Bioresource Technology 443 (2026) 133835

Contents lists available at ScienceDirect

Bioresource Technology

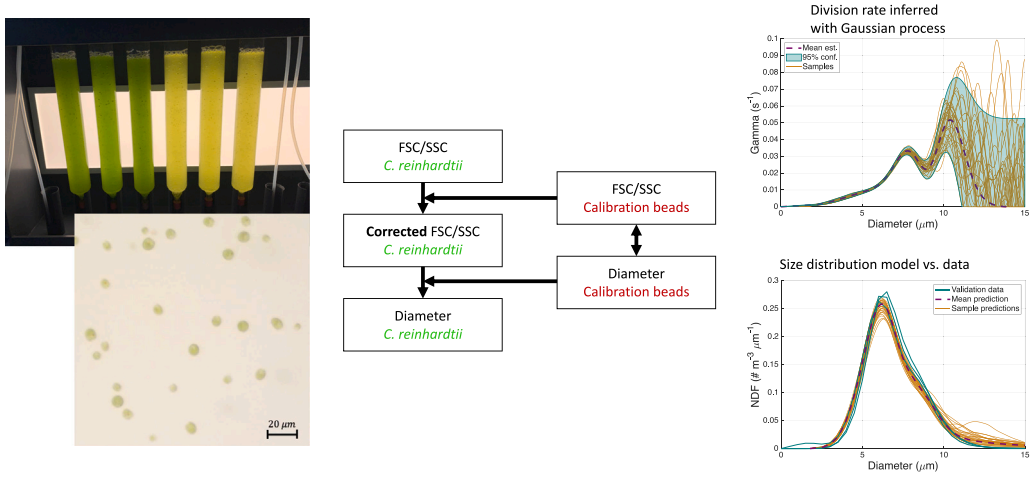
journal homepage: www.elsevier.com/locate/biortech

Inferring cell division kinetics in *Chlamydomonas reinhardtii* from flow cytometry with Gaussian process regression

Michiel Busschaert [a,b], Michael Schagerl [b], Christian Griebler [b], Florence H. Vermeire [a], Steffen Waldherr [b,*]

[a] KU Leuven, Department of Chemical Engineering, Celestijnenlaan 200F, 3001 Leuven, Belgium
[b] University of Vienna, Department of Functional and Evolutionary Ecology, Djerassiplatz 1, Vienna, Austria

GRAPHICAL ABSTRACT



ARTICLE INFO

Dataset link: https://github.com/MichielBusschaert/DataInferringCellDivision

Keywords:
Microalgae
Mie theory
Cell division rate
Single-cell techniques
Population balance model
Cell size estimation

ABSTRACT

Green algae, such as *Chlamydomonas reinhardtii*, are promising candidates for various industrial applications. For microorganisms, cell size relates to metabolic activity and cellular composition, among others, and is thus a highly relevant property for bioprocess operations. Besides time-consuming microscopy measurements, cell sizes can be estimated indirectly for example with flow cytometry based on light scattering, resulting in a measured size distribution. However, calibration for absolute cell sizes is obstructed by the optical properties of the involved particles, resulting in less accurate size distributions, which in turn can hinder applications such as model development or process monitoring. In this study, a novel approach is proposed to improve the estimation accuracy of the cell size distribution by utilizing a physical model on light scattering around a sphere. Benchmarked with microscopy image analysis, the model shows substantial improvement. Using the corrected size distribution, the cell division rate is inferred with an extended Gaussian process regression used upon a population balance model. The resulting model is able to accurately describe the observed

* Corresponding author.
E-mail addresses: michiel.busschaert@kuleuven.be (M. Busschaert), michael.schagerl@univie.ac.at (M. Schagerl), christian.griebler@univie.ac.at (C. Griebler), florence.vermeire@kuleuven.be (F.H. Vermeire), steffen.waldherr@univie.ac.at (S. Waldherr).

https://doi.org/10.1016/j.biortech.2025.133835
Received 9 October 2025; Received in revised form 28 November 2025; Accepted 17 December 2025
Available online 18 December 2025
0960-8524/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

size distribution with the estimated division kinetics. The approach is tested using data obtained from the cultivation of *C. reinhardtii* as a model organism. The results provide mechanistic insight into *C. reinhardtii* cell division and cell size heterogeneity.

1. Introduction

Organisms performing oxygenic photosynthesis can roughly be classified into embryophytes prevailing in terrestrial ecosystems, and algae dominating aquatic systems. The latter group indicates a high potential for industrial-scale exploitation, as a consequence of their higher growth rate and the fact that they are subject to less competition for land use (Usai et al., 2023; Leong and Chang, 2022; Hempel et al., 2011). Among the various algae phyla, green algae are a promising group for industrial applications (Schagerl, 2022). Besides a renewable alternative to fossil-based feedstock, green algae are candidates for, among others, hydrogen production (Jiao et al., 2024; Tamburic et al., 2011), treatment of wastewater (Usai et al., 2023; Abdelfattah et al., 2023), and they offer unique properties for production of high-quality compounds, which are used in food, agriculture and pharmaceutical industry (Mehariya et al., 2021; Bernard, 2011). In particular, *Chlamydomonas reinhardtii* is a widely studied model organism within the unicellular green algae.

To enable economic competitiveness of algae-based biotechnology, appropriate modeling, monitoring (Schagerl et al., 2022) and control tools need to be developed (Bernard, 2011). A key advancement in modeling algae growth is the introduction of the Droop equation (Droop, 1968), which expresses the growth rate as function of internal cell quota. This model has been further extended to take environmental factors such as light intensity (including photoinhibition), temperature, and nutrient supply (Lee et al., 2015) into account. Recent modeling approaches have considered the use of population balance models (PBM) with the objective of describing cell mass (or related cell size) distribution during cultivation (Usai et al., 2023; Atzori et al., 2021; Tronci and Schaum, 2022; Bertucco et al., 2015). Differently sized cells have varying cellular composition, which in turn displays different dynamic rates (e.g., CO_2 fixation rate), resulting in different growth and cell division kinetics. These rates are often unknown in advance and are dependent on various external factors. In order to obtain a description of these kinetic rates, accurate experimental data on the distribution within a cell population are essential for model development, which can then be used via the PBM framework to simulate the cell size distribution in the process. Cell size distribution plays an essential role in microalgae processing, affecting physical processing, for example filtration, centrifugation and sedimentation (Usai et al., 2023).

To fully exploit these modeling advances, they are ideally combined with on-line monitoring techniques to enable more accurate predictions with respect to cell-to-cell variability (Waldherr, 2018). Three main measurement techniques for data collection on cell size distribution are currently applied (Gayen and Venkatesh, 2008; Gong and Min, 2022): microscopic image analysis, Coulter counting, and flow cytometry (FCM). Of these, FCM appears the most practical approach due to its high-throughput, rapid data collection, and versatility in measuring additional cell properties. The core operating principle of FCM (Adan et al., 2017) is based on measuring the scattering of light from interaction with a particle. Distinct particles are (ideally) separated via a microfluidic system to prevent interference from scattering by multiple particles. The measured scattering intensity can subsequently be related to the size of that specific particle, typically requiring calibration with reference particles. However, as experimentally observed (Chioccioli et al., 2014; Agagliate et al., 2018), and as predicted by light scattering theory (Bohren and Huffman, 1998), calibration between light scattering and particle size is influenced by other optical properties, quantified as the refractive index of the particle and surrounding fluid. These effects can be explained by the Mie theory, which is a physical model originating from classical electromagnetism. In this paper, a novel method to adjust the derivation of the cell size distribution based on FCM-data is proposed. The method relies on Mie scattering theory, and is restricted to particles that are approximately homogeneous and of spherical shape, although the underlying principle could be extended to other geometries with modifications. The measured size distribution is validated with microscopic image analysis, which is assumed as reference for the *true* size distribution.

Based on the corrected FCM-based size distribution, a model is developed that describes the cell size distribution at (quasi-)steady-state conditions by using the PBM framework to structure its dynamic behavior into the effects of cell growth and cell division. Whereas cell growth can be measured relatively easily, cell division is less direct. In the model development, the kinetics of cell division are approximated from the size distribution, using Gaussian process regression (Rasmussen and Williams, 2006). This regression technique incorporates the model structure describing cell size distribution dynamics. This specific approach allows quantifying uncertainty in predicting the division rate. Such a prediction can be used to describe the cell size distribution as a mathematical model, which in turn can be applied for characterizing cell size dependent processes.

Thus, this paper aims to address two challenges. First, estimating the cell size distribution from FCM data, taking into account Mie scattering. Second, inferring the cell division rate from the estimated size distribution. The remainder of the paper is structured as follows: Section 2 provides an overview of the experimental conditions under which the data were collected. Next, two methodologies are discussed, along with a brief overview of the supporting theory. The first method, introduced in Section 3, is related to the correction of FCM data for size distribution estimation (Agagliate et al., 2018; Moutier et al., 2016). The second methodology (Section 4) provides a manner to infer division kinetics from experimentally measured size distributions. Both procedures are applied in Section 5 to a cultivation experiment of *C. reinhardtii*, which was set up to validate the applicability and results of the developed approaches. To conclude, a brief discussion on the developed methodology and obtained results is provided in Section 6.

2. Cultivation & flow cytometry measurements

In this subsection, an overview of the experimental procedure and conditions employed for data collection is outlined. This includes (*i*) the cultivation of the green alga *C. reinhardtii* and describing its growth curve, and (*ii*) the collection of light scattering properties using FCM, and microscopic image analysis.

