Model selection in deterministic models of mRNA transfection

Modelauswahl in deterministischen Modellen zur mRNA Transfektion

Laura Lechtenberg

München 2015

Model selection in deterministic models of mRNA transfection

Modelauswahl in deterministischen Modellen zur mRNA Transfektion

Laura Lechtenberg

Masterarbeit
an der Fakultät für Physik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Laura Lechtenberg
aus Flensburg

München, den 8. April 2015

Gutachter: Prof. Dr. Joachim Rädler

Contents

1 Summary

In this master thesis the mRNA transfection dynamics of single cells were studied via time-lapse fluorescence microscopy. For this purpose six different models were designed that are based on ordinary differential equations. They were fitted on the mean fluorescence trajectories of single cells using a maximum likelihood approach. The optimization was done for multiple starting points in the parameter space to find the global maximum of the likelihood function and the associated model parameters. To compare the fitting performance of the different models the Akaike and Bayesian Information Criteria were used that balance between under- and overfitting and enable a ranking of all models within the set.

In the first part of this work the behaviour of both model selection criteria was studied on artificially created datasets. For this purpose each model was used once to generate an associated dataset of 100 single trajectories. Afterwards each model of the set was fitted once to each trajectory per dataset. The models were compared by both selection criteria for the whole population of trajectory fits as well as for each fit separately. This way it was found out that the model that was selected as best on the population scale is less complex than the model that was used to generate the data. In addition it was shown that the selection criteria selected different models as best for subgroups within the population. For some cases it was discovered that the model that was selected most frequently for the individual fits differs from the best model on the population scale.

In the second part of this work the mRNA transfection dynamics were studied for Human Embryonic Kidney (HEK) cells via time-lapse fluorescence microscopy. Two experiments were performed in parallel with cells that were transfected with mRNA either coding for eGFP or d2eGFP proteins. The cells were seeded on microstructured single-cell arrays for high-throughput measurements. The six different models were fitted on the trajectories of the mean fluorescence intensities per cell for each population per experiment. The model selection revealed that three models that share the ribosomal translation pathway were highest ranked on the scale of both populations and selected most frequently for the individual fits. The parameter uncertainty analysis was done for one individual fit per highest ranked model and fortunately the protein degradation rates and the measurement noise were indicated as practical identifiable. These two parameters were compared for the highest ranked models and yielded very similar statistical quantities on the logarithmic scale. However, the mean values and the standard deviations of the highest ranked models differed significantly for the back transformed parameter distributions. Indeed, the median values of the parameter distributions were almost equal and comparable to the transformed median values of the logarithmic parameter distributions. This result was expected, as the median is more robust towards outliers. For this reason the median values were compared with a literature reference obtained by flow cytometry analysis of Chinese Hamster Ovary cells (CHO) that were

transfected with plasmid DNA. In this work the following median values of the protein degradation rates were obtained: $\beta_{eGFP} = 0.19$ and $\beta_{d2eGFP} = 0.35$. As expected, the degradation rate of the destabilized d2eGFP proteins was estimated higher than of eGFP. The rate of d2eGFP proteins is close to the literature reference of $\beta_{d2eGFP,FC} = 0.35$ [1]. On the contrary the degradation rate of the eGFP proteins was estimated higher than the literature reference of $\beta_{eGFP,FC} = 0.04$ [1]. On the one hand the literature values can only be guiding values as the transfection dynamics of plasmid DNA and mRNA are different and on the other hand the measurement noise was estimated significantly higher for the cells that expressed eGFP than d2eGFP proteins. The mean value for eGFP constructs is given by $\sigma_{eGFP} = 43.92$ and for d2eGFP constructs by $\sigma_{d2eGFP} = 16.95$. The cells that expressed eGFP proteins were more vital during the observation which led to more movement on the single-cell array and could cause higher fluctuations in the trajectories of the mean fluorescence intensities that were fitted. The global optimization of the likelihood function would probably lead to even more precise results after an improvement of the image analysis methods.

2 Introduction

The dynamics of mRNA transfection were previously studied by Leonhardt et al. via time-lapse microscopy for high-throughput measurements on single cell arrays [2]. The cells were transfected with mRNA that either coded for eGFP or d2eGFP proteins using the transfection reagent *Lipofectamin2000®*. The mean fluorescence intensities of single cells were extracted and fitted by a simple model to obtain the parameter distributions of the degradation rates of transfected mRNA and reporter proteins. This model assumed direct reaction pathways of the mRNA translation as well as for the mRNA and protein degradation. The fitting of the single cell trajectories was done by the commercial software *IGOR-PRO* that used the Levenberg-Marquardt algorithm to minimize a loss function of the squared distances of fitted and observed data points. As only one initial guess was used as a starting point for the algorithm, the estimated model parameters probably correspond to sub-optimal solutions of the minimization of the loss-function. In addition it was discovered that the estimated mRNA and protein degradation rates were highly correlated and interchangeable in the analytic solution of the rate equations. Furthermore the residuals of the fits showed systematic deviations over time from the observed data. This indicated that perhaps even more complex models are needed to describe the mRNA transfection dynamics. [2]

In this work the mRNA transfection dynamics are analysed for six different models that are based on systems of ordinary differential equations. Five new models are designed in addition to the simple model. They differ in their mRNA translation and degradation pathways and share the direct degradation of reporter proteins. As in the study of Leonhardt et al. cells were seeded on microstructured single-cell arrays to provide high-throughput measurements via time-lapse microscopy. By fitting the six models on the mean fluorescence trajectories of single cells the question is addressed which models are better suited to cover the dynamics of mRNA transfection than the simple model. To compare their fitting quality, the Akaike and Bayesian Information Criteria are used that enable a ranking of all models in the set and a comparison of their fitting quality at a glance. Instead of minimizing the loss function of squared distances between the observed and fitted data, the probability to observe the data given the model parameters, called likelihood function, is maximized. To find the model parameters that are probably related to the global maximum of the likelihood function, multiple starting points are randomly chosen by latin hypercube sampling. This global optimization method was established in collaboration with the Institute of Computational Biology of the Helmholtz Center Munich.

This work is organized as follows: All basic concepts that were used in this work are explained in Chapter 3. The six models that were used for fitting are introduced in Chapter 4.1. To find out whether the model selection criteria select models that are closest to the underlying reaction dynamics, the model selection is studied on artificial datasets in Chapter 4. The six models are compared by both selection criteria on the population scale per dataset in Chapter 4.2.3. The model selection is studied as well for the individual trajectory fits in Chapter 4.2.4 to obtain a complete picture of models that were selected as best for the datasets. The mRNA transfection dynamics were studied for the six models via lime-lapse fluorescence microscopy. HEK-cells were transfected with mRNA that either codes for eGFP or d2eGFP and observed in parallel. To find out which model is best to describe the mRNA transfection dynamics the models are compared on the population scale in Chapter 5.2.1 and in detail for the individual trajectory fits per population in Chapter 5.2.2. The parameter distributions of the best models per population are analysed in Chapter 5.3 to compare the protein degradation rates of eGFP and d2eGFP.

3 Fundamental concepts

3.1 Experimental concepts

3.1.1 Transfection

In the past years the delivery of nucleic acids such as DNA, RNA or plasmids into eukaryotic cells, which is called transfection, has become an interesting field for clinical studies. The use of nucleic acids for genetherapy is facilitated by their low toxicity and high target specification. However, the breakthrough for drug approval is hindered by poor cellular uptake and low biological stability due to fast degradation rates. [3, 4]
During the past decades several techniques have been developed to enhance gene delivery including mechanical and electrical methods such as microinjection, particle bombardment, electroporation or magnetofection as well as viral and non-viral vector-assisted methods [3,5]. Viral delivery methods are most frequently used worldwide as they exhibit high transfection efficiencies in several different human tissues [3]. Though non-viral methods received increasing attention due to their lower risk of immunology response [3,4,6,7]. Delivery systems based on cationic helper lipids that form complexes with the negatively charged nucleic acids are commonly used. These complexes are called lipoplexes and the delivery is also called lipofection. As mRNA can directly be translated into proteins without entering the nucleus, thereby reducing the risk of chromosomal integration, its usage instead of DNA is preferred [4, 7]. This was limited by fast degradation rates and immunology response in the past which was reduced by the development of long-lived, non-immunogenic RNA [8]. Figure 1 shows the basic principle of mRNA lipofection: [6, 9]

1. Cationic helper lipids and the negatively charged mRNA are mixed within the transfection medium and self-assemble to lipoplexes.
2. The cells are incubated with the solution as long as enough lipoplexes have sedimented onto the slightly anionic cell membrane.
3. They are incorporated via endocytosis and released into the cytosol. After unpacking of lipoplexes, the mRNA is translated into unfolded GFP proteins.
4. The latter are activated after folding and maturation of the chromophore and can be observed via fluorescence microscopy.

In this work the dynamics of translation and protein expression were studied by mRNA transfection using the commercial liposome-based transfection reagent *Lipofectamin2000* ®. As mRNA can directly be translated after entering the cell, short onset times of expression are achieved [4]. Oppositely the beginning of observation with time-lapse microscopy is limited by previous preparation steps, e.g. by the incubation time of cells with the lipoplex solution. Protocols of the experiments in this work can be found in the Appendix 7.2.

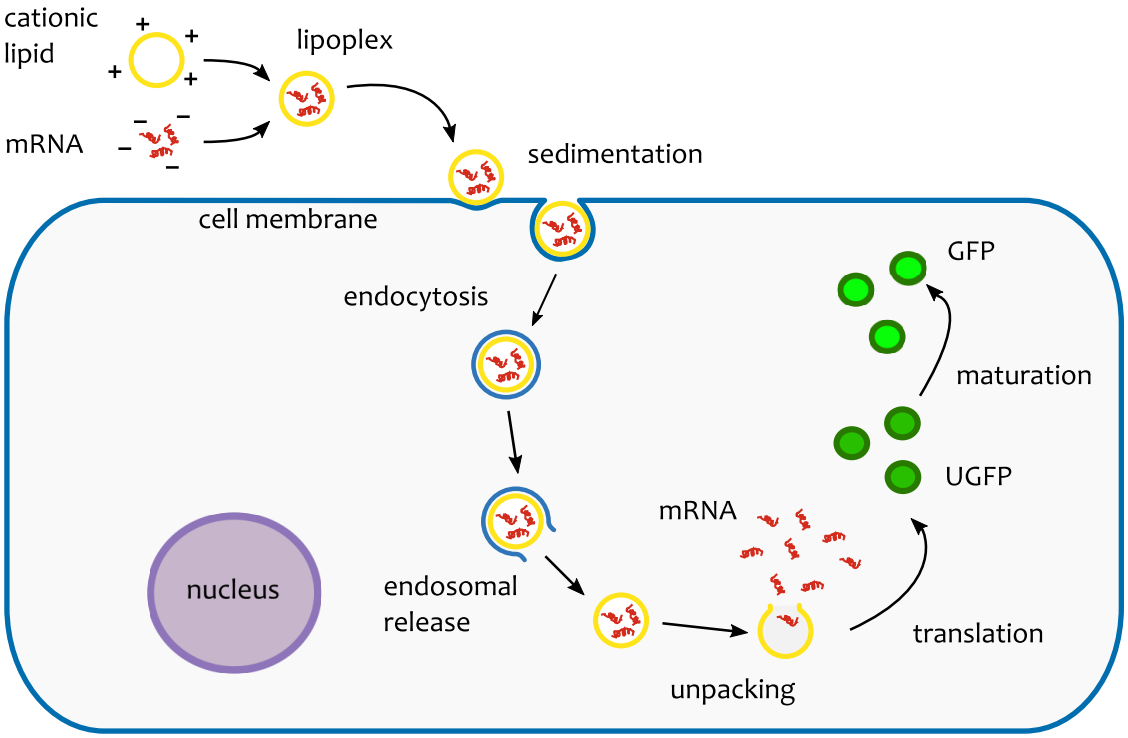


Figure 1: Basic principle of mRNA Lipofection: Lipoplexes are formed by self-assembly of cationic lipids and negatively charged mRNA. After incubation of cells with transfection medium, lipoplexes sediment on the cell membrane. They are incorporated via endocytosis. After endosomal release and unpacking of lipoplexes the mRNA is translated into unfolded GFP proteins. After maturation the active GFP can be observed via fluorescence microscopy.

3.1.2 Microstructured single-cell arrays

The analysis of single-cell arrays provides a deeper understanding of cell-to-cell variability and stochasticity of biochemical networks and parameters for example for gene expression and cell death [10,11]. Therefore several techniques for the fabrication of microstructures for single as well as multiple cells have been developed that offer well-defined microenvironments with comparable growth conditions. This includes approaches with physical barriers that restrict the cells in a small region as well as methods based on selective attachment by adhesion molecules named metals, proteins or polymers. [12–16]
In this work the same method as described in [17] was used to create microstructered single-cell arrays. The fabrication is shown in Figure 2. First a master is produced by negative photolithography and used subsequently to produce PDMS-stamps. The stamp is placed on a μdish (ibidi) and treated with oxygen plasma to obtain hydrophilic surfaces on exposed regions. Afterwards they are filled with cell-repellent PLL-g-PEG polymer. The PDMS stamp is removed and the embedded squares are functionalized with the extracellular matrix protein fibronectin, which enables cell adhesion [18].

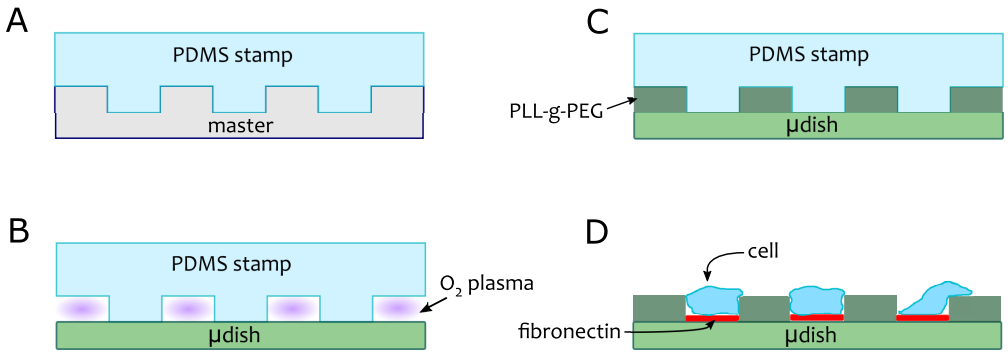


Figure 2: Fabrication of microstructured single-cell arrays: A master is used to create PDMS stamps (A) that are placed on a μdish. Exposed surfaces are treated with oxygen plasma (B) and passivated with the cell-repellent polymer PLL-g-PEG (C). After removal of the PDMS stamp the embedded squares are functionalized with the extracellular matrix protein fibronectin that is preferred by cells after seeding (D).

3.1.3 Time-lapse fluorescence microscopy

Time-lapse fluorescence microscopy provides an insight into the metabolic dynamics of cells by the use of fluorescent dyes or proteins. In contrast to electron microscopy and flow cytometry it enables the observation in spatial as well as temporal resolution [19]. The use of almost non-toxic fluorescent dyes and proteins like GFP and its derivatives enables the study of complex intracellular processes in intact living cells on the single-cell level or even subcellular scale [20]. For this reason fluorescenc microscopy is commonly used to study cellular processes like gene expression, the principle of interference RNA (iRNA) and transfection [12, 21, 22].
The emission of a photon caused by a transition of an electron from an excited singlet state orbital into the ground state is called fluorescence. The fluorophore is typically an aromatic molecule that can be excited by the absorption of a photon. Subsequently a photon of lower energy is emitted. For this reason the emission spectrum is red-shifted compared to the absorption spectrum. By a process called fluorescence quenching the intensity decreases by an interaction of the fluorophores with solvent molecules (quenchers). The fluorophore can return to the ground state by collision with the solvent molecules without emission of a photon or can be inactivated by complex-formation with the quencher. [23]
In addition fluorophores can loose there capability of absorption and emission by an irreversible conformational change or reaction. This effect is called photobleaching and is enhanced for molecules in higher excited states [24].

In this work an inverted epifluorescence microscope was used, more precisely the *Nikon TI Eclipse* fluorescence microscope that was equipped with an objective lens (CFI PlanFluor DL-10 x, Phase 1, N.A. 0.30, Nikon). A LED lamp (*SOLA SE II lumencor*®) was used to ensure homogeneous illumination of the probe. The multispectral light passes through an

excitation filter, is reflected on a dichroic mirror and illuminates the probe. This leads to excitations of the fluorescent proteins. The emitted light passes through the objective lens and an emission filter before reaching the CCD camera (*Andor Clara-E*). The time-lapse movies were obtained by snap-shots images that were taken in an interval of 10 min. To ensure optimal growth conditions for the observed cells, a heating chamber was used that provides a temperature close to 37 ℃. A motorised stage was used to scan multiple positions of the probe automatically. The problem of the inherent focus drift during long-term observation was solved by the commercial hardware *Nikon Perfect Focus System*.

In this work cells were transfected with mRNA that is translated within the cell into variants of the green fluorescent protein (GFP), more precisely eGFP and d2eGFP. The eGFP variant that was used in this work has a major excitation peak at 488 nm and an emission peak at 507 nm [25]. GFP consists of 238 amino acids that form a cylinder like structure by eleven β-sheets. The central α-helix contains the chromophore that is protected from the environmental surrounding by the cylinder. During protein folding the fluorophore is formed by an autocatalytic process which is called maturation. [26]

3.1.4 Image Analysis

The main challenges of image analysis of single-cell time-lapse movies are the uneven illumination of the probe, the photobleaching of reporter proteins and the autofluorescence of cells and culture medium. In this work experimental data was collected from a large number of time-lapse movies that show the evolution of fluorescence intensities of single cells in time intervals of 10 min. Each movie contains 180 TIFF-images of 1392 $\times$ 1040 pixel. As manual analysis of each image would be error-prone and tedious, an automatic image analysis was carried out by in-house-written ImageJ plug-ins in two steps: background correction and well analysis.

Background correction

The method used for background correction is adapted from Schwarzfischer et al. [27]. The raw image $I(x,t)$ at time point t with space coordinates $x = (x_1, x_2)$ is decomposed as follows:

- cellular signal $s(x,t)$
- homogeneous background signal $b(t)$ that decreases over time due to photobleaching
- coordinate-specific uneven illumination function $g(x)$
- camera offset $o(x)$

They are summarized by the following equation:

$$I(x,t) = s(x,t) \cdot g(x) + b(t) \cdot g(x) - o(x) \tag{1}$$

By rearranging the formula, one obtains the equation of the cellular signal:

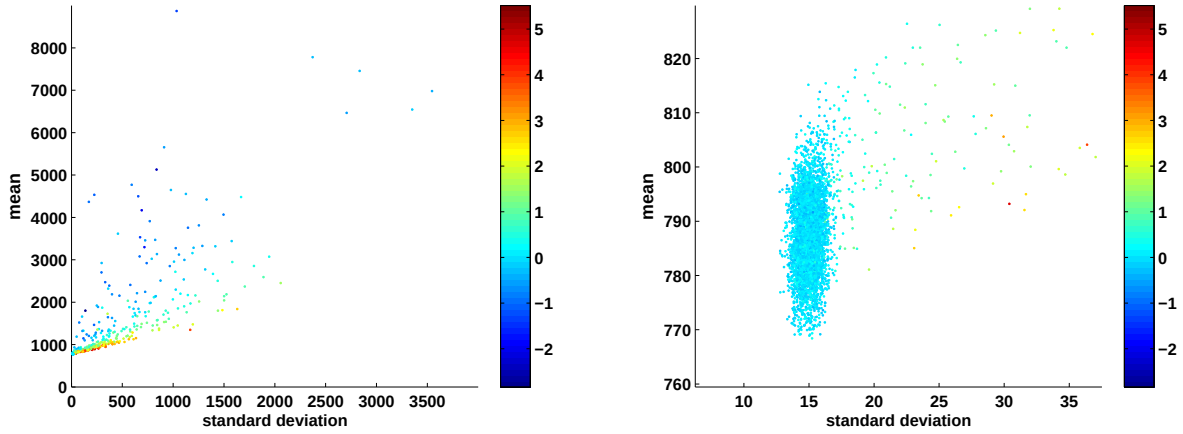
$$s(x,t) = \frac{I(x,t) - b(t) \cdot g(x) - o(x)}{g(x)} \tag{2}$$

In this work images were corrected by the illuminated background signal, denoted as $B(x,t) = b(t) \cdot g(x) + o(x)$. The background was reconstructed for each image per time-lapse movie separately and subtracted from its original image:

$$\tilde{s}(x,t) = I(x,t) - B(x,t) \tag{3}$$

As the illumination was nearly homogeneous due to the LED lamp and to reduce the time needed for image preprocessing, the calculation of the coordinate-specific illumination function $g(x)$ was omitted and remains for further improvement of the background correction method. By estimating $g(x)$ the bleaching of each pixel by the uneven illumination can be taken into account [27].

To reconstruct the background image $B(x,t)$, cellular and background signals were separated by a density-based clustering of moments. First each image $I(x,t)$ was divided into small segments with areas of 16×16 pixel and four moments of their intensity distributions were calculated: mean, variance, skewness and kurtosis. Each segment was represented as a point in the four-dimensional space of distribution moments. As an example the mean is displayed in dependence on the standard deviation for all segments of one image in Figure 3a. Their skewness is displayed in different colours as shown by the colour-bar. Points that correspond to cellular segments spread widely, whereas those related to background segments accumulate in a dense region with nearly the same colour as shown by Figure 3b. The intensity distributions of background segments have skewnesses of nearly zero as depicted in light blue and are therefore symmetric. To distinguish between cellular and background points in the space of moments, a density based clustering was done using the DBSCAN-algorithm (Density-Based Spatial Clustering of Applications with Noise).



(a) Points in moment space

(b) Background cluster

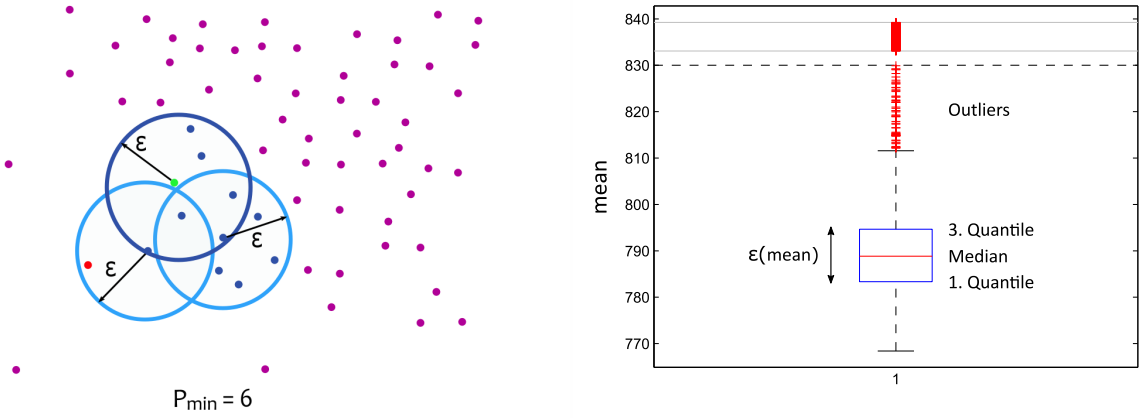
Figure 3: Points in the space of moments defined by mean and standard deviation: The example image was divided into 16 × 16 pixel segments and three moments of their intensity distributions were calculated. The skewness is colour-marked as displayed in the colour-bar. Point that correspond to cellular signals spread widely in space and colour as depicted in Figure 3a whereas the points that are related to background segments accumulate with similar moments in a dense region as shown by Figure 3b.

The algorithm starts with an arbitrary point in moment space and searches for neighbouring points N within a distance called ϵ. If N is larger than a threshold, denoted as P_{min}, they are assigned to the background cluster together with the starting point. Otherwise they are assigned to noise which is related to cellular signals. Afterwards the algorithm repeats the procedure for each neighbour point, thereby looking for next nearest neighbours within the cluster. By this way the starting point, all neighbours and next neighbours are marked as visited (Figure 4a) [28]. The process continues by choosing an unvisited point as the new starting point. The scan repeats until all points have been marked as visited. Points that are assigned to the cluster are related to background segments whereas the residual points are assumed to correspond to segments with cellular signals.

The clustering success depends on the choice of the interpoint distance ϵ and the threshold of minimal required points P_{min}. In this work it was assumed that P_{min} depends on the logarithm of the total number of segments N given by $P_{min} = \ln(N)$. This approach was adapted from Briant et. al. [29]. The distance ϵ is represented by a vector, whose elements consist of the distances for each moment. As the mean values of the segment intensitiy distributions decay over time due to photobleaching, the cluster is moving in space within the timespan of one video. For this reason ϵ was calculated for each image separately. For example ϵ_{mean} stands for the distance of the first moment. It is adjusted by the interquantile distance of the first (25 %) and third (75 %) quantiles of the distribution of segment means as shown

by the boxplot in Figure 4b. The segment means are depicted by the y-axis and the red line in the box represents the median. The blue lines above and beneath the median correspond to the quantile values. The black lines mark the boundaries whose values of 768 and 812 cover the mean limits of the background cluster of Figure 3b well. The red asterix stand for outliers in the distribution of means that are mainly related to the mean values for segments that contain cellular signals. In this example the interquantile distance $\epsilon_{mean} = 11.4$ was determined. This method was used for each moment separately to obtain ϵ_{mean}, $\epsilon_{variance}$, $\epsilon_{skewness}$ and $\epsilon_{kurtosis}$. During the search of the algorithm a point and a neighbour point are compared by the differences of their moments. If they are all smaller than their related ϵ, the points fulfil the first condition. If more or equal than P_{min} candidates are found for the point, they are assigned to the background cluster.

The background image is reconstructed by replacing the segments that are member of the cluster by the median values of their intensity distributions. Segments that were not assigned to the cluster by the DBSCAN algorithm are set to zero. Hence the background image consists of gaps where cellular segments were detected as depicted in Figure 5b. The Figures are all contrast adjusted for better visualization. To fill in the gaps an iterative interpolation was done that is based on the mean of an 8-neighbourhood of zero-valued segments. Most gaps



(a) DBSCAN search

(b) Interpoint distance ϵ

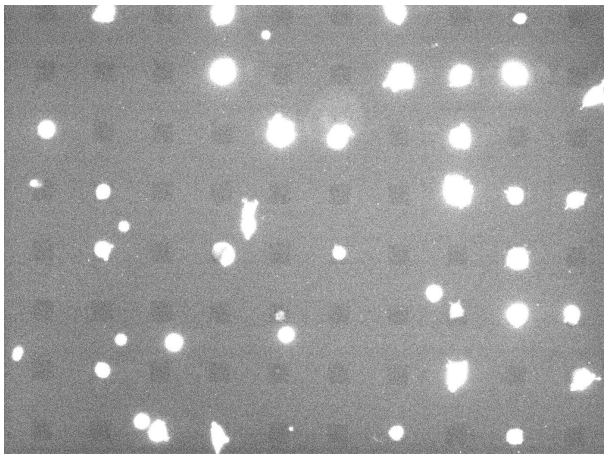
Figure 4: Principle of DBSCAN: As shown by figure 4a the algorithm starts at an arbitrary point in moment space (green) and searches for points within the distance ϵ (dark blue). If their number exceeds P_{min} they are assigned to the cluster together with the initial point otherwise they are marked as Noise (red). The search is repeated for the next neighbours (light blue) and starts again till all points have been marked as visited. The distance ϵ is calculated for each moment separately by the interquantile distance of the distribution of moments. This is shown for all segment means of the example image by Figure 4b.

are surrounded by slightly brighter segments, that were not detected as cellular during the clustering. For this reason the interpolated regions deviate to higher intensities that differ from the expected background. This effect is dependent on the segment size. Smaller segments result in a higher resolution of cellular gaps but exhibit the risk of higher interpolation errors as the 8 neighbourhood shrinks. For this reason the interpolation method has to be improved in the future. The background correction in Figure 6b is finished by subtracting the reconstructed background image that is depicted in Figure 6a from the original shown by Figure 5a.

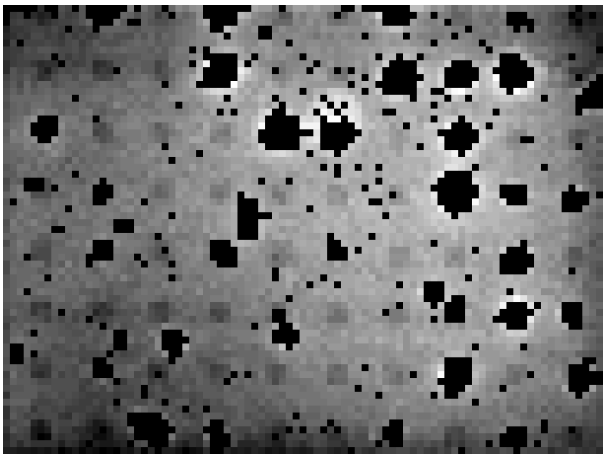
Well Analysis

To gather the mean fluorescence intensity per cell, an in-house ImageJ-plug-in written by Christian Meggele was used. A grid of rectangular regions, called wells, is overlapped with the images of one time-lapse movie, such that all cells on the microarray fit into the wells. The well positions, their width and height can be adjusted by the user. Cells of interest are selected by a click on the corresponding well which is thereupon colour-marked as shown by Figure 7. This selection is done manually, as dead or migrating cells have to be excluded from the analysis. By running the program statistical quantities per colour-marked well such as the mean are calculated over the well area for all image frames of the movie. This way a dataset is generated that contains the mean fluorescence of all coloured wells over distinct time points. Afterwards the dataset is stored as an Excel-file for further analysis.

As the mean fluorescence is calculated over the well area, the shrinking and expansion of cells during the observation is not taken into account. For example in Figure 7 some cells do not fill in their wells and many low values corresponding to the dark surrounding of the cell are included in the calculation of the mean. This problem can be solved by calculating the mean over the cell area, that is given by the gaps in the image after clustering in Figure 5b. By overlapping the well and the cell area the union indicates which pixel has to be included for calculation of the mean. Furthermore the interpolation method has to be improved to reduce the error caused by the reconstructed background that is individually shifted towards higher values as explained above and shown by Figure 6a. An automatic detection of migrating and dead cells would reduce the time needed for the image analysis and yield probably result that are independent of the user. In addition the experimental setup could be improved by well shapes with adequate depth that are more suitable to the round shapes of the cell bodies and prevent movement and migration beyond the well area. Perhaps the nutrient supply and removal of metabolic degradation products could be improved by an automatic, slow fluid flow during the long-term observation via fluorescence microscopy.

