency, 0.071 for the amplitude and 0.288 for the phase. After filtering, with the same thresholds as for human cells, we have an average of 0.007 for the frequency and 0.177 for the amplitude. Converting this to the period gives us 143 seconds. This compares to the average period found in human T cells.

7.4. Types of Activated Cells

It is interesting to see whether there are different types of activated cells. We find an answer by looking at the activated cells only. We have two data sets of activated cells, one from mouse cells and one from human cells. Naturally there will be differences between the two. This was further explored in section 7.2. By looking at only one data set of the positive

control at a time we generate the images seen in figure 7.3. To reduce the dimensionality once again Principal Component Analysis is used. From figure 7.3, it is not apparent that there are different types of activated cells.

Lastly we can have a look whether particles differ in the oscillation. Using the approximation of the oscillation and the filtering done in section 7.3 we can for example differ between oscillating T cells, marked by high amplitude and period around the expected 125 seconds, and non-oscillating T cells, marked by low amplitude or different period.

We note that out of 139 particles in the positive control of human T cells only 14 have the behaviour of a heavily oscillating T cell. In the mouse positive control it is 192 out of 1058.

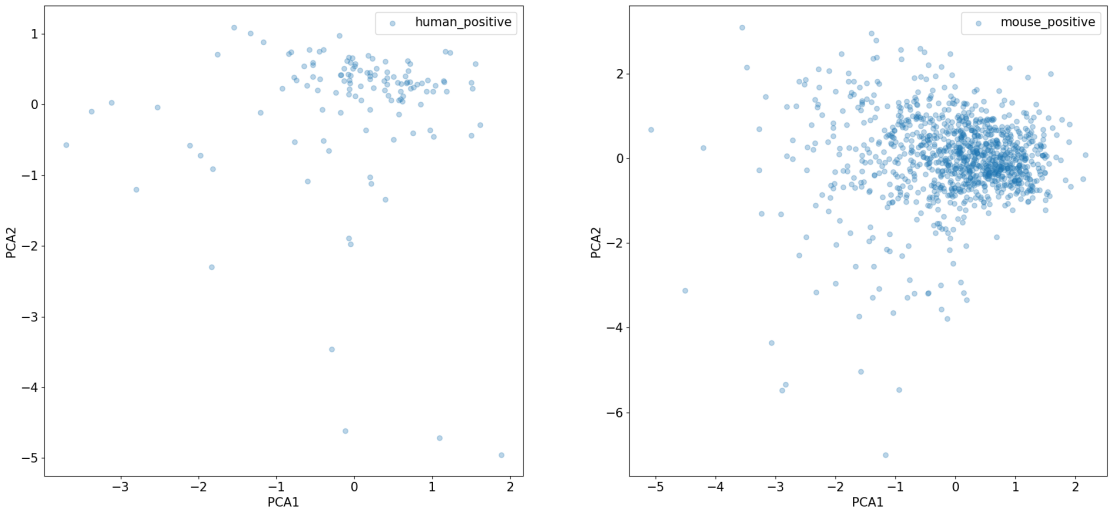

Figure 7.3.: First two axes of the Principal Component Analysis of activated human and mouse T cells. There appear not be any major differences between the two types of T cells.

Using the approximation, all the research questions have been investigated. The resulting algorithm has versatile applications.

8. Discussion

This chapter gives answers to the questions from chapter 1 using the results from chapter 7 and gives an outlook which aspects could be improved in further work.

- Which criteria can distinguish between unactivated, activated and pre-activated cells?

  Pre-activated cells can be detected by having a parameter u that is bigger than $\mu + 0.5\sigma$, with μ being the mean and σ being the standard deviation of u. Distinguishing between activated and unactivated T cells can be done by the algorithm proposed in section 7.1.

- Do different types of activated cells exists? How are they different?

  There are no apparent differences between different activated T cells. The parameters found behave according to a normal distribution. Differences can be found in how pronounced the oscillations are.

- With which frequencies does the Calcium concentration repeat after activation?

  The oscillations appear to have a frequency of about 0.0075 which corresponds to a period of 133 seconds in both human and mouse T cells.

- Is there a difference between mouse and human cells?

