Inference on an interacting diffusion system with application to in vitro glioblastoma migration

Mathematical Medicine and Biology: A Journal of the IMA (2024) **41**, 250–276
https://doi.org/10.1093/imammb/dqae010
Advance Access publication on 13 August 2024

Inference on an interacting diffusion system with application to *in vitro* glioblastoma migration

Gustav Lindwall*
Chalmers tvärgata 3, 412 58 Gothenburg, Sweden
*Corresponding author. Email: lindwall@evolbio.mpg.de

and

Philip Gerlee
Chalmers tvärgata 3, 412 58 Gothenburg, Sweden

[Received on 31 August 2021; revised on 16 July 2024; accepted on 9 August 2024]

Glioblastoma multiforme is a highly aggressive form of brain cancer, with a median survival time for diagnosed patients of 15 months. Treatment of this cancer is typically a combination of radiation, chemotherapy and surgical removal of the tumour. However, the highly invasive and diffuse nature of glioblastoma makes surgical intrusions difficult, and the diffusive properties of glioblastoma are poorly understood. In this paper, we introduce a stochastic interacting particle system as a model of *in vitro* glioblastoma migration, along with a maximum likelihood-algorithm designed for inference using microscopy imaging data. The inference method is evaluated on *in silico* simulation of cancer cell migration, and then applied to a real data set. We find that the inference method performs with a high degree of accuracy on the *in silico* data, and achieve promising results given the *in vitro* data set.

Keywords: agent based modelling; mathematical biology; glioblastoma; diffusion; statistical inference.

1. Introduction

Glioblastoma multiforme is a brain tumour characterized by a diffuse and highly invasive growth. In particular, a high degree of variability in the invasive behaviour has been observed in different patients, leading to a hypothesis that there is a genetic component that determines whether the tumour exhibits a diffuse or solid morphology. Urbańska *et al.* (2014) Since the 1980s, the use of mathematical modelling to aid in the analysis of tumours have steadily grown as a field, and mathematical oncology is today an important tool for physicians treating cancer patients Anderson & Maini (2018); Hamis *et al.* (2019). On a macroscopic level, a glioblastoma tumour is characterized by two main features; the proliferation rate and the cell diffusivity. Both of these features are emergent phenomena stemming from complex dynamics at the cell level Skog *et al.* (2008). Up until recently single cell studies were hard to conduct, but with further advancements in microscopy technology and image analysis individual cells can now be tracked with a high degree of precision. Cell tracking technology has made it possible to fit agent based models (also referred to as individual-based) to *in vitro* and *in vivo* data. Historically, partial differential equations (PDEs) of the Fisher–Kolmogorov type have been the starting point when discussing models of tumour growth Tracqui *et al.* (1995). In its most basic form, the normalized density of a tumour $u(\mathbf{x}, t)$ at a point $\mathbf{x} \in \mathbf{R}^3$ and a time $t \geq 0$ is determined by