Chlamydomonas reinhardtii (strain SAG 11-32b) was grown in vertically oriented bubble column photobioreactors (Schagerl et al., 2022) with 1 L working volume and diameter of approximately 60 mm. Sterile air is supplied at the bottom of the column. This results in a two-phase, bubble flow reactor which in turn promotes mixing and gas transfer. The columns are continuously illuminated via a white LED-panel on a single side (measured light intensity of 80 μmol photons m^{-2} s^{-1}). The continuous illumination causes asynchronous growth of *C. reinhardtii* cells, in contrast to natural growth conditions or simulated day/night cycles that induce synchronization of the cell population (Vítová et al., 2011). As a consequence, the relative size distribution remains more or less unchanged over time. The sterilized growth medium used for cultivation is a modification of the desmidiacean medium described by Pringsheim (Pringsheim, 2016). It consists of water (80% milli-Q and 20% sterilized bog water) enriched with KNO_3 (1 mM), $(NH_4)_2HPO_4$ (0.15 mM), $MgSO_4 \cdot 7\,H_2O$ (0.08 mM), $CaSO_4 \cdot 2\,H_2O$ (0.17 mM),

$Fe_{III}Na-EDTA$ (0.09 mM), and vitamin B12 in trace quantity. Cultivation took place over approximately seven days. Biomass growth was tracked by regular measurement of dry mass, optical density at 750 nm, and *in vivo* chlorophyll-a according to the protocols of Schagerl et al. (2022). A specific growth rate of $0.0215\,h^{-1}$ was observed during the exponential growth phase from dry mass measurements.

In addition, light scattering properties of particles were measured with FCM (Cytek CellStream). Samples of the *C. reinhardtii* culture were harvested from the air-fed columns, and measured shortly after. The cell culture samples were diluted tenfold prior to FCM measurements. The forward scatter intensity (FSC) was measured at 456/51 nm (at 30% power), the side scatter intensity (SSC) at 773/56 nm (at 3% power), and the fluorescence intensity at multiple wavelengths (excitation at 488 nm at 50% power, and emissions at 773/56, 528/46, 583/24, 611/31, 702/87 nm). For calibration, spherical polystyrene size reference beads (ThermoFischer, F-13838) of six different diameters were used (1, 2, 4, 6, 10 and 15 µm). Sample throughput for both the *C. reinhardtii*-culture and polystyrene particles was at a rate of $14.64\,\mu L\,min^{-1}$.

Finally, widefield microscopy (Zeiss Axio Imager.M1) was used for microphotographs at 40× magnification with a pixel size of 2584 × 1936 pixels from a cell culture as described above. The images were analyzed using the Zeiss ZEN 2011 software to estimate the size distribution as an alternative method to FCM. The cells on the 2D images were automatically annotated with a circle of equal surface area, and the image annotations were manually curated within the software to adjust any errors. Five images were analyzed in that way, yielding data for a total of 59 cells. The diameters of the annotated circles were used to compute a size distribution with a kernel density estimator.

3. Size distribution reconstruction

This section describes the procedure to characterize the cell size distribution from FCM data. First, an overview of Mie scattering theory is given, which provides a mathematical framework to describe FCM data. Second, a novel approach is introduced which uses Mie scattering theory to compensate for calibration issues in FCM.

3.1. Mie scattering theory

This section highlights the implications of the optical phenomena for particle size distribution characterization by means of FCM measurements. The Mie scattering theory describes the interactions between incident electromagnetic radiation and a particle. For a rigorous and comprehensive overview of optical scattering theory, refer to, e.g., Bohren and Huffman (1998), Nieto-Vesperinas (2020). Please also refer to the Supplementary Material for a more detailed description of the calculations provided here (Bohren and Huffman, 1998; Agagliate et al., 2018).

In essence, the scattering of light with wavelength λ on a spherical, homogeneous particle depends on two parameters: the particle diameter D, a geometric property, and the refractive index n, a material property that may depend on the incident wavelength λ. The latter is conventionally expressed as a complex number, where the real part corresponds to light obstruction, and the imaginary part corresponds to absorption by the material. Given these parameters, one can model the FSC and SSC in a flow cytometer as function of these parameters, i.e.

$$I_j(D,n) = function(D,n;\lambda_j,\theta_{1,j},\theta_{2,j}) \tag{1}$$

where $j \in \{\mathrm{FSC},\mathrm{SSC}\}$. Here, θ_1 and θ_2 are two geometric properties of the flow cytometer device, and are estimated via reference particle calibration (Agagliate et al., 2018). More details on the evaluation of Eq. (1) are provided in the Supplementary Material.

3.2. Adjusting scattering data in flow cytometry

In this section, the procedure to compute the size distribution from FCM data is explained. Part of the procedure is derived from Mie scattering theory discussed in the previous subsection. Two datasets from FCM measurements are used: one dataset $\mathcal{D}_C$ corresponding to the measurements of a *C. reinhardtii* culture, and one dataset $\mathcal{D}_{PS}$ corresponding to the reference polystyrene particles. Each dataset includes forward scattering, side scatter, and fluorescence measurements, i.e., $\mathcal{D}_j = \{FSC_j, SSC_j, F_j\}$ with $j \in \{C, PS\}$.

As a first step, the data are preprocessed. Regarding the *C. reinhardtii* measurements, noise (minerals, dust particles, etc., other than cells) is distinguished from cells by applying a Gaussian mixture model to the features in $\mathcal{D}_C$, which separates the measurements into two clusters. The dataset $\mathcal{D}_C$ is then reduced to the category with the highest light scattering intensity, to exclude noise from the data.

Next, the scattering measurements involving reference polystyrene particles FSC_{PS} and SSC_{PS} are associated to the six different diameters of the reference particles. Local maxima in the frequency of the dataset $\mathcal{D}_{PS}$ are linked to the known reference particle sizes. Polynomial regression is performed to interpolate particle diameters at different measured FSC,

$$D(FSC_{PS}) = \begin{pmatrix} 1 & \log_{10}(FSC_{PS}) & \log_{10}(FSC_{PS})^2 & \cdots \end{pmatrix} \cdot \boldsymbol{\beta}, \tag{2}$$

where the weight vector $\boldsymbol{\beta}$ is determined by minimizing the squared error of the diameter and the measured FSC of the six reference particles for a certain polynomial degree. Continuing with the preprocessed *C. reinhardtii* dataset $\mathcal{D}_C$, the next step would be to apply a regression similar to Eq. (2) to calculate the approximate cell diameter. However, as discussed in Section 3.1 and as illustrated in Eq. (1), scattering intensity depends on the refractive index. Therefore, an additional operation is included, which converts the FSC_C measurements into *pseudo*-measurements. The term *pseudo*-measurement is used to describe (according to Mie theory) the FSC intensity that one would measure for a hypothetical polystyrene sphere, which has the same diameter as the real *C. reinhardtii* cell for which the FSC intensity is actually measured. In other words, for two identically sized spheres but distinct optical properties (one *C. reinhardtii* and one polystyrene), and if the FSC intensity is measured for one of these particles, the term *pseudo*-measurement refers to the FSC intensity one would expect to measure for the other sphere. This *pseudo*-measurement is obtained by evaluating $I_{\mathrm{FSC}}(D,n)$ with Eq. (1) for different values of D and for the refractive indices for *C. reinhardtii* and polystyrene. The angles $\theta_{1,j}$ and $\theta_{2,j}$, and the incident intensity $I_{\mathrm{inc,j}}$, are calibrated with the reference polystyrene particles. For polystyrene, the refractive index is valued at $n_{PS} = 1.591$ as specified by the supplier. For *C. reinhardtii*, the refractive index at measured wavelength is taken as $n_C = 1.3527 + i \cdot 4.2176 \times 10^{-3}$ according to Qi et al. (2016). This provides the equivalence curve which transforms the measurements FSC_C into *pseudo*-measurements FSC_{PS}, which is calibrated according to the regression in Eq. (2) to obtain an improved value for cell diameter. Applying this procedure to each entry in the dataset $\mathcal{D}_C$ returns a set of cell diameters that can be visualized as a histogram based on occurrence frequency. Here, instead of a histogram, the cell size distribution is described by a continuous *number density function* (NDF) via a kernel density estimate. This NDF statistically quantifies size distribution in a cell population as a density function. As such, it is always nonnegative, and the integration over a certain range of cell sizes represents the number of cells with size in that range. As the cell size distribution may change over time, due to, e.g., cell growth and division, the NDF is time and size dependent, denoted as $n(t,x)$, with $x \in [0,\infty)$ (Küper et al., 2019; Waldherr, 2018).

4. Modeling division mechanism from experimental data

Based on the identified NDFs from experiments, a model for the cellular division rate is inferred. This is based on the framework of

population balance modeling (PBM), which relates cell division and cell growth to the NDF, where growth rate is estimated directly from dry mass measurements or optical density. This is integrated with a *Gaussian process regression* (GPR) model to represent the cell division rate.

4.1. Population balance modeling

In order to predict the changes in the NDF $n(t,x)$ size distribution over time, the PBM concept provides a mathematical framework to describe the expected size distribution evolution in a deterministic manner, even if some variability between individual entities exists (Ramkrishna, 2000). In a well-mixed bioreactor without biomass removal, the two main processes taking place at the cellular level are cell growth following substrate uptake, and cell division. The former leads to an increase in average cell size, the latter in a decrease thereof. The cumulative effect of cell growth and division is described by the population balance equation (PBE),

$$\frac{\partial n(t,x)}{\partial t} = -\frac{\partial \mu(x) x n(t,x)}{\partial x} - \gamma(x) n(t,x) \quad + \int_x^\infty \beta(x,\xi)\gamma(\xi) n(t,\xi) d\xi. \tag{3}$$

On the right-hand side, $\mu(x)$ represents the specific growth rate, and the partial derivative reflects the continuous nature of cell growth.