  There are differences between the two different types of T cells. The biggest differences are present in the parameter u of the Ca^{2+} concentration before activation.

As the first question is the most relevant we want to discuss the answer further. To give a value for accuracy achieved by the proposed algorithm we have to know how many T cells from the experiment dataset are activated. We do this by using some of the particles from the positive and negative control. Using random sampling we select some particles from each control data set to be the experiment data. Of course, we then have to remove these particles from the control data to avoid training on the data we test with. By varying the number of particles chosen from the two control data sets we can vary the percentage of activated cells in this generated experiment data. The results from this test can be found in table 8.1.

We now discuss possible improvements. The proposed algorithm and subsequent analysis of the clustering performed can be improved by applying the discussed algorithm to enough data points to achieve a good approximation for average and standard deviation that hold for arbitrary data points from the same T cell line.

Another improvement can be made in the frequency analysis. Exploring what causes these oscillations can give a better understanding as to what type of oscillation is to be expected and how these might differ between different T cells. The assumption made in

percentage activated	k-means output	gaussian mixture output
0%	5.5%	4%
25%	27.5%	26%
50%	49.5%	49.5%
75%	72%	70.5%
100%	96%	97.5%

Table 8.1.: Percentage activated output by k-means and gaussian Mixture for different true values of percentages of activated T cells.

this work, that the oscillations can be modelled by a single sine wave, might prove wrong. In this case a different modelling approach has to be chosen. Alternatively, the output of the FFT can be used directly.

Lastly, the performance of the Python implementation given in the appendix A is sub-optimal. Changing to other programming languages or improving the performance of the code in Python desirable.

In summary this work proposed an algorithm for detecting activated T cells in Ca^{2+} concentration recordings, without a need for user specified criteria for activation detection. The accuracy of this algorithm is acceptable, and the implementation is given in Python.

List of Figures

List of Tables

List of Algorithms

Bibliography

[AH24] K Maude Ashby and Kristin A Hogquist. "A guide to thymic selection of T cells". In: *Nature Reviews Immunology* 24.2 (2024), pp. 103–117.

[AW04] Robert T Abraham and Arthur Weiss. "Jurkat T cells and development of the T-cell receptor signalling paradigm". In: *Nature reviews immunology* 4.4 (2004), pp. 301–308.

[BCL99] Mary Ann Branch, Thomas F Coleman, and Yuying Li. "A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems". In: *SIAM Journal on Scientific Computing* 21.1 (1999), pp. 1–23.

[Bjö96] Åke Björck. "The Nonlinear Least Squares Problem". eng. In: *Numerical Methods for Least Squares Problems*. siam, 1996, pp. 339–358. ISBN: 978-0-898713-60-2.

[BV18] Stephen Boyd and Lieven Vandenberghe. "Levenberg–Marquardt algorithm". eng. In: *Introduction to Applied Linear Algebra*. Cambridge University Press, 2018, pp. 391–399. ISBN: 781316518960.

[CA22] Geoffrey M Cooper and Kenneth Adams. "The Nucleus". eng. In: *The cell: a molecular approach*. 19. edition. Oxford University Press, 2022, pp. 336–364. ISBN: 9780197583722.

[Cor+09] Thomas H Cormen et al. "Polynomials and the FFT". eng. In: *Introduction to Algorithms*. Massachusetts Institute of Technology, 2009, pp. 898–925. ISBN: 978-0-262-53305-8.

[CŻ13] Edwin K. P. Chong and Stanislaw H. Żak. "Gradient Methods". eng. In: *An Introduction to Optimization*. Wiley, 2013, pp. 131–159. ISBN: 9781118279014.

[Gan12] William F. Ganong. "Overview of Cellular Physiology in Medical Physiology". eng. In: *Review of medical physiology*. 24. edition. Stamford, Conn: McGraw-Hill, 2012, pp. 35–66. ISBN: 9780071780032.

[Gan97] William F. Ganong. "Circulating Body Fluids". eng. In: *Review of medical physiology*. 18. ed. Stamford, Conn: Appleton & Lange, 1997, pp. 486–488. ISBN: 9780838584439.