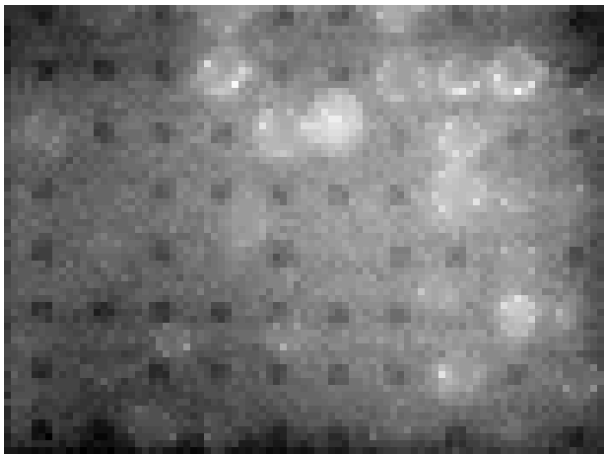


(a) Original image

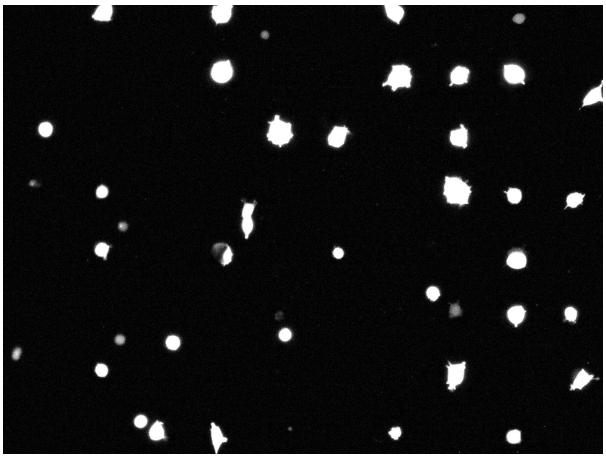


(b) Segmented, clustered image

Figure 5: Image after clustering: The original image as shown by Figure 5a is segmented into areas of 16×16 pixel. After detection of background related segments by the DBSCAN algorithm, they are replaced by their median value as shown by Figure 5b. The black gaps are related to segments with cellular signals. Both images are contrast adjusted for better visualization.



(a) Interpolated background image



(b) Background corrected image

Figure 6: Image after interpolation and background correction: The background image was reconstructed by an iterative interpolation by the mean of the 8-neighbourhood of each zero valued segment as displayed in Figure 6a. The background corrected image was obtained after subtraction of the interpolated image from the original as shown by Figure 6b. As above both images are contrast adjusted.

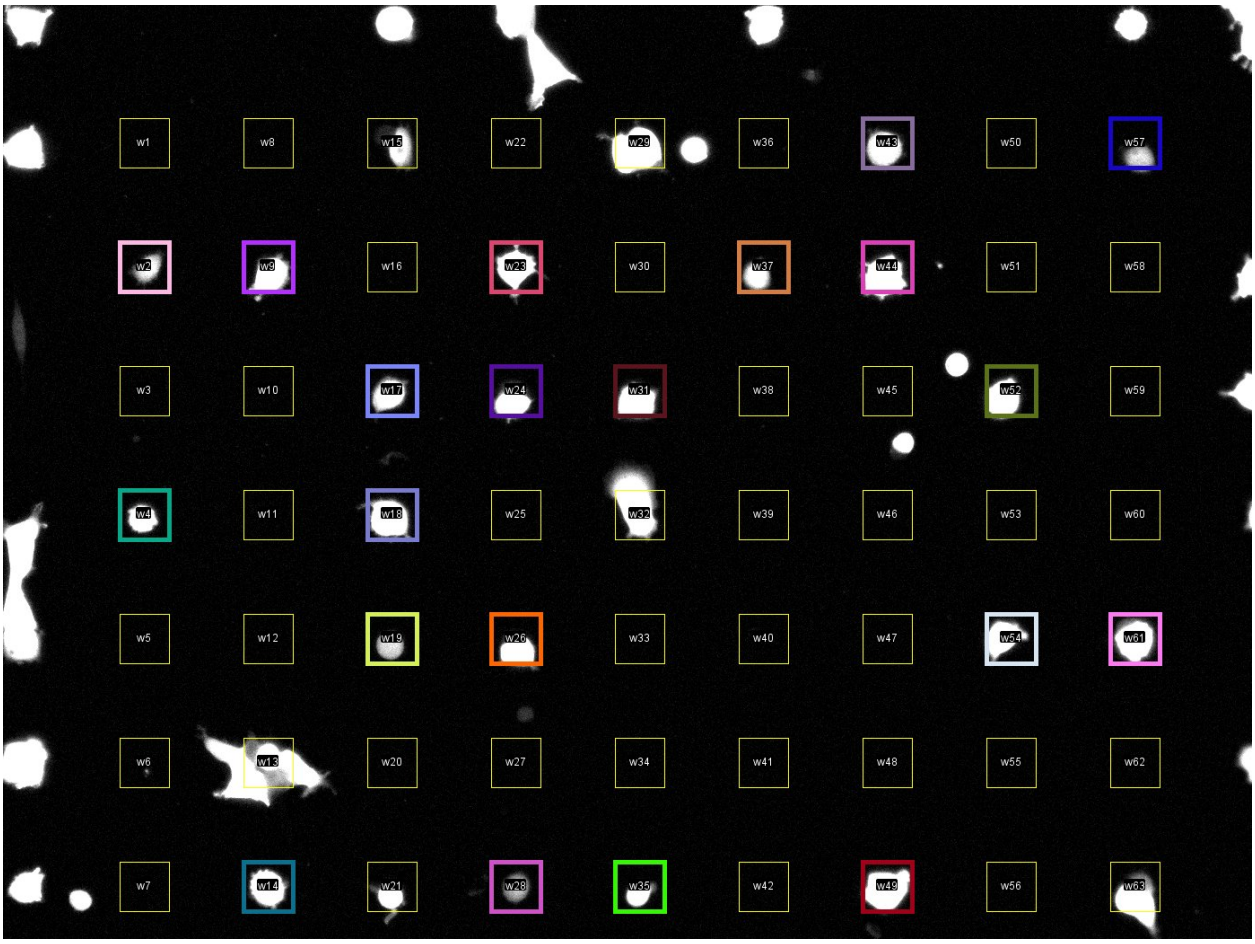


Figure 7: Well analysis: An in-house-written ImageJ plug-in was used to calculate the mean fluorescence per well. For this purpose a grid of user adjusted wells is overlapped with the image of one time-lapse movie. Cells of interest are manually selected by colour-marked wells as migrating and dead cells have to be excluded from the analysis. As most of the cells do not fill in their well, the mean fluorescence is individually biased towards lower intensity values.

3.2 Theoretical concepts

3.2.1 Deterministic rate equations

In Systems biology cellular processes are commonly modelled by reaction rate equations based on systems of ordinary differential equations (ODE) denoted as $\Sigma(\theta)$ that describe the underlying network of biochemical reactions [30]:

$$\Sigma(\theta) = \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}x(t,\theta) = f(x(t,\theta),\theta) & x(0,\theta) = x_0(\theta) \\ y(t_i,\theta) = h(x(t_i,\theta),\theta) \end{cases} \tag{4}$$

Molecular concentrations are represented by the dynamical state variables x whereas reaction rates are described by the model parameters θ. Initial conditions are given by $x_0(\theta)$. Usually only a part of the state variables x are experimentally accessible which is modelled by so called model observables $y(t_i,\theta)$. Indeed the experimentally measured data $\tilde{y}(t_i)$ is collected at discrete time points t_i and often relative to the model observable. In this work the fluorescence intensity and not the absolute concentration of reporter proteins are measured by time-lapse microscopy and linear relationship between both quantities is assumed. For this purpose scaling sc and offset κ parameters are used to modify $y(t_i,\theta)$ and are additionally assigned to the parameter vector θ of the model:

$$y(t_i,\theta) = sc \cdot h(x(t_i,\theta)) + \kappa \tag{5}$$

The observed data $\tilde{y}(t_i)$ is corrupted by measurement noise ϵ that was assumed to be additive and normally distributed $\epsilon(t_i) \sim N(0,\sigma(t_i;\theta))$ with zero mean and unknown standard deviation $\sigma(t_i,\theta)$: [30]

$$\tilde{y}(t_i) = y(t_i,\theta) + \epsilon(t_i) \tag{6}$$

The standard deviation $\sigma(t_i,\theta)$ was estimated as an additional parameter in θ. Systems that are modelled by reaction rate equations are assumed to be well-mixed and to exhibit a large amount of molecules such that their diffusion is fast compared to their reaction rates which provides spatial homogeneity. [30]

3.2.2 Global optimization

After experimental measurements and given a model of interest one usually seeks for the parameter values of the model that yield the best fit of the data [31]. This procedure is called parameter estimation and Maximum Likelihood Estimation is widely used for this purpose. It is based on the frequentist perspective which assumes that the values of θ are deterministic, but uncertain variables. In contrast, the Bayesian perspective assumes that parameter values are stochastic variables that possess a distribution [32].

The maximum likelihood method is stated as an optimization problem: Given a model of interest one searches for parameter values θ such that they maximize the probability $p(D|\theta)$ to fit the observed data $D = \{\tilde{y}(t_i)\}_{i=1}^{N}$. The probability $p(D|\theta) = L_D(\theta)$ is called likelihood function and the obtained parameter values are named maximum likelihood estimates θ^{MLE}. Neither existence nor uniqueness is guaranteed for them. By assuming that the likelihood function is sufficiently smooth, the parameter estimates have to fulfil the following condition [31]:

$$\left.\frac{\partial L_D(\theta)}{\partial \theta_i}\right|_{\theta^{MLE}} = 0 \tag{7}$$

To ensure that the obtained optimum is a maximum and not a minimum, more precisely that the shape in the vicinity of the likelihood estimates is convex, the Hessian has to be negative definite [31]. In this work it was assumed that the measurement noise is normally distributed and independent for different time points. In this case the likelihood function provides a distance measure for the experimental data $\tilde{y}(t_i)$ and the model observable $y(t_i, \theta)$ [30]:

$$L_D(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2(t_i,\theta)}} \cdot \exp\left(-\frac{1}{2\sigma^2(t_i,\theta)} \cdot \left(\tilde{y}(t_i) - y(t_i,\theta)\right)^2\right) \tag{8}$$

Deterministic rate equations in system biology models are often highly non-linear such that it is not possible to obtain the analytic solution. In this case non-linear optimization algorithms are applied to search for the maximum likelihood estimates. For numerical reasons it is convenient to consider the equivalent minimization problem using the negative log-likelihood function $J(\theta)$. The optimization is done by changing all unknown parameter values such that the model output $y(t_i, \theta)$ covers the experimental data $\tilde{y}(t_i)$ best. Unfortunately the algorithm often finds multiple sub-optimal solutions for the parameter estimates that are called local optima. Several methods have been developed to find the global optimum instead that are based on deterministic, stochastic or hybrid optimization algorithms [30, 33]. Raue et al. compared 15 different algorithms and recommended a deterministic multi-start

method using sensitivity equations [30]. This approach was used in this work.

The parameter values are adjusted stepwise in an iterative manner to find the most likely value of the negative log-likelihood function $J(\theta)$. The search converges when a predefined stopping criterion is fulfilled. This could include the maximum allowed number of iteration steps as well as the strength of the parameter adjustment [31]. The derivatives pointing to the minimum of $J(\theta)$ guide the iteration procedure and can be calculated reliably by sensitivity analysis. For this purpose the inner derivatives $\frac{dx(t,\theta)}{d\theta}$ were calculated numerically by additional equations simultaneously to the model system $\Sigma(\theta)$ [30]:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}x(t,\theta)}{\mathrm{d}\theta} &= \frac{\partial f}{\partial x}\frac{\mathrm{d}x(t,\theta)}{\mathrm{d}\theta} + \frac{\partial f}{\partial \theta} \\ \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{d}y(t,\theta)}{\mathrm{d}\theta} &= \frac{\partial h}{\partial x}\frac{\mathrm{d}x(t,\theta)}{\mathrm{d}\theta} + \frac{\partial h}{\partial \theta} \end{aligned} \tag{9}$$

For this purpose the stiff ODE-solver CVODES was used that is coded in C for efficiency in combination with a Matlab mex interface. Multiple starting points for the optimization were drawn by latin hypercube sampling as different starting points could yield different local optima. This way it was ensured that they cover the whole region in the parameter space [30]. In this work 150 different starting points have been selected for a parameter space that ranges from 10^{-5} to 10^{5} for each considered parameter.

3.2.3 Parameter uncertainty analysis

The aim of systems biology is to find a model that is suitable to make predictions and to explain the unknown biological processes of complex phenomena. As parameters have to be estimated with a finite accuracy from experimental noisy data and only a subset of model state variables are measured experimentally, parameter uncertainty analysis is important. Though, its methods are not commonly used in the field of systems biology [32].

The issue of parameter uncertainty is closely connected to parameter identifiability that addresses the question whether the estimated parameter is unique within a defined range [32]. The uncertainty of parameter estimates can be characterized by structural and practical non-identifiability. The further is caused by the model structure whereas the latter is related to experimental data and measurement noise [34]. If model parameters are structural non-identifiable, one could try to improve the identifiability by reducing the number of state and parameter variables or by fixing parameter values that are less important for prediction.

On the other hand one could also change the experimental design to measure additional model observables [35]. If correlations between model parameters are observed, the non-identifiability can sometimes be improved by replacing them with a combination of them. However, finding the right combination is a difficult task.

Asymptotic confidence intervals

A confidence interval $[\theta^{i,-}, \theta^{i,+}]$ to the level α provides a range for the parameter estimate θ^i that is located within this interval with probability α [36]. The asymptotic confidence interval points $\theta^{i,\pm}$ are calculated for the parameter $\theta^{i,MLE}$ by the following equation:

$$\theta^{i,\pm} = \theta^{i,MLE} \pm \sqrt{\frac{\Delta_\alpha \cdot \Sigma_{ii}}{2}} \tag{10}$$

The confidence level is given by Δ_α and related to the $\alpha' th$ percentile of the Gaussian distribution with the covariance matrix Σ. The estimated parameters θ^{MLE} converge to a Gaussian distribution in the asymptotic limit of large observation points N. If the model parameters are identifiable, their covariance matrix Σ is given by the inverse of the Fisher Information Matrix F^{-1} [32]:

$$F(\theta^{true}) = \sum_{i=1}^{N} \frac{1}{\sigma(t_i)} \cdot (\nabla_\theta y|_{t_i,\theta^{true}})^T (\nabla_\theta y|_{t_i,\theta^{true}}) \tag{11}$$

As the true parameters θ^{true} are unknown, F is evaluated at the estimated parameters θ^{MLE}. If the standard deviations of the measurement noise $\sigma(t_i)$ are stated as an additional parameter, the Fisher Information Matrix is approximated by the Hessian matrix H. For small datasets and parameters that are non-linearly correlated or exhibit multiple optima more reliable confidence intervals can be calculated by the Profile Likelihood method [32].

Profile Likelihood Analysis

Parameter uncertainty for datasets with large measurement noise is preferably studied with Profile Likelihood Analysis. The profile likelihood function $PL(\theta_i)$ of parameter θ_i is calculated by fixing θ_i and optimizing the likelihood function by varying all other model parameters θ_j along the profile interval [37].

$$PL(\theta_i) = \max_{\theta_{j\neq i}} L_D(\theta) \tag{12}$$

If several local maxima have been found during the optimization, the profiles should be evaluated for each of them to obtain an complete impression of the structure of the likelihood function in the vicinity of the estimates. [32] The profile ration $R(\theta_i)$ of θ_i is obtained by dividing the profile likelihood function of θ_i by the maximum likelihood function along the profile interval [37]:

$$R(\theta_i) = \frac{PL(\theta_i)}{L(D|\theta^{MLE})} \tag{13}$$

A ration above 1 would indicate, that the profile likelihood function is higher valued than the maximum likelihood function. The confidence intervals for the parameter θ^i are given by [37]:

$$C_i = \left\{ \theta_i |\, R(\theta_i) < \exp\left(-\frac{\delta_\alpha}{2}\right) \right\} \tag{14}$$

A parameter is called practical non-identifiable if the likelihood ration $R(\theta_i)$ is not lower than the threshold $\delta_\alpha = \chi(\alpha, 1)$ for increasing and decreasing values of θ_i on the profile interval [37].

3.2.4 Model selection

Model selection is philosophically split into two different viewpoints concerning the question if there exists a true model that fully reflects the unknown reality or if all models are merely approximations and could never be true. The latter is assumed in an information theoretic perspective. George Box made a famous statement: "All models are wrong but some are useful" [38,39]. Every model should be as simple as possible only including what is necessary. On the one hand a model with high complexity including a large number of parameters is able to cover the observed data best but it is often bad for prediction as too many spurious effects, e.g. noise, are covered as well. One the other hand a model with too few parameters leads to high bias from the data and its underlying process. Thus model selection is faced with the problem to find a model complexity that balances between under- and over-fitting. [40] Model selection should include the principle of "multiple working hypotheses" by using a set of models. During analysis of the observed data some models will be less often chosen as best than others. By rejecting models that were less favoured during model selection, adding new models to the set and compare them with the models that were selected as best previously, a set of models can be obtained that evolves such that the suitability to fit the data increases [40].

In this work two widely used model selection criteria were used that enable to obtain an easy comparison and ranking of models: the Akaike and Bayesian Information Criterion.

Akaike information criterion

The Akaike information criterion (AIC) is based on the Kullback-Leibler (KL) divergence which is a measure of how much information $I(f, m)$ is lost by model m that approximates the reality f [38]:

$$I(f, g) = \int f(x) \log\left(\frac{f(x)}{m(x|\theta)}\right) dx \tag{15}$$

$$= C - E_f\left[\log\left(m(x|\theta)\right)\right] \tag{16}$$

The best model that loses the least information is obtained by minimizing $I(f, m)$ over all models m in the set M. The KL information cannot be used directly as the full reality f is unknown and model parameters θ_m often exhibit uncertainty. As C is an unknown constant, that does not depend on the data or the model, one is restricted to estimate the relative KL-information $E_f\left[\log\left(m(x|\theta)\right)\right]$ for each model m in the set. Akaike found out, that the maximized log-likelihood function is a biased estimator of the relative KL-information for large sample sizes N and that the bias is approximately equal to the number of estimable parameters n_θ [39]. This relationship is expressed in the AIC value as follows: [38]

$$AIC = -2 \cdot \log\left(L_D(\theta)\right) + 2 \cdot n_\theta \tag{17}$$

By calculating the AIC for each model m_i in a set of models M, a ranking is obtained by sorting the values in ascending order. The best model is then the one with minimal AIC value. For a large sample size more complex models will be selected as the different numbers of adjustable parameters are negligible compared to the likelihood function [39]. As the AIC is defined for a large sample size, a small sample size version should be used, when the number of parameters n_θ is relative large compared to the number of data points N, namely $(\frac{n}{n_\theta} \leq 40)$ [38]:

$$AIC_c = -2 \cdot \log\left(L_D(\theta)\right) + 2 \cdot n_\theta + \frac{2 \cdot n_\theta \cdot (n_\theta + 1)}{N - n_\theta - 1} \tag{18}$$

The AIC values themselves are not easily interpretable and it is common to rescale them to the minimum value [38]:

$$\Delta_{AIC,i} = AIC_i - AIC_{min} \tag{19}$$

Instead one could also calculate the AIC weights $w_{AIC,i}$ that are related to the normalized model likelihoods and interpretable as the probability of model m_j to be the best model in the set $M = \{m_i\}_{i=1}^{n_M}$ in the sense of the KL-divergence [38,41]. The weights can be obtained as follows:

$$w_{AIC,j} = \frac{\exp\left(-\frac{1}{2}\Delta_{AIC,j}\right)}{\sum_{i=1}^{n_M} \exp\left(-\frac{1}{2}\Delta_{AIC,i}\right)} \tag{20}$$

Bayesian information criterion

The Bayesian information criterion (BIC) is based on an approximation of the posterior probability $p(m|D,\theta_m)$ that model m is the true model given the observed data D and model parameters θ_m:

$$p(m|D,\theta_m) = \frac{p(D|m,\theta_m)\cdot p(m)}{p(D)} = \frac{\int p(D|\theta_m)\cdot p(\theta_m|m)\cdot p(m)d\theta_m}{p(D)} \tag{21}$$

The approximation assumes non-informative prior probabilities $p(\theta_m|m) = 1$ and integrates out all model parameters θ_m by a Taylor expansion of $p(D|\theta_m)$ at the maximum likelihood estimate θ_m^{ML}. For this purpose identifiable model parameters are assumed and constant terms as the prior probability $p(m)$ for the model m are neglected in the limit of a large number of data points N. [42] The result is given by the BIC as follows:

$$BIC = -2\cdot\log\left(p(m|D)\right) \approx -2\cdot\log\left(L_D(\theta_m^{ML})\right) + n_\theta\cdot\log(N) \tag{22}$$

$$\Delta_{BIC,i} = BIC_i - BIC_{min} \tag{23}$$

It is common to rescale the values with respect the minimum value to obtain $\Delta_{BIC,i}$ per model m_i. Even though the equations of BIC and AIC are similar, they are based on different approaches. As with AIC the model with the smallest BIC is selected as the best model in the set M. The selected model is not meant to be the model that has generated the data. If the number of data points N increases to infinity, the BIC will select the quasi-true, most parsimonious model with probability 1 assuming a finite model set and independent and identically distributed data points. The quasi-true model is closest to the true model in the sense of the KL-divergence and the best that can be used for interference if the sample size is very large. In practice the number of data points is limited and the selected model is of lower complexity than the quasi-true model. The property that BIC will select the quasi-true model if the sample size grows to infinity is named asymptotic consistency. [38] However, in contrast to AIC, the BIC tends to select less complex models as the penalty term depends on both the number of parameters n_θ and the sample size N [43].
The weights w_{BIC} are related to the model posterior probabilities $p(m|D)$ and can be calculated as follows [38]:

$$p(m_j|D) \approx w_{BIC,j} = \frac{\exp\left(-\frac{1}{2}\Delta_{BIC,j}\right)}{\sum_{i=1}^{M} \exp\left(-\frac{1}{2}\Delta_{BIC,i}\right)} \tag{24}$$

4 Deterministic modelling of mRNA transfection

In this work six models based on ordinary differential equations were used to study the dynamics of mRNA transfection. All models are introduced in Chapter 4.1. The Akaike and Bayesian Information Criteria were used to compare the fitting performance of the different models. To find out, whether these criteria select models as best, that are closest related to the underlying network of biochemical reactions, the model selection was studied on artificial datasets in Chapter 4.2. Each model was used once to generate an associated dataset as described in Chapter 4.2.1. Afterwards each model was fitted once on the trajectories of each dataset as explained in Chapter 4.2.2. The model selection was studied on the population scale per dataset in Chapter 4.2.3 and for the individual fits per population in Chapter 4.2.4. This way the question is addressed, whether the generating models will be selected as best by the model selection criteria.

4.1 Introduction of considered models

The modelling of cellular processes such as gene expression is limited by the complexity of biochemical reaction networks and by the various interactions between several cellular compartments. For this reason a model can only be an approximation of the natural process of interest and the modelling process should be closely related to the experiment [44]. In this work the mRNA transfection dynamics were modelled via rate equations as explained in Chapter 3.2.1. The fluorescence intensities of reporter proteins were measured via time-lapse fluorescence microscopy. For this reason the concentration of fluorescent proteins was used as the model observable. In addition to the simple model that was used in the previous study of Leonhardt et al. [2], five new models were designed to find out which model is best suitable to describe the dynamics of mRNA transfection. All models that were used in this work are introduced in the following.

The *Trivial-Model* $\mathcal{M}_1$

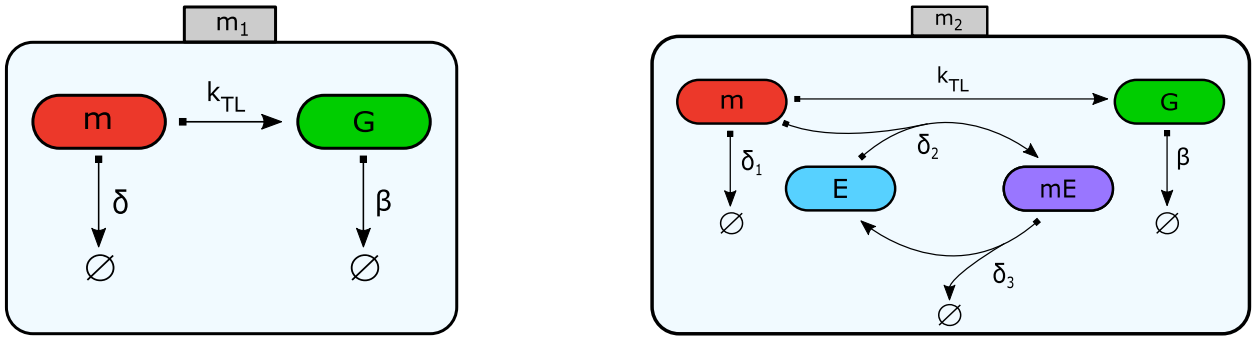
The least complex model $\mathcal{M}_1$ includes direct degradation pathways for the transfected mRNA, named m, with the rate δ and reporter proteins, named G, with the rate β. As shown by Figure 8a the mRNA is translated directly into proteins with the rate k_{TL}. The initial concentration of mRNA, denoted as m_0, was assumed to be unknown and added to the parameter vector $\theta_1 = (\delta, \beta, k_{TL}, m_0)$. The *Trivial-Model* $\mathcal{M}_1$ is the least complex model in the set with a parameter number of $n_{\theta,1} = 4$. The rate equations of $\mathcal{M}_1$ are given as follows:

$$\begin{aligned} \dot{m} &= -\delta \cdot m \\ \dot{G} &= -\beta \cdot G + k_{TL} \cdot m \end{aligned} \tag{25}$$

The *Two-Degradation-Model* $\mathcal{M}_2$

The *Two-Degradation-Model* incorporates two pathways for mRNA degradation as displayed in Figure 8b. On the one hand the mRNA, denoted as m, decays directly with the rate δ_1 and on the other hand it binds to enzymes E with the rate δ_2 and degrades with the rate δ_3 via release of the mRNA-enzyme complex mE. As for the *Trivial-Model* $\mathcal{M}_1$ the mRNA, m, is translated with the rate k_{TL} into proteins, G, which decay directly with the rate β. The initial concentrations of mRNA and degradation enzymes, denoted as m_0 and E_0, are unknown and added to the parameter vector $\theta_2 = (\delta_1, \delta_2, \delta_3, \beta, k_{TL}, m_0, E_0)$. Thus, the number of parameters of $\mathcal{M}_2$ is equal to $n_{\theta,2} = 7$ and the rate equations are given as follows:

$$\begin{aligned} \dot{m} &= -\delta_1 \cdot m - \delta_2 \cdot m \cdot E \\ \dot{G} &= -\beta \cdot G + k_{TL} \cdot m \\ \dot{E} &= -\delta_2 \cdot m \cdot E + \delta_3 \cdot mE \\ \dot{mE} &= -\dot{E} \end{aligned} \tag{26}$$



(a) *Trivial-Model* $\mathcal{M}_1$ (b) *Two-Degradation-Model* $\mathcal{M}_2$

Figure 8: In Figure 8a the model $\mathcal{M}_1$ describes the translation of mRNA (m) into GFP (G) with the rate k_{TL} and the degradation of both with the rates δ and β. Model $\mathcal{M}_2$ shown by Figure 8b includes an additional decay pathway of mRNA by enzymatic binding with the rate δ_2 and release with the δ_3.

The *Ribosome-Model* $\mathcal{M}_3$

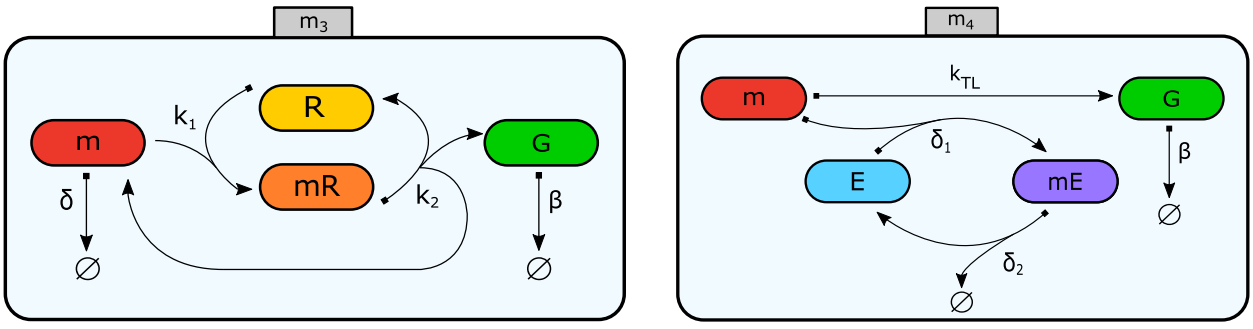
In comparison to the *Trivial-Model* $\mathcal{M}_1$ the *Ribosome-Model* $\mathcal{M}_3$ is modified by a translation pathway using ribosomal binding with the rate k_1 and release with the rate k_2 as displayed in Figure 9a. The ribosomes are denoted as R and the mRNA-ribosome complex as mR. The degradation pathways of mRNA and proteins are modelled directly with the rates δ and β respectively. The initial concentrations of mRNA, m_0, and ribosomes, R_0, are unknown and added to the parameter vector of $\mathcal{M}_3$ given by $\theta_3 = (\delta, \beta, k_1, k_2, m_0, R_0)$. The number of model parameters for the *Ribosome-Model* equals $n_{\theta,3} = 6$ and the rate equations are stated as follows:

$$\begin{aligned} \dot{m} &= -\delta \cdot m - k_1 \cdot m \cdot R + k_2 \cdot mR \\ \dot{G} &= -\beta \cdot G + k_2 \cdot mR \\ \dot{R} &= -k_1 \cdot m \cdot R + k_2 \cdot mR \\ \dot{mR} &= -\dot{R} \end{aligned} \tag{27}$$

The *Enzyme-Degradation-Model* $\mathcal{M}_4$

As shown by Figure 9b the mRNA in the *Enzyme-Degradation-Model* degrades via an enzymatic pathway. After binding to the enzymes with the rate δ_1 the mRNA is depleted with the rate δ_2 by the release of the mRNA-enzyme complex mE. The translation pathway as well as the degradation pathway of the proteins remain unchanged in comparison to the *Trivial-Model* $\mathcal{M}_1$. The parameter vector of $\mathcal{M}_4$ is therefore given by $\theta_4 = (\delta_1, \delta_2, \beta, k_{TL}, m_0, E_0)$ and includes the initial concentrations of mRNA and enzymes, denoted as m_0 and E_0. The *Enzyme-Degradation-Model* has the same complexity as the *Ribosome-Model* with a parameter number of $n_{\theta,4} = 6$. The rate equations are given as follows:

$$\begin{aligned} \dot{m} &= -\delta_1 \cdot m \cdot E \\ \dot{G} &= -\beta \cdot G + k_{TL} \cdot m \\ \dot{E} &= -\delta_1 \cdot m \cdot E + \delta_2 \cdot mE \\ \dot{mE} &= -\dot{E} \end{aligned} \tag{28}$$



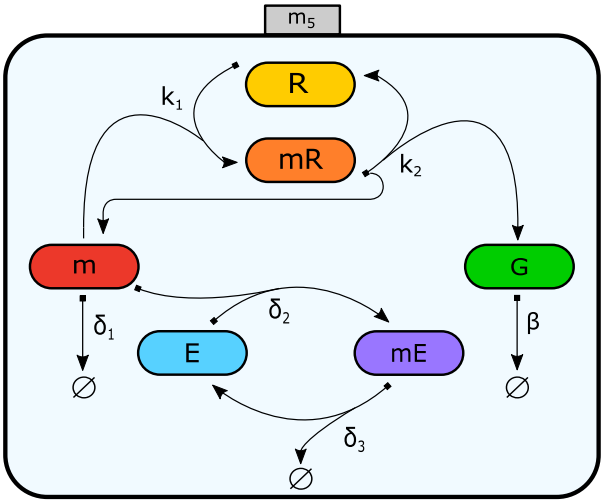
(a) *Ribosome-Model* $\mathcal{M}_3$ (b) *Enzyme-Degradation-Model* $\mathcal{M}_4$

Figure 9: In model $\mathcal{M}_3$ in Figure 9a the mRNA is translated into GFP via ribosomal binding and release with the rates k_1 and k_2. The mRNA degrades with the rate δ and the proteins with the rate β. Model $\mathcal{M}_4$ in Figure 9b describes a direct translation of mRNA with the rate k_{TL} and a direct decay of proteins with the rate β. The mRNA degrades via an enzymatic pathway with the rates δ_1 and δ_2.