The division rate is represented by $\gamma(x)$, while $\beta(x,x')$ denotes the division kernel, further discussed below. For more details on these terms, and more generally on the derivation and use of PBMs, refer to, e.g., Ramkrishna (2000), Mantzaris and Daoutidis (2004).

Under constant growth conditions, the population may reach an exponential growth phase. This implies that the relative shape of the NDF remains fixed, only its L_1-norm (total cell density) increases. Under this assumption, the accumulation term in Eq. (3) satisfies

$$\frac{\partial n(t,x)}{\partial t} = \bar{\mu} n(t,x), \tag{4}$$

where $\bar{\mu} = \frac{1}{\int_0^\infty \xi n(t,\xi) d\xi} \int_0^\infty \mu(\xi)\xi n(t,\xi) d\xi$ is the population-averaged growth rate (Busschaert et al., 2023). Note that both sides of Eq. (4) can be multiplied with x and integrated over x from 0 to ∞, which results in the relation $\frac{dm(t)}{dt} = \bar{\mu} m(t)$, where $m(t) = \int_0^\infty x n(t,x) dx$ is the biomass concentration. This is the well-known equation describing exponential increase of biomass. Thus, the population-averaged growth rate $\bar{\mu}$ is obtainable from experimental dry mass or OD measurements.

4.2. Simulation of the size distribution

From Eq. (4), the NDF $n(t,x)$ can be decomposed into the product of a time-dependent and size-dependent factor, i.e., $n(t,x) := n_t(t) \cdot n_x(x)$, and the PBM can be solved via separation of variables. The time-dependent $n_t(t)$ satisfies Eq. (4) and describes exponential growth. The size-dependent $n_x(x)$ satisfies Eq. (3), substituting the left-hand side with Eq. (4). As constant growth conditions are assumed for the remainder of this paper, and as the focus is on the relative size distribution, the notation $n_x(x)$ is simplified to $n(x)$.

To solve this size-dependent factor $n(x)$, Eqs. (3) and (4) are solved with forward simulation by means of a finite volume method with first-order upwind scheme (Motz et al., 2002). The simulated NDFs (under a choice of division rate γ') are compared against the experimentally determined NDFs. To quantify the mismatch between these two density functions, the *Earth Mover's Distance* (EMD), also known as Wassersteins distance, is used (Orlova et al., 2016), which in one dimension is defined as

$$EMD(\gamma') = \int_0^\infty |c_{\exp}(x) - c_{\sim}(x;\gamma')| dx. \tag{5}$$

Here, $c_j(x) = \int_0^x n(\xi) d\xi$ represents the cumulative density function (CDF) of the NDF.

4.3. Division kernel

As mentioned in previous subsections, the division kernel $\beta(x,x')$ in Eq. (3) represents the expected number of daughter cells and their size distribution following the division of a mother cell. Thereby, the sizes of daughter cells are denoted by x and the size of the mother cell by x'.

For the microalgal species *C. reinhardtii*, it is widely observed that a mother cell can undergo multiple simultaneous cell divisions with the number of daughter cells always being a power of 2 (Harris et al., 2009; Donnan et al., 1985). The number of divisions is generally correlated to the size of the mother cell (Usai et al., 2023; Liu et al., 2023; Harris et al., 2009). According to the prevailing explanation (Liu et al., 2023; Harris et al., 2009; Craigie and Cavalier-smith, 1982), the division mechanism is initiated once a single cell reaches a certain size (*commitment*). After a delay period in which the cell still grows, alternate binary fission occurs within the cell wall, until the daughter cell sizes cross a different threshold. Finally, the cell wall bursts, and the daughter cells are released into the medium. The daughter cells then continue to grow. It is noted that alternative mechanisms have been proposed (e.g., see Shelton et al., 2017). One approach to describe this mechanism implements the division kernel $\beta(x,x')$ as a weighted sum of generalized Hill-Ng distributions (Hill and Ng, 1996; Diemer and Olson, 2002; Concas et al., 2016). The weights denoted are by $w_j(x')$, such that (Usai et al., 2023; Concas et al., 2016)

$$\beta(x,x') = \sum_{j=1}^{\infty} \frac{w_j(x') k_j}{B(p_j,q_j)} \frac{1}{x'} \left(\frac{x}{x'}\right)^{p_j-1} \left(1-\frac{x}{x'}\right)^{q_j-1}, \tag{6}$$

with $q_j = p_j(k_j - 1)$. In the above, $B(p_j,q_j)$ is the Euler beta function, k_j is the number of daughter cells (always a power of 2), and p_j is an integer parameter describing the variance in daughter cell sizes within a certain division number. For larger p_j, the daughter cell size range becomes more narrow. In the limit of p_j approaching infinity, this distribution approximates the Dirac delta distribution which indicates identical daughter cells upon division. On the condition that the sum of weights equals 1 for all x', the division kernel in Eq. (6) satisfies the usual requirement of conservation of mass (Ramkrishna, 2000). The expected number of daughter cells for division of a mother cell of size x' is given by

$$\bar{k}(x') = \sum_{j=1}^{\infty} = w_j(x') k_j, \tag{7}$$

such that the expected number of divisions is equal to $\log_2(\bar{k}(x'))$. Eqs. (6) and (7) are applicable for any valid choice of weights $w_j(x')$, i.e. $\sum_{j=1}^{\infty} w_j(x') = 1$ and $w_j(x') \geq 0$ for all x'. The infinite sum in Eq. (6) can be truncated to only a few terms to account for the expected number of divisions in practice, equivalent to a zero weighting at the respective number of daughter cells k_j.

The selection of weights $w_j(x')$ eventually determines the number of divisions taking place. In this text, two models are assumed, and further utilized to define a mathematical operator used to infer the division rate (Section 4.4.2). The first model of interest is the *binary fission model*, which only assumes a single division into two daughter cells. This corresponds to the division kernel in Eq. (6) with $k_1 = 2$ and $w_1(x') = 1$ for all x'. The only remaining parameter is p_1, describing the variance in the daughter cell sizes. The second model considered in this text is the *Noisy Power Law model* described by Liu et al. (2023), and models the mechanism described in previous paragraph. Here the weights $w_j(x')$ are given as

$$w_j(x') = \frac{1}{2}\left(\mathrm{erf}\left(\frac{\log_2(x'/x_c) - \log_2(k_j)}{\sqrt{2}\sigma_d}\right) - \mathrm{erf}\left(\frac{\log_2(x'/x_c) - \log_2(k_{j+1})}{\sqrt{2}\sigma_d}\right)\right), \tag{8}$$

where erf(·) is the error function, x_c is a critical commitment cell mass, and σ_d represents the standard deviation of the *noise* in number of cell divisions, and induces overlap in different regimes of cell division. The critical cell mass x_c acts as a threshold in the number of cell divisions based on the ratio $\frac{x'}{x_c}$, while σ_d softens this threshold in a probabilistic sense.

4.4. Division rate inference

In this section, a novel method to model the division rate from quasi-steady-state size distributions is presented. This is achieved by adapting standard Gaussian process regression (GPR) to the current problem setting, using the PBM introduced in Section 4.1. The mathematical details of the novel method are provided in the Supplementary Material. Below, a condensed overview is given, based on the works by Rasmussen and Williams (2006), Särkkä (2011), Solin and Särkkä (2020), Usai et al. (2023).

4.4.1. Gaussian process regression

GPR is a Bayesian regression technique to infer an unknown function $f(x)$ from (noisy) measurements $\mathbf{y}^* = f(\mathbf{x}^*) + \epsilon$, where ϵ corresponds to measurement noise, and is often taken as uncorrelated Gaussian distributed noise with variance σ_n^2. With this, GPR is evaluated as

$$f(x)|(\mathbf{y}^*, \mathbf{x}^*) \sim \mathcal{GP}\left(\tilde{m}(x;\boldsymbol{\theta}), \tilde{k}(x,x';\boldsymbol{\theta})\right). \tag{9}$$

Here, $\tilde{m}(x;\boldsymbol{\theta})$ and $\tilde{k}(x,x';\boldsymbol{\theta})$ are functions implicitly depending on $(\mathbf{y}^*, \mathbf{x}^*)$ (see Supplementary Material). The symbol $\boldsymbol{\theta}$ denotes so-called *hyperparameters* which are discussed below. Eq. (9) is to be interpreted as follows: based on the given data and chosen hyperparameters, the estimated function $f(x)$ is represented as a Gaussian process. This means that for a certain value of x at which to evaluate $f(x)$, the *mean* (best) estimate is obtained by evaluating $\tilde{m}(x;\boldsymbol{\theta})$. The *variance* (uncertainty) in this estimate is obtained by evaluating $\tilde{k}(x,x;\boldsymbol{\theta})$. Furthermore, the *covariance* between two estimated function values $f(x)$ and $f(x')$ is represented by $\tilde{k}(x,x';\boldsymbol{\theta})$.