[Hag18] Kimberly Hagel. *Positive and Negative Selection of T Cells*. 2018. URL: https://immunobites.com/2018/08/20/positive-and-negative-selection-of-t-cells/ (visited on 06/21/2024).

[JRB14] Noah Joseph, Barak Reicher, and Mira Barda-Saad. "The calcium feedback loop and T cell activation: how cytoskeleton networks control intracellular calcium flux". In: *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1838.2 (2014), pp. 557–568.

[KCF18] Brahma V Kumar, Thomas J Connors, and Donna L Farber. "Human T cell development, localization, and function throughout life". In: *Immunity* 48.2 (2018), pp. 202–213.

[Lew01] Richard S Lewis. "Calcium Signaling Mechanisms in T Lymphocytes". In: *Annual Review of Immunology* 19.Volume 19, 2001 (2001), pp. 497–521. ISSN: 1545-3278.

[ML23] Jonathan Morgan and Alan E Lindsay. "Modulation of antigen discrimination by duration of immune contacts in a kinetic proofreading model of T cell activation with extreme statistics". In: *PLOS Computational Biology* 19.8 (2023), e1011216.

[MMS17] Magdiel Martínez, Namyr A Martínez, and Walter I Silva. "Measurement of the intracellular calcium concentration with Fura-2 AM using a fluorescence plate reader". In: *Bio-protocol* 7.14 (2017), e2411–e2411.

[NW99] Jorge Nocedal and Stephen J. Wrigh. "The Dogleg Method". eng. In: *Numerical Optimization*. Springer, 1999, pp. 73–76.

[Rey+09] Douglas A Reynolds et al. "Gaussian mixture models." In: *Encyclopedia of biometrics* 741.659-663 (2009).

[RLH90] MW Roe, JJ Lemasters, and B Herman. "Assessment of Fura-2 for measurements of cytosolic free calcium". In: *Cell calcium* 11.2-3 (1990), pp. 63–73.

[Rog24] Kara Rogers. *endoplasmic reticulum*. 2024. URL: https://www.britannica.com/science/endoplasmic-reticulum (visited on 06/23/2024).

[SB16] Dianne S. Schwarz and Michael D. Blower. "The endoplasmic reticulum: structure, function and response to cellular signaling". In: *Cellular and Molecular Life Sciences* 73 (2016), pp. 79–94.

[SI13] Peter B Stathopulos and Mitsuhiko Ikura. "Structural aspects of calcium-release activated calcium channel function". In: *Channels* 7.5 (2013). PMID: 24213636, pp. 344–353.

[SKJ09] Jennifer E Smith-Garvin, Gary A Koretzky, and Martha S Jordan. "T cell activation". In: *Annual review of immunology* 27 (2009), pp. 591–619.

[SSB77] Ulrich Schneider, Hans-Ulrich Schwenk, and Georg Bornkamm. "Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma". In: *International journal of cancer* 19.5 (1977), pp. 621–626.

[VL04] C Voglis and IE Lagaris. "A rectangular trust region dogleg approach for unconstrained and bound constrained nonlinear optimization". In: *WSEAS International Conference on Applied Mathematics*. Vol. 7. 2004.