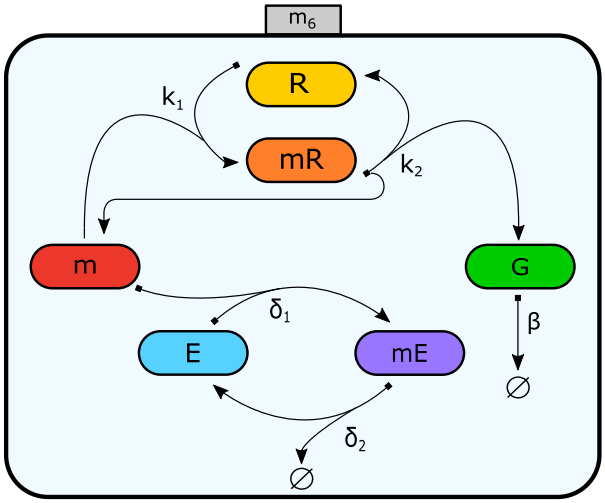
The *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$

The *Two-Degradation-Ribosome-Model* is the most complex model in the set as it includes a translation pathway of mRNA into proteins via ribosomes and two degradation pathways of the mRNA as shown by Figure 10a. The first is modelled by a direct decay and the second by an enzymatic degradation. The *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ is therefore a combination of the *Ribosome-Model* $\mathcal{M}_3$ and the *Two-Degradation-Model* $\mathcal{M}_2$ with a parameter vector of $\theta_5 = (\delta_1, \delta_2, \delta_3, \beta, k_1, k_2, m_0, E_0, R_0)$ with $n_{\theta,5} = 9$ elements. The rate equations are given as follows:

$$\begin{aligned}
\dot{m} &= -\delta_1 \cdot m - \delta_2 \cdot m \cdot E - k_1 \cdot m \cdot R + k_2 \cdot mR \\
\dot{G} &= -\beta \cdot G + k_2 \cdot mR \\
\dot{R} &= -k_1 \cdot m \cdot R + k_2 \cdot mR \\
\dot{mR} &= -\dot{R} \\
\dot{E} &= -\delta_1 \cdot m \cdot E + \delta_2 \cdot mE \\
\dot{mE} &= -\dot{E}
\end{aligned} \tag{29}$$



(a) *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$



(b) *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$

Figure 10: In Figure 10a the mRNA is translated by ribosomes with the rates k_1 and k_2 and degraded by enzymes with the rates δ_2 and δ_3. In addition the mRNA decays with the rate δ_1. The proteins degrade directly with the rate β. Model $\mathcal{M}_6$ in Figure 10b exhibits the same pathways as model $\mathcal{M}_5$ except of the simple decay of mRNA with the rate δ_1.

The *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$

The *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ includes a ribosomal translation of mRNA into proteins with the binding rate k_1 and release rate k_2. As shown by Figure 10b the transfected mRNA degrades via an enzymatic pathway with the binding rate δ_1 and the release rate δ_2. As for the *Trivial-Model* $\mathcal{M}_1$ the proteins degrade directly with the rate β. The *Enzyme-Degradation-Ribosome-Model* is thus a combination of the *Enzyme-Degradation-Model* $\mathcal{M}_4$ and the *Ribosome-Model* $\mathcal{M}_3$. The parameter vector $\theta_6 = (\delta_1, \delta_2, \beta, k_1, k_2, m_0, E_0, R_0)$ includes the rates and the unknown initial concentrations of ribosomes R_0, enzymes E_0 and transfected mRNA m_0. The number of model parameters is equal to $n_{\theta,6} = 8$ and the rate equations are given as follows:

$$\begin{aligned}
\dot{m} &= -\delta_1 \cdot m \cdot E - k_1 \cdot m \cdot R + k_2 \cdot mR \\
\dot{G} &= -\beta \cdot G + k_2 \cdot mR \\
\dot{R} &= -k_1 \cdot m \cdot R + k_2 \cdot mR \\
\dot{mR} &= -\dot{R} \\
\dot{E} &= -\delta_1 \cdot m \cdot E + \delta_2 \cdot mE \\
\dot{mE} &= -\dot{E}
\end{aligned} \tag{30}$$

4.2 Model selection for artificial datasets

In this chapter the question is addressed, whether a model that is closest to the underlying network of biochemical reactions is selected as best model by AIC and BIC. For this purpose each introduced model in the set is used once to generate an associated dataset of 100 trajectories which is explained in Chapter 4.2.1. Afterwards each model in the set is fitted on the trajectories per dataset as explained in Chapter 4.2.2 and the AIC and BIC values are calculated for each individual fit as described in Chapter 3.2.4. To compare the models on the level of the population of the 100 trajectories, the AIC and BIC population weights were calculated and sorted in decreasing order to obtain a ranking. The results on the population scale are analysed in Chapter 4.2.3. Furthermore the model selection is studied in detail for each individual fit per population in Chapter 4.2.4 to find out, whether multiple models were selected as best within the population.

4.2.1 Generation of artificial data

For each model $\mathcal{M}_i$ an artificial dataset $\mathcal{D}_j$ was generated by simulating the corresponding ODE-System $\Sigma(\theta_{\mathcal{M}_i})$ with Matlab. To capture cell-to-cell-variability the parameters of the deterministic system have to differ among each other.

For this purpose fixed parameter values $\theta_{i,fixed}$ are altered as follows:

$$\theta_{i,noise} = \theta_{i,fixed} \cdot e^{\sigma_\theta \cdot r} \tag{31}$$

The value r is a pseudorandom number drawn from the standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ and is multiplied by a value $\sigma_\theta = 0.1$. To cover the different onset times of protein expression, the starting time $t_n(0)$ of the simulation per trajectory n was given as an additional model parameter. To generate artificial measurement errors for each trajectory n, the data $D_{j,n}$ was multiplied by normal noise with $\sigma_{noise} = 0.7$:

$$\mathcal{D}_{j,n,stochastic} = \mathcal{D}_{j,n} \cdot e^{\sigma_{noise} \cdot r} \tag{32}$$

This way $N = 100$ trajectories per dataset $\mathcal{D}_j$ were generated. The data that was simulated for time points below 1.9 h was excluded from the artificial datasets to be in accordance with the experimental datasets which can not be observed during the time of approximately 1 to 2 h when the cells are incubated with the transfection medium.

4.2.2 Fitting of artificial data

Each model $\mathcal{M}_i$ in the set $M = \{\mathcal{M}_i\}_{i=1}^6$ was fitted once on each trajectory in the dataset $\mathcal{D}_j$ of the set $D = \{\mathcal{D}_j\}_{j=1}^6$. A sigmoid function τ was used to model the different expression onset times:

$$\tau = \frac{t^{10}}{t^{10} + t_0^{10}} \tag{33}$$

This function was multiplied either by the translation rate k_{TL} of the models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_4$ or by the ribosomal binding rate k_1 of the models $\mathcal{M}_3$, $\mathcal{M}_5$ and $\mathcal{M}_6$. The parameter t_0 was added to the parameter vector together with the unknown measurement noise σ and the scaling and offset parameters denoted as sc and κ as explained in Chapter 3.2.1. The parameters were estimated by the global optimization of the likelihood function as explained in Chapter 3.2.2 using 150 different starting points that were chosen by latin hypercube sampling.

4.2.3 Model selection on the scale of populations

In this chapter the question is addressed whether the model that generated the data is selected as best model on the population level. For this purpose artificial datasets were generated that consists of $N = 100$ individual trajectories as explained in Chapter 4.2.1. Each model $\mathcal{M}_i$ in the set was used once to fit the trajectories of dataset $\mathcal{D}_j$. Afterwards, the AIC values of the N trajectory fits were summed up for each model $\mathcal{M}_i$ in dataset $\mathcal{D}_j$ to obtain the AIC population values $AIC^{i,j}_{pop.}$. They were used to calculate the differences $\Delta^{i,j}_{AIC,pop.}$ for all six models with respect to the best model that is given by the lowest AIC population value $\min_i AIC^{i,j}_{pop.}$:

$$\begin{aligned} AIC^{i,j}_{pop.} &= \sum_{n=1}^{N} AIC_{n,i,j} \\ \Delta^{i,j}_{AIC,pop.} &= AIC^{i,j}_{pop.} - \min_i AIC^{i,j}_{pop.} \end{aligned} \tag{34}$$

As the differences $\Delta^{i,j}_{AIC,pop.}$ are not easily interpretable, the AIC population weights $w^{i,j}_{AIC,pop.}$ were calculated for each model $\mathcal{M}_i$ in the dataset $\mathcal{D}_j$ as follows:

$$w^{i,j}_{AIC,pop.} = \frac{\exp\left(-\frac{1}{2} \cdot \Delta^{i,j}_{AIC,pop.}\right)}{\sum_{i=1}^{6} \exp\left(-\frac{1}{2} \cdot \Delta^{i,j}_{AIC,pop.}\right)} \tag{35}$$

In accordance to the AIC weights that were explained in Chapter 3.2.4 the AIC population weights refer to the probability that model $\mathcal{M}_i$ is the best model on the population scale of dataset $\mathcal{D}_j$. The same method was used to calculate the BIC population weights $w^{i,j}_{BIC,pop.}$. A ranking of models was obtained for each dataset $\mathcal{D}_j$ by sorting the population weights in decreasing order. The rankings are analysed in the following for each dataset dependent on either AIC or BIC as model selection criterion.

Model selection based on AIC population weights

The matrix pattern in Figure 11a shows a ranking of all models $\mathcal{M}_i$ based on the AIC population weights for each dataset $\mathcal{D}_j$. The six ranking positions are represented by different colours as displayed in the colour-bar. For example model $\mathcal{M}_1$ was selected as the best model for dataset $\mathcal{D}_1$ in the last row, followed by the models $\mathcal{M}_3$, $\mathcal{M}_4$, $\mathcal{M}_6$, $\mathcal{M}_2$ and $\mathcal{M}_5$.
It was expected that for each dataset the model that generated the data would be selected as best model. In this case the ranking pattern would show a brown-red coloured diagonal of first ranks from the bottom left corner to the upper right in Figure 11a. This assumption is obviously not confirmed by the ranking. Except for the datasets $\mathcal{D}_4$ and $\mathcal{D}_6$ the *Trivial-Model* $\mathcal{M}_1$ was selected as best model. Thus, the least complex model with the fewest number of

parameters $n_{\theta,1} = 8$ was able to cover the data not only for dataset $\mathcal{D}_1$ but also for datasets that were generated by more complex models. Oppositely the most complex model $\mathcal{M}_5$ with the highest number of parameters $n_\theta = 13$ was low ranked for all datasets except of $\mathcal{D}_6$. Only for datasets $\mathcal{D}_6$ and $\mathcal{D}_4$ similar complex models than those that generated the data were highest ranked.

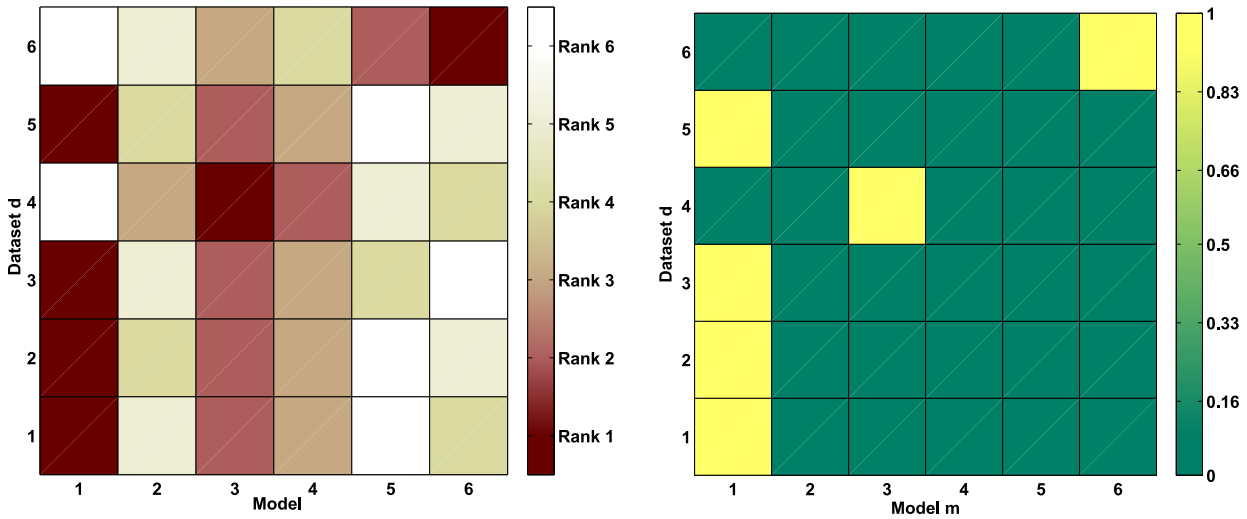
In completion to the ranking in Figure 11a the matrix pattern in Figure 11b shows the probability of model $\mathcal{M}_i$ to be the best model on the population scale of dataset $\mathcal{D}_j$. The corresponding weights within specific ranges are shown in colours from dark green to yellow as displayed in the colour-bar. As their values are related to probabilities, they sum up to 1 in a row of one dataset. The *Trivial-Model* $\mathcal{M}_1$ shows high values above 0.83 for the datasets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ and $\mathcal{D}_5$ which is in agreement with the first rank in Figure 11a. Dataset $\mathcal{D}_6$ shows the highest weight for its generating model $\mathcal{M}_6$. Astonishingly model $\mathcal{M}_3$ has the highest weight in dataset $\mathcal{D}_4$ even though model $\mathcal{M}_4$ was used to generate the data and has the same number of parameters $n_{\theta,4} = 10$.

Model selection based on BIC population weights

In comparison to AIC less complex models are selected by BIC as the calculation of BIC values includes the number of observation points in the penalty term for model complexity. This property is reflected in the ranking matrix based on BIC population weights shown by Figure 12a. Most datasets show a ranking pattern that is similar to the order of the model complexity in increasing order as follows: $\mathcal{M}_1$, $\mathcal{M}_3$, $\mathcal{M}_4$, $\mathcal{M}_2$, $\mathcal{M}_6$ and $\mathcal{M}_5$. In agreement with the result obtained by AIC population weights the ranking deviates from this tendency for the datasets $\mathcal{D}_4$ and $\mathcal{D}_6$. The BIC population weights are displayed in Figure 12b and exhibit the same pattern as for the AIC population weights 11b.

Conclusion

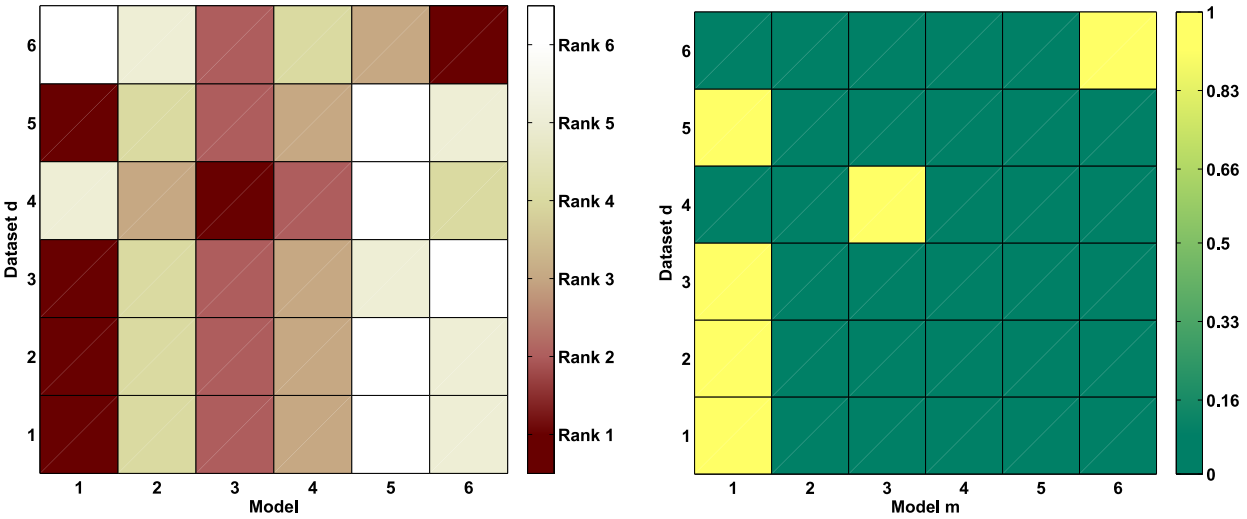
The expectation that the generating model will be selected as best model for its associated dataset was not confirmed by the model selection based on AIC and BIC population weights. On the contrary both selection criteria favoured less complex models. As explained in chapter 3.2.4 the BIC selects less complex models than its target model when the number of data points per trajectory is small. In this work a trajectory per dataset consists of 116 data points. The AIC depends on the number of data points as well and would select more complex models with an increasing number of data points. Both model selection criteria are defined in the limit of an infinite sample size. In practice this is not the case and less complex models are favoured as best models than the target models. In addition both criteria assume identifiable parameters, which is not ensured for the models that were used in this work.



(a) Ranking matrix of AIC population weights

(b) Matrix of AIC population weights

Figure 11: Model selection based on AIC population weights: The matrix pattern in Figure 11a depicts the ranking of models for each dataset that are represented by different colours as displayed in the colour-bar. The matrix coloured from dark green to yellow in Figure 11b shows the probability of model $\mathcal{M}_i$ to be the best model for dataset $\mathcal{D}_j$ in the sense of AIC. They are divided into ranges as displayed in the colour-bar.



(a) Ranking matrix of BIC population weights

(b) Matrix of BIC population weights

Figure 12: Model selection based on BIC population weights: In comparison to AIC less complex models were favoured by model selection based on BIC population weights and the same models were highest ranked as shown by Figure 12a. The same pattern is obtained in the matrix of BIC population weights in Figure 12b.

4.2.4 Model selection of individuals

The model selection on the population scale yields a model that is able to describe the dynamics best for an ensemble of cells. This way the method ignores the potential presence of subgroups within the population that favour different models as best for their fits. In this work the cell-to-cell variability was modelled by fixed parameters that were altered by random numbers drawn from a normal distribution as explained in Chapter 4.2.1. This may cause subgroups and different model selection patterns within the population of a dataset. These model selection substructures are analysed in the following by considering the model selection of individual fits.

Artificial dataset $\mathcal{D}_1$

The artificial dataset $\mathcal{D}_1$ was generated by model $\mathcal{M}_1$ and fitted by each model in the set as shown by Figure 13. The data points that were generated below the time 1.9 h were removed from the dataset to be in accordance with the experimental datasets, that consists only of measurement points after the incubation time of the transfection medium. For this reason the data points below 1.9 h were removed for all datasets $\mathcal{D}_j$ in this work. By considering the AIC and BIC weights for each fit within the population of 100 trajectories, the model selection pattern of individuals is obtained.

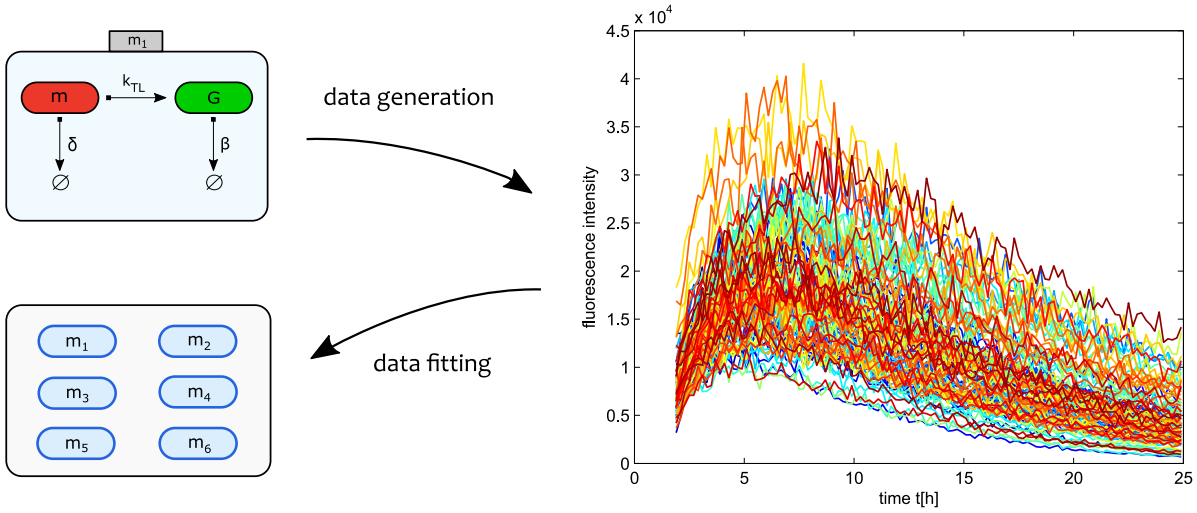
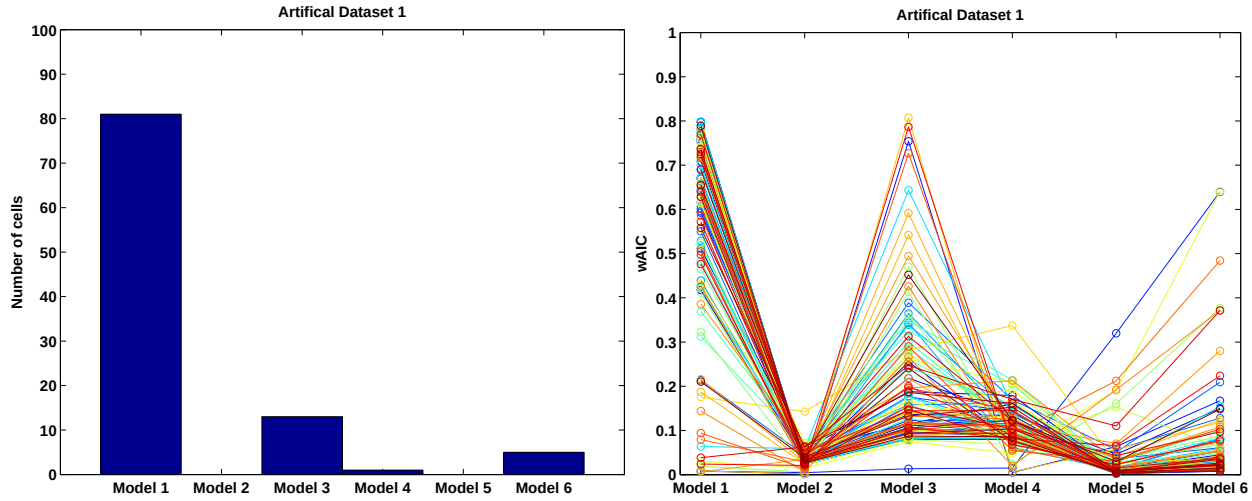


Figure 13: Artificial dataset $\mathcal{D}_1$: The data was generated by the *Trivial-Model* $\mathcal{M}_1$ with parameters $\log_{10}(\theta) = \log_{10}(m_0, \delta, \beta, k_{TL}, t_0, sc) = (3, -0.5, -1, 1, 0, 0)$. Afterwards each model $\mathcal{M}_i$ in the set $M = \{\mathcal{M}_i\}_{i=1}^{6}$ was used once to fit the individual trajectories. For each fit the model selection was done using AIC and BIC weights.

In Chapter 4.2.3 the model selection based on AIC population weights yields a ranking pattern as follows: $\mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_6, \mathcal{M}_2, \mathcal{M}_5$ (Figure 11a). Less complex models were thus able to fit the data best which was expected as the least complex model $\mathcal{M}_1$ was used to generate the data. Figure 14a shows, how frequently each model was selected as best model within the population based on the highest AIC weight. The highest ranked models $\mathcal{M}_1$ and $\mathcal{M}_3$ were chosen most frequently followed by $\mathcal{M}_6$ and $\mathcal{M}_4$. Interestingly model $\mathcal{M}_6$ was selected more often as best model than $\mathcal{M}_4$ which was higher ranked on the population scale. In Figure 14b the AIC weights per model are displayed as empty circles for each colour-marked individual trajectory fit in the population of $N = 100$ fits. The weights of model $\mathcal{M}_1, \mathcal{M}_3$ and $\mathcal{M}_6$ spread over a wide range whereas those of $\mathcal{M}_2, \mathcal{M}_4$ and $\mathcal{M}_5$ are more densely distributed at low values. However, the majority of individual fits in the population shows higher weights for model $\mathcal{M}_4$ than $\mathcal{M}_6$. This explains, why $\mathcal{M}_4$ was higher ranked. $\mathcal{M}_4$ has higher weights and lower AIC values than $\mathcal{M}_6$. The AIC population weights were obtained by summing up all AIC values per model over all individuals. For this reason the AIC population value is smaller for $\mathcal{M}_4$ than $\mathcal{M}_6$ which leads to a higher population weight and ranking position.

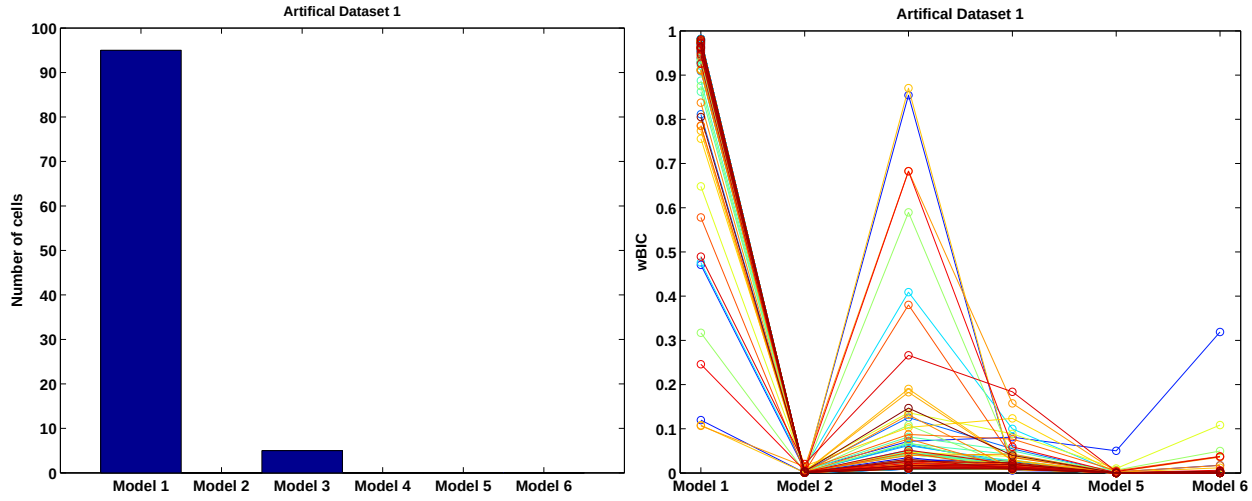
The ranking pattern based on BIC population weights differs from the ranking obtained by AIC population weights in an interchange of model $\mathcal{M}_2$ and $\mathcal{M}_6$ as shown by Figure 12a. The models are in the same order as it would be obtained by sorting them with increasing complexity. Figure 15a shows the frequency of best models that were selected by the highest BIC weight. Model $\mathcal{M}_1$ was chosen most frequently followed by $\mathcal{M}_3$. In comparison to the best models that were selected by the highest AIC weight in Figure 14a the frequency of $\mathcal{M}_1$ as best model is increased and reduced for $\mathcal{M}_3$ whereas $\mathcal{M}_4$ and $\mathcal{M}_6$ are excluded as best model. In Figure 15b only the models $\mathcal{M}_1$ and $\mathcal{M}_3$ exhibit BIC weights on a wide range whereas those of the other models are more restricted at low values. The property of BIC to select less complex models than AIC is reflected on the individual as well as on the population scale.

In conclusion the generating model $\mathcal{M}_1$ was highest ranked on the population scale and selected most frequently for the individual fits independent of AIC or BIC as model selection criterion. Only a few individuals selected the *Ribosome-Model* $\mathcal{M}_3$ as best model that was ranked on the second position on the population scale.



(a) Frequency of selected best models (b) Individual AIC weights

Figure 14: Model selection based on AIC weights: Figure 14a shows the frequency of individuals that selected the displayed models as best based on the highest AIC weight. Figure 14b shows the AIC weight per model as empty circles for each colour-marked individual fit within the population of 100 trajectories. The models $\mathcal{M}_1$ and $\mathcal{M}_3$ were selected most frequently as best model and exhibited the highest weights.



(a) Frequency of selected best models (b) Individual BIC weights

Figure 15: Model selection based on BIC weights: As shown by Figure 15a the frequency as best model based on the highest BIC weight is increased for the least complex model $\mathcal{M}_1$ and reduced for all other models in comparison to AIC. As displayed in Figure 15b the BIC weights of model $\mathcal{M}_1$ are significantly higher than for all other models except for a small minority that shows high values for $\mathcal{M}_3$.

Artificial dataset $\mathcal{D}_2$

As shown by Figure 16 the dataset $\mathcal{D}_2$ was generated by the *Two-Degradation-Model* $\mathcal{M}_2$. All models in the set were fitted once to each trajectory in the dataset as explained in Chapter 4.2.2. By this way the AIC and BIC weights were calculated for each fit individually.

As shown by Figure 11a the model selection based on AIC population weights yielded a ranking of models for the dataset $\mathcal{D}_2$ as follows: $\mathcal{M}_1$, $\mathcal{M}_3$, $\mathcal{M}_4$, $\mathcal{M}_2$, $\mathcal{M}_6$, $\mathcal{M}_5$. Even though the *Two-Degradation-Model* $\mathcal{M}_2$ was used to generate the dataset, less complex models were favoured as best model on the population scale. The model selection based on BIC population weights shows the same ranking as depicted in Figure 12a. In comparison to the population scale the models $\mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_4$ were as well selected most frequently as best models for the individual fits within the population as shown by Figure 17a. On the other hand the complex model $\mathcal{M}_6$ was selected more often by highest AIC weights than the generating model $\mathcal{M}_2$ which was higher ranked on the population scale. In Figure 17b the AIC weights of the colour-marked individual fits are shown in detail for each model. Two substructures are displayed: The first group shows highest weights close to 1 for model $\mathcal{M}_1$ and nearly zero valued weights for all other models. The second group shows approximately highest weights in the range from 0.3 to 0.5 for model $\mathcal{M}_3$ closely followed by model $\mathcal{M}_4$ with weights in the range from 0.2 to 0.4. Only a few individuals deviate from both groups by exhibiting highest weights for model $\mathcal{M}_2, \mathcal{M}_5$ or $\mathcal{M}_6$.