The inference strongly depends on the choice of hyperparameters $\boldsymbol{\theta}$, which implicitly define the functions $\tilde{m}$ and $\tilde{k}$. These hyperparameters typically comprise only a few parameters. In this work, $\boldsymbol{\theta}$ consists of only 3 hyperparameters: the noise hyperparameter σ_n, and two hyperparameters associated to the chosen squared exponential kernel. A common criterion to select an appropriate set of hyperparameters is the *maximum a posteriori* (MAP) likelihood. This corresponds to the maximization of the *a posteriori* likelihood of the hyperparameters $\boldsymbol{\theta}$, specified as

$$p(\boldsymbol{\theta}|\mathbf{y}^*, \mathbf{x}^*) \propto p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}). \tag{10}$$

The first factor on the right-hand side is the *marginal likelihood* (i.e., how likely is it to have observed $\mathbf{y}^*$ if a specific set of hyperparameters $\boldsymbol{\theta}$ were used), and the second factor is the prior hyperparameter likelihood (*hyperprior*). Conventionally, the prior likelihood $p(\boldsymbol{\theta})$ is assumed uniform, such that MAP is equivalent to maximizing the marginal likelihood $p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta})$ (Rasmussen and Williams, 2006). Numerically, this is often achieved via gradient descent. In case the hyperprior is not uniform, this numerically corresponds to a type of regularization. Numerical optimization via gradient descent can be applied on the condition that the gradient of the prior hyperparameter likelihood with respect to $\boldsymbol{\theta}$ exists.

4.4.2. Gaussian process regression for division rate inference

Next, it is shown how this GPR framework can be adapted to infer the division rate $\gamma(x)$ given experimentally determined size distributions $n_{\text{exp}}^{(i)}(x)$ (where the superscript (i) denotes one measured NDF), average growth rate $\bar{\mu}_{\text{exp}}$, and a choice for division kernel $\beta(x,x')$. To achieve this, the PBM in Eqs. (3) and (4) is used and reformulated. First, a training data function $y^{(i)}(x)$ is defined as

$$y^{(i)}(x) := \bar{\mu}_{\text{exp}} n_{\text{exp}}^{(i)}(x) + \frac{\partial \mu_{\text{exp}}(x) x n_{\text{exp}}^{(i)}(x)}{\partial x}. \tag{11}$$

Second, a (linear) *division operator* $\mathcal{G}_x^{(i)}$ acting on the division rate $\gamma(x)$ is defined as

$$\mathcal{G}_x^{(i)}\gamma(x) := -n_{\text{exp}}^{(i)}(x)\gamma(x) + \int_x^\infty \beta(x,\xi) n_{\text{exp}}^{(i)}(\xi)\gamma(\xi) d\xi. \tag{12}$$

At least theoretically, in absence of noise and numerical errors, the training data should be identical to the division operator by definition, i.e. $y^{(i)}(x) = \mathcal{G}_x^{(i)}\gamma(x)$. In addition, one may have multiple measurements of the NDF. Here, GPR can be applied to infer the division operator $\mathcal{G}_x^{(i)}\gamma(x)$ from the training data $y^{(i)}(x)$. In turn, the division rate $\gamma(x)$ can be derived as $\mathcal{G}_x^{(i)}$ is a linear operator with respect to $\gamma(x)$, see Särkkä (2011) and the Supplementary Material for further information.

It remains to evaluate the hyperparameters $\boldsymbol{\theta}$ with this setup. For this, the MAP criterion in Eq. (10) is adapted. Unlike most GPR applications, where the prior hyperparameter likelihood $p(\boldsymbol{\theta})$ is assumed uniform, it is empirically found that a two-stage procedure is necessary to avoid overfitting. In the first stage, a prior hyperparameter likelihood is identified. This is obtained by first treating each experimentally measured NDF separately, and identify the hyperparameters $\boldsymbol{\theta}^{(i)}$ via maximum marginal likelihood. In the second stage, the prior hyperparameter likelihood $p(\boldsymbol{\theta})$ is defined as a multivariate normal distribution, with mean and covariance calculated from the numerical values of the different $\boldsymbol{\theta}^{(i)}$. The covariance is additionally scaled with a regularization constant. This corresponds to quadratic regularization. Then, the MAP criterion in Eq. (10) is evaluated using the entire dataset of experimentally measured NDFs.

5. Results

In this section, the results of the procedures described above are provided. First, the size distribution reconstruction is reviewed, including the regression of the calibration curve, and the consequent correction for the refractive index. The determined cell size distribution is validated with microscopy data. Second, the division rate $\gamma(x)$ inference is achieved by evaluating the GPR under different forms of the division kernel $\beta(x,x')$. A simulation utilizing the inferred division rate is compared with experimental data.

5.1. Size correction in flow cytometry

The main methodological results for the reconstruction of the size distribution are shown in Fig. 1. An example of the measured FCM data for *C. reinhardtii* and the reference polystyrene beads is depicted in Fig. 1a. The scattering measurements for *C. reinhardtii* are preprocessed using a Gaussian mixture model and assigned into two clusters, which are shown in yellow and green in Fig. 1a. Only the green datapoints (mostly clustered at higher FSC and SSC) are assumed to correspond to actual cells, and only these are used to construct the size distribution. In the purple data, corresponding to the reference polystyrene particles, six distinct peaks are identified (indicated with transparent purple circles). These correspond with the six different uniform reference sizes (see Section 2). The scattering intensity is positively correlated with particle diameter.

Regarding the calibration of the reference particles, Fig. 1c shows the corresponding result of the polynomial regression defined in Eq. (2). The figure indicates that the quadratic degree provides a sufficient regression accuracy, with a mean squared error of 5.9×10^{-2} μm^2.

Concerning the correction for the refractive index n, the parameters $I_{\text{inc,FSC}}$, $\theta_{1,FSC}$, and $\theta_{2,FSC}$ in Eq. (1) are calibrated to the polystyrene reference particle measurements. The FSC measurement angle range is found as $\left[\theta_{1,\text{FSC}}, \theta_{2,\text{FSC}}\right] = [2.6°, 9.0°]$ (initial range: [2.0°, 9.2°], as

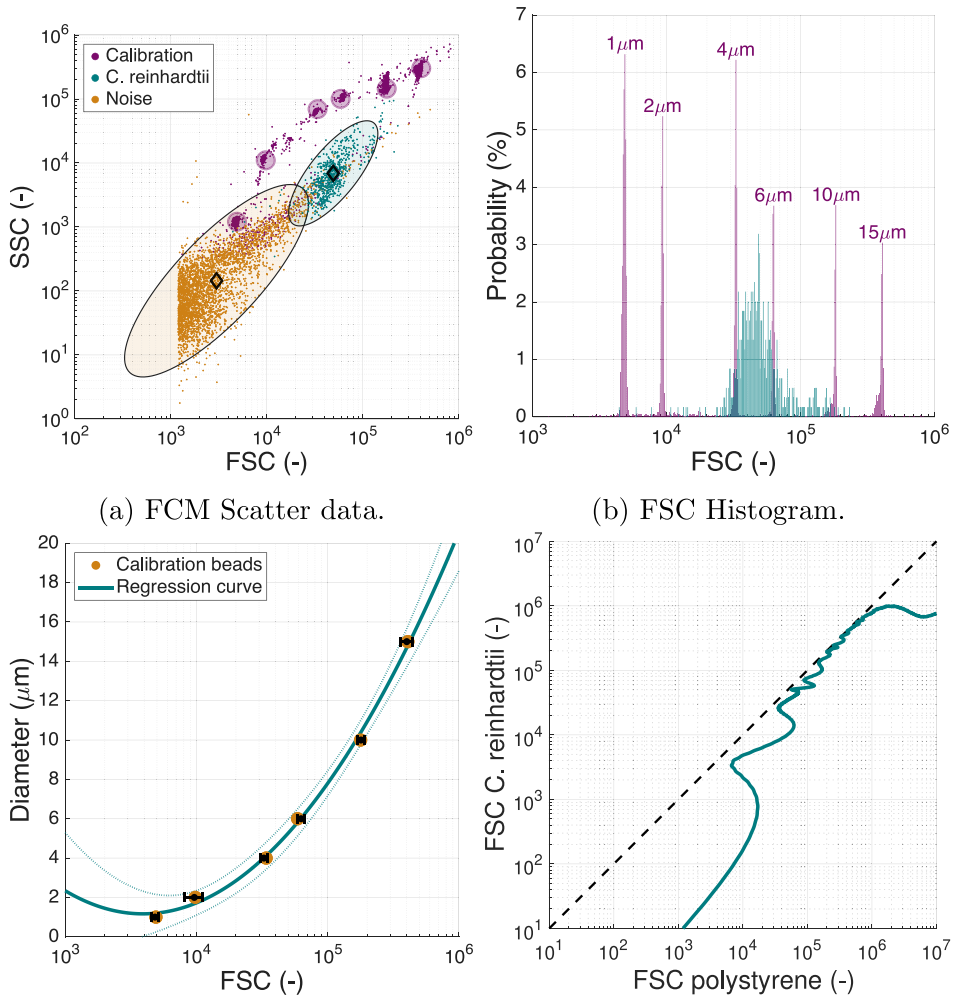


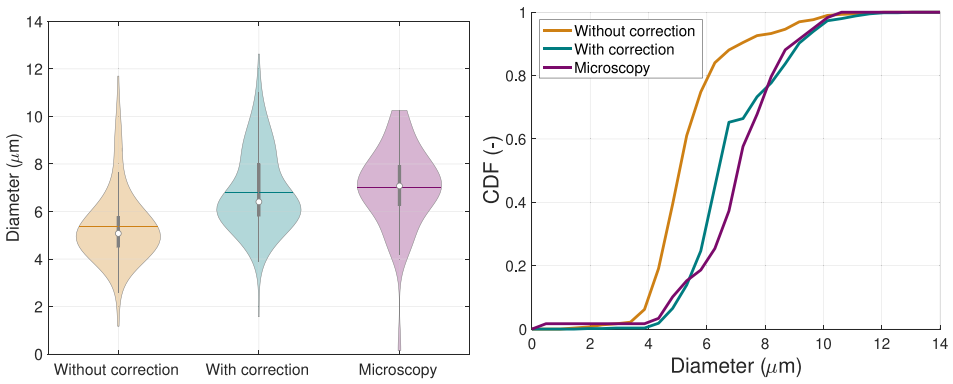
Fig. 1. Processing of flow cytometer data. **(a)** Measured forward and side scatter (FSC and SSC) of polystyrene size reference beads (purple), and on *C. reinhardtii*-culture. The latter is clustered into noise (yellow) and cells (green) using a Gaussian mixture model (GMM). The six purple larger transparent circles indicate identified peaks, representing the reference sizes. The yellow and green ellipses indicate the 95%-confidence region from the GMM clusters. **(b)** Histogram of FSC for calibration particles (purple) and *C. reinhardtii* cells (after GMM clustering). The six peaks in calibration FSC correspond to the six reference diameters. **(c)** Calibration curve (green solid line) and 95%-confidence (green dotted line) relating FSC of polystyrene beads (circles) to diameter, obtained via quadratic regression, as in Eq. (2). The horizontal error bars indicate the bounds of ± one standard deviation from the mean of calibration FSC data. **(d)** Equivalence curve (green) for spherical particles with refractive index n_C (y-axis) and n_{PS} (x-axis), obtained from Mie theory. The diagonal (black dashed line) is indicated as reference.