A. Python Implementation

We give a minimal version of the algorithm and code described in this work in Python.

```python
import time
import math
import scipy
import numpy as np
import pandas as pd
from sklearn.mixture import GaussianMixture
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler


def progress_bar(iterable, prefix=""):
    start = time.time()
    for i, item in enumerate(iterable):
        yield item
        x = i * 10 // len(iterable)
        print(f'\r{prefix}|{x * "*"}{(10 - x) * " "}| {i}/{len(iterable)}',
              end='', flush=True)
    minutes, sec = divmod(time.time() - start, 60)
    print(f"\r{prefix} took {int(minutes): 02}min {sec: 03.1f}s")


def approximate(dataframe):
    def calc_t(w, k, alpha=0.99):
        tmp = 1 / alpha - 1
        if tmp < 0.0001:
            return None
        try:
            return w - math.log(tmp) / k
        except ValueError:
            return None

    def approx_func_(x, w1, t, w2, a, d, u, k1, k2):
        if t is None:  # transition point lies outside datapoints
            return u
        elif x <= t:  # logistic function before transition point
            tmp = -k1 * (x - w1)
            if tmp <= 32:
                res = (a - u) / (1 + math.exp(tmp))
            else:
                res = 0
            return res + u
        else:  # logistic function after transition point
            tmp = -k2 * (x - w2)
            if tmp <= 32:
                res = (a - d) / (1 + math.exp(tmp))
```

```
            else:
                res = 0
            return res + d

    def approx_func(x_arr, w1_start, w2_w1, a_d, d_u, u, k1, k2):
        w1 = w1_start + start
        w2 = w2_w1 + w1
        d = d_u + u
        a = a_d + d
        transition_point = calc_t(w1, k1)
        return (np.vectorize(approx_func_)
                (x_arr, w1, transition_point, w2, a, d, u, k1, k2))

    min_val = np.min(dataframe['ratio'])
    median_val = np.median(dataframe['ratio'])
    max_val = np.max(dataframe['ratio'])
    start, end = min(dataframe['frame']), max(dataframe['frame'])

    lower_bounds = (0, 0, 0, 0, min_val, 0.05, -1)
    upper_bounds = (end - start, end - start, max_val, max_val,
                    max_val, 3, -0.01)
    p0 = (0, (end - start) / 2, max_val - median_val, median_val - min_val,
          min_val, 0.1, -0.03)

    popt, *_ = scipy.optimize.curve_fit(approx_func, dataframe['frame'],
                                        dataframe['ratio'], p0=p0,
                                        method='trf',
                                        bounds=(lower_bounds, upper_bounds))

    w1_start, w2_w1, a_d, d_u, u, k1, k2 = popt
    w1 = w1_start + start
    w2 = w2_w1 + w1
    d = d_u + u
    a = a_d + d
    t = calc_t(w1, k1)
    return {"a": a, "u": u, "d": d, "k1": k1, "k2": k2, "w1": w1, "w2": w2}


def approximation_loop(file_name):
    data = pd.DataFrame(pd.read_hdf(f"../data/{file_name}.h5"))

    # filter data
    data = data[np.isfinite(data["ratio"])]
    data = data[np.less(data["ratio"], np.full((len(data["ratio"])), 5))]
    data = data[np.greater(data["ratio"], np.full((len(data["ratio"])), 0))]
    if "incl_act" in data.columns:
        data = data[data["incl_act"]]
    if "activated" in data.columns:
        if "pos" in file_name:
            data = data[data["activated"] == "activated"]
        elif "neg" in file_name:
            data = data[data["activated"] == "non_act"]

    all_parameters = list()
    parameters_saved = ["idx", "a", "u", "d", "k1", "k2", "w1", "w2"]
```

```
101
102     for particle_idx in progress_bar(set(data['particle']),
103                                      f"approx {file_name}: "):
104         single_particle_data = (
105             data.loc)[data['particle'] == particle_idx][['frame', 'ratio']]
106
107         # skip if too few datapoints
108         if len(single_particle_data['frame']) < 300:
109             continue
110
111         try:  # throws error if no best fit was found
112             parameters = approximate(single_particle_data)
113
114             parameters["idx"] = str(particle_idx) + file_name
115             all_parameters.append([parameters[e] for e in parameters_saved])
116
117         except Exception as e:
118             print(f"\nerror in particle {particle_idx}: {e}")
119
120     return pd.DataFrame(all_parameters, columns=parameters_saved)