$$\partial_t u = D\Delta u + ru(1-u). \tag{1.1}$$

Here D is a diffusion coefficient, related to the macroscopic cell migration rate, and r is a growth rate, related to the cellular proliferation rate. We see that as $u \to 1$, the growth rate vanishes, and the equation locally becomes of the pure diffusion type. This equation is usually solved on a compact domain Ω representing the brain anatomy featuring Neumann boundary conditions and has an initial distribution u_0, which is of compact support. An interesting emergent phenomena is travelling wave solutions to (1.1). In the context of glioblastoma modelling, such travelling wave solutions have been studied by e.g. Gerlee and Nelander in Gerlee & Nelander (2016). Equations such as Fisher–Kolmogorov and the wider class of convection–reaction–diffusion equations it belong to, are not the only partial differential equation approach taken when modelling tumour growth. Multiple authors have have modelled cell migration phenomena using Boltzmann-like equations, where the evolution of cell velocity is considered the driving factor behind migration. For examples of such approaches, see Othmer & Hillen (2000) and Painter & Hillen (2013). While essentially phenomenological in nature, (1.1) has a deep connection as a limit result in the theory of random walks Oelschläger (1989). Thus, to study tumour growth and how the diffusivity depend on microscopic dynamics in greater detail, agent based models are a natural extension of the continuum approach of (1.1). Agent based models can be divided into two chief paradigms; lattice and off-lattice models. Historically, lattice-based models have been common within mathematical oncology, see e.g. Johnston *et al.* (2017). The discrete nature of lattices make for relatively easy study *in silico* and by considering the limit of an infinitesimal lattice, one can establish a natural connection to Fisher–Kolmogorov type equations under suitable circumstances Davies *et al.* (2014). Off-lattice models are usually based on the theory of Brownian random walks, and are usually modelled using stochastic differential equations (SDEs) Turelli (1977). The calculus of stochastic differentials is well established, and the framework has been used in physics and finance to study continuous and noisy phenomena since the turn of the century Lemons & Gythiel (1997); Einstein (1905) and 1970s Black & Scholes (2019), respectively. The greatest breakthrough was perhaps when stochastic calculus was made rigorous with the invention of Itô calculus in the 1940s Itô (1944). The connection between stochastic calculus and diffusion equations is established through the Fokker–Planck equation and Itô calculus Klebaner (2012).

1.1 *Earlier work in inference on tumour models*

The problem we consider is if the migration of glioblastoma cells cultured *in vitro* can be described using a system of coupled SDEs, and if the parameters of the model can be inferred from time-lapse microscopy data. This type of tracking-inference problem has been studied to great lengths in the Bayesian community Arulampalam *et al.* (2002), but few one-size-fits-all solutions exists to anything but the most basic problems. One of the earliest studies into estimating the parameters of the Fisher's equation for glioblastoma growth was conducted by Swanson *et al.* Swanson *et al.* (2000) in 2000, but that study did not consider single-cell tracking, but rather used population density data to fit the model. In a fairly early study from 2009 Tremel *et al.* (2009), Tremel *et al.* used microscopy data to fit a modified Fisher equation, utilizing single cell tracking. The population growth parameters were fitted using cell counting, and the diffusion term was inferred by measuring cell speed, comparing it to the characteristic speed given by the travelling wave solution of the Fisher equation. In a study from 2015 Johnston *et al.* (2015), Johnston *et al.* used microscopy data to fit parameters of the Fisher's equation directly, bypassing the tracking of individuals in favor of looking at population level dynamics. In the same research group, Lagergren *et al.* recently used a neural network approach to fit an extended Fisher-type equation to data from a scratch essay Lagergren *et al.* (2020). For a modern review of machine learning and nonlinear mixed effect models for Fisher-like equations, we refer to Everett *et al.* (2020). Individual based, on-lattice models have however been studied in great lengths. In 2014 Johnston *et al.* (2014), Approximate

Bayesian Computation (ABC) Tavaré *et al.* (1997) was used to derive posterior distributions of D and r in (1.1), using the radial distribution function (see (3.1)–(3.2)) as summary statistics. In 2017 Browning *et al.* (2018), Browning *et al.* used a lattice-free approach to fit a random walk model of cancer migration. The underlying stochastic process in this case was however a Poisson point process, and not the Brownian motion approach typically employed. Once again, ABC was the inference method of choice, and the summary statistics considered was the radial distribution function. However, the authors of Johnston *et al.* (2014) and Browning *et al.* (2018) have recently considered SDE driven models Browning *et al.* (2020), and cite a number of works on inference of SDEs. Among them we find the work conducted by Brückner *et al.* (2020) regarding inference in Langevin-type equations (see (1.3)–(1.4) below for a sketch of such equations), and the work of Schnoerr *et al.* Schnoerr *et al.* (2016) on stochastic reaction–diffusion processes, which has a connection to stochastic partial differential equations in the mean field limit. The data on which we base this study is microscopy imaging where single cells are clearly distinguishable, and so we chose to model the cell population as a system of stochastic differential equations, and make use of the Fokker–Planck equation for this system to find an expression for the likelihood function. We then find the maximum of this likelihood using an SMC-within-Gibbs Schön & Lindsten (2015) approach, using the mode of our posterior distribution as a point estimate of the parameters in our system. A fully Bayesian evolution of this algorithm can be adopted with some additional work, and is the subject of future research.