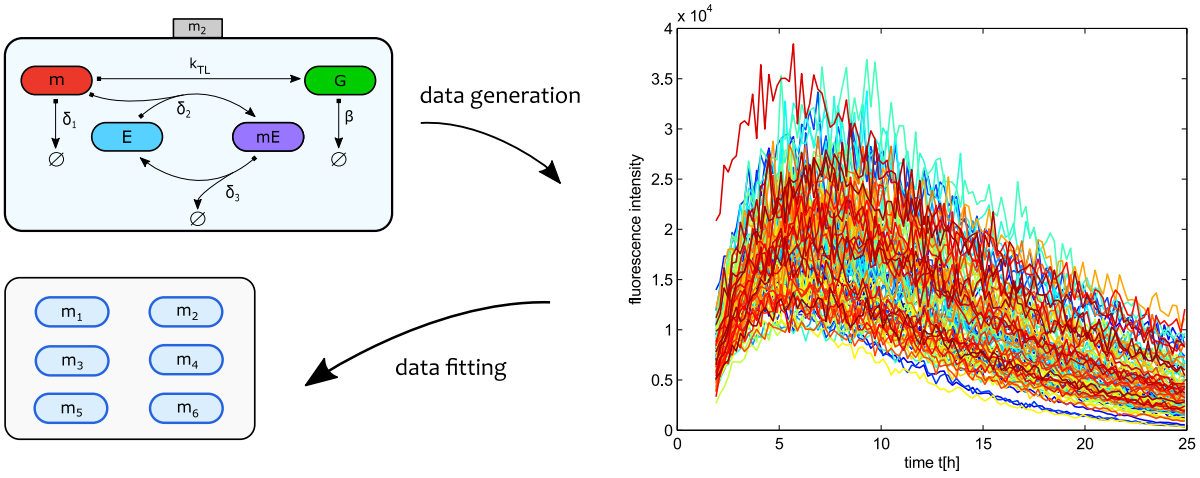
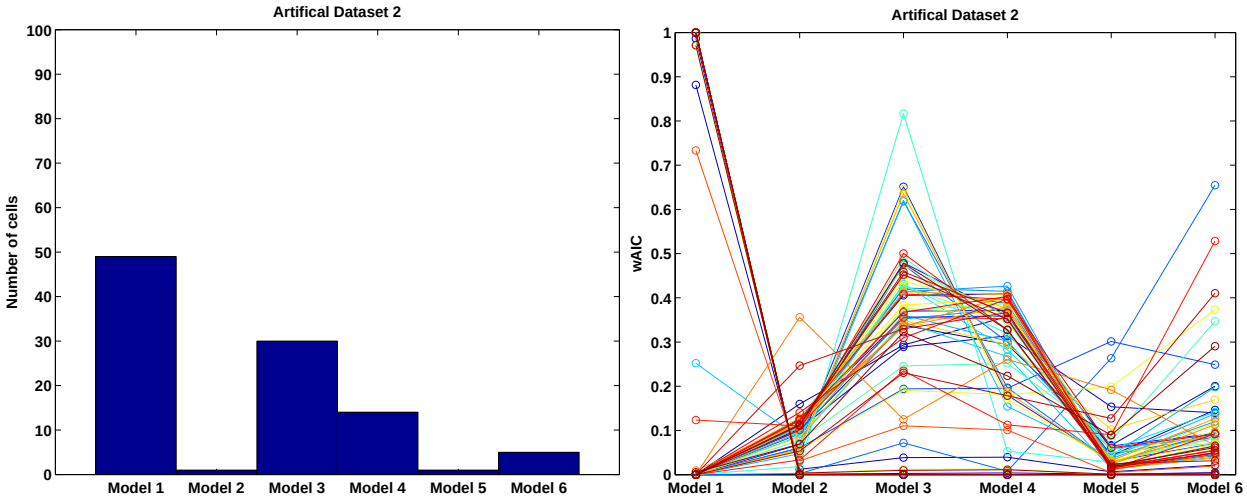
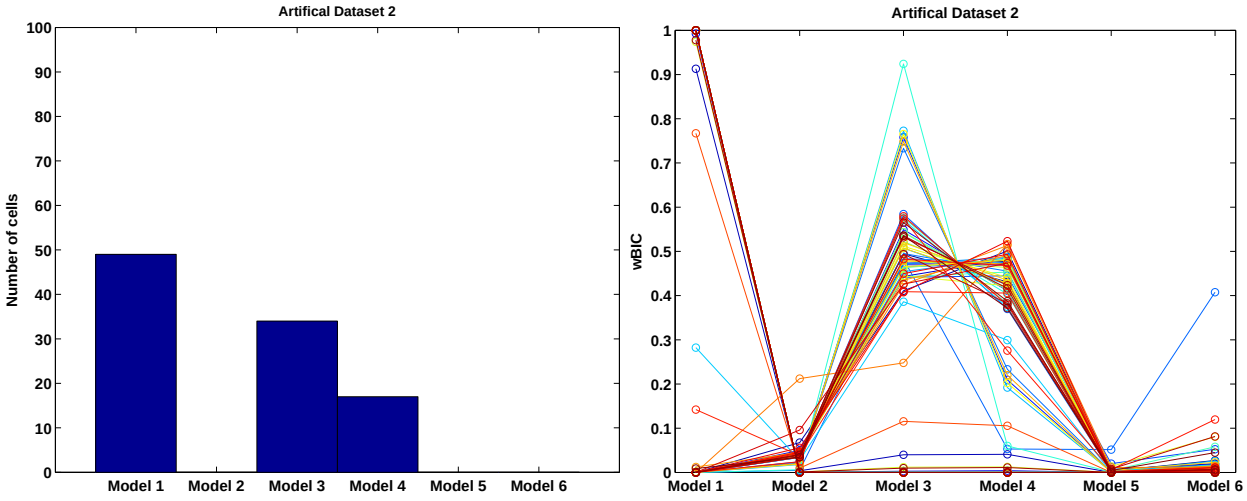


Figure 16: Artificial dataset $\mathcal{D}_2$: The dataset was generated by model $\mathcal{M}_2$ with logarithmic parameter values of $\log_{10}(\theta) = \log_{10}(m_0, E_0, \delta_1, \delta_2, \delta_3, \beta, k_{TL}, t_0, sc) = (2, 0.5, -1, 2, -1, -0.5, 2, 0, 0)$ as explained in Chapter 4.2.1. All models in the set $M = \{\mathcal{M}_i\}_{i=1}^{6}$ were fitted once to the dataset as explained in Chapter 4.2.2. Thereby the AIC and BIC weights per fit were calculated.



(a) Frequency of selected best models (b) Individual AIC weights

Figure 17: Model selection based on AIC weights: The histogram 17a shows that most individuals selected model $\mathcal{M}_1$ as best model based on their highest AIC weight followed by $\mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_6, \mathcal{M}_2$ and $\mathcal{M}_5$. Figure 17b shows two substructures: The first shows highest weights for $\mathcal{M}_1$ whereas the second exhibits highest weights for $\mathcal{M}_3$ and $\mathcal{M}_4$.



(a) Frequency of selected best models (b) Individual BIC weights

Figure 18: Model selection based on BIC weights: As shown in Figure 18a the models $\mathcal{M}_1, \mathcal{M}_3$ and $\mathcal{M}_4$ were selected as best models with the highest BIC weights per individual fit. In comparison to the individual model selection by AIC weights the frequencies of the models $\mathcal{M}_3$ and $\mathcal{M}_4$ are increased and reduced for all other models as shown by Figure 18b.

Figure 18a shows the best models that were selected for the individual fits based on the highest BIC weights. In comparison to AIC the models $\mathcal{M}_2, \mathcal{M}_5$ and $\mathcal{M}_6$ were not selected as best models whereas the frequencies of $\mathcal{M}_3$ and $\mathcal{M}_4$ were slightly increased. In Figure 18b the weights of $\mathcal{M}_3$ and $\mathcal{M}_4$ were shifted towards higher values whereas those of $\mathcal{M}_2, \mathcal{M}_5$ and $\mathcal{M}_6$ were further decreased.

Summarized less complex models than the generating model $\mathcal{M}_2$ were able to fit the data and were selected as best models on the population scale as well as for the individual fits. By considering the model selection results in detail for each fit two substructures were found. The first exhibits highest weights for the *Trivial-Model* $\mathcal{M}_1$ whereas the second shows high weights for the *Ribosome-Model* $\mathcal{M}_3$ closely followed by the *Enzyme-Degradation-Model* $\mathcal{M}_4$.

Artificial dataset $\mathcal{D}_3$

The artificial dataset $\mathcal{D}_3$ was generated by model $\mathcal{M}_3$ as shown by Figure 19. Each model in the set was fitted once to each trajectory in the dataset as explained in Chapter 4.2.2. This way the AIC and BIC weights for each model per individual fit were calculated.

The model selection based on the AIC population weights revealed a ranking pattern in decreasing order as follows: $\mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5, \mathcal{M}_2, \mathcal{M}_6$ (Figure 11a). The result based on BIC population weights only differs in an exchange of the model $\mathcal{M}_5$ by $\mathcal{M}_2$ (Figure 12a). Astonishingly the model $\mathcal{M}_5$ was selected more frequently as best model based on the highest AIC weight than $\mathcal{M}_1$ or $\mathcal{M}_3$ as displayed in Figure 20a. In Figure 20b the AIC weights for the individual fits in the population are shown in detail for each model. Two substructures were found. The first group selected model $\mathcal{M}_5$ as best model with highest AIC weights close to 1 and weights close to zero for all other models. The second group shows highest weights for the models $\mathcal{M}_1$ or $\mathcal{M}_3$ and weights close to zero for the model $\mathcal{M}_5$. The substructures indicate that only $\mathcal{M}_5$ was able to fit the data of the second group whereas the models $\mathcal{M}_1$ and $\mathcal{M}_3$ are similar good to fit the data of the first group. In comparison the model selection based on AIC weights the frequency of the *Trivial-Model* $\mathcal{M}_1$ is higher for model selection based on BIC weights. On the contrary the models $\mathcal{M}_3$ and $\mathcal{M}_4$ were selected less frequently. This result reflects the property of BIC to favour less complex models. However, the frequency of $\mathcal{M}_5$ that was selected as best model for the second group remained unchanged.

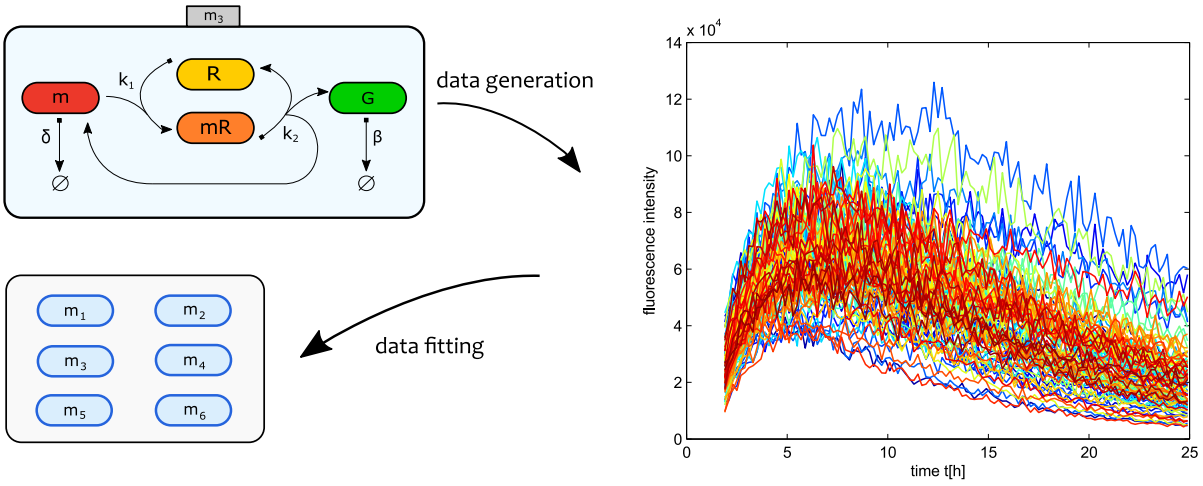
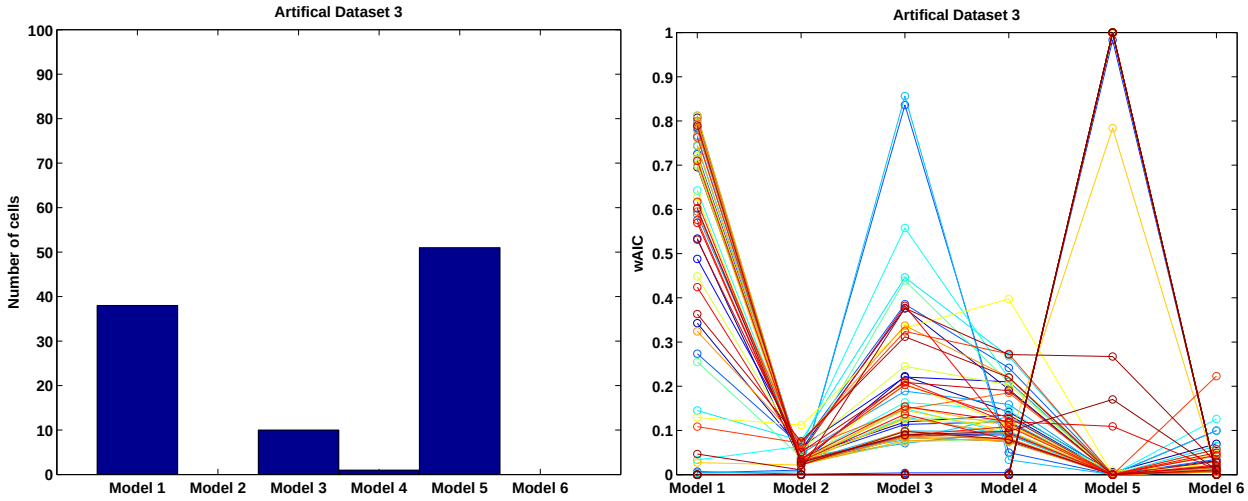
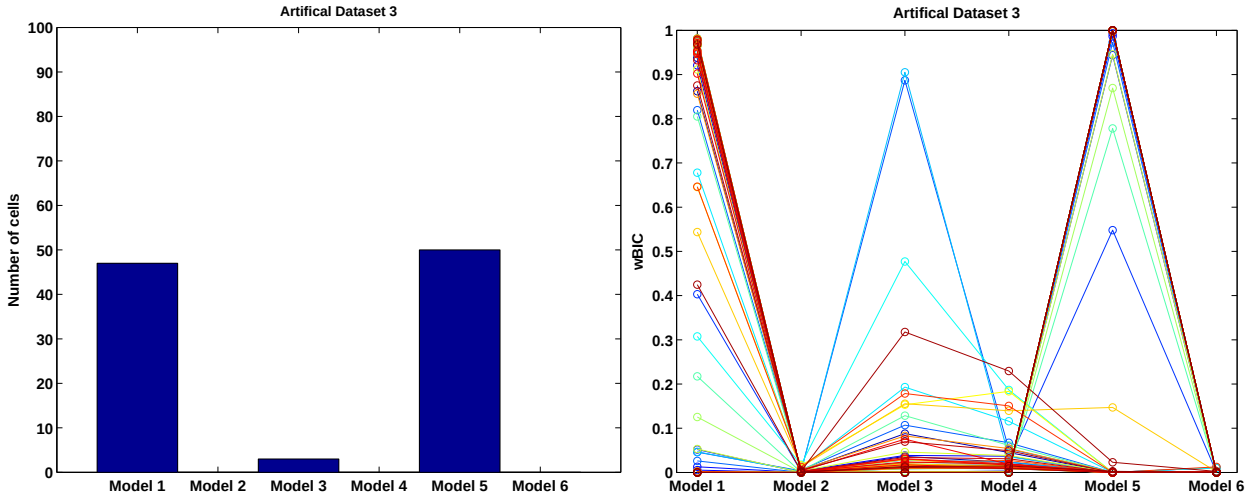


Figure 19: Artifical dataset $\mathcal{D}_3$: Model $\mathcal{M}_3$ was used to generate the data as explained in Chapter 4.2.1. The fixed logarithmic parameters to base 10 were chosen as follows $\log_{10}(\theta) = \log_{10}(m_0, R_0, \delta, \beta, k_1, k_2, t_0, sc) = (3, 2, 0.5, -0.75, 2, 2.5, 0, 0)$. Afterwards each model in the set was fitted once to the trajectories of dataset $\mathcal{D}_3$ and the individual AIC and BIC weights were calculated.



(a) Frequency of selected best models (b) Individual AIC weights

Figure 20: Model selection based on AIC weights: Figure 20a shows that the most complex model $\mathcal{M}_5$ was selected most frequently as best model followed by $\mathcal{M}_1$, $\mathcal{M}_3$ and $\mathcal{M}_4$. Figure 20b shows two substructures. The first exhibits highest AIC weights for model $\mathcal{M}_5$ and low values for all other models. The second shows high AIC weights for $\mathcal{M}_1$ and $\mathcal{M}_3$ followed by $\mathcal{M}_4, \mathcal{M}_2$ and $\mathcal{M}_6$.



(a) Frequency of selected best models (b) Individual BIC weights

Figure 21: Model selection based on BIC weights: In comparison to model selection based on AIC weights more individuals selected $\mathcal{M}_1$ as best model whereas the frequencies of $\mathcal{M}_3$ and $\mathcal{M}_4$ are decreased. Model $\mathcal{M}_5$ is selected most frequently. In Figure 21b the models $\mathcal{M}_1$ and $\mathcal{M}_5$ show highest BIC weights for most of the individuals.

Artificial dataset $\mathcal{D}_4$

The dataset was generated by the *Enzyme-Degradation-Model* $\mathcal{M}_4$ and fitted by each model in the set as shown by Figure 22. The AIC and BIC weights were calculated for each individual fit separately. Model $\mathcal{M}_3$ was highest ranked and exhibits population weights above 0.83 for AIC as well as BIC. The ranking obtained by AIC population weights $\mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_2, \mathcal{M}_6, \mathcal{M}_5, \mathcal{M}_1$ only differs from the ranking by BIC population weights in an exchange of the most complex model $\mathcal{M}_5$ by the least complex model $\mathcal{M}_1$.

In Figure 23a the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ was selected most frequently based on the highest AIC and BIC weights even though it was ranked on the fourth position on the population scale as shown by Figure 11a. However, the number of individuals that chose the highest ranked model $\mathcal{M}_3$ is close to the frequency of $\mathcal{M}_6$. In Figure 23b two substructures can be found. The first group shows high AIC weights close to 1 for model $\mathcal{M}_6$ whereas those of all other models are close to zero. The second group exhibits highest AIC weights for the *Ribosome-Model* $\mathcal{M}_3$ and for the *Enzyme-Degradation-Model* $\mathcal{M}_4$. The model selection based on BIC weights yields similar frequencies for the models that were selected as best by AIC. As shown by Figure 24a only a few more individuals selected the *Trivial-Model* $\mathcal{M}_1$ as best. As shown by Figure 24b the BIC weights of $\mathcal{M}_1$ are increased but only a few individuals show higher weights for $\mathcal{M}_1$ than for $\mathcal{M}_3$.

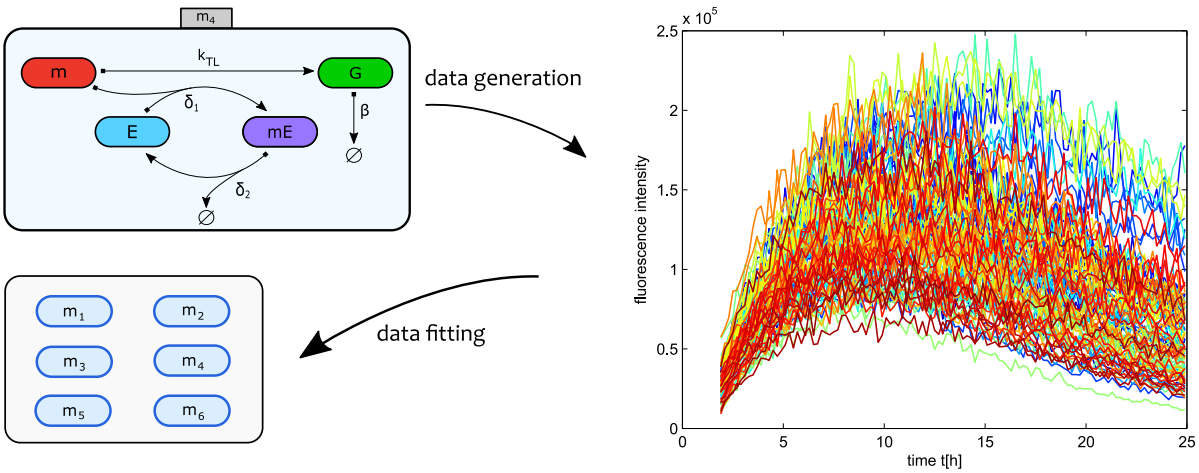
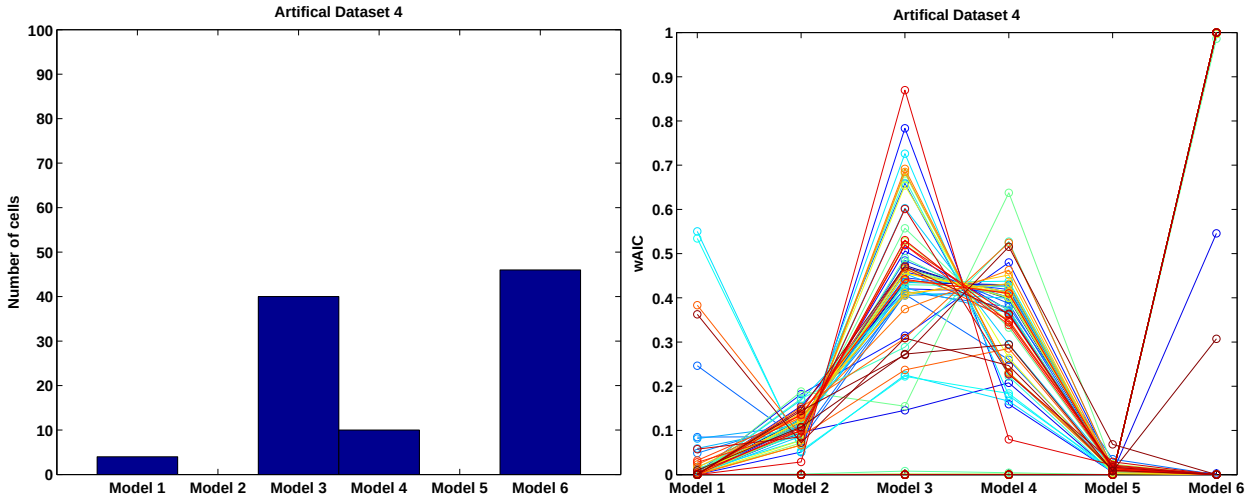
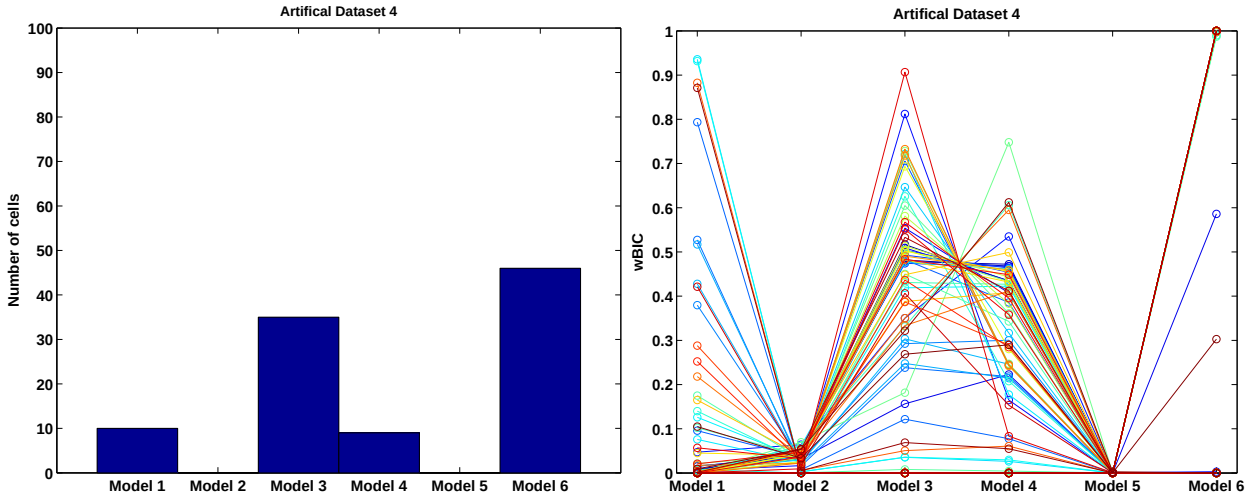


Figure 22: Artificial dataset $\mathcal{D}_4$: The dataset was generated by model $\mathcal{M}_4$ using logarithmic parameters to base 10 of $\log_{10}(\theta) = \log_{10}(m_0, E_0, \delta_1, \delta_2, \beta, k_{TL}, t_0, sc) = (3, -1.5, 1, 3.5, -1, 1.5, 0, 0)$ as explained in Chapter 4.2.1. Each model in the set $M = \{\mathcal{M}_i\}_{i=1}^{6}$ was fitted once to each trajectory of the dataset. This way the AIC and BIC weights were calculated for each fit.



(a) Frequency of selected best models (b) Individual AIC weights

Figure 23: Model selection based on AIC weights: As depicted in histogram 23a model $\mathcal{M}_6$ was selected most frequently as best model based on the highest AIC weights followed by $\mathcal{M}_3, \mathcal{M}_4$ and $\mathcal{M}_1$. The Figure 23b shows two substructures. On the one hand model $\mathcal{M}_6$ exhibits high AIC weights close to 1 and on the other hand model $\mathcal{M}_3$ and $\mathcal{M}_4$ show similar high weights that are most densely located at values close to 0.45.



(a) Frequency of selected best models (b) Individual BIC weights

Figure 24: Model selection based on BIC weights: In comparison to model selection based on the highest AIC weights the same models were selected as best by highest BIC weights as shown in Figure 24a. The BIC weights of model $\mathcal{M}_1$ shown by Figure 23b are higher than the AIC weights in Figure 23b.

Artificial dataset 5

Dataset $\mathcal{D}_5$ was generated with the *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ as shown by Figure 25 and explained in Chapter 4.2.1. Each model $\mathcal{M}_i$ in the set $M = \{\mathcal{M}_i\}_{i=1}^{6}$ was used once to fit the $N = 100$ trajectories of the dataset as explained in Chapter 4.2.2. The AIC and BIC weights were calculated for each fit separately to study the model selection for the individual fits.

As shown by Figure 11a and 12a the model selection on the population scale yields the same ranking for AIC and BIC as follows: $\mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_2, \mathcal{M}_6, \mathcal{M}_5$. This is the same order that is obtained by sorting the models with increasing complexity. Indeed, the *Trivial-Model* $\mathcal{M}_1$ is selected most frequently for the individual fits based on the highest AIC and BIC weight as shown by Figure 26a and 27a. In contrast to BIC the models $\mathcal{M}_3$ and $\mathcal{M}_4$ were selected as best for a few individual fits by AIC. Their AIC weights are more widely distributed to higher values as shown by Figure 26b than their BIC weights that are displayed in Figure 27b. In conclusion the *Trivial-Model* $\mathcal{M}_1$ with the fewest number of parameters in the model set was sufficient to fit the data that was generated by the *Two-Degradation-Ribosome-Model* that is the most complex model in the set.

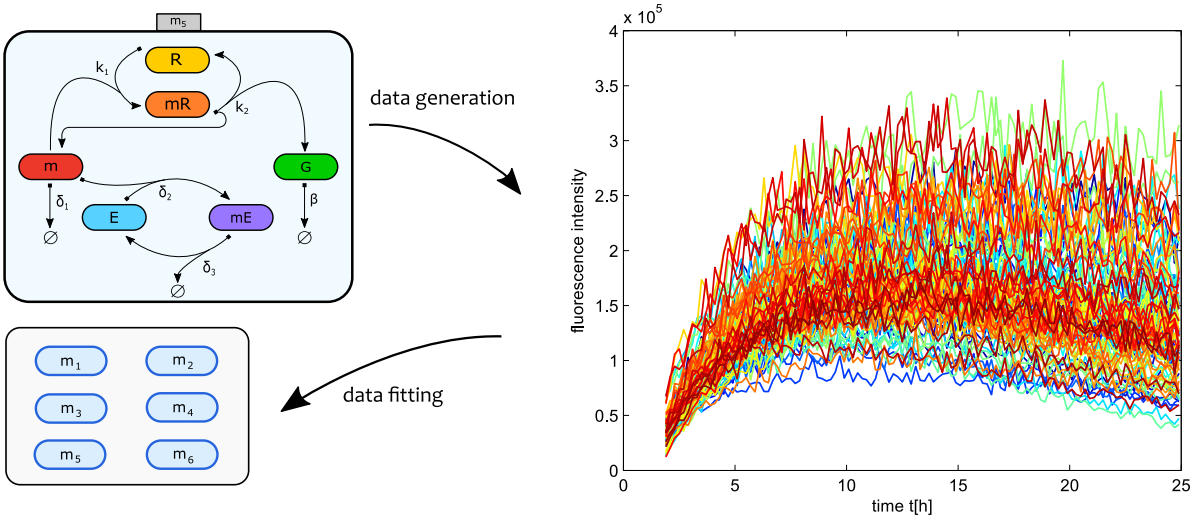
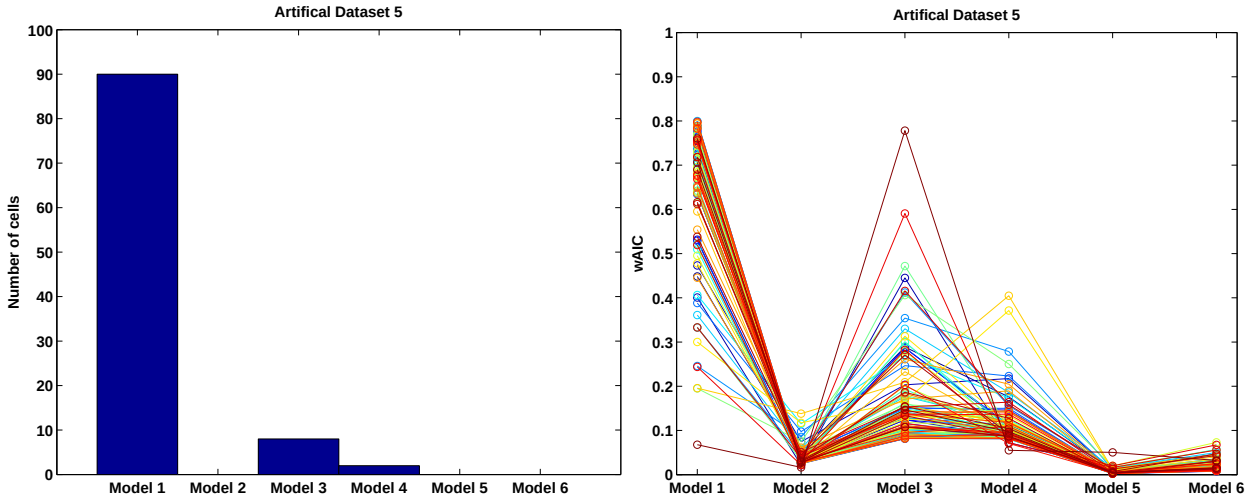
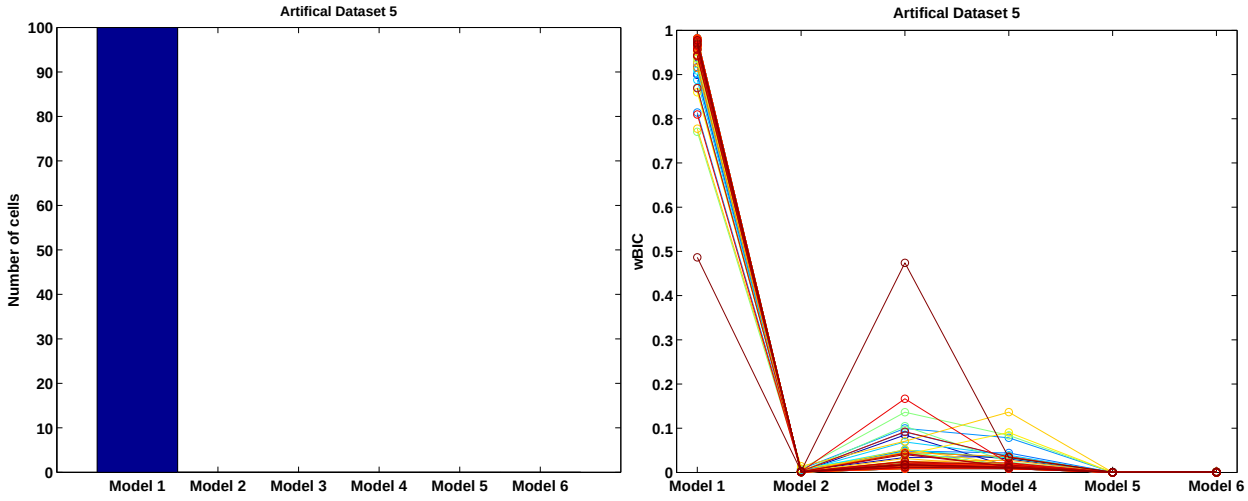


Figure 25: Artificial dataset $\mathcal{D}_5$: Model $\mathcal{M}_5$ was used to generate the data with logarithmic parameters of $\log_{10}(\theta) = \log_{10}(m_0, R_0, E_0, \delta_1, \delta_2, \delta_3, \beta, k_1, k_2, t_0, sc) = (3, 2.3, 2, 0.1, 3, -2, -1, 0.2, 1.8, 0, 0)$ as explained in Chapter 4.2.1. Each model in the set $M = \{\mathcal{M}_i\}_{i=1}^{6}$ was fitted once the individual trajectories. This way the AIC and BIC values were calculated for each model per individual fit.