121
122
123 def load_from_file_or_approx(file_name):
124     try:
125         df = pd.read_hdf(f"intermediate/par_{file_name}.h5", "parameters")
126     except FileNotFoundError:
127         df = approximation_loop(file_name)
128         df.to_hdf(f"intermediate/par_{file_name}.h5", key="parameters")
129     return df
130
131
132 def normalize(neg_df, pos_df, exp_df, normalized_columns):
133     all_data = pd.concat([neg_df, pos_df])
134
135     scaler = StandardScaler()
136     all_data[normalized_columns] = (scaler.fit_transform(
137         all_data[normalized_columns]))
138
139     neg_df = all_data[all_data["idx"].isin(neg_df["idx"])]
140     pos_df = all_data[all_data["idx"].isin(pos_df["idx"])]
141     exp_df[normalized_columns] = (
142         scaler.transform(exp_df[normalized_columns]))
143
144     return neg_df, pos_df, exp_df
145
146
147 def remove_outliers(data, width, par_used):
148     for par in set(par_used):
149         mean, std = data[par].mean(), data[par].std()
150         data = data.drop(data[(data[par] <= mean - std * width) &
151                               (data[par] >= mean + std * width)].index)
152     return data
153
154
155 def separate(neg_df, pos_df, prediction_parameters, clustering):
```

```
156     neg_df["activation"] = "negative"
157     pos_df["activation"] = "positive"
158     data = pd.concat([neg_df, pos_df])
159 
160     data["predicted"] = clustering.fit_predict(data[prediction_parameters])
161 
162     # find association between predicted clusters and files
163     neg_0 = len(data[(data['predicted'] == 0) &
164                      (data['activation'] == "negative")])
165     neg_1 = len(data[(data['predicted'] == 1) &
166                      (data['activation'] == "negative")])
167     pos_0 = len(data[(data['predicted'] == 0) &
168                      (data['activation'] == "positive")])
169     pos_1 = len(data[(data['predicted'] == 1) &
170                      (data['activation'] == "positive")])
171 
172     permutation = (0, 1) if neg_0 + pos_1 > neg_1 + pos_0 else (1, 0)
173 
174     return permutation, clustering
175 
176 
177 if __name__ == "__main__":
178     FILE_NAME_NEG = "mouse_negative"
179     FILE_NAME_POS = "mouse_positive"
180     FILE_NAME_EXP = "mouse_experiment"
181 
182     USED_COLUMNS = ["a", "u", "d", "k1", "k2"]
183     CLUSTERING_METHODS = [GaussianMixture(covariance_type="diag",
184                                           n_components=2, n_init=10),
185                           KMeans(n_clusters=2, n_init=10)]
186 
187     print(f"Clustering files {FILE_NAME_NEG}, "
188           f"{FILE_NAME_POS} and {FILE_NAME_EXP} with "
189           f"parameters {USED_COLUMNS}.\n")
190 
191     neg_df = load_from_file_or_approx(FILE_NAME_NEG)
192     pos_df = load_from_file_or_approx(FILE_NAME_POS)
193     exp_df = load_from_file_or_approx(FILE_NAME_EXP)
194 
195     neg_df = remove_outliers(neg_df, 3, USED_COLUMNS)
196     pos_df = remove_outliers(pos_df, 3, USED_COLUMNS)
197 
198     neg_df, pos_df, exp_df = normalize(neg_df, pos_df, exp_df, USED_COLUMNS)
199 
200     n = min(len(neg_df), len(pos_df))
201     neg_df, pos_df = neg_df.sample(n), pos_df.sample(n)
202 
203     for clustering_method in CLUSTERING_METHODS:
204         print(f"CLUSTERING_METHOD: {clustering_method}")
205 
206         per, clustering = separate(neg_df, pos_df, USED_COLUMNS,
207                                    clustering_method)
208 
209         neg_df["predicted"] = clustering.predict(neg_df[USED_COLUMNS])
210         pos_df["predicted"] = clustering.predict(pos_df[USED_COLUMNS])
```

```
211         exp_df["predicted"] = clustering.predict(exp_df[USED_COLUMNS])
212
213         neg_act = len(neg_df[neg_df['predicted'] == per[1]])
214         pos_act = len(pos_df[pos_df['predicted'] == per[1]])
215         exp_act = len(exp_df[exp_df['predicted'] == per[1]])
216         neg_len, pos_len, exp_len = len(neg_df), len(pos_df), len(exp_df)
217
218         print("file: activated     out of     percentage")
219         print(f"neg:  {neg_act:<13} {neg_len:<10} "
220               f"{neg_act / neg_len * 100:.3f}")
221         print(f"pos:  {pos_act:<13} {pos_len:<10} "
222               f"{pos_act / pos_len * 100:.3f}")
223         print(f"exp:  {exp_act:<13} {exp_len:<10} "
224               f"{exp_act / exp_len * 100:.3f}\n")
```