per Agagliate et al., 2018), and incident intensity $I_{\text{inc,FSC}} = 8.1 \times 10^{14}$ (initial guess: 1×10^{10}). Consequently, the equivalence curve is calculated from Eq. (1). This serves to convert FSC measurements from one material (*C. reinhardtii* cells) into *pseudo*-measurements of the other material (polystyrene), as shown in Fig. 1d. As the equivalence curve is the ratio of the $I_{\text{FSC}}(D, n)$ of two different values of n, the incident illumination $I_{\text{inc,FSC}}$ does not influence the equivalence curve and further processing. As the equivalence curve remains entirely to the right of the diagonal line, the FSC intensity for the *C. reinhardtii* cells needs to be increased for the calibration using Fig. 1c. In other words, without the current correction for refractive index, the particle size distribution would be underestimated. After data preprocessing, the *C. reinhardtii* measurements have an FSC intensity between 4.4×10^4 to 2.4×10^5.

Finally, it remains to derive the cell size distribution. For this, the FSC measurements are adjusted to an equivalent polystyrene FSC after preprocessing *pseudo*-measurement (Fig. 1d), and are then converted into an approximated particle diameter using the calibration (Fig. 1c). A histogram is composed of the occurrence frequency of particle diameters. This histogram is subsequently smoothened with a kernel density function estimation. This results in the experimentally constructed number density function $n_{\text{exp}}(x)$, which characterizes the size distribution. As can be seen from Fig. 2, applying the additional correction for FSC data improves the quantification of the size distribution, with the microscopy data as a reference for validation (based on 59 cell sizes identified in microscopy data versus 596 values in FCM). The NDF is shown in Fig. 2a, the CDF in Fig. 2b. This improvement is quantified using the Kullback–Leibler divergence (Kullback and Leibler, 1951), for which one obtains 6.5×10^{-1} nats and 0.9×10^{-1} nats for the uncorrected and corrected distribution, respectively. Similarly, the Earth Mover's Distance (EMD), defined as the integrated absolute error of the CDF in Fig. 2b, equals 1.7 μm and 0.5 μm, respectively.

5.2. Division rate inference

The procedure described in the previous section is repeated over the available datasets, resulting in a total of twelve NDFs $n_{\text{exp}}^{(i)}$ over a span



(a) Violin plot of number density functions. (b) Cumulative density functions.

Fig. 2. Comparison of reconstructed size distribution with size correction for refractive index (green) based on Mie theory, and without this correction (yellow), from $N = 596$ datapoints, validated with a size distribution measured from microscopy as reference (purple), from $N = 59$ datapoints. **(a)** The violin plot indicates the number density function vertically. The center line represents the mean value, the gray vertical lines indicate 25%, 50% (white circle) and 75%-percentiles. **(b)** Cumulative density plot of the size distributions. The Earth Mover's Distance between two distributions is interpreted as the area enclosed by their cumulative density functions.

of two days. Exponential growth is observed with a specific growth rate $\bar{\mu}_{\text{exp}} = 0.0215\,\text{h}^{-1}$.

In order to validate the inferred division rate, a forward simulation of the PBM is performed by solving Eqs. (3) and (4), substituting $\gamma(x)$ with the inferred mean of the GPR. In addition, the uncertainty in the GPR is propagated in the forward simulation, by random collection of a finite number of samples of $\gamma(x)$ and solving the PBM.

Two cases are discussed in the following. In the first, a binary division kernel is fixed to illustrate the general approach. In the second, a more comprehensive division kernel is used, of which the parameters are estimated by a grid-search technique.

5.2.1. Binary fission

First, the division kernel $\beta(x, x')$ is considered as a binary fission model (see Section 4.1), i.e., the general distribution in Eq. (6) with $k_1 = 2$ and $w_1(x') = 1$ for all x'. The spread parameter p_1 is fixed at 60. In total, twelve FCM measurements of the same culture are available. For each, the experimental NDF $n_{\text{exp}}^{(i)}$ is evaluated using the correction procedure as described and illustrated in the previous subsection. Of these twelve measurements, nine are randomly selected as training data, while the remaining three are reserved for validation.

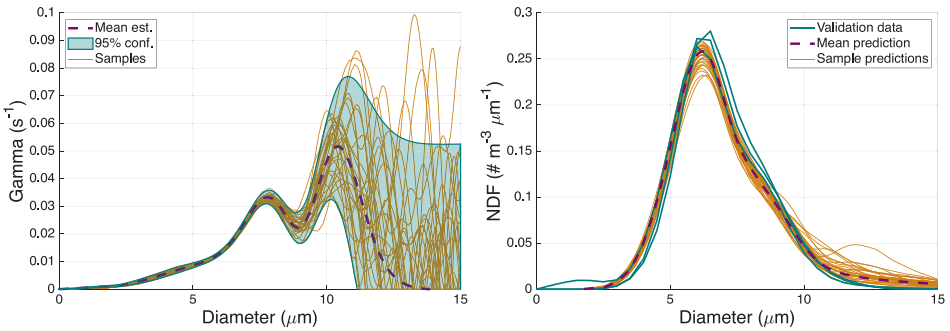
Using the corrected NDFs in the training data $n_{\text{exp}}^{(i)}$, the measured specific growth rate $\bar{\mu}_{\text{exp}}$, and the binary division kernel $\beta(x, x')$, the linear operators $\mathcal{G}_x^{(i)}$ and the training $y^{(i)}$ are calculated and used to define a GPR over the division rate $\gamma(x)$, as described in Section 4.4.2. Consequently, the hyperparameters are optimized in a two-step procedure. The result for the inferred division rate $\gamma(x)$ is shown in Fig. 3a. As GPR returns a (Gaussian) distribution over functions rather than a single best-fit description, this is reflected in the symmetric uncertainty intervals indicated in Fig. 3b. As visualization of the covariance over the cell mass x, a few random samples from the distribution are shown as well. It can be observed that the uncertainty rapidly increases for larger cell diameters, which follows from the NDF being close to 0 (which is, where the slope of the cumulative density function in Fig. 2b is close to 0). This reflects the fact that not much information on cell division beyond approximately 10 μm diameter can be derived from that data.

The inferred division rate $\gamma(x)$ is used in forward simulation, this is, the solution of Eqs. (3) and (4). The uncertainty in the division rate $\gamma(x)$ is propagated via finite random sampling from the inferred distribution, and evaluating the forward simulation for each sample. Fig. 3b illustrates the ability to model the validation data using the inferred division rate, and illustrates a relatively small variance in the simulated NDFs. This indicates that the high uncertainty in the division rate for larger cell sizes seen in Fig. 3a does not significantly propagate into the forward simulation. The fit between validation data and simulation results is quantified using the Earth Mover's Distance (EMD), defined in Eq. (5). The calculated EMD is averaged over the validation NDFs, for each sample simulation, resulting in a mean value of 3.1×10^{-1} μm with a standard deviation of 1.4×10^{-1} μm. The deviation in EMD can be seen qualitatively from Fig. 3b, as the samples appear as a *bundle* around the mean prediction. This corresponds to a certain degree of uncertainty quantification in the forward simulation. Finally, it is emphasized that the relative uniformity between experimentally measured NDFs can be interpreted as evidence for the assumption of stationary size distribution characterized by Eq. (4).

5.2.2. Selecting division kernel

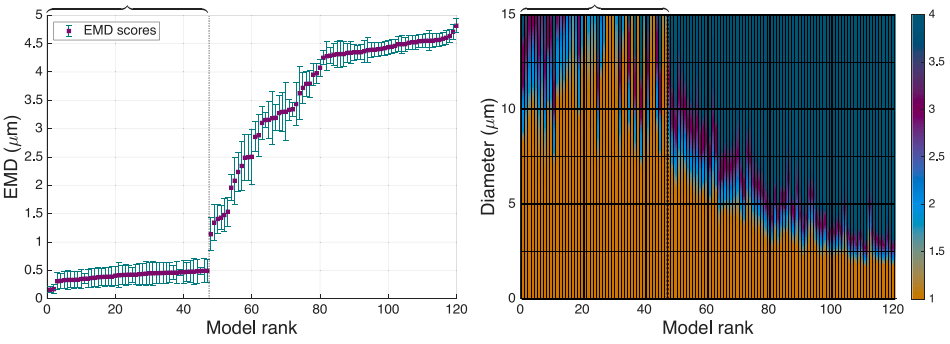
The previous results in this text assume a fixed binary division kernel. To consider the effect of the choice of division kernel, the approach presented above is applied to the division kernel according to Eqs. (6) and (8) for different parameters, with division into 2, 4, 8 or 16 daughter cells. Two parameters, x_c and σ_d, need to be determined in the weights defined in Eq. (8). With respect to the generalized Hill-Ng distribution spread parameters p_i, one could define a value for each number of divisions. Here, to reduce the number of free parameters only the spread for binary division p_1 is varied. For a higher division number, the value of p_i is chosen to match the variance of the binary division distribution according to $p_i = \left[\frac{1}{k_i}\left(\frac{k_1^2}{k_i^2}\frac{k_i-1}{k_1-1}(p_1 k_1 + 1) - 1\right)\right]$, where the square brackets denote the rounding to the nearest integer operation. Thus, in total three parameters need to be estimated: x_c, σ_d, and p_1. To define appropriate parameters, random sets are sampled over a Latin Hypercube Space. The EMD over the validation data set from 30 samples is used to evaluate performance. In addition, the two-step hyperparameter optimization is slightly modified; the covariance of hyperparameters calculated in the first step, is scaled down with a factor 50, as it is found necessary for hyperparameter optimization. This corresponds to tightening the soft regularization constraint induced by the hyperprior.