(a) Frequency of selected best models (b) Individual AIC weights

Figure 26: Model selection based on AIC weights: The histograms 26a shows that the least complex model in the set $\mathcal{M}_1$ was selected as best model based on the highest AIC weights followed by $\mathcal{M}_3$ and $\mathcal{M}_4$. In Figure 26b the AIC weights of model $\mathcal{M}_1$ and $\mathcal{M}_3$ are distributed over a wide range, but the majority of weights shows values above 0.6 for $\mathcal{M}_1$ and close to 0.1 for $\mathcal{M}_3$ and $\mathcal{M}_4$.



(a) Frequency of selected best models (b) Individual BIC weights

Figure 27: Model selection based on BIC weights: In histogram 27a the *Trivial-Model* $\mathcal{M}_1$ was solely selected as best model based on the highest BIC weight. Figure 27b illustrates that all BIC weights of $\mathcal{M}_3$ and $\mathcal{M}_4$ are decreased whereas those of $\mathcal{M}_1$ are increased in comparison to the individual AIC weights in Figure 26b.

Artificial dataset $\mathcal{D}_6$

The *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ generated the data of dataset $\mathcal{D}_6$ as shown by Figure 28 and explained in Chapter 4.2.1. Each model of the set was fitted to the data and the AIC and BIC weights for each fit were calculated. The model selection based on AIC population weights as shown in Figure 11a yielded a ranking as follows: $\mathcal{M}_6$, $\mathcal{M}_5$, $\mathcal{M}_3$, $\mathcal{M}_4$, $\mathcal{M}_2$, $\mathcal{M}_1$. The order based on BIC population weights differs only in an exchange of $\mathcal{M}_5$ by $\mathcal{M}_3$ as shown by Figure 12a.

In agreement with the highest ranked model on the population scale the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ was selected most frequently for the individual fits based on the highest AIC weights as shown in Figure 29a. On the contrary the *Ribosome-Model* $\mathcal{M}_3$ was selected more frequently as best model based on the highest BIC weights than $\mathcal{M}_6$ as shown by Figure 30a. The AIC and BIC weights of both models are widely distributed in comparison to the other models in the set as displayed in Figure 29b and 30b. The BIC tends to select less complex models than AIC, as the penalty term for the model complexity includes the number of observation points per trajectory. For this reason the frequency of $\mathcal{M}_3$ as best model is increased in Figure 30a whereas the frequency of the more complex model $\mathcal{M}_6$ is reduced. Even though $\mathcal{M}_3$ was selected most frequently it was not highest ranked by BIC on the population scale.

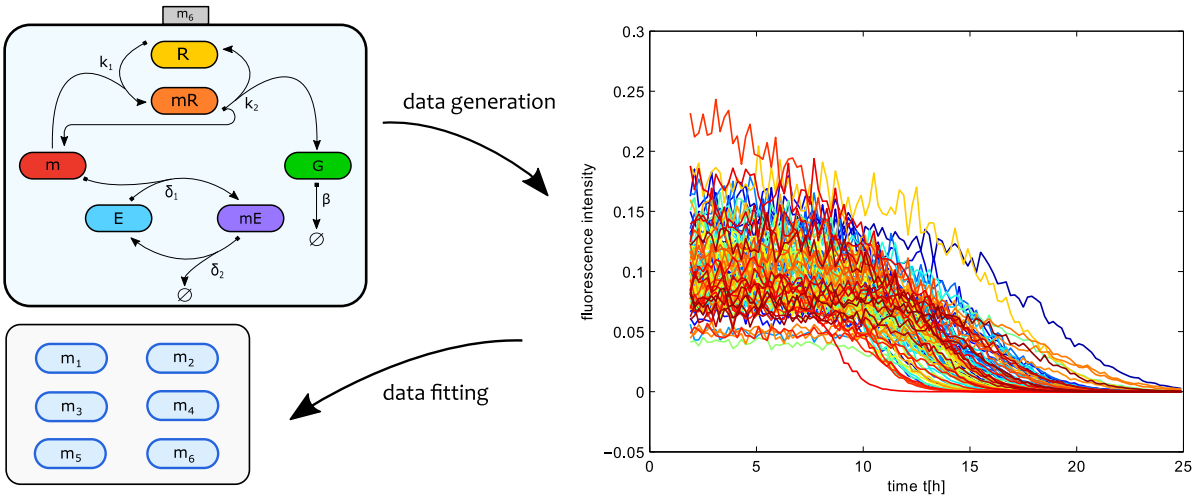
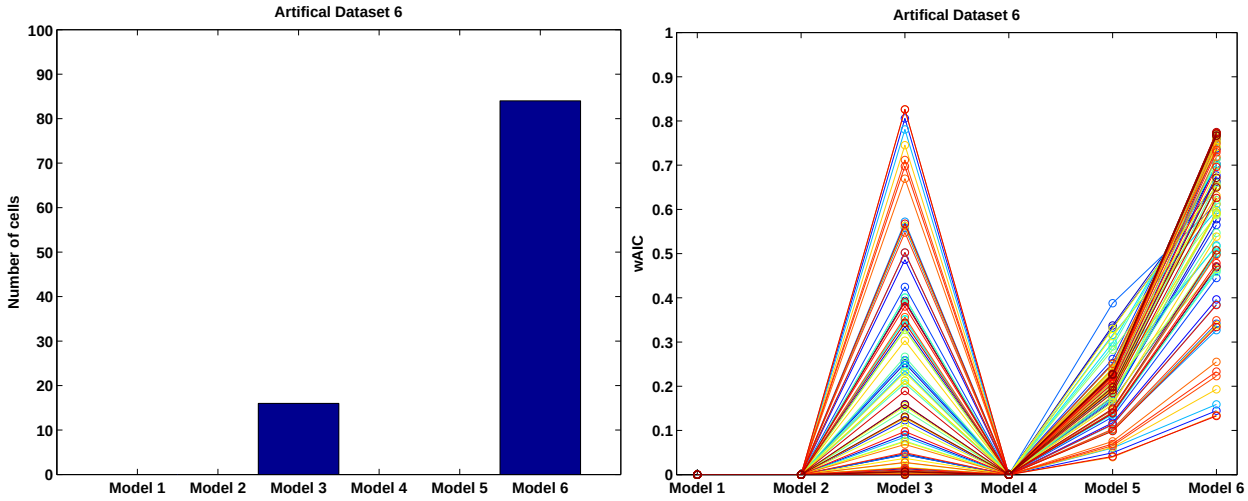
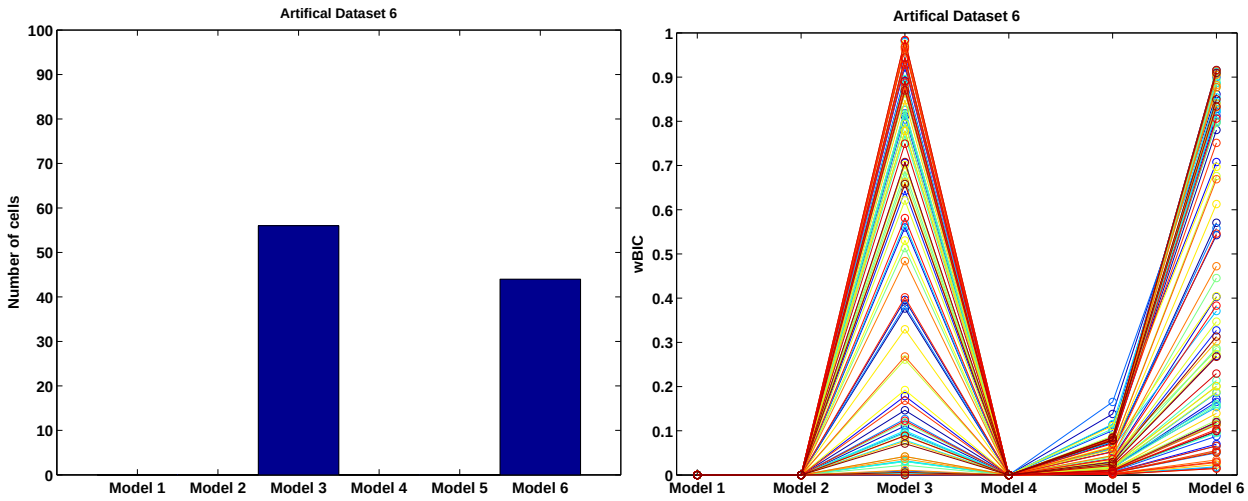


Figure 28: Artificial dataset $\mathcal{D}_6$: The model $\mathcal{M}_6$ was used to generate the data with logarithmic parameters of $\log_{10}(\theta) = \log_{10}(m_0, R_0, E_0, \delta_1, \delta_2, \beta, k_1, k_2, t_0, sc) = (2.5, 1.4, 2.4, -2.1, -1.2, -0.4, 2.5, 2.7, 0, 0)$. As explained in Chapter 4.2.2 each model in the set $M = \{\mathcal{M}_i\}_{i=1}^{6}$ was used once to fit the trajectories. This way the AIC and BIC values for each model per individual fit were calculated.



(a) Frequency of selected best models (b) Individual AIC weights

Figure 29: Model selection based on AIC weights: Figure 29a shows that model $\mathcal{M}_6$ was selected most frequently as best model based on the highest AIC weights followed by $\mathcal{M}_3$. In Figure 29b the models $\mathcal{M}_3, \mathcal{M}_6$ and $\mathcal{M}_5$ exhibit AIC weights above zero in contrast to $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_4$. The weights for most of the individual fits are given in an order that is equal to the first three ranks on the population scale by AIC as follows: $\mathcal{M}_6, \mathcal{M}_5$ and $\mathcal{M}_3$.



(a) Frequency of selected best models (b) Individual BIC weights

Figure 30: Model selection based on the BIC weights: As shown by Figure 30a model $\mathcal{M}_3$ was selected most frequently as best model followed by $\mathcal{M}_6$. In comparison to the selection based on AIC the individual fits exhibit higher BIC weights for $\mathcal{M}_3$ and lower weights for $\mathcal{M}_5$ as shown by Figure 30b.

4.3 Conclusion

The model selection on the population scale reveals that less complex models were favoured as best models for the artificial datasets and not the generating models. By considering the model selection for individual fits, several subgroups were found within the population that favoured different models in comparison to the highest ranked models on the population scale. For example the model $\mathcal{M}_3$ was highest ranked for the artificial dataset $\mathcal{D}_4$ even though the generating model $\mathcal{M}_4$ has the same complexity as $\mathcal{M}_3$. By considering the model selection of the individual trajectory fits, the more complex model $\mathcal{M}_6$ was selected most frequently as best model.

On the one hand a part of the artificial data was removed to be in accordance with the experimental data but on the other hand the model selection could perhaps lead to different results given a complete dataset. In addition the AIC and BIC information criteria are based on the assumption that the size of the observed data points is large in the sense of an asymptotic limit. Burnham et al. demonstrated the different target models and philosophies of AIC and BIC [38]. With increasing sample size, the complexity of the model that is selected by AIC increases. Its target model is the best approximating model with lowest discrepancy to the truth in the Kullback-Leibler information. In contrast to BIC the AIC is not consistent. For the latter the probability that the quasi-true model is selected converges to one for an increasing sample size. The target model is in addition not assumed to be the data generating model, as it is defined as the model that is most parsimonious. In practice the sample size is smaller than theoretically required and the BIC selects less complex models than its target model. This property of BIC may explains why the BIC selects less complex models than the generating models. Furthermore identifiable model parameters are assumed. In this work the artificial data consists of 116 data points and the identifiability of model parameters is not ensured. However, AIC and BIC are commonly used in practice and enable a comparison of all models within the set at a glance.

5 Model selection for single-cell mRNA transfection

This chapter is organized as follows: After fitting each model of the set on the datasets of the cells that either expressed eGFP or d2eGFP proteins, the model selection is studied on the population scale in Chapter 5.2.1 to obtain a ranking of models. In Chapter 5.2.2 the model selection is done for each individual fit to obtain a complete picture of favoured models within the two populations. For the two highest ranked models the parameter distributions are analysed in Chapter 5.3.2 for the cells that expressed eGFP and in Chapter 5.3.3 for the cells that expressed d2eGFP proteins. To find out whether the parameters of the two highest ranked models could be practical identifiable, parameter uncertainty is studied for one individual fit per population using the profile likelihood method and the local approximation of the negative log-likelihood function as explained in Chapter 3.2.3. The parameters that are indicated as practical identifiable are compared for the highest ranked models for each dataset separately in chapter 5.3.2 and 5.3.3. In the last Chapter 5.3.4 the protein degradation rates of the two populations are compared to each other.

5.1 Experimental methods

In this work HEK-293 cells that show high transfection efficiencies were seeded on microstructured single-cell arrays. A multi-channel slide was used that consists of six microstructured channels, to study the protein expression dynamics for different mRNA constructs in parallel. Furthermore the fluorescence microscope was equipped with a heating chamber to provide optimal temperature conditions of approximately 37 °C for the long-term observation of cells with a duration of almost 30 h. However, temperature gradients have been observed at the edges of the multi-channel slide. For this reason the first and the sixth channel were excluded from the experiment. After seeding and growing of the cells on the single cell-arrays they were transfected with mRNA using *Lipofectamin2000* ®. The transfection medium with mRNA coding for eGFP was filled in the second and third channel whereas the medium with mRNA coding for d2eGFP was added to the fourth and fifth channel. A protocol of the transfection process can be found in the Appendix 7.2.

5.2 Model selection

In this chapter the six models that are introduced in Chapter 4.1 are compared in their suitability to fit the experimental data that was gathered by the mRNA transfection using eGFP and d2eGFP as reporter proteins. To study whether a single model can describe the underlying dynamics of mRNA transfection for both datasets, the model selection was done on the scale of populations as explained in Chapter 4.2.3. The models were ranked for each dataset in decreasing order of their probability to be the best model in the set. For this purpose the Akaike (AIC) and Bayesian (BIC) Information Criteria were used that are explained in Chapter 3.2.4. This way the model architectures of highly ranked models of both datasets can be compared to find common pathways that are independent of the used mRNA coding sequence. The results are described in Chapter 5.2.1.
Cell-to-cell variability and inhomogeneous culture conditions as for example temperature gradients or variable nutrient concentrations could lead to variations of best models within a population. As shown in Chapter 4.2.4 this inhomogeneity in model selection can be studied by considering the best models for the individual trajectory fits. The two populations that either expressed eGFP or d2eGFP proteins are analysed this way in Chapter 5.2.2.

5.2.1 Model selection for populations

For each population that either expresses eGFP or d2eGFP proteins the ranking was obtained by sorting the AIC and BIC population weights in decreasing order. The weights are related to the probability of being the best model in the set in the sense of AIC or BIC. They were calculated in the same manner as for the artificial datasets which is described in Chapter 4.2.3. The rankings are displayed in Figure 31 in colours from red brown to white as shown by the colour-bar. Additionally the population weights are illustrated in colours in the range from dark green to yellow in steps of 0.16 intervals as depicted by the second colour-bar.

The population that was transfected with mRNA coding for eGFP includes 236 cells. The ranking for this population is displayed in Figure 31 and reveals that the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ is selected as best model by AIC and the *Ribosome-Model* $\mathcal{M}_3$ by BIC. Both models include a ribosomal translation of mRNA into proteins and a direct degradation of reporter proteins. Model $\mathcal{M}_6$ is more complex than $\mathcal{M}_3$ as the mRNA degradation pathway involves enzymatic binding and release whereas model $\mathcal{M}_3$ assumes a direct degradation of mRNA. As expected the BIC selects a model of lower complexity than the AIC.

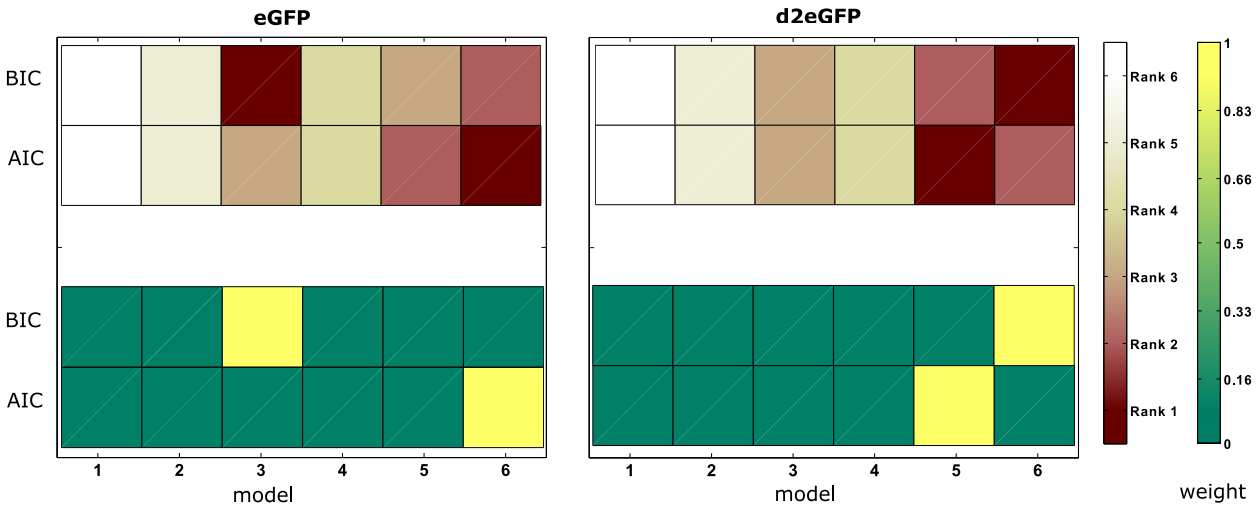


Figure 31: Model selection on the population scale: The ranking of models is represented in colors from red brown to white for each population and model selection criterion separately. The associated ranking positions are shown by the colour-bar. Additionally the population weights are illustrated in colours from dark green to yellow in interval steps of 0.16 which are displayed in the second colour-bar.

The population weights of all models except of $\mathcal{M}_3$ and $\mathcal{M}_6$ are below 0.16 which indicates that they were less able to describe the underlying dynamics on the population scale. The population of 394 cells that expressed d2eGFP proteins selects the model $\mathcal{M}_5$ by AIC and model $\mathcal{M}_6$ by BIC. These models are the most complex models in the set with the highest number of model parameters. They include a ribosomal translation and enzymatic degradation of mRNA and differ only by an additional pathway of direct mRNA degradation. All other models exhibit population weights beneath a value of 0.16 as illustrated in dark green.

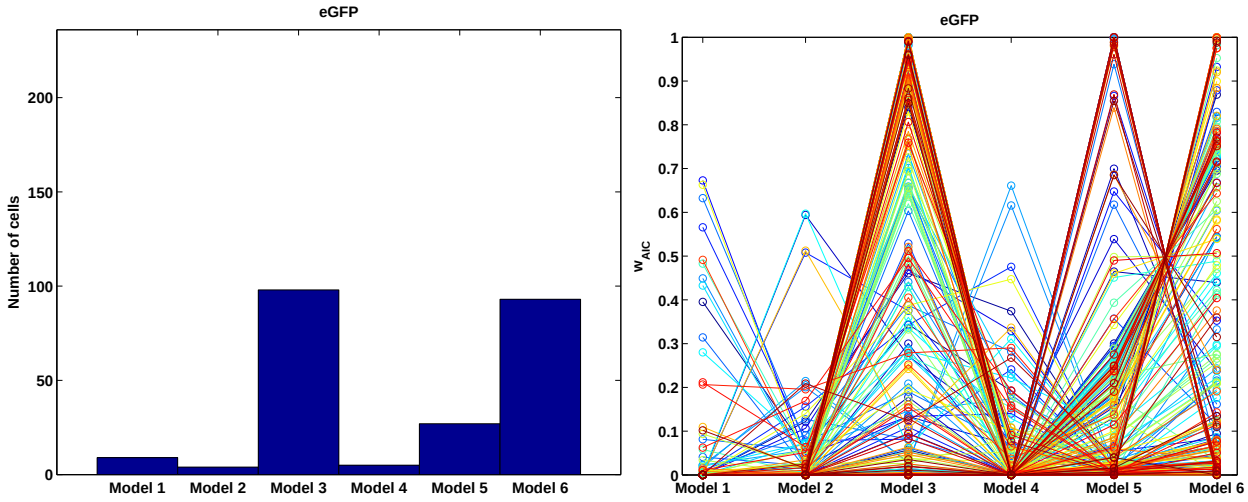
Even though the first ranking positions of both population differ and depend on the model selection criterion, they have in common that model $\mathcal{M}_3$, $\mathcal{M}_5$ and $\mathcal{M}_6$ are placed on the first three ranking positions. They share the pathway of ribosomal translation into mRNA and the direct degradation of reporter proteins and only differ in their mRNA degradation pathways. In contrast the *Trivial-Model* $\mathcal{M}_1$ with the fewest number of parameters is ranked on the last position. According to this result even more complex models could be designed and compared with $\mathcal{M}_3$, $\mathcal{M}_6$ and $\mathcal{M}_5$ to find a model that is even more suitable to describe the underlying dynamics. On the other hand the quality of the data has a crucial influence on the model selection. For this reason all previous steps including the observation via time-lapse fluorescence microscopy and image analysis should be improved before increasing the complexity of considered models in the set.

5.2.2 Individual model selection

In the following the individual model selection is analysed for both populations separately. To address the question whether these populations incorporates subgroups that favour models that are different from the highest ranked model on the population scale, the AIC and BIC weights were calculated for each fit separately. The best model for an individual fit is given by the highest AIC or BIC weight.

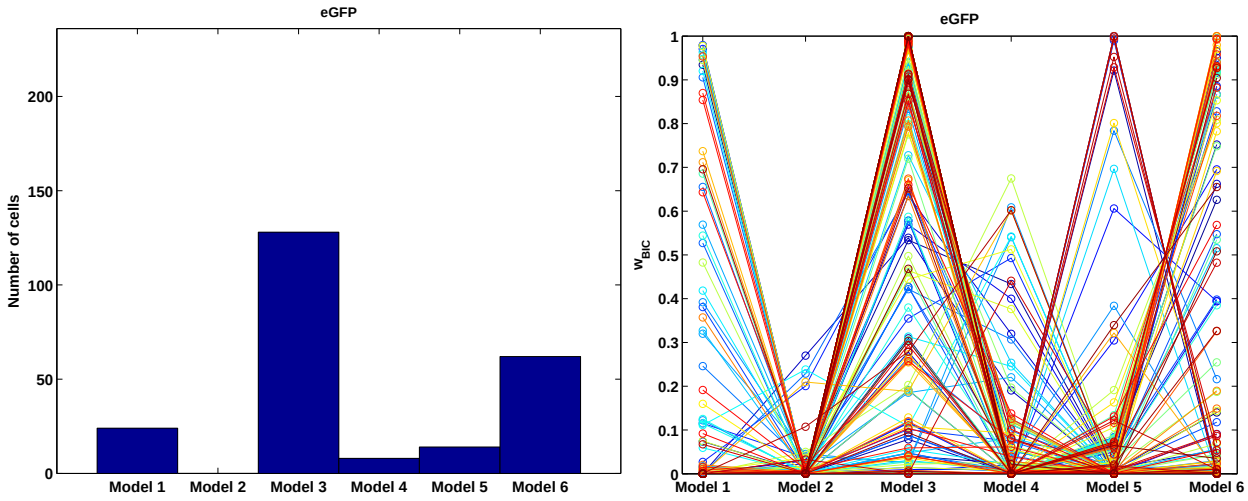
Figure 32a shows the frequencies of best models for the individual fits given by the highest AIC weight. The individual trajectory fits correspond to the population of 236 cells that expressed eGFP proteins. Interestingly model $\mathcal{M}_3$ is selected more frequently than $\mathcal{M}_6$ which is highest ranked on the first position by AIC population weights in Figure 31. For this reason the sum of AIC values of $\mathcal{M}_3$ has to be higher than for $\mathcal{M}_5$ and $\mathcal{M}_6$. Figure 32b shows the individual AIC weights for each model in detail. The weights are represented by empty circles for each colour-marked trajectory fit. At least three substructures can be identified. The first group shows high weights for the model $\mathcal{M}_6$, low values for $\mathcal{M}_5$ and nearly zero values for all other models. The second group exhibits high weights for $\mathcal{M}_5$ and values close to zero for all other models and the third group shows high weights for $\mathcal{M}_3$ and low values for all other models.
As shown by Figure 33a the models $\mathcal{M}_3$, $\mathcal{M}_6$ and $\mathcal{M}_5$ are frequently selected as best models by BIC with frequencies in decreasing order. This order reflects the ranking of models by BIC on the population scale in Figure 31. As expected the weights of less complex models are increased by BIC in comparison to AIC as shown by Figure 33b.

In conclusion the model selection of individuals reveals that the model that is selected most frequently as best for the individual fits can differ from the highest ranked model of the population. On the population scale all AIC or BIC values contribute to the population weights and therefore to the ranking. On the contrary the best models that are selected frequently for the individual fits are solely based on the highest weights of AIC or BIC respectively. For this reason both methods can yield different results. For the population of cells that expressed eGFP proteins the *Ribosome-Model* $\mathcal{M}_3$ and the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ are highest ranked on the population scale and selected most frequently as best models for the individual fits. In Chapter 5.3.2 they are used to analyse the parameter estimates for the population of 236 single cells that expressed eGFP proteins.



(a) Frequency of selected best models (b) Individual AIC weights

Figure 32: Model selection by AIC weights for eGFP constructs: Figure 32a shows the frequency of selected best models based on the highest AIC weight for 236 cells. Figure 32b displays all AIC weights of the colour-marked individual fits as empty circles. Model $\mathcal{M}_3, \mathcal{M}_6$ and $\mathcal{M}_5$ are selected most frequently as best model and exhibit the highest AIC weights.



(a) Frequency of selected best models (b) Individual BIC weights

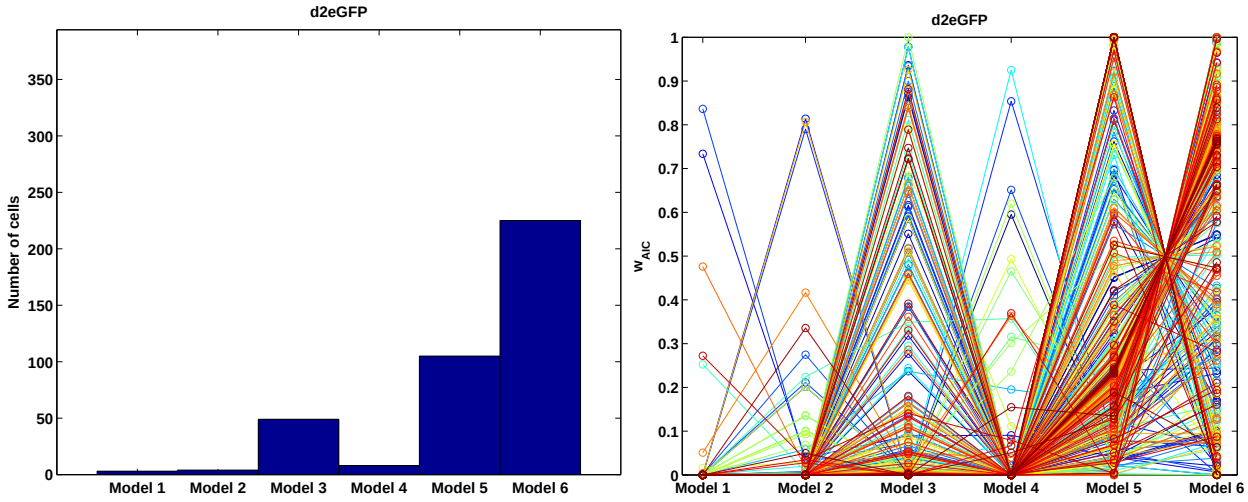
Figure 33: Model selection based on BIC weights for eGFP constructs: The models $\mathcal{M}_3$ and $\mathcal{M}_6$ are selected most frequently as best models based on the highest BIC weights as shown by Figure 33a. Three substructures can be found in Figure 33b that show high weights for $\mathcal{M}_3$, $\mathcal{M}_6$ or $\mathcal{M}_5$ and low values for all other models.

Figure 34a shows the frequencies of best models based on the highest AIC weight in the population of 394 cells that expressed d2eGFP proteins. Model $\mathcal{M}_6$ is selected most frequently followed by model $\mathcal{M}_5$ and $\mathcal{M}_3$. They are ranked at the first three positions on the population scale in the same order. Figure 34b shows the same substructure of individual AIC weights as the individuals of the population that expressed eGFP proteins. Only the tendency to select model $\mathcal{M}_6$ is even more emphasized for individuals that expressed d2eGFP in comparison to those that expressed eGFP proteins.
In Figure 35a most of the individuals select $\mathcal{M}_6$ as best model based on their highest BIC weights followed by $\mathcal{M}_3$ and $\mathcal{M}_5$. Oppositely $\mathcal{M}_5$ is higher ranked than $\mathcal{M}_3$ on the population level by BIC in Figure 31. For this reason the sum of BIC values which was used to obtain the ranking has to be lower for $\mathcal{M}_5$ than for $\mathcal{M}_3$. Indeed, the weights of $\mathcal{M}_5$ are more widely distributed over the entire range from 0 to 1 in comparison to $\mathcal{M}_3$ as shown by Figure 35b. However, more individuals exhibit highest weights for $\mathcal{M}_3$ than for $\mathcal{M}_5$ which explains the contradiction to the population ranking.

Summarized the most complex models in the set are able to fit the experimental data best that was gathered from the population of cells that expressed d2eGFP proteins. The *Ribosome-Model* $\mathcal{M}_3$, the *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ and the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ are selected most frequently as best models for the individual fits. In Chapter 5.3.3 the parameter estimates are analysed for the highest ranked models $\mathcal{M}_5$ and $\mathcal{M}_6$ for the population of 394 cells that expressed d2eGFP proteins.

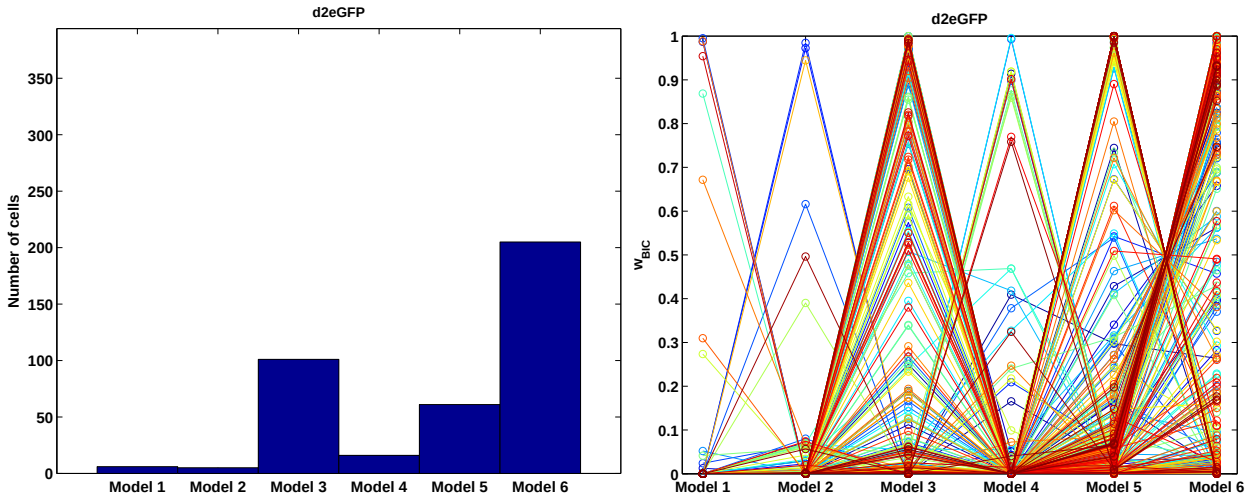
5.2.3 Conclusion

The model selection by AIC and BIC population weights reveals that more complex models than the *Trivial-Model* $\mathcal{M}_1$ are selected as best models for both populations. The *Enzym-Degradation-Ribosome-Model* $\mathcal{M}_6$ is highest ranked by AIC and the *Ribosome-Model* $\mathcal{M}_3$ by BIC for the population of 236 cells that expressed eGFP proteins. On the other hand the *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ was highest ranked by AIC and the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ by BIC for the population of 236 cells that expressed d2eGFP proteins.
The model selection reveals that the models that were most frequently chosen as best for the individual fits can differ from the highest ranked models on the population scale. This is caused by the different approaches of calculating the associated weights. By taking into account the results of both methods it was determined that the models $\mathcal{M}_3$, $\mathcal{M}_5$ and $\mathcal{M}_6$ are best suited to describe the dynamics of mRNA transfection. They share the ribosomal translation of mRNA and the simple degradation of fluorescent proteins and differ only in the complexity of the mRNA degradation pathways.



(a) Frequency of selected best models (b) Individual AIC weights

Figure 34: Model selection based on AIC weights for d2eGFP constructs: As shown by the histogram 34a model $\mathcal{M}_6$ is selected most frequently as best model for the 394 individual fits, followed by $\mathcal{M}_5$ and $\mathcal{M}_3$. The weights are shown in detail in Figure 34b. The majority of individuals exhibits high weights for $\mathcal{M}_6$ in the range from 0.7 to 1 and low values for $\mathcal{M}_5$ around 0.2.



(a) Frequency of selected best models (b) Individual BIC weights

Figure 35: Model selection based on BIC weights for d2eGFP constructs: As shown by Figure 35a the models $\mathcal{M}_6$, $\mathcal{M}_3$ and $\mathcal{M}_5$ are selected most frequently as best models for the individual fits based on the highest BIC weight. The weights are shown in detail in Figure 35b. At least three substructures can be found that exhibit high weights for $\mathcal{M}_6$, $\mathcal{M}_3$ or $\mathcal{M}_5$.

5.3 Parameter estimation

In this chapter the parameter distributions are analysed for each highest ranked model separately. Parameter uncertainty analysis is used for one fit of a single cell trajectory to find the model parameters that could be practical identifiable. In Chapter 5.3.2 the analysis is done for the population of cells that expressed eGFP proteins and in Chapter 5.3.3 for those that expressed d2eGFP proteins. First, the *Trivial-Model* is analysed as a case study in Chapter 5.3.1 to emphasize the importance of parameter uncertainty. Finally, the degradation rates of the mRNA constructs that either code for eGFP or d2eGFP proteins are compared in Chapter 5.3.4.

5.3.1 The Trivial-Model as a case study

In the following the parameter estimates of the *Trivial-Model* $\mathcal{M}_1$ that was fitted on the trajectories of the cells that expressed eGFP proteins are examined to point out the importance of parameter uncertainty analysis. The model $\mathcal{M}_1$ is the least complex model in the set as it only includes direct reaction pathways. The mRNA is translated into proteins with the rate k_{TL} and degrades with the rate δ. The proteins degrade directly with the rate β as shown by Figure 8a. In this case study the initial concentration of transfected mRNA m_0 was unknown and estimated as an additional parameter by global optimization of the likelihood function as explained in Chapter 3.2.2. The rate equations of $\mathcal{M}_1$ can be solved analytically and the solutions $m(t)$ for the mRNA and $G(t)$ for the reporter proteins are given by the following equations:

$$m(t) = m_0 \cdot e^{-\delta \cdot (t - t(0))} \tag{36}$$

$$G(t) = \frac{k_{TL} \cdot m_0}{\delta - \beta} \cdot e^{-\beta \cdot (t - t(0))} \cdot \left(1 - e^{-(\delta - \beta) \cdot (t - t(0))}\right) \tag{37}$$

Unfortunately the rates δ and β can be interchanged without leading to different values of the solution $G(t)$ of the reporter proteins which is used as the model observable y for the global optimization of the likelihood function. Both parameters are not distinguishable which is caused by the model structure. For this reason they are called structural non-identifiable. This problem is not visible in the parameter distributions that are shown by Figure 36 on the logarithmic scale to base 10. The degradation rates $log_{10}(\delta)$ and $\log_{10}(\beta)$ are both densely distributed within the range from 0 to -2.

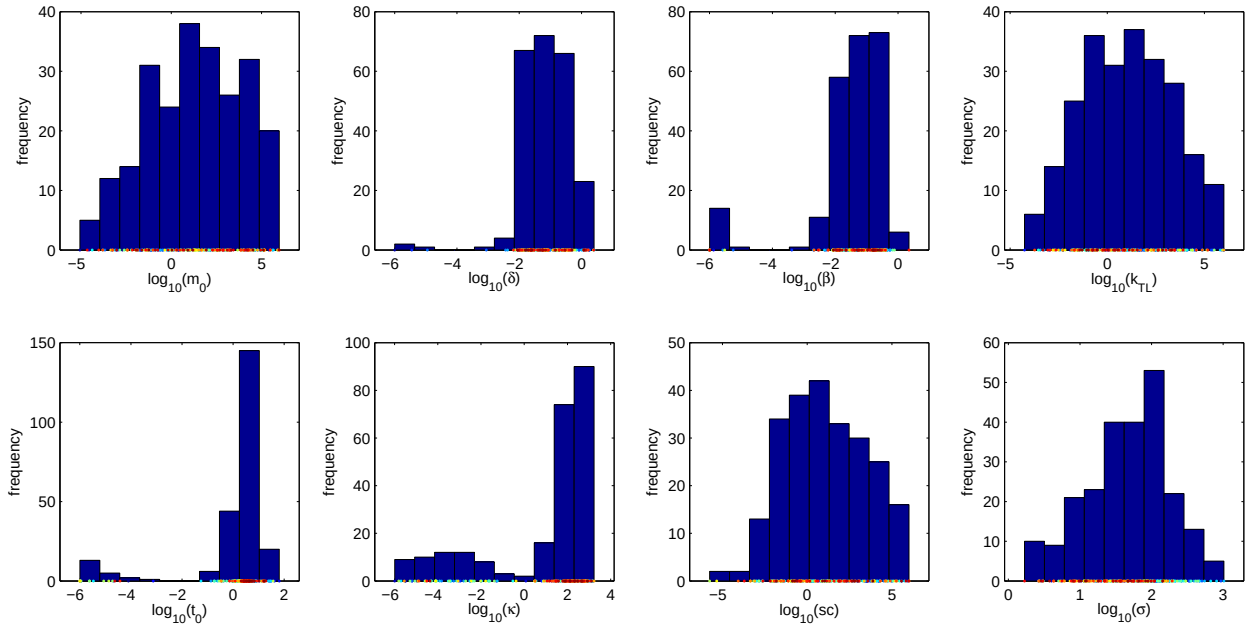


Figure 36: Parameter estimates of model $\mathcal{M}_1$ without outliers: The degradation rates δ of transfected mRNA and β for eGFP proteins are densely distributed whereas the translation rate k_{TL} and the initial concentration of mRNA m_0 spread over a wide range of the predefined parameter interval from -6 to 5.

As explained in chapter 3.2.3 two methods are chosen in this work to study parameter uncertainty: The profile likelihood analysis and the local approximation of the objective function $J(\theta)$ based on the Hessian matrix. The profile likelihood analysis leads to more reliable results for non-linear rate equations [37]. For all models that are discussed in this work, the parameter uncertainty analysis was done for one individual fit of a single-cell trajectory within a population. In this case the population consists of cells that expressed eGFP proteins. This fit exhibits the same maximum of the likelihood function for at least 10 multistarts in comparison to the 150 multistarts. This is illustrated in Figure 37 by the Subfigure in the the bottom left: The log-likelihood function shows the same value of $log(L_D(\theta)) = -991.1$ for ten multistarts. As shown by the Subfigure in the top left several lower valued maxima were found by the global optimization of the likelihood function that correspond to local maxima in the parameter space. Each maximum is related to one combination of model parameters of the *Trivial-Model* $\mathcal{M}_1$ as shown by the Subfigure on the right-hand side. The red line corresponds to the combination of parameters that yield the best, probably global maximum of the likelihood function with a value of $log(L_D(\theta)) = -991.1$. The black lines represent combinations that lead to local maxima that are close to the highest value. With increasing greying of the lines, the parameters correspond to sub-optimal values of the maximum likelihood function.

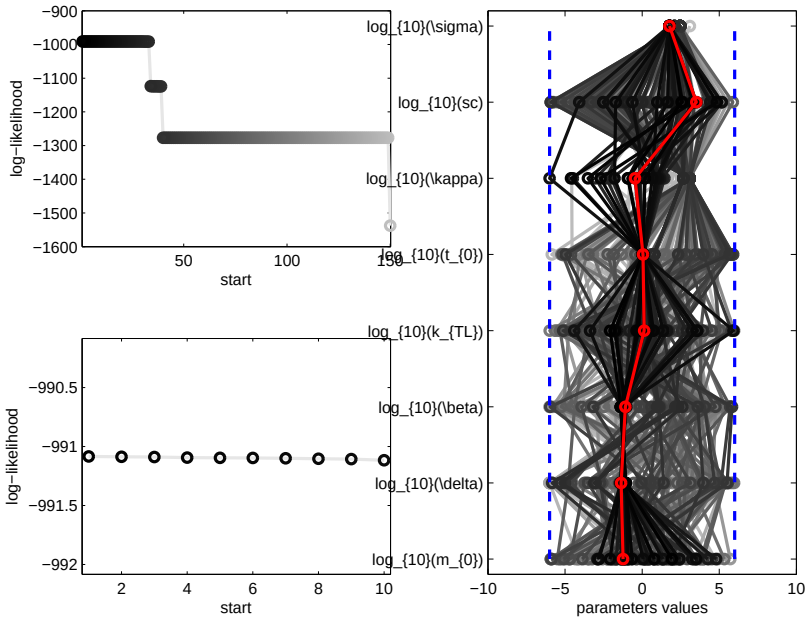


Figure 37: Parameter uncertainty analysis of model $\mathcal{M}_1$ for one individual fit: As shown by the Subfigure in the bottom left a maximum of the likelihood function with a value of $log(L_D(\theta)) = -991.1$ was found for at least 10 multistarts. Several local maxima were found by the global optimization that exhibit lower values as shown by the Subfigure in the top left. The red line in the Subfigure on the right-hand side corresponds to a parameter combination of the *Trivial-Model* $\mathcal{M}_1$ that yielded probably the global maximum of the likelihood function. The lines coloured from dark to grey are related to parameters of lower valued, local maxima of the likelihood function.

The uncertainty of the parameters of the *Trivial-Model* that correspond to the red line in Figure 37 was studied by the profile likelihood analysis and by the second-order local approximation of the negative log-likelihood function $J(\theta)$. The profile rations are displayed by the red lines and the local approximations are shown by the blue lines in Figure in 38. The profile rations are calculated as explained in chapter 3.2.3. A parameter is practical identifiable if the ration values are below the threshold for increasing and decreasing values of the profile interval. The threshold is given by the black, dashed line in Figure 38 and is related to the 95 % confidence interval. The red circles in Figure 38 correspond to all optima that have been found by the global optimization of $\mathcal{M}_1$ using the multi-start method that are within the 95 % confidence interval. In Figure 38 the profile rations are bounded for the degradation rates δ and β as well as for the measurement noise σ and the expression onset parameter t_0, which indicates that they are practical identifiable. On the other hand one can observe two different optima for δ and β at values close to -1.4 and -1.1 that are identical for both parameters. Additionally the arrangements of the red-circled optima that have been found to be within the 95% confidence interval are completely symmetric and located at these optima. To obtain a complete picture of the likelihood function one could perform the profile likelihood analysis

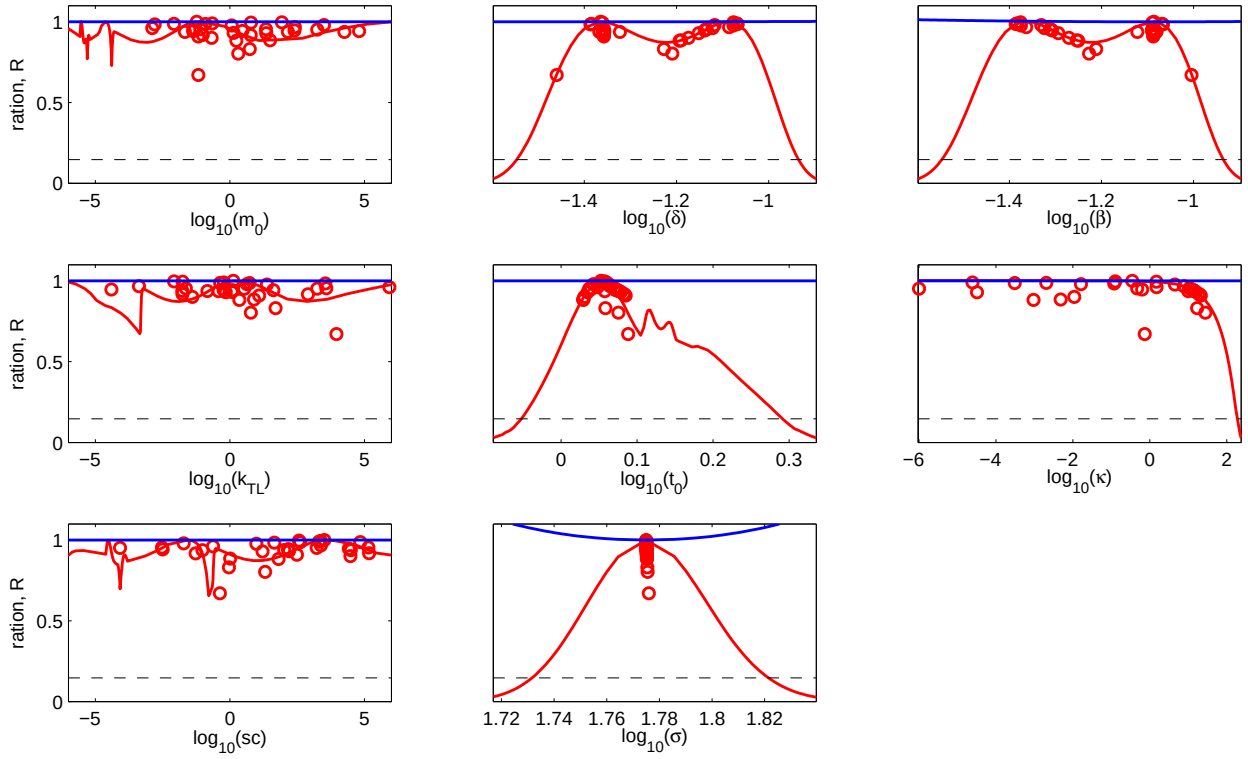


Figure 38: Parameter uncertainty analysis of model $\mathcal{M}_1$ for one individual fit: The profile rations are illustrated by the red lines whereas the local approximation of the objective function is shown by the blue lines. The black, dashed threshold is related to the 95 % confidence interval. Practical identifiability is indicated by bounded lines with increasing and decreasing parameter values.

for each of the red-circled optima within the confidence interval. The symmetry in the profile rations of δ and β reflects the structural non-identifiability of both parameters caused by their interchangeability which was explained above. However, the local approximation of the likelihood function displayed in blue only yields that both parameters are practical non-identifiable. This result further emphasizes that the profile likelihood analysis yields a deeper insight into the structure of the likelihood function in the parameter space than the local approximation. Both methods indicate practical non-identifiability for all model parameters except for δ, β, σ and t_0 as shown by the unbounded red and blue lines in Figure 38.

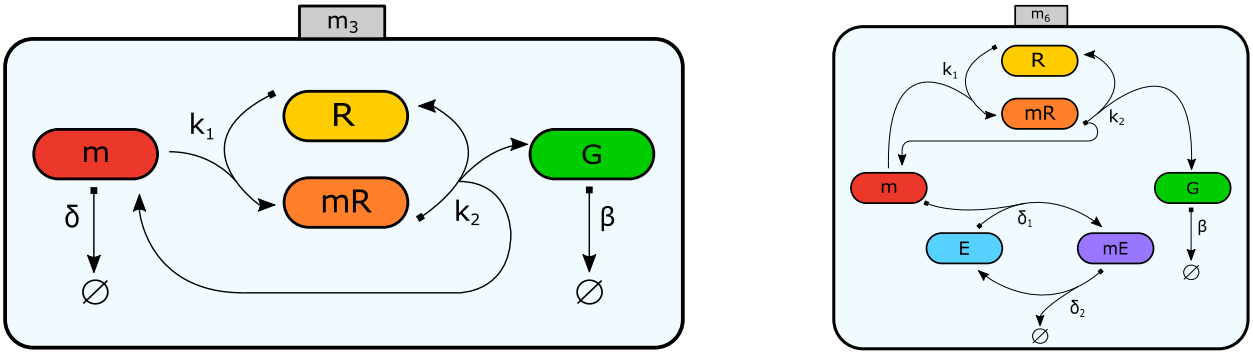
In conclusion it was shown by this case study that parameter uncertainty analysis is important to obtain reliable parameter estimates. The profile likelihood analysis and the local approximation of the objective function for one fit yield an impression of practical identifiability of the model parameters of the *Trivial-Model* which can not be obtained by solely considering the parameter distributions and the associated statistical quantities like the mean and standard deviation.

5.3.2 Parameter estimation for eGFP constructs

The models $\mathcal{M}_3$ and $\mathcal{M}_6$ were highly ranked on the population scale by AIC as well as BIC and most frequently selected for the individual trajectory fits as described in chapter 5.2. They share the rates of ribosomal binding and release, denoted as k_1 and k_2, as well as the degradation rate β of the eGFP proteins as shown by Figure 39. In this chapter the parameter histograms of both models are analysed and the question is addressed whether their common model parameters are probably identifiable and comparable in the sense that the values of their median, mean and standard deviation are approximately equal.

Parameter estimates of the *Ribosome-Model* $\mathcal{M}_3$

The histograms of the estimated parameters of the *Ribosome-Model* $\mathcal{M}_3$ are shown by Figure 40. The degradation rate of transfected mRNA δ and of eGFP proteins β as well as the parameters for t_0 that were used to cover the different onset times of protein expression are most densely distributed. Taking into account that the parameter values are shown on the logarithmic scale to base 10, the other parameters spread widely over the logarithmic parameter interval to base 10 from -5 to 5 in several orders of magnitude. To study parameter uncertainty, the profile likelihood rations were calculated for an individual fit that yielded the highest maximum likelihood function for at least 10 multistarts as explained in Chapter 5.3.1 for the *Trivial-Model* $\mathcal{M}_1$.



(a) *Ribosome-Model* $\mathcal{M}_3$ (b) *Ribosome-Enzyme-Model* $\mathcal{M}_6$

Figure 39: Best models of the set for eGFP coding mRNA constructs: Model $\mathcal{M}_6$ was selected as best model based on AIC population weights and model $\mathcal{M}_3$ based on BIC population weights. Both models were selected most frequently with respect to the highest AIC and BIC weights of the individual fits. The models share the pathways of ribosomal binding and release with rates k_1 and k_2 as well as the pathway of eGFP degradation with the rate β.

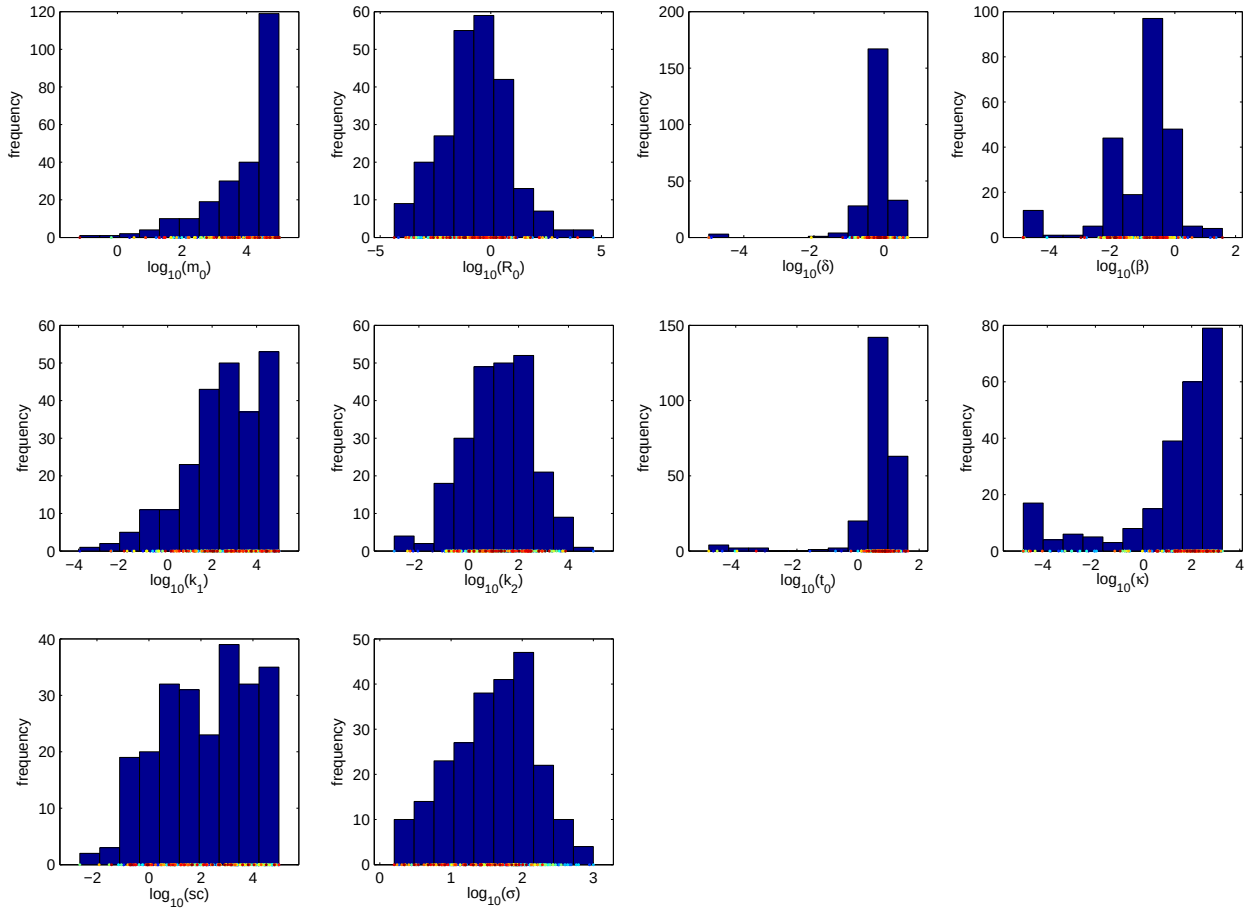


Figure 40: Parameter estimates of model $\mathcal{M}_3$ without outliers: As shown by the histograms most of the parameters are widely distributed over the parameter interval from -5 to 5 on the logarithmic scale to base 10. Only the degradation rates δ of the transfected mRNA and β of the eGFP proteins as well as the parameter t_0 that was used to cover the different onset times of protein expression are densely distributed.

The profile likelihood analysis indicates that the mRNA degradation rate δ, the eGFP degradation rate β and the measurement noise σ are practical identifiable. Their values are below the threshold with decreasing and increasing values of the profile interval as shown by the red lines in Figure 41. The threshold is given by the black, dashed line and is related to the 95 % confidence interval. The expression onset parameter t_0 could be practical identifiable as well which is not obviously given by the profile rations. In contrast the local approximation shows that only the measurement noise σ is practical identifiable as displayed by the bounded blue line in Figure 41. In contrast all other model parameters are indicated as practical non-identifiable by this method. All other parameters except for those discussed above are indicated as practical non-identifiable by the profile rations and by the local approximation.

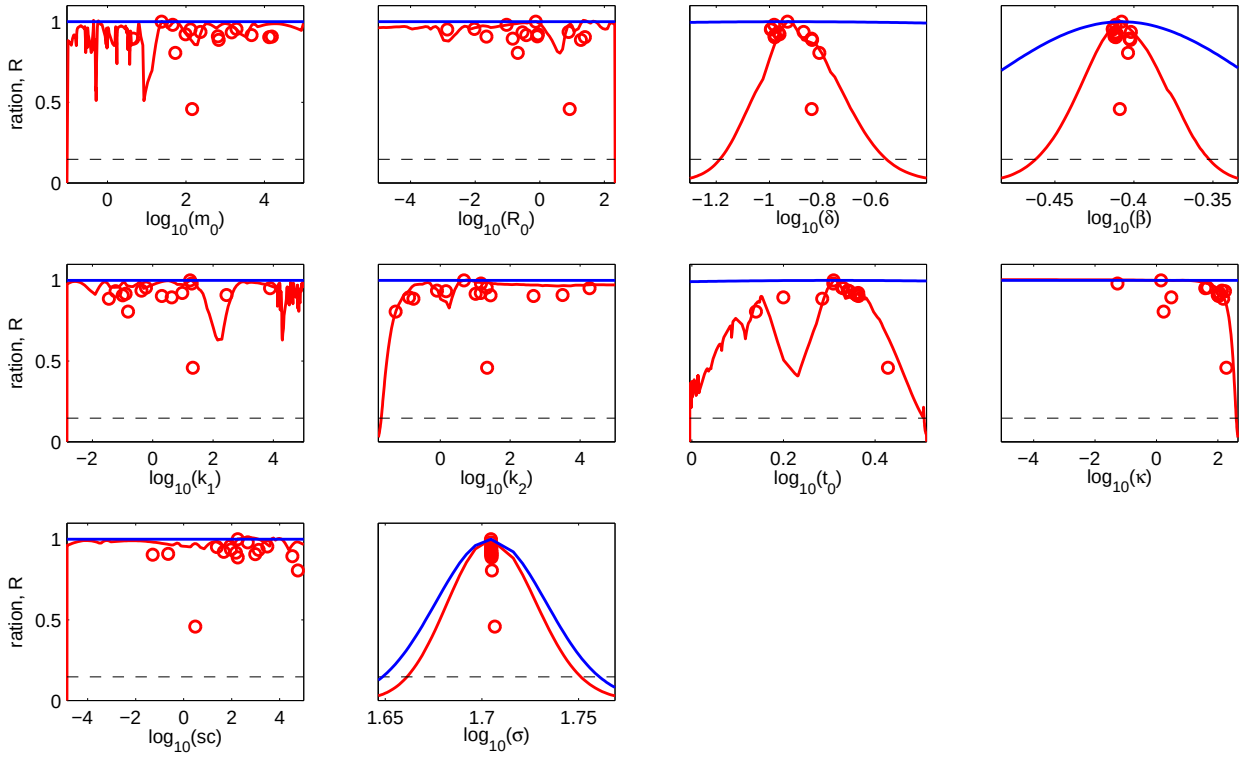


Figure 41: Parameter uncertainty analysis of model $\mathcal{M}_3$ for one individual fit: The profile rations displayed in red are bounded for the mRNA degradation rate δ, the eGFP degradation rate β and the measurement noise σ. The threshold depicted as the black dashed line is related to the 95 % confidence interval. The local approximation of the objective function displayed in blue indicates that only σ is practical identifiable. However, both methods yield practical non-identifiable for all other model parameters.

In conclusion, the *Ribosome-Model* $\mathcal{M}_3$ could have practical identifiable parameters for the individual trajectory fit for the mRNA and eGFP degradation rates, denoted as δ and β, as well as for the measurement noise σ. The model $\mathcal{M}_3$ was selected as best model by BIC population weights for the dataset that is related to the cells that expressed eGFP proteins. In the end of this chapter statistical quantities like the mean, standard deviation and median are calculated for these parameters over all single-cell trajectories of the population and are compared with the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ that was selected as best model by AIC population weights.

Parameter estimates of the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$

The *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ was selected as best model by AIC popuation weights and similar frequently as the *Ribosome-Model* $\mathcal{M}_3$ as best model for the individual fits. In contrast to $\mathcal{M}_3$ it includes an enzymatic degradation pathway of the transfected mRNA with binding and release rates denoted as δ_1 and δ_2. Figure 42 shows the histograms of parameter estimates of $\mathcal{M}_6$. The degradation rate β of eGFP proteins and the expression onset parameter t_0 are most densely distributed. However, this does not reveal that they are identifiable as explained in Chapter 5.3.1. The other parameters are widely distributed over the parameter interval form -5 to 5. The parameter uncertainty was studied for one individual fit with at least 10 multistarts that yielded the same best maximum of the likelihood function. The results for each parameter are shown by Figure 43.