The resulting model performance in the forward simulation is quantified using the EMD, shown in Fig. 4a for different division kernel parameter values, and sorted according to their value. Fig. 4b indicates the expected number of cell divisions per diameter, as sorted per Fig. 4a. As indicated, the first 47 out of 120 correspond to a lower EMD value, and can qualitatively estimate the experimental NDFs well.



(a) Inferred division rate $\gamma(x)$ (b) Forward simulation of the number density function.

Fig. 3. Results from division rate inference using a binary division kernel. **(a)** Inferred division rate $\gamma(x)$ as Gaussian process with mean (purple dashed line) and 95%-confidence region (green shaded area). Several samples (yellow solid lines) are drawn from the distribution. **(b)** Simulated number density functions compared with experimentally constructed number density functions (green solid line). The mean function (purple dashed line) utilizes the mean division rate from (a), the sampled functions (yellow solid line) utilize the sampled division rates.



(a) EMD scores (b) Expected number of divisions

Fig. 4. Simulation results using randomized parameters in the division kernel (Eqs. (6) and (8)). **(a)** Sorted Earth Mover's Distance (EMD) scores (± one standard deviation) for the forward simulation with division rate constructed under different evaluated parameters in the division kernel. The first 47 models (indicated by the accolade and dotted line) correspond to a relative lower EMD value, and qualitative well fit. **(b)** Expected number of divisions as function of cell size under different values for division kernel parameters, sorted according to the corresponding EMD values in Fig. 4a, with the 47 *best* models indicated by the accolade and thick dotted line. The expected number of divisions corresponds to $\log_2 \bar{k}(x')$, specified in Eq. (7).

Based on Fig. 4b, this seemingly corresponds to division kernels in which the transition from binary to multiple divisions occurs at larger cell diameters. In contrast, parameters that correspond to a poor EMD score are associated to division kernels favoring multiple divisions at smaller cell sizes.

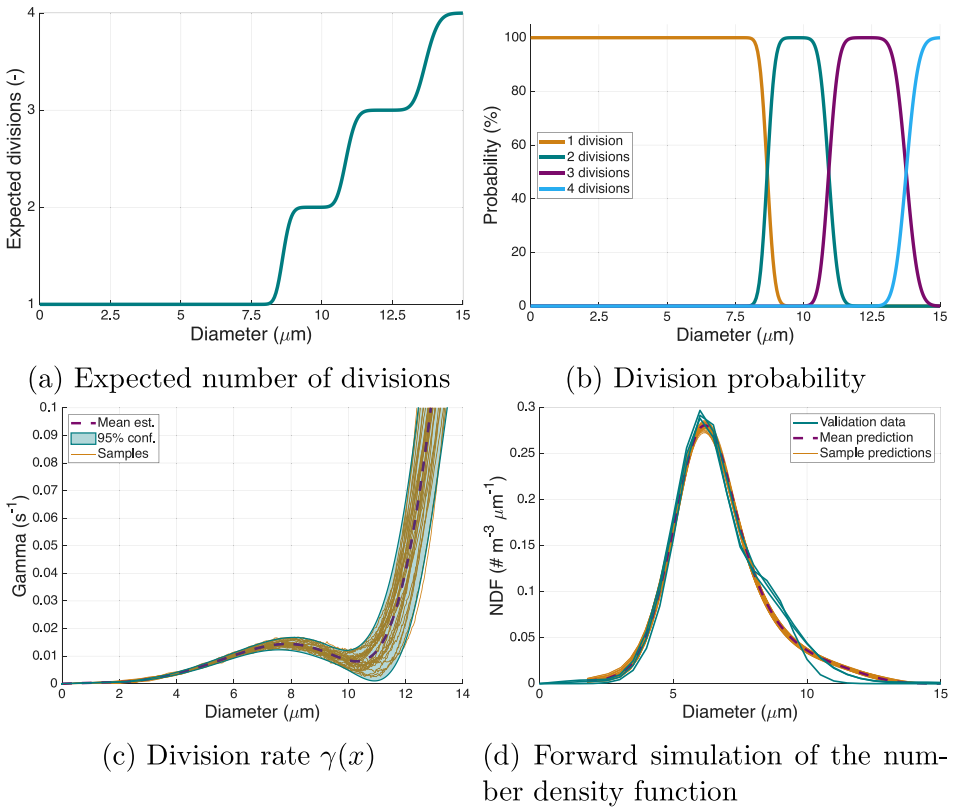
For illustration, the best performing division kernel parameters (as in Fig. 4a) correspond to the parameter values: $x_c = 98\,\text{pg}$, $\sigma_d = 0.11$, and $p_1 = 51$, and some results are shown in Fig. 5. In Fig. 5a, the expected number of divisions $\log_2 \bar{k}(x')$ is given, and as evidently, increases with an increasing cell diameter. Related to this, Fig. 5b shows the weight values $w_j(x')$ as defined in Eq. (8) for the set of parameters, and explains the pattern seen in Fig. 5a. Fig. 5c shows the corresponding inferred division rate. Its associated forward simulation of the NDF is given in Fig. 5d. Similar to the binary division case in Fig. 3a, the division rate is characterized by an increased prediction with respect to cell diameter. In addition, for larger cell sizes, the uncertainty in the inferred division rate increases due to the NDF approaching zero. For reference, with respect to the forward simulation, the EMD calculated for the binary fission model lies between the second and third best performing models in Fig. 4a. However, these models are characterized by a relative degree of variation; the mean EMD for the binary fission model lies within one standard deviation of the 43 division kernel models from Fig. 4a. In other words, the performance improvement of the best ranked model in Fig. 4a compared to the other 43 best ranked models is marginal.

6. Discussion

The overall goal of this work is twofold. First, the problem of obtaining the size distribution an algae cell culture is addressed (Section 3). This is shown to be achievable by adjusting measured FSC data with a correction based on Mie scattering theory. Second, the cell division rate is inferred based on the estimated size distribution (Section 4), for which the framework of Gaussian process regression (GPR) is used.

6.1. Size distribution reconstruction

Following Section 5.1, it is demonstrated that (*i*) the refractive index plays an important role in scattering measurements in FCM, and (*ii*) that this effect can partially be accounted for by considering an optical model based on Mie scattering. Essentially, the model describes expected scattering measurement as function of particle diameter D and refractive index n. Former studies (Agagliate et al., 2018), which serve as the basis of the current work, have used both FSC and SSC measurements to directly generate a lookup-table as function of particle diameter D and refractive index n, calibrated using the reference particles. This is, a pair of measured FSC and SSC is associated to a pair of D and n. However, this approach did not improve the calculated size distribution as compared to the method discussed here. This could be due to the use of SSC measurements, which are influenced by intracellular structures and compounds to a greater extent compared to



(a) Expected number of divisions

(b) Division probability

(c) Division rate $\gamma(x)$

(d) Forward simulation of the number density function

Fig. 5. Simulation results of the best performing model (i.e. rank 1) as in Fig. 4. **(a)** Expected number of divisions $\log_2 \bar{k}(x')$ as function of cell diameter. **(b)** Probability of a cell undergoing a certain number of divisions, corresponding to the weights $w_j(x')$ in Eq. (8). **(c)** Division rate $\gamma(x)$ as Gaussian process utilizing the division kernel corresponding to the maximum *a posteriori* likelihood. Mean (purple dashed line), 95%-confidence (green shaded area) and samples (yellow solid lines) are shown. **(d)** Simulated number density function using the mean function of division rate $\gamma(x)$ (purple dashed line) and samples (yellow solid lines), compared with experimentally determined number density functions (green solid lines), using the division rate shown in Fig. 5c.

FSC measurements (Adan et al., 2017; Tzur et al., 2011). In this aspect, by eliminating the reliance on SSC data, and taking fixed values for the refractive index n instead, the procedure is somewhat simplified. It is also more robust by removal of a degree of freedom and would be less affected by intracellular heterogeneity. In addition, the quantification of the diameter follows from the regression over the measurements, rather than inversion of Eq. (1). It is noted that microscopy data is used as validation benchmark, although this data in turn may not perfectly capture the true size distribution. An example of potential lack of representation is the sharper cutoff at larger cell sizes as seen in Fig. 2a.