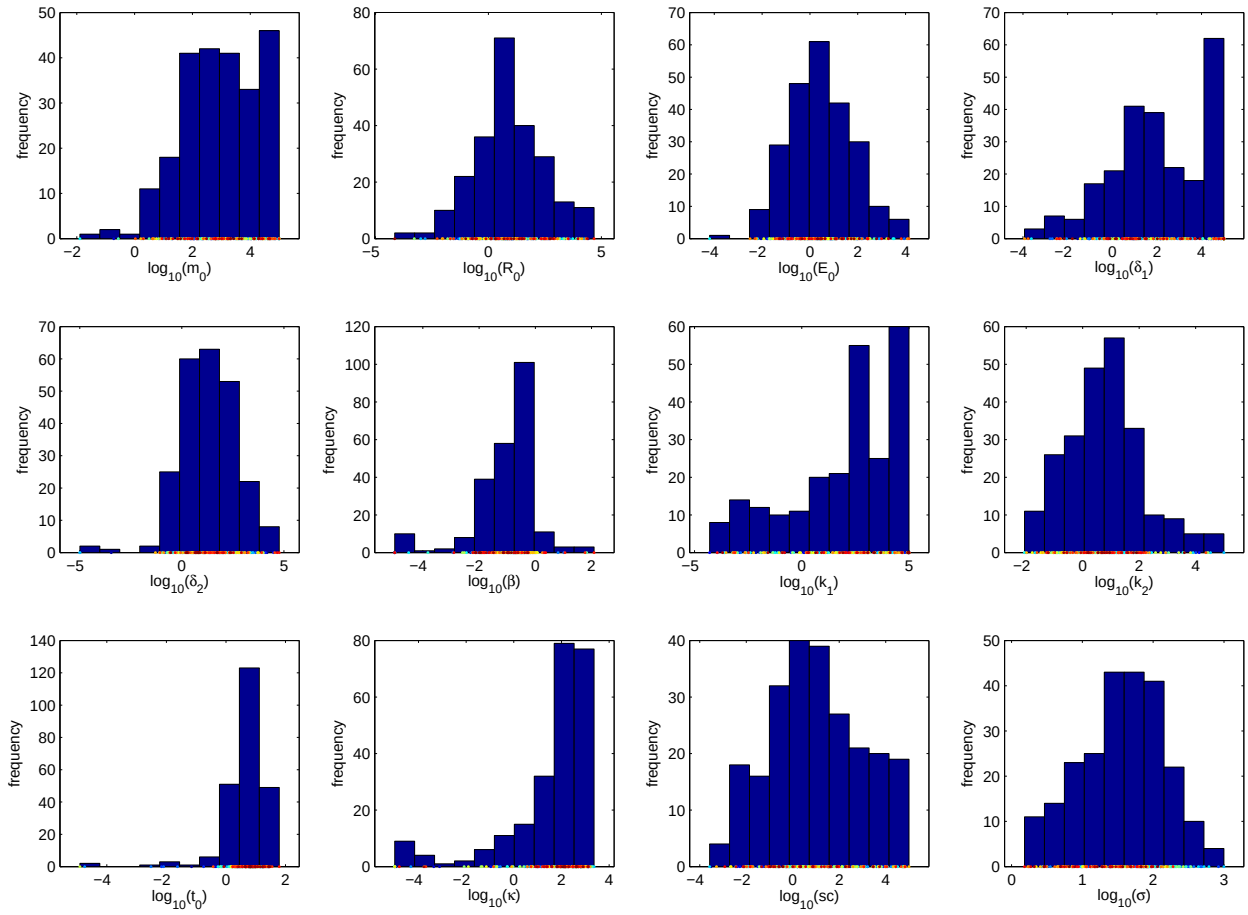


Figure 42: Parameter estimates of model $\mathcal{M}_6$ without outliers: The degradation rate β of eGFP proteins and the parameter t_0 that was used to cover the different onset times of protein expression are most densely distributed. The other model parameters are widely distributed over the parameter interval from -5 to 5 on the logarithmic scale to base 10.

The profile rations of the degradation rate β of the eGFP proteins are bounded below the threshold with decreasing and increasing parameter values. Additionally the measurement noise σ exhibits almost bounded rations as well. For this reason it was assumed that they are practical identifiable for this individual fit. The profile rations of the other model parameters show various structures and sudden steep decrements. The latter could be caused by a failure of numerical integration of the rate equations. To obtain a complete picture over the entire profile interval one has to perform the profile likelihood analysis for each of the red-circled optima above the black, dashed threshold that refers to the 95 % confidence interval. The second-order local approximation of the objective function $J(\theta)$ indicates as well that β and σ could be identifiable even though the blue lines are not bounded. However, both methods indicate practical non-identifiability for all other model parameters.

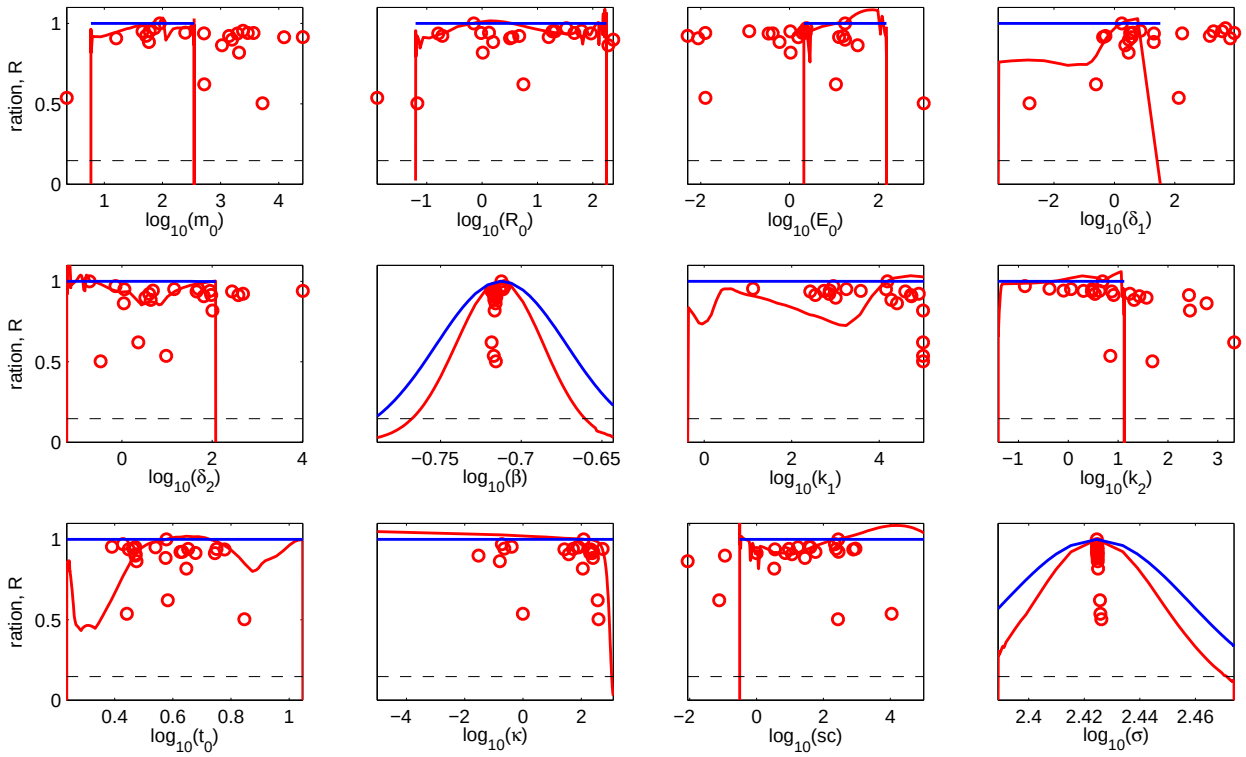


Figure 43: Parameter uncertainty analysis of model $\mathcal{M}_6$ for one individual fit: The profile likelihood analysis reveals that the degradation rate β of eGFP proteins is practical identifiable. This is confirmed by the local approximation of the objective function as the values displayed by the blue line are almost bounded. In the same manner the measurement noise σ could be practical identifiable. However, both methods yield practical non-identifiability for all other model parameters.

In conclusion the degradation rate β of the eGFP proteins and the measurement noise σ were indicated as practical identifiable for the individual fit of the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$. It was selected as best model on the population scale by AIC and frequently as best model for the individual fits. In the end of this chapter statistical quantities like the median, mean and standard deviation are calculated over the population of all single-cell fits and compared with the results of the *Ribosome-Model* $\mathcal{M}_3$.

Protein and mRNA degradation rates for eGFP constructs

The degradation rate β of eGFP proteins and the measurement noise σ are indicated as practical identifiable by the profile likelihood analysis for a single fit using the *Ribosome-Model* $\mathcal{M}_3$ and the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$. Both models were selected as best models in the set by AIC or BIC population weights and most frequently for the individual fits. In addition the degradation rate δ of the transfected mRNA was indicates as practical identifiable for an individual fit by model $\mathcal{M}_3$. On the contrary the degradation rates δ and β of the *Trivial-Model* are structural non-identifiable as they are interchangeable in the analytic solution of the proteins as explained in Chapter 5.3.1. The model $\mathcal{M}_1$ was ranked on the last position by AIC as well as BIC population weights as shown by Figure 31. The median, mean and standard deviation of the degradation rates β and δ and the measurement noise σ are summarized in Table 1.

Table 1: Median, mean and standard deviation of the degradation rate of transfected mRNA δ and eGFP proteins β and the parameter of the measurement noise σ.

rate	**median**			**mean**			**standard deviation**		
	$\mathcal{M}_1$	$\boldsymbol{\mathcal{M}_3}$	$\boldsymbol{\mathcal{M}_6}$	$\mathcal{M}_1$	$\boldsymbol{\mathcal{M}_3}$	$\boldsymbol{\mathcal{M}_6}$	$\mathcal{M}_1$	$\boldsymbol{\mathcal{M}_3}$	$\boldsymbol{\mathcal{M}_6}$
$\boldsymbol{\delta}$ $[h^{-1}]$	0.08	0.71	-	0.24	0.87	-	0.44	0.73	-
$\boldsymbol{\beta}$ $[h^{-1}]$	0.05	0.20	0.18	0.13	0.69	1.43	0.20	3.21	9.87
$\boldsymbol{\sigma}$	54.49	45.36	42.47	99.55	85.68	82.41	143.06	128.16	123.62
$\boldsymbol{\log_{10}(\delta)}$	-1.17	-0.14	-	-1.16	-0.23	-	0.82	0.64	-
$\boldsymbol{\log_{10}(\beta)}$	-1.24	-0.70	-0.74	-1.52	-1.11	-1.03	1.28	1.21	1.13
$\boldsymbol{\log_{10}(\sigma)}$	1.74	1.65	1.63	1.67	1.58	1.57	0.57	0.59	0.59

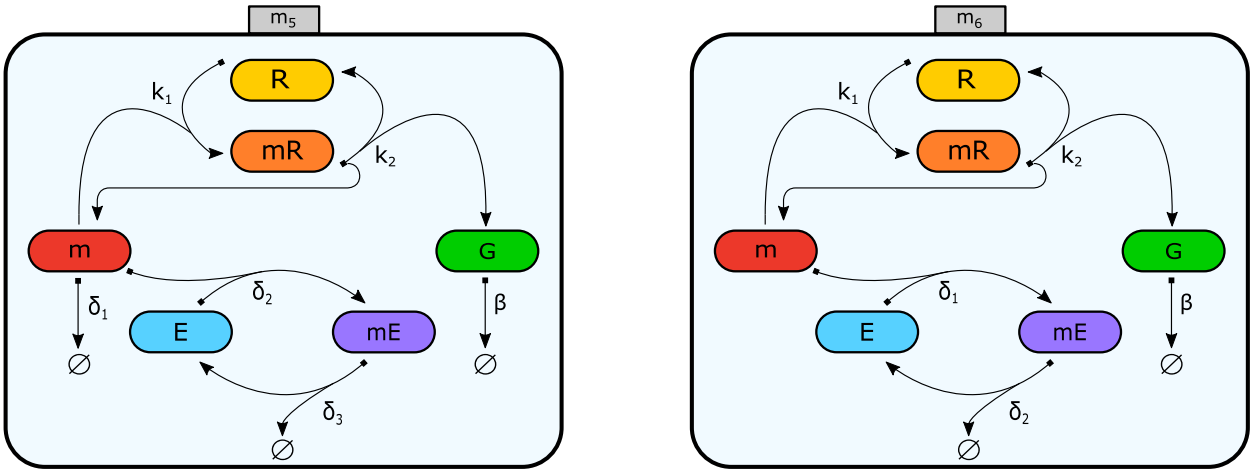
Most logarithmic parameters to base 10 were almost normally distributed as shown by the histograms in Figure 40 and 42. On the other hand the parameters themselves are log-normally distributed. For this reason the mean of the parameter distributions is shifted considerably towards higher values in comparison to the mode. In contrast the median is more robust towards outliers. As shown by Table 1 the logarithmic degradation rates β of the eGFP proteins and the logarithmic values of the measurement noise σ are similar for the highest ranked models $\mathcal{M}_3$ and $\mathcal{M}_6$. On the contrary they are only similar in their median values for the back transformed parameters. In comparison to the *Trivial-Model* $\mathcal{M}_1$ both models yield higher degradation rates δ and β but lower values for the measurement noise σ. This result emphasizes that they were more able to fit the trajectories of the population of cells that expressed eGFP proteins. Interestingly the degradation rate δ of transfected mRNA was estimated higher by model $\mathcal{M}_3$ than the protein degradation rate β which confirms the assumption that the mRNA is less stable than the proteins.

5.3.3 Parameter estimation for d2eGFP constructs

The most complex models $\mathcal{M}_5$ and $\mathcal{M}_6$ were highest ranked by AIC and BIC population weights and selected most frequently as best models for the individual fits together with $\mathcal{M}_3$ as described in Chapter 5.2. As shown by Figure 44 the models $\mathcal{M}_5$ and $\mathcal{M}_6$ share all pathways except of the additional direct mRNA degradation pathway in model $\mathcal{M}_5$. The binding and release rates of the enzymatic degradation pathways are denoted differently.

Parameter estimates of the *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$

The *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ was selected as best model by considering the AIC population weights for the cells that expressed d2eGFP proteins as shown by Figure 31. The model $\mathcal{M}_5$ was frequently selected as best model for the individual fits together with $\mathcal{M}_6$ and $\mathcal{M}_3$ as described in Chapter 5.2.2. The histograms of parameter estimates of model $\mathcal{M}_5$ are shown by Figure 45. The degradation rate β of d2eGFP proteins, the measurement noise σ, the offset parameter κ and the expression onset parameter t_0 are most densely distributed. All other parameters spread widely over the parameter interval from -5 to 5 on the logarithmic scale to base 10. The profile likelihood rations and the local approximations of the maximum likelihood function are studied for one individual fit that yielded the same maximum likelihood values for at least 10 multistarts. The results of the parameter uncertainty analysis are shown by Figure 46.



(a) *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ (b) *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$

Figure 44: Best models of the set for d2eGFP coding mRNA construct: Model $\mathcal{M}_5$ was selected as best model based on AIC population weights and model $\mathcal{M}_6$ based on BIC population weights. They were frequently chosen as best models for the individual fits together with $\mathcal{M}_3$. Model $\mathcal{M}_5$ and $\mathcal{M}_6$ differ in the degradation pathways of mRNA as $\mathcal{M}_5$ includes an additional pathway of degradation without enzymes.

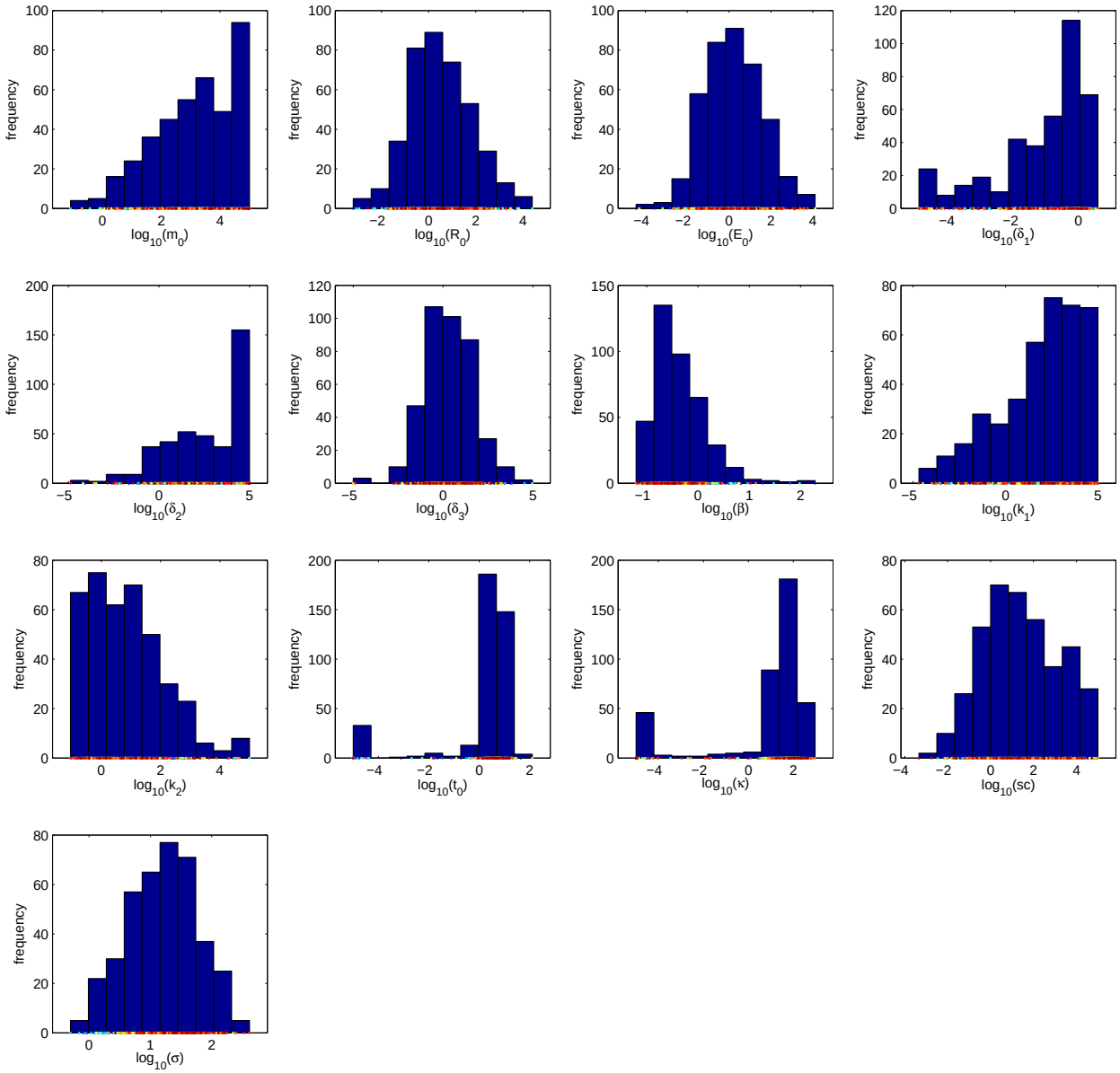


Figure 45: Parameter histograms of model $\mathcal{M}_5$ without outliers: The d2eGFP decay rate β, the offset parameter κ, the measurement noise σ as well as the parameter t_0 that was used to cover the different onset times of protein expression are more densely distributed than the other model parameters.

The profile rations of model $\mathcal{M}_5$ are shown for each parameter separately by the red lines in Figure 48 together with the second-order local approximation of the negative log-likelihood function $J(\theta)$ based on the Hessian given by the blue lines. The profile rations are bounded for the decay rate β of d2eGFP proteins and the measurement noise σ with increasing and decreasing parameter values. The threshold that is given by the black, dashed line is related to the 95 % confidence interval. On the contrary the local approximation in blue reveals practical identifiability only for the measurement noise σ. However, both methods yield practical non-identifiability for all other model parameters. As the observation exhibits only 116 time points and the model parameters could be highly non-linear correlated, the profile likelihood analysis yields more reliable results than the second-order local approximation of the negative log-likelihood function. For this reason it was assumed that both β and σ are practical identifiable for this single fit.

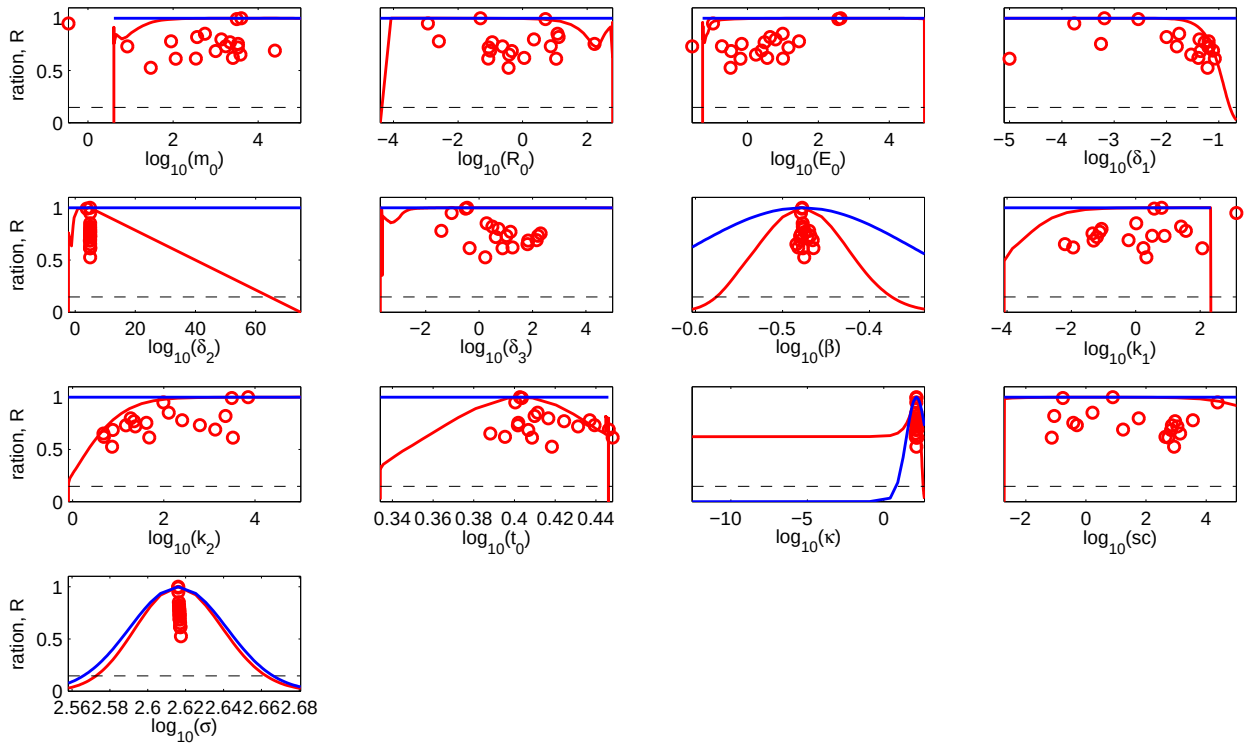


Figure 46: Parameter uncertainty analysis of model $\mathcal{M}_5$ for one individual fit: The profile rations are displayed in red are bounded for the d2eGFP decay rate β and the measurement noise σ. The threshold given by the dashed black line is associated to the 95 % confidence intervals. In contrast the rations obtained from the second-order local approximation of the objective function $J(\theta)$ reveals practical identifiability only for the measurement noise σ. Both methods indicate practical non-identifiability for the other model parameters.

Estimated parameters of the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$

In addition to the *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ was highly ranked by BIC population weights and was frequently selected as best model for the individual fits. The parameter histograms of $\mathcal{M}_6$ are shown by Figure 47. The degradation rate of d2eGFP proteins β, the offset parameter κ, the measurement noise σ and the expression onset parameter t_0 are most densely distributed. On the other hand all other parameters spread over several orders of magnitude in the parameter interval from -5 to 5 on the logarithmic scale to base 10. The parameter uncertainty analysis was done for one individual that exhibits the same best maximum of the likelihood function for at least 10 multistarts. The results are shown by Figure 48.

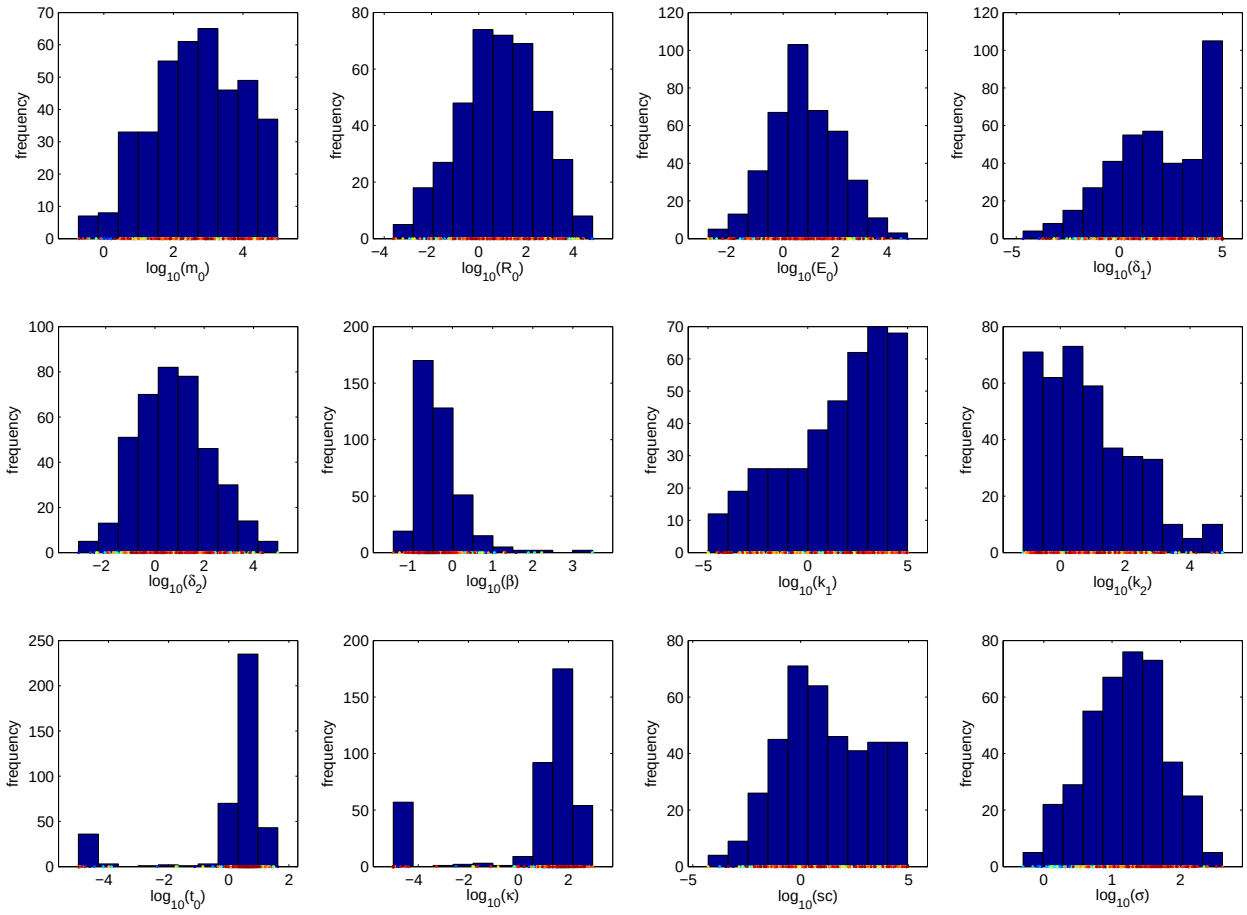


Figure 47: Parameter histograms of model $\mathcal{M}_6$ without outliers: The d2eGFP decay rate β, the offset parameter κ, the measurement noise σ and the expression onset parameter t_0 are more densely distributed than the other model parameters. The parameter uncertainty analysis was done for one individual trajectory fit that exhibits the same optimum for at least 10 multistarts.

The profile rations that are represented by the red lines in Figure 48 are bounded for the d2eGFP degradation rate β and the measurement noise σ. This indicates that they are practical identifiable. The threshold that is given by the dashed black line is related to the 95 % confidence interval. The profiles of other model parameters show various de- and increases even above the ration of 1 which indicates that higher maxima were found by the optimization of the profile likelihood than given by the maximum likelihood function. In addition the calculation of the profiles were interrupted as displayed by the sudden steep decrements of the red lines. Maybe the numerical integration of the rate equations was not possible at this point. To obtain a complete picture of the profile rations one could perform the optimization of the profile likelihood function for each of the illustrated red-circled optima that are above the threshold. In this case it was assumed that the profile rations indicate practical non-identifiability for all parameters except of β and σ. The second-order local approximations of the objective function $J(\theta)$ given by the blue lines yield practical non-identifiability for all model parameters except for the measurement noise σ.

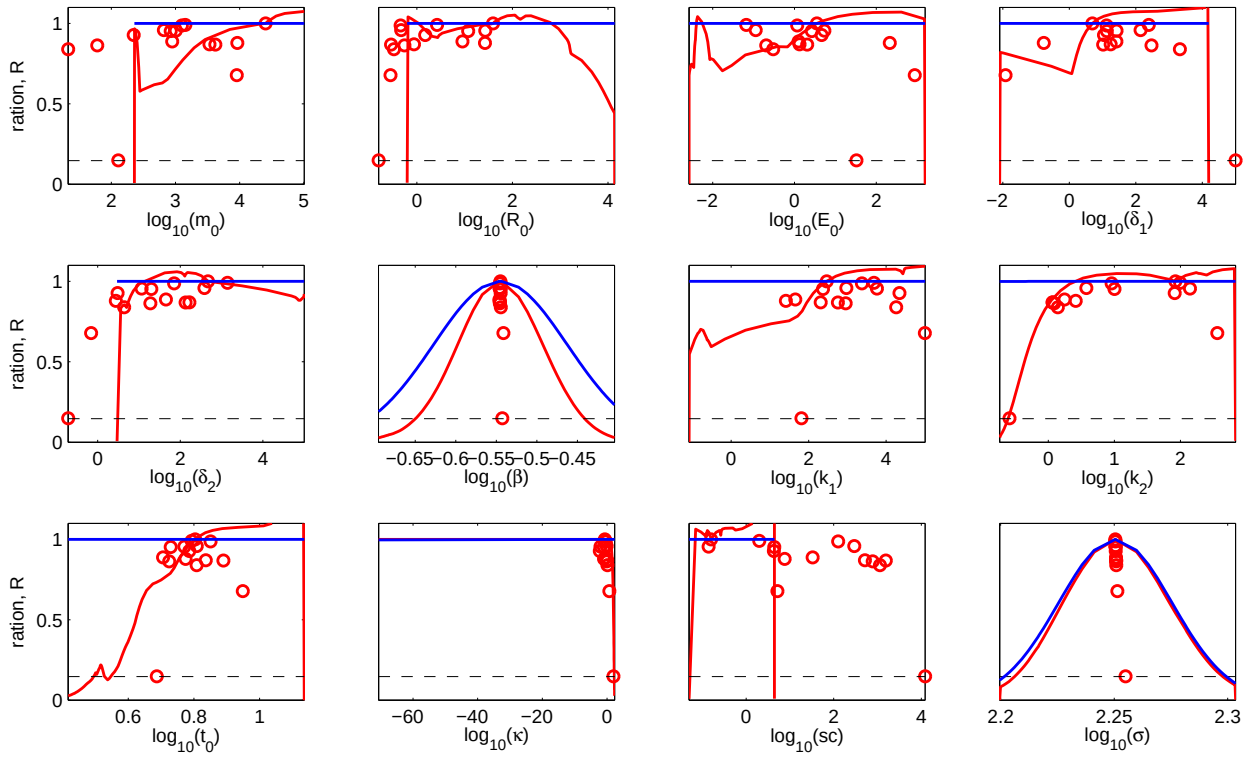


Figure 48: Parameter uncertainty analysis of model $\mathcal{M}_6$ for one individual fit: The profile rations displayed in red are bounded for the d2eGFP degradation rate β and the measurement noise σ. The second-oder local approximation of the negative log-likelihood $J(\theta)$ indicates that only the measurement noise σ is practical identifiable. Both methods yield practical non-identifiability for the other model parameters.

Protein degradation rates for d2eGFP constructs

The Profile likelihood analysis yield practical identifiability for one single fit for the eGFP degradation rate β and the measurement noise σ for both models $\mathcal{M}_5$ and $\mathcal{M}_6$. They were selected as best models in the set on the population scale by AIC and BIC respectively. In contrast the *Trivial-Model* was ranked on the last position by both model selection criteria. To compare the parameters β and σ of these three models, the median, mean and standard deviation were calculated for the log-10 as well as the back transformed parameter distributions. The results are summarized in Table 2. As the log-parameters are almost normally distributed, the back transformed parameters are log-normally distributed. The median is more robust towards outliers and closer to the mode than the mean of the parameter distributions.

The *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ and the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ yield very similar rates on the logarithmic scale to base 10. On the contrary the logarithmic rates of the *Trivial-Model* $\mathcal{M}_1$ differ significantly from the results of $\mathcal{M}_5$ and $\mathcal{M}_6$. On the one hand $\mathcal{M}_1$ was ranked on the last position by AIC and BIC population weights as shown by Figure 31 and one the other hand the degradation rate β of the reporter proteins is structural non-identifiable as explained in Chapter 5.3.1. For this reason the rates obtained by $\mathcal{M}_5$ and $\mathcal{M}_6$ are more reliable. The median values of the back transformed parameter distributions are very similar for $\mathcal{M}_5$ and $\mathcal{M}_6$ whereas the mean values and standard deviations differ significantly. As already explained above the median is more reliable as it is more robust towards outliers and closer to the mode of the parameter distribution.

Table 2: Median, mean and standard deviation of the degradation rate of d2eGFP proteins β and the parameter of the measurement noise σ.

rate	**median**			**mean**			**standard deviation**		
	$\mathcal{M}_1$	$\boldsymbol{\mathcal{M}_5}$	$\boldsymbol{\mathcal{M}_6}$	$\mathcal{M}_1$	$\boldsymbol{\mathcal{M}_5}$	$\boldsymbol{\mathcal{M}_6}$	$\mathcal{M}_1$	$\boldsymbol{\mathcal{M}_5}$	$\boldsymbol{\mathcal{M}_6}$
$\boldsymbol{\beta}$ $[h^{-1}]$	0.19	0.34	0.35	35.36	1.79	17.54	489.57	11.40	219.5
$\boldsymbol{\sigma}$	35.49	16.94	16.95	68.20	33.05	33.27	96.39	46.58	46.62
$\mathbf{log_{10}}(\boldsymbol{\beta})$	-0.73	-0.47	-0.46	-0.59	-0.35	-0.33	0.60	0.50	0.59
$\mathbf{log_{10}}(\boldsymbol{\sigma})$	1.55	1.23	1.23	1.49	1.19	1.20	0.59	0.56	0.56

5.3.4 Comparison of eGFP and d2eGFP degradation rates

As described in Chapter 5.3.2 and 5.3.3 the median, mean and standard deviation were calculated for the parameter distributions of the protein degradation rates β and of the measurement noise σ for the highest ranked models. For the population of cells that expressed eGFP proteins the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ was selected as best model by AIC and the *Ribosome-Model* $\mathcal{M}_3$ by BIC. For the population of cells that expressed d2eGFP proteins the *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ was chosen as best model by AIC and the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$ was highest ranked by BIC. To compare the degradation rates and the values of measurement noise for the eGFP and d2eGFP constructs, the mean and median values and standard deviations were averaged over the highest ranked models per population. The results are summarized in Table 3.

The average over the highest ranked models is only justified for the statistical quantities of the logarithmic parameter distributions as their values were very similar for both models. On the contrary the average is not reasonable for the mean and standard deviation of the back transformed parameter distributions as they were significantly different for both models. As expected the degradation rate of the d2eGFP proteins is higher than of eGFP which is confirmed by the mean as well as the median values of $\log_{10}(\beta)$ for the two different constructs in Table 3. Interestingly the measurement noise $\log_{10}(\sigma)$ shows higher median and mean values for eGFP than for d2eGFP. As both constructs have been used in the same experimental setup, this difference can only be traced back to the image analysis. The cells that were transfected with mRNA coding for eGFP proteins were more vital than those expressing d2eGFP proteins. They showed more movement within their wells and migration over

parameter	**median**		**mean**		**standard deviation**	
	eGFP	**d2eGFP**	**eGFP**	**d2eGFP**	**eGFP**	**d2eGFP**
$(\boldsymbol{\beta})$	0.19	0.35	1.06	9.67	6.54	115.45
$(\boldsymbol{\sigma})$	43.92	16.95	84.05	33.16	125.89	46.60
$\mathbf{log_{10}}(\boldsymbol{\beta})$	-0.72	-0.47	-1.07	-0.34	1.17	0.55
$\mathbf{log_{10}}(\boldsymbol{\sigma})$	1.64	1.23	1.58	1.20	0.59	0.56

Table 3: Median, mean and standard deviation of the logarithmic parameters of the protein degradation β and the measurement noise σ that were averaged over the highest ranked models per population.

the single-cell-array. Migrating cells were manually excluded from the image analysis but the movements of the cellular body on the well edges led to increased fluctuations of mean fluorescence intensities that were calculated over the well area. In addition the cell division often starts with a shrinkage of the cell body followed by an expansion at later observation points. As the mean fluorescence intensities were calculated over the well area, too much background pixel were included in the calculation. This leads to an temporary decrease of the protein expression trajectory until the two daughter cells expand. This effect as well as the fluctuations of the cell beyond the well area increased the measurement noise. As the cells expressing eGFP proteins were more vital than those expressing d2eGFP proteins, this could explain why the values of the estimated measurement noise differ significantly between both constructs. In addition the statistics were obtained for different cell numbers as many of the migrating cells had to be excluded from the population of 236 that expressed eGFP proteins. In contrast the population that expressed d2eGFP proteins consists of 394 cells.

In the literature the fluorescence of reporter proteins were measured by flow cytometry for Chinese Hamster Ovary cells (CHO) that were transfected with plasmid DNA to express d2eGFP and eGFP proteins. The analysis yielded approximate half-lives of 2h for d2eGFP and 17 h for eGFP proteins [1]. By assuming exponential decay of the proteins, the half-lives τ can be transformed into rates as follows:

$$rate = \frac{\ln(2)}{\tau} \tag{38}$$

This way the degradation rates for d2eGFP with $\beta = 0.35$ and for eGFP with $\beta = 0.04$ are calculated. These associated logarithmic degradation rates are given by $\log_{10}(\beta) = -0.45$ for d2eGFP proteins and by $\log_{10}(\beta) = -1.40$ for eGFP proteins. The degradation rate of d2eGFP proteins is well reflected by the median value of the logarithmic distribution given by $log_{10}(\beta) = -0.47$ in Table 3. In contrast the logarithmic degradation rates for the eGFP proteins were estimated higher than the literature value with a median value of $\log_{10}(\beta) = -0.72$. On the one hand this difference could be caused by the high measurement noise obtained for eGFP and on the other hand the transfection dynamics of mRNA and plasmid DNA are not directly comparable. In contrast to mRNA the DNA has to enter the nucleus before the translation into proteins. For this reason the onset times of protein expression are more widely distributed and shifted to later time points for plasmid DNA than for mRNA [45]. The broad distribution of expression onset times of plasmid DNA should lead to a higher variability of fluorescence intensities at given observation points than for mRNA in the flow cytometry analysis. This affects the half-lives given in the literature as they were obtained by calculating the mean fluorescence intensities over all cells at different time points. For this reason the rates that are associated to the half-lives can only be guiding values.

In the previous study of Leonhardt et al. [2] the degradation rates were estimated by fitting the *Trivial-Model* $\mathcal{M}_1$ on single-cell trajectories of populations of different cell lines. The optimization was done by the software IGOR-PRO that used the Levenberg-Marquart algorithm that minimizes the squared distances of observed and fitted data. As only one initial guess was given as a starting point for the optimization per trajectory fit the parameter estimates probably correspond to local maxima of the likelihood function and not to the global maximum in parameter space. In addition the model was structural non-identifiable in the degradation rates of transfected mRNA δ and proteins β as explained in Chapter 5.3.1.

Leonhardt et al. yielded protein degradation rates for d2eGFP of $\beta_{d2eGFP} = 0.165 \pm 0.142$ for HU7-cells and 0.124 ± 0.082 for A459 cells. These values are roughly close to the median value $\beta_{d2eGFP,\mathcal{M}_1} = 0.19$ for d2eGFP that have been obtained by the *Trivial-Model* $\mathcal{M}_1$ in this work. As shown in Chapter 5.3.3 the results of the *Trivial-Model* are significantly different from the statistical values that were obtained by the *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ and the *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$. The average over these highest ranked models yield a median value of $\beta_{d2eGFP} = 0.35$ instead that is close to the literature reference.

6 Conclusion and Outlook

Within my master thesis the dynamics of mRNA transfection of single cells were studied using six deterministic models that are based on rate equations. For this purpose the cells were seeded on microstructured single-cell arrays that provide high-throughput measurements for time-lapse microscopy. An improved fitting method was established in collaboration with the Institute of Computational Biology of the Helmholtz Center Munich that is based on a global optimization of the likelihood function using a multistart approach. To compare the fitting quality of the different models, the Akaike and Bayesian Information Criteria were applied that enable a ranking of models at a glance.

In the first part of this work model selection was studied on artificially generated datasets that were closely related to the experimental conditions. For this purpose each model was used once to create an associated dataset of single cell trajectories that started at randomly chosen time points. As the observation of cells is not possible during the incubation time of the transfection medium within the first hours, the data was removed from the artificial datasets for this timespan. By fitting each model once on the trajectories of the datasets, the question was addressed whether the generating models will be selected as best during model selection. This expectation was not confirmed by this master thesis. In most of the cases less complex models were favoured as best model for the whole population of single cell trajectories. By considering the model selection in detail for each trajectory fit, it was shown that multiple models were selected as best within the population per artificial dataset. To gain an deeper understanding of the observed model selection patterns, it could be beneficial to preserve the data that was removed for the first hours. This would implicate a modification of the experimental setup such that the dynamics can be observed in time with the incubation of the transfection medium. Furthermore the model selection could be studied by criteria that are more suitable for small datasets and models that exhibit parameter non-identifiability.

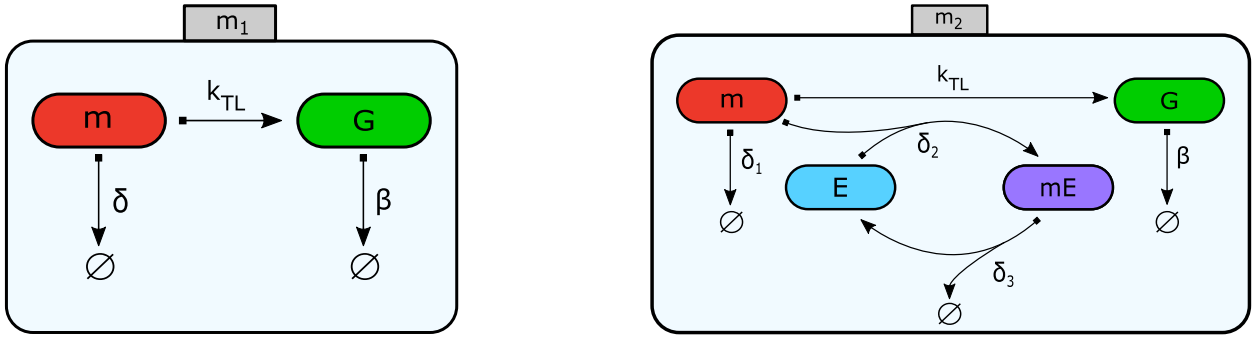
In the second part of this work the model selection was studied for experimental datasets. For this purpose HEK-cells were transfected with mRNA that either coded for eGFP or d2eGFP proteins. After observation via time-lapse microscopy the mean fluorescence intensities per single cell were gathered by an automatic image analysis. This way the data was obtained for each of the two populations. Afterwards, the six models were fitted on the experimental data by the global optimization method and the model selection was done on the population scale as well as for the individual trajectory fits. This way it was observed that models were selected as best that include a pathway of ribosomal translation of the transfected mRNA and only differed in the degradation pathways of the mRNA. Parameter uncertainty analysis was done for one individual fit per dataset for the highest ranked models on the population scale. This way an impression was obtained which model parameters could be practical

identifiable. To obtain a deeper insight into the identifiability of the model parameters, the analysis should be done for each single-cell trajectory fit per population. Furthermore it was observed that the profile likelihood method was sometimes interrupted during the optimization. In this cases the analysis should be done for each optimum of the likelihood function that was found to be within the 95 % confidence interval.
The models that were selected as best for the description of mRNA transfection dynamics in this work yielded very similar statistical quantities for the distributions of the logarithmic protein degradation rates β that were indicated as practical identifiable by the parameter uncertainty analysis. For this reason they were averaged over the highest ranked models to facilitate the comparison of the degradation rates of the eGFP and d2eGFP reporter proteins. It was shown that the mean and median degradation rates of the d2eGFP proteins were estimated significantly higher than of eGFP. Additionally the median degradation rate of d2eGFP, given by $\beta_{d2eGFP} = 0.35$, was found to be very close to the literature value of $\beta_{d2eGFP,FC} = 0.35$ that was gathered by flow cytometry analysis of CHO-cells that were transfected with plasmid DNA [1]. However, the median degradation rate of eGFP given by $\beta_{eGFP} = 0.19$ was estimated significantly higher than the literature reference of $\beta_{eGFP,FC} = 0.04$. On the one hand the transfection dynamics of plasmid DNA and mRNA are not direclty comparable but on the other hand this difference could also be caused by the measurement noise that was found to be higher for eGFP than for d2eGFP.

Within my master thesis the background correction of time-lapse movies was improved. A density based clustering algorithm was used to detect cellular signals on each image frame separately. They were removed from the image and the resulting gaps were reconstructed by an iterative interpolation method based on an 8-neighbourhood of the segmented image. As the surrounding of the cell is often brighter than the background due to the light scattering of the cellular fluorescence, the interpolation led to a bias towards higher values. This effect depends on the individual cell and on the image segmentation and reduced the data quality in this work. In addition the mean fluorescence intensities were gathered over the area of their confinements on the single-cell array. The movements of the cell body within this confinement, called well, caused high fluctuations in the mean intensities. This is especially increased during the cell division due to the shrinkage and subsequent expansion of the cell. An improvement of the interpolation method as well as of the well analysis is needed to reduce the measurement noise of the image analysis.

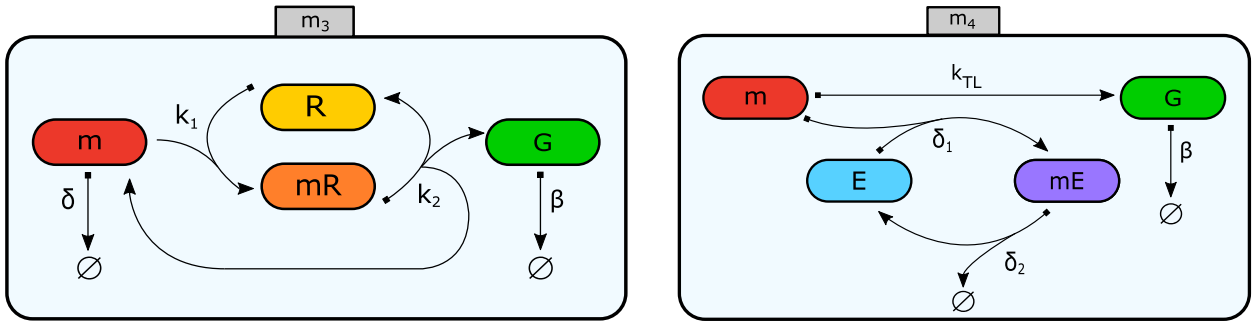
7 Appendix

7.1 The model set at a glance



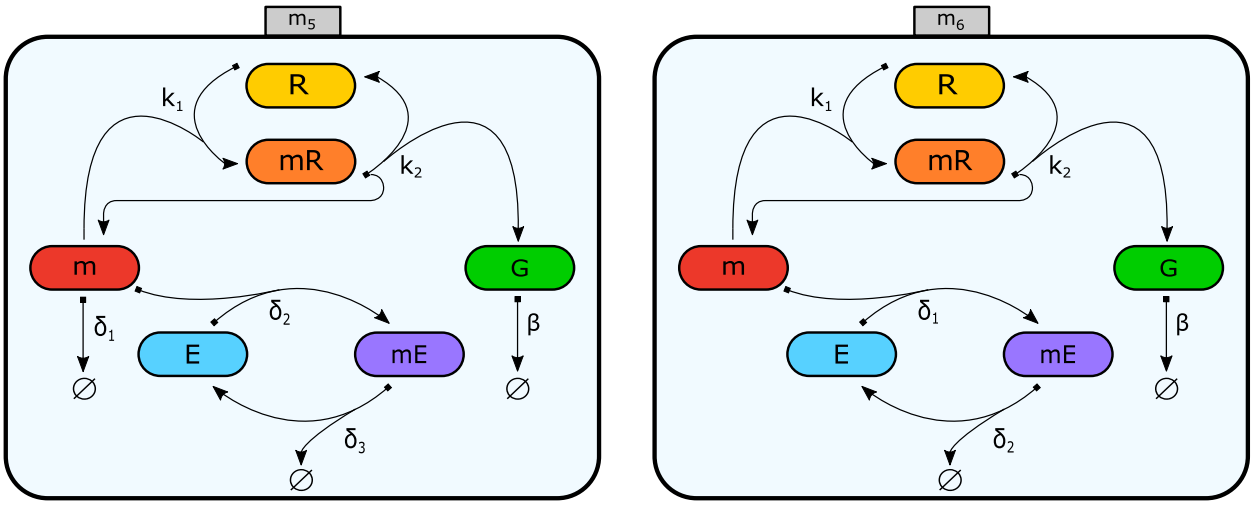
(a) *Trivial-Model* $\mathcal{M}_1$ (b) *Two-Degradation-Model* $\mathcal{M}_2$

Figure 49: In Figure 49a the model $\mathcal{M}_1$ describes the translation of mRNA (m) into GFP (G) with the rate k_{TL} and the degradation of both with the rates δ and β. Model $\mathcal{M}_2$ shown by Figure 49b includes an additional decay pathway of mRNA by enzymatic binding with the rate δ_2 and release with the δ_3.



(a) *Ribosome-Model* $\mathcal{M}_3$ (b) *Enzyme-Degradation-Model* $\mathcal{M}_4$

Figure 50: In model $\mathcal{M}_3$ in Figure 50a the mRNA is translated into GFP via ribosomal binding and release with the rates k_1 and k_2. The mRNA degrades with the rate δ and the proteins with the rate β. Model $\mathcal{M}_4$ in Figure 50b describes a direct translation of mRNA with the rate k_{TL} and a direct decay of proteins with the rate β. The mRNA degrades via an enzymatic pathway with the rates δ_1 and δ_2.



(a) *Two-Degradation-Ribosome-Model* $\mathcal{M}_5$ (b) *Enzyme-Degradation-Ribosome-Model* $\mathcal{M}_6$

Figure 51: In Figure 51a the mRNA is translated by ribosomes with the rates k_1 and k_2 and degraded by enzymes with the rates δ_2 and δ_3. In addition the mRNA decays with the rate δ_1. The proteins degrade directly with the rate β. Model $\mathcal{M}_6$ in Figure 51b exhibits the same pathways as model $\mathcal{M}_5$ except of the simple decay of mRNA with the rate δ_1.

7.2 Transfection protocol

Date of experiment: 13.01.2015

- SNIM RNA given by Mehrje: d2eGFP - eGFP II
- cells: HEK293
- transfection medium: Lipofectamine2000
- medium: RPMI + NaPyruvat+Hepes + 10 % FCS + Glutamin; OptiMEM; L-15 Medium + 10 % FCS

HEK293-cells were disseminated in growth medium in prepared slides (μ slide 6 channel) and flushed with 60μ medium. The cell seeding started at 8:45 and was stopped after at 12:15. Afterwards the medium was removed and the channels were washed with PBS. OptiMEM and 50ng transfection medium per channel were added. After incubation for 2 hours, the medium was exchanged for L15-Medium and the slides were moved to the microscope. The ratio of mRNA to lipoplexes was $1\mu g : 2.5\mu l$.

Preparation of Lipoplexes

1. dilute of 50 ng mRNA in OptiMEM-medium; GV = 25μl
2. dilute of 0.125 μl Lipofectamine 2000 in OptiMEM-medium and incubation at room temperature for 5 min; GV = 25μl
3. mix the mRNA solution and lipoplexes at room temperature for 20 min
4. add lipoplexes to cell slides: remove 100 μl, add 50 μl transfection medium, add 80 μl L15-medium and flush twice

Cells were observed for 30 h under the Nikon microscope, in which one frame was taken each 10 min, in the beginning transmitted light, later eGFP-Fluorescence scanning.

References

[1] Karl-Heinz Hoffmann, editor. *Coupling of Biological and Electronic Systems*. Springer Science & Business Media, 2002.

[2] C. Leonhardt. *On Quantitative mRNA Transfection*. PhD thesis, LMU Munich, 2014.

[3] S.D. Patil, D.G. Rhodes, and D.J. Burgess. Dna-based therapeutics and dna delivery systems: A comprehensive review. *The AAPS Journal*, 7(9):61–77, 2005.

[4] S. Zou, K. Scarfo, M.H. Nantz, and J.G. Hecker. Lipid-mediated delivery of rna is more efficient than delivery of dna in non-dividing cells. *Int J Pharm*, 389:232–243, 2010.

[5] F. Scherer, M. Anton, U. Schillinger, J. Henke, C. Bergemann, A. Krueger, B. Gaensbacher, and C. Plank. Mangetofection: enhancing and targeting gene delivery by magnetic force in vitro and in vivo. *Gene Therapy*, 9:102–109, 2002.

[6] Y. Gao, X. Liu, and X. Li. Research process on sirna delivery with nonviral carriers. *International Journal of Nanomedicine*, 6:1017–1025, 2011.

[7] A. Yamamoto, M. Kornmann, J. Roseneckek, and C. Rudolph. Current prospects delivery gene delivery. *European Journal of Pharmaceutics and Biopharmaceutics*, 2008.

[8] M.S.D Kormann, G. Hasenpusch, M.K. Aneja, G. Nica, A.W. Flemmer, S. Herber-Jonat, M. Huppmann, L.E. Mays, M. Illenyi, A. Schams, M. Griese, I. Bittmann, R. Handgretinger, D. Hartl, J. Rosenecker, and C. Rudolph. Expression of therapeutic proteins after delivery of chemically modified mrna in mice. *Nature Biotechnology*, 2010.

[9] P.L. Felgner, T.R. Gadek, M. Holm, R. Roman, H.W. Chan, M. Wenz, J.P. Northrop, G.M. Ringold, and M. Danielsen. Lipofection: A highly efficient, lipid-mediated dna-transfection procedure. *Biochemistry*, 1987.

[10] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell Press*, 135(2):216–226, 2008.

[11] D. Wlodkowic, S. Faley, M. Zagnoni, J.P. Wikswo, and J.M. Cooper. Microfluidic single cell array cytometry for the analysis of tumor apoptosis. *Anal Chem.*, 81:5517–5523, 2009.

[12] K. King, Sihong Wang, Daniel Irimia, A. Jayaraman, M. Toner, and M. Yarmush. A high-throughput microfluidic real-time gene expression living cell array. *Lab Chip*, 7(1):77–85, January 2007.

[13] D.D. Carlo, L.Y. Wu, and L.P. Lee. Dynamic single cell culture array. *Lab on a Chip*, 6:1445–1449, 2006.

[14] E. Ostuni, C.S. Chen, D.E. Ingber, and G.M. Whitesides. Selective deposition of proteins and cells in arrays of microwells. *Langmuir*, 17:2828–2834, 2001.

[15] A. Azioune, M. Storch, M. Bornens, M. Thry, and M. Piel. Simple and rapid process for single-cell micro-patterning. *Lab on a Chip*, 9:1640–1642, 2008.

[16] A. Folch and M. Toner. Microengineering of cellular interactions. *Annual Review of*

Biomedical Engineering, 2:227–256, 2000.

[17] P. J. Roettgermann, A. Alberola, and J. O. Raedler. Cellular self-organization on microstructured surfaces. *Soft Matter*, 10, 2014.

[18] S.K. Akiyama, S.S. Yamada, W. Chen, and K.M. Yamada. Analysis of fibronectin receptor function with monoclonal antibodies: Roles in cell adhesion, migration, matrix assembly, and cytoskeletal organization. *The Journal of Cell Biology*, 109:863–875, 1989.

[19] D. Muzzey and A. van Oudenaarden. Quantitative time-lapse fluorescence microscopy in single cells. *Annual Review of Cell and Developmental Biology*, 2009.

[20] R. Pepperkok and J. Ellenberg. High-throughput fluorescence microscopy for systems biology. *Nature*, 7:690–696, September 2006.

[21] M. Tigges, T. Marquez-Lago, J. Stelling, and M. Fussenegger. A tunable synthetic mammalian oscillator. *Nature*, 457:309–312, January 2009.

[22] B. Neumann, M. Held, U. Liebel, H. Erfle, P. Rogers, R. Pepperkok, and J. Ellenberg. High-throughput rnai screening by time-lapse imaging of live human cells. *Nature Methods*, 3(5):385–390, April 2006.

[23] Joseph R. Lakowicz. *Principles of Fluorescence Spectroscopy*. Springer Science & Business Meida, 3 edition, 2006.

[24] Markus Sauer, Johan Hofkens, and Jrg Enderlein. *Handbook on Fluorescence Spectroscopy and Imaging - From Ensemble to Single Molecules*. WILEY-VCH Verlag GmbH&Co. KGaA, 2011.

[25] G.H. Patterson, Knobel S.M., W.D. Sharif, S.R. Kain, and D.W. Piston. Use of the green fluorescent protein and its mutants in quantitative fluorescence microscopy. *Biophysical Journal*, 73(5):2782–2790, 1997.

[26] Roger Y. Tsien. The green fluorescent protein. *Annual Review of Biochemistry*, 67:509–544, 1998.

[27] M. Schwarzfischer, C. Marr, J. Krumsiek, P.S. Hoppe, T. Schroeder, and F.J. Theis. Efficient fluorescence image normalization for time-lapse movies. In *Proc. Microscopic Image Analysis with Applications in Biology*, 2014.

[28] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings*, 1996.

[29] D. Briant and A. Kut. St-dbscan: An algorithm for clustering spatial-temporal data. *Data & Knowledge and Engineering*, 60:208–221, 2007.

[30] A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelke, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F.J. Theis, U. Klingmller, and J. Timmer. Lessons learned from quantitative dynamic modeling in systems biology. *PLoS ONE*, 8(9), 2013.

[31] J. Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47:90–100, 2003.

[32] J. Vanlier, C.A. Tiemann, P.A.J. Hilbers, and N.A.W. van Riel. Parameter uncertainty

in biochemical models described by ordinary differential equations. *Mathematical Biosciences*, 246:305–314, 2013.

[33] M. Ashyraliyev, Y. Fomekong-Nanfack, J.A. Kaandorp, and J.G. Blom. Systems biology: parameter estimation for biochemical models. *the FEBS Journal*, 276:886–902, 2009.

[34] W.W. Chen, M. Niepel, and P.K. Sorger. Classic and contemporary approaches to modeling biochemical reactions. *Genes and Development*, 24:1861–1875, 2010.

[35] O.T. Chis, J.R. Banga, and E. Balsa-Canto. Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, 6(11), 2011.

[36] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[37] S. Hock, J. Hasenauer, and F. Theis. Modeling of 2d diffusion processes based on microscopy data: parameter estimation and practical identifiability analysis. *BMC Bioinformatics*, 14, 2013.

[38] K. Burnham and D. Anderson. Multimodel interference: Understanding aic and bic in model selection. *Sociological Methods Research*, 33:261–304, 2004.

[39] D. Posada and T.R. Buckley. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.*, 53(5):793–808, 2004.

[40] K.P. Burnham and D.R. Anderson. Kullback-leibler information as a basis for strong interference in ecological studies. *Wildlife Research*, 28:111–119, 2001.

[41] S.T. Buckland, K. Burnham, and N.H. Augustin. Model selection: An integral part of interference. *Biometrics*, 53(2):603–618, 1997.

[42] A.E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

[43] S. Vrieze. Model selection and psychological theory: A discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological Methods*, 17:228–243, 2012.

[44] J. Hasty, D. McMillen, F. Isaacs, and J.J. Collins. Computational studies of gene regulatory networks in numero molecular biology. *Nature Reviews Genetics*, 2:268–279, 2001.

[45] C. Leonhardt, G. Schwake, T.R. Stögbauer, S. Rappl, J. Kuhr, T. Ligon, and J.O. Rädler. Single-cell mrna transfection studies: Delivery, kinetics and statistics by numbers. *Nanomedicine: NBM*, 2014.

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich bei der Entstehung meiner Masterarbeit unterstützt haben. Insbesondere gilt mein Dank...

Prof. Dr. Joachim O. Rädler für die Möglichkeit an seinem Lehrstuhl meine Masterarbeit zu einem interessanten Thema schreiben zu dürfen sowie fr die gute Betreuung und für die hilfreichen Anregungen und Ideen.

Dr. Jan Hasenauer für die gute Betreuung meiner Masterarbeit am Helmholtz Zentrum München und den vielen Ideen zur Verbesserung und Gestaltung meiner Masterarbeit.

Insbesondere gilt mein Dank **Fabian Fröhlich**, der mir mit all seinem Wissen über die Theorie beiseite Stand und mir jede noch so kleine Frage mit viel Geduld beantwortet hat.

Gerlinde Schwake für all ihre Hilfe zur Umsetzung der Experimente und für die Beantwortung all meiner Fragen in diesem Kontext.

Dr. Carolin Leonhardt für die gute Betreuung in den ersten Monaten meiner Masterarbeit.

dem gesamten Lehrstuhl für die nette Arbeitsatmosphäre am Lehrstuhl.

meiner Familie und meinen Eltern die mich immer dazu ermutigt haben, nach den Sternen zu greifen und neue Herausforderungen anzugehen. Ich danke ihnen auch für ihre Geduld während des Schreibens meiner Arbeit, das Korrekturlesen und vor allem für die viele Unterstützung während meines Studiums.

Erklärung

Hiermit erkläre ich, die vorliegende Arbeit selbständig verfasst zu haben und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel benutzt zu haben.

München, den 8.04.2015