Furthermore, Mie scattering theory assumes a perfectly spherical and perfectly homogeneous particle, neither of which applies exactly to *C. reinhardtii* cells. In particular, optically distinct cell compartments and compounds result in more convoluted scattering patterns. Addressing this would involve more advanced solution methods to evaluate the function in Eq. (1), and more advanced measurement techniques to characterize these compartments. In addition, this implies that the described methodology is restricted to (nearly) spherical microorganisms. For different cell shapes, not only does the scattering geometry change, but asymmetric cell orientation within the FCM has an additional impact. Still, the method can be applied in the same way to other broad-elliptical to spherical microalgae, which includes other biotechnologically important species such as *Chlorella vulgaris*, *Dunaliella salina*, or *Haematococcus pluvialis* (Borowitzka, 1999).

6.2. Division rate inference

As illustrated in Section 5.2, the approach of utilizing GPR with the appropriate modifications is able to infer a distribution over the division rate $\gamma(x)$. Moreover, the method enables one to accurately describe the observed size distribution using a forward simulation of the PBM. The inherent uncertainty quantification in GPR is propagated via random sampling. The performance of forward simulation is quantified using the EMD. Some additional remarks are discussed below.

As described in Section 4.4.2, the hyperparameter optimization procedure is performed in two steps. The first step considers each measured NDF individually, and produces a so-called hyperprior (a prior distribution on the hyperparameters). In the second step, this prior distribution is utilized in optimizing the posterior distribution. This approach varies somewhat from the more commonly used log marginal likelihood maximization (Rasmussen and Williams, 2006), in the sense that an additional regularization term is added to the log marginal likelihood objective. This additional regularization ensures that optimal hyperparameters are obtained, which perform well in forward simulation. Without the additional hyperprior regularization, multiple local optima are observed in the log marginal likelihood, but which may result in an inaccurate prediction in the forward simulation. With respect to the influence of the hyperprior, the degree of regularization induced by this term can be controlled by scaling the hyperparameter covariance. This has been found to be necessary for more complex division kernels (6) to prevent overfitting hyperparameters that are optimal in marginal likelihood but suboptimal in forward simulation performance. The advantage of working with the MAP objective lies in its computational efficiency due to the practical use of gradient descent.

For best performance of the division rate estimation, an appropriate division kernel needs to be selected. As illustrated, this can to some extent be optimized for by computing the Gaussian process under different candidate division kernels. It is shown that the model performance in the forward simulation can (in some cases) be greatly affected by

the choice of division kernel. At the same time, and at least with the current experimental observations, a good model–data fit can be obtained for different division kernels, within a certain range. Within this range, identifiability of the *correct* division kernel could be considered challenging. This can to some extent be attributed to the black-box nature of Gaussian process regression. As other studies (Liu et al., 2023) have shown, one can experimentally determine parameters allowing to construct suitable division kernels, in particular for multiple-dividing organisms as *C. reinhardtii*, from single cell tracking data.

The modeled division weights $w_j(x')$ are shown in Fig. 5b for the best performing model, and more generally with the expected number of divisions in Fig. 4b. These results can be compared with the experimental findings reported by Liu et al. (2023), who inferred division probabilities from single-cell tracking data with a stochastic cell division model. Despite the differences in the modeling approach which prevent a direct comparison, there is at least some quantitative overlap between these results: in the study done here, a division in 4 daughter cells occurs mostly between around 8 to 11 μm diameter, corresponding to a sphere volume of 270 to 700 μm^3, while in Liu et al. (2023) a division in 4 daughter cells was inferred to be most likely at 400 μm^3, with non-zero probability from 200 to 800 μm^3. On the other hand, Liu et al. (2023) observe basically no divisions for volumes less than 200 μm^3, with only a small probability peak for binary division centered at around 270 μm^3, whereas the estimation results here already indicate a non-negligible cell division rate from around 4 μm diameter (ca. 35 μm^3). Moreover, in Liu et al. (2023) a division in 4 daughter cells seems the most frequent case, whereas the estimation performed here would indicate binary fission as the most common case, since most cells in the dataset are between 4 and 7 μm in diameter for which the expected number of divisions is 1. Possibly the differences are due to different growth media and strains, as well as a diurnal light cycle being used in Liu et al. (2023), whereas this work uses a permanent light supply.

6.3. Relevance to algal bioprocesses

In recent years, it has become clear that cellular heterogeneity, cell size distributions, and cell cycle and division mechanisms are crucial factors for systems analysis and engineering of algal bioprocesses (Bernard, 2011; Concas et al., 2016; Usai et al., 2023; Chioccioli et al., 2014). In addition to heterogeneity in cell sizes, a wide heterogeneity in cellular concentrations of interesting target products such as lipids or central carbohydrate metabolites has been observed (Garz et al., 2012; Di Caprio et al., 2018). Even if in these studies, the heterogeneity in the target metabolite seemed not to be correlated to cell size, it is important to distinguish variability due to changes in cell size, with mostly unchanged volumetric concentrations in each cell, from variability in the actual target concentration. Obviously, the determination of an accurate target product concentration per cell requires an absolute cell size measurement, which in the above mentioned studies have been obtained from a Coulter counter. Here, a method for absolute cell size measurements based on flow cytometry is developed, which is of interest since flow cytometry can assess additional parameters on the single cell level in a simultaneous measurement, for example additional cell morphology through side scatter or fluorescence signals from chlorophyll autofluorescence or fluorescently labeled target products. Generally, flow cytometry has been proposed as a suitable tools also for online monitoring of algal bioprocesses, and the methods developed in this study provide a promising way to link the obtained data to quantitative models including single-cell heterogeneity.

7. Conclusions

In this study, two relevant challenges in algal biotechnology are addressed. First, a method to improve the estimation of cell size distribution from flow cytometry data is outlined. The measured scattering intensity is adjusted based on Mie scattering theory, and further combined with regression to obtain cell size. It is shown that the cell size approximation can be improved significantly by this approach. Second, using experimental measurements in combination with the size distribution reconstruction procedure, a Gaussian process regression based on a population balance model is implemented to model the division rate. The inferred division rate matches experimental data at good accuracy. Two different division kernel models are compared, one assuming binary fission and the other permitting multiple fission as is common in microalgae. Although the estimation results show a good fit for different division kernels, in the considered process setup they favor models with mostly binary fission.

CRediT authorship contribution statement

Michiel Busschaert: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis. **Michael Schagerl:** Writing – review & editing, Supervision, Resources, Methodology. **Christian Griebler:** Writing – review & editing, Supervision, Resources, Methodology. **Florence H. Vermeire:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Steffen Waldherr:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been funded by the Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO) with grant numbers G066621N, V439223N, and K204325N.

In addition, we would like to express our gratitude to Alexander Kaptejna and Grit Rasch (both at the University of Vienna) for their assistance with algae cultivation and flow cytometry, respectively.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.biortech.2025.133835.

Data availability

The microscopy and flow cytometry data used in this manuscript are available at https://github.com/MichielBusschaert/DataInferringCellDivision.

References

Abdelfattah, A., et al., 2023. Microalgae-based wastewater treatment: Mechanisms, challenges, recent advances, and future prospects. Environ. Sci. Ecotechnology 13, http://dx.doi.org/10.1016/j.ese.2022.100205, Art. no. 100205.

Adan, A., Alizada, G., Kiraz, Y., Baran, Y., Nalbant, A., 2017. Flow cytometry: Basic principles and applications. Crit. Rev. Biotechnol. 37 (2), 163–176. http://dx.doi.org/10.3109/07388551.2015.1128876.

Agagliate, J., Röttgers, R., Twardowski, M.S., McKee, D., 2018. Evaluation of a flow cytometry method to determine size and real refractive index distributions in natural marine particle populations. Appl. Opt. 57 (7), 1705–1716. http://dx.doi.org/10.1364/AO.57.001705.

Atzori, F., Jerono, P., Schaum, A., Baratti, R., Tronci, S., Meurer, T., 2021. Identification of a cell population model for algae growth processes. IFAC-PapersOnLine 54 (7), 132–137. http://dx.doi.org/10.1016/j.ifacol.2021.08.347.

Bernard, O., 2011. Hurdles and challenges for modelling and control of microalgae for CO2 mitigation and biofuel production. J. Process Control 21 (10), 1378–1389. http://dx.doi.org/10.1016/j.jprocont.2011.07.012.

Bertucco, A., Sforza, E., Fiorenzato, V., Strumendo, M., 2015. Population balance modeling of a microalgal culture in photobioreactors: Comparison between experiments and simulations. AIChE J. 61 (9), 2702–2710. http://dx.doi.org/10.1002/aic.14893.

Bohren, C.F., Huffman, D.R., 1998. Absorption and Scattering of Light by Small Particles. John Wiley & Sons, ISBN: 978-3-527-61815-6, http://dx.doi.org/10.1002/9783527618156.

Borowitzka, M.A., 1999. Commercial production of microalgae: ponds, tanks, tubes and fermenters. J. Biotech. 70 (1–3), 313–321. http://dx.doi.org/10.1016/S0168-1656(99)00083-8.

Busschaert, M., Vermeire, F.H., Waldherr, S., 2023. Modeling cell size distribution with heterogeneous flux balance analysis. IEEE Control. Syst. Lett. 7, 1903–1908. http://dx.doi.org/10.1109/LCSYS.2023.3282699.

Chioccioli, M., Hankamer, B., Ross, I.L., 2014. Flow cytometry pulse width data enables rapid and sensitive estimation of biomass dry weight in the microalgae *Chlamydomonas reinhardtii* and *Chlorella vulgaris*. PLoS One 9 (5), http://dx.doi.org/10.1371/journal.pone.0097269, Art. no. e97269.

Concas, A., Pisu, M., Cao, G., 2016. A novel mathematical model to simulate the size-structured growth of microalgae strains dividing by multiple fission. Chem. Eng. J. 287, 252–268. http://dx.doi.org/10.1016/j.cej.2015.11.021.

Craigie, R.A., Cavalier-smith, T., 1982. Cell volume and the control of the *chlamydomonas* cell cycle. J. Cell Sci. 54 (1), 173–191. http://dx.doi.org/10.1242/jcs.54.1.173.

Di Caprio, F., Pagnanelli, F., Wijffels, R.H., Van Der Veen, D., 2018. Quantification of *Tetradesmus obliquus* (Chlorophyceae) cell size and lipid content heterogeneity at single-cell level. In: Graham, L. (Ed.), J. Phycol. 54 (2), 187–197. http://dx.doi.org/10.1111/jpy.12610.

Diemer, R., Olson, J., 2002. A moment methodology for coagulation and breakage problems: Part 3—Generalized daughter distribution functions. Chem. Eng. Sci. 57 (19), 4187–4198. http://dx.doi.org/10.1016/S0009-2509(02)00366-4.

Donnan, L., Carvill, E.P., Gilliland, T.J., John, P.C.L., 1985. The cell cycles of *Chlamydomonas* and *Chlorella*. New Phytol. 99 (1), 1–40. http://dx.doi.org/10.1111/j.1469-8137.1985.tb03634.x.

Droop, M.R., 1968. Vitamin B_{12} and marine ecology. IV. The kinetics of uptake, growth and inhibition in *Monochrysis lutheri*. J. Mar. Biol. Assoc. UK 48 (3), 689–733. http://dx.doi.org/10.1017/S0025315400019238.

Garz, A., Sandmann, M., Rading, M., Ramm, S., Menzel, R., Steup, M., 2012. Cell-to-cell diversity in a synchronized *Chlamydomonas* culture as revealed by single-cell analyses. Biophys. J. 103 (5), 1078–1086. http://dx.doi.org/10.1016/j.bpj.2012.07.026.

Gayen, K., Venkatesh, K., 2008. Quantification of cell size distribution as applied to the growth of *Corynebacterium glutamicum*. Microbiol. Res. 163 (5), 586–593. http://dx.doi.org/10.1016/j.micres.2006.07.017.

Gong, A., Min, M., 2022. Measuring the size and growth of single cells. Biophys. Rep. 8 (3), 150–157. http://dx.doi.org/10.52601/bpr.2022.210036.

Harris, E.H., Stern, D.B., Witman, G., 2009. The *Chlamydomonas* Sourcebook, second ed. Elsevier, ISBN: 978-0-12-370873-1.

Hempel, F., et al., 2011. Microalgae as bioreactors for bioplastic production. Microb. Cell Factories 10, http://dx.doi.org/10.1186/1475-2859-10-81, Art. no. 81.

Hill, P.J., Ng, K.M., 1996. Statistics of multiple particle breakage. AIChE J. 42 (6), 1600–1611. http://dx.doi.org/10.1002/aic.690420611.

Jiao, H., et al., 2024. Recent advances in sustainable hydrogen production from microalgae: Mechanisms, challenges, and future perspectives. Ecotoxicol. Environ. Safety 270, http://dx.doi.org/10.1016/j.ecoenv.2023.115908, Art. no. 115908.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86. http://dx.doi.org/10.1214/aoms/1177729694.

Küper, A., Durr, R., Waldherr, S., 2019. Dynamic density estimation in heterogeneous cell populations. IEEE Control. Syst. Lett. 3 (2), 242–247. http://dx.doi.org/10.1109/LCSYS.2018.2847905.

Lee, E., Jalalizadeh, M., Zhang, Q., 2015. Growth kinetic models for microalgae cultivation: A review. Algal Res. 12, 497–512. http://dx.doi.org/10.1016/j.algal.2015.10.004.

Leong, Y.K., Chang, J.-S., 2022. Bioprocessing for production and applications of bioplastics from algae. In: Algae-Based Biomaterials for Sustainable Development. Elsevier, ISBN: 978-0-323-96142-4, pp. 105–132. http://dx.doi.org/10.1016/B978-0-323-96142-4.00008-7.

Liu, D., Vargas-García, C.A., Singh, A., Umen, J., 2023. A cell-based model for size control in the multiple fission alga *Chlamydomonas reinhardtii*. Curr. Biology 33 (23), 5215–5224. http://dx.doi.org/10.1016/j.cub.2023.10.023.

Mantzaris, N.V., Daoutidis, P., 2004. Cell population balance modeling and control in continuous bioreactors. J. Process Control 14 (7), 775–784. http://dx.doi.org/10.1016/j.jprocont.2003.12.001.

Mehariya, S., Goswami, R.K., Karthikeysan, O.P., Verma, P., 2021. Microalgae for high-value products: A way towards green nutraceutical and pharmaceutical compounds. Chemosphere 280, 130553. http://dx.doi.org/10.1016/j.chemosphere.2021.130553.

Motz, S., Mitrovic, A., Gilles, E.-D., 2002. Comparison of numerical methods for the simulation of dispersed phase systems. Chem. Eng. Sci. 57 (20), 4329–4344. http://dx.doi.org/10.1016/S0009-2509(02)00349-4.

Moutier, W., et al., 2016. Scattering of individual particles from cytometry: Tests on phytoplankton cultures. Opt. Express 24 (21), http://dx.doi.org/10.1364/OE.24.024188, Art. no. 24188.

Nieto-Vesperinas, M., 2020. Fundamentals of Mie scattering. In: Dielectric Metamaterials. Elsevier, ISBN: 978-0-08-102403-4, pp. 39–72. http://dx.doi.org/10.1016/B978-0-08-102403-4.00007-4.

Orlova, D.Y., et al., 2016. Earth mover's distance (EMD): A true metric for comparing biomarker expression levels in cell populations. PLoS One 11 (3), http://dx.doi.org/10.1371/journal.pone.0151859, Art. no. e0151859.

Pringsheim, E.G., 2016. Pure Cultures of Algae: Their Preparation and Maintenance. Cambridge University Press, ISBN: 978-1-316-61320-7.

Qi, H., He, Z.-Z., Zhao, F.-Z., Ruan, L.-M., 2016. Determination of the spectral complex refractive indices of microalgae cells by light reflectance - transmittance measurement. Int. J. Hydrog. Energy 41 (9), 4941–4956. http://dx.doi.org/10.1016/j.ijhydene.2016.01.112.

Ramkrishna, D., 2000. Population Balances: Theory and Applications to Particulate Systems in Engineering. Academic Press, ISBN: 978-0-08-053924-9, http://dx.doi.org/10.1016/B978-0-12-576970-9.X5000-0.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, ISBN: 0-262-18253-X.

Särkkä, S., 2011. Linear operators and stochastic partial differential equations in Gaussian process regression. In: Honkela, T., Duch, W.o.a., Girolami, M., Kaski, S. (Eds.), Artificial Neural Networks and Machine Learning, Vol. 6792. ICANN 2011, Springer Berlin Heidelberg, ISBN: 978-3-642-21738-8, pp. 151–158. http://dx.doi.org/10.1007/978-3-642-21738-8_20.

Schagerl, M., 2022. Algae: Their world explored. In: El-Sheekh, M.M., Abdullah, N., Ahmad, I. (Eds.), Practice, Progress, and Proficiency in Sustainability. IGI Global, ISBN: 978-1668424384, pp. 1–31. http://dx.doi.org/10.4018/978-1-6684-2438-4.ch001.

Schagerl, M., Siedler, R., Konopáčová, E., Ali, S.S., 2022. Estimating biomass and vitality of microalgae for monitoring cultures: A roadmap for reliable measurements. Cells 11 (15), http://dx.doi.org/10.3390/cells11152455, Art. no. 2455.

Shelton, D.E., Leslie, M.P., Michod, R.E., 2017. Models of cell division initiation in *Chlamydomonas*: A challenge to the consensus view. J. Theoret. Biol. 412, 186–197. http://dx.doi.org/10.1016/j.jtbi.2016.10.018.

Solin, A., Särkkä, S., 2020. Hilbert space methods for reduced-rank Gaussian process regression. Stat. Comput. 30 (2), 419–446. http://dx.doi.org/10.1007/s11222-019-09886-w.

Tamburic, B., Zemichael, F.W., Maitland, G.C., Hellgardt, K., 2011. Parameters affecting the growth and hydrogen production of the green alga *Chlamydomonas reinhardtii*. Int. J. Hydrog. Energy 36 (13), 7872–7876. http://dx.doi.org/10.1016/j.ijhydene.2010.11.074.

Tronci, S., Schaum, A., 2022. Modeling microalgae cell mass distributions using the Fokker–Planck equation. IFAC-PapersOnLine 55 (26), 78–83. http://dx.doi.org/10.1016/j.ifacol.2022.10.380.

Tzur, A., Moore, J.K., Jorgensen, P., Shapiro, H.M., Kirschner, M.W., 2011. Optimizing optical flow cytometry for cell volume-based sorting and analysis. PLoS One 6, http://dx.doi.org/10.1371/journal.pone.0016053, Art. no. e16053.

Usai, A., Theodoropoulos, C., Di Caprio, F., Altimari, P., Cao, G., Concas, A., 2023. Structured population balances to support microalgae-based processes: Review of the state-of-art and perspectives analysis. Comput. Struct. Biotechnol. J. 21, 1169–1188. http://dx.doi.org/10.1016/j.csbj.2023.01.042.

Vítová, M., et al., 2011. *Chlamydomonas reinhardtii*: Duration of its cell cycle and phases at growth rates affected by light intensity. Planta 233, 75–86. http://dx.doi.org/10.1007/s00425-010-1282-y.

Waldherr, S., 2018. Estimation methods for heterogeneous cell population models in systems biology. J. R. Soc. Interface 15 (147), http://dx.doi.org/10.1098/rsif.2018.0530, Art. no. 20180530.