1.2 *Biological mechanism behind cell migration*

One can divide the means of locomotion in cell migration into two categories, external and individual factors. External factors include chemotaxis and the extra-cellular matrix (ECM), both of which have been studied extensively Hillen & Painter (2009); Chauviere *et al.* (2007). Cell migration involves processes at several length and time scales and is therefore inherently difficult to describes succinctly. While single cells are by default the simplest form of life, there is still a complex set of chemical and physical factors behind their migration, many that are poorly understood. For example, when a cell is close to division (mitosis), the motility of the cell has been observed to decrease. This has led to concepts such as the go-or-grow model Gerlee & Nelander (2012). The mode of migration of a cell has biomechanical explanations on the individual cell level that is a field of research in its own Malik & Gerlee (2019); Bodor *et al.* (2020). In practice, however, the migrating behaviour on a larger scale is described by some reasonably tractable stochastic process, as is commonplace in ecological models. Common choices of stochastic processes include different kinds of persistent random walks, and their diffusion limit in Brownian motion Othmer & Hillen (2000). If we denote a single cells location at time t as $x(t)$, Brownian motion is expressed on SDE form as

$$\mathrm{d}x(t) = \sigma \mathrm{d}W_t. \tag{1.2}$$

In a persistent random walk, the stochastic component is instead applied to the velocity of the cell, $v(t)$. It obeys the system of SDEs

$$\mathrm{d}x(t) = v(t)\,\mathrm{d}t, \tag{1.3}$$

$$\mathrm{d}v(t) = -av(t)\,\mathrm{d}t + b\,\mathrm{d}W(t), \tag{1.4}$$

where $a > 0$ is to be interpreted as friction, acting as a braking influence on the velocity process. Persistent random walks have a strong foundation in statistical mechanics, and the pathwise behaviour

TABLE 1 *Most of the notation used in this paper*

Variable	Explanation
$\mathbf{x}(t)$	The locations of all cancer cells at time t
i, j	Used as indices for individual cells
$\mathbf{x}_i(t)$	The location of cell i at time t
θ	A vector containing the parameters for the interaction potential (2.1).
σ	A vector containing the diffusion coefficients for every cell.
K	The total number of images in our data set.
k	Used to talk about individual images
t_k	The time image k was taken
$\mathbf{x}_k$	Location of all detected cells in the k:th image
$\mathbf{x}_{ik}$	The location of cell i in image k
N_t/N_k	The total number of cells at time t/ in image k
$\mathbf{X}$	Capital $\mathbf{X}$ is reserved for real data
S	In the particle filter introduced in 2.4, S is the number of particles.
s	In the particle filter introduced in 2.4, s is used as the index for individual particles.
L	In the particle filter introduced in 2.4, L is the time resolution used for simulation.
l	In the particle filter introduced in 2.4, l is used to index time steps.
A	In the particle swarm introduced in 4.3, A is the number of agents.
(a)	In the particle swarm introduced in 4.3, (a) is used to index individual agents.

of realizations of (1.3)–(1.4) do not suffer from the non-differentiability of the cell paths, which is a pathological feature of Brownian motion. An article covering using such models in a 3D setting was recently written by Scott *et al.* Scott *et al.* (2021).

1.3 *Notation*

In this paper, we will attempt to fit a flexible SDE model to *in vitro* imaging data using a simulation-based maximum likelihood approach. Given the large number of variables in the model we supply the reader with the below table (Table 1), which lists all variables and their meanings.

2. Model

Regardless of modeling paradigm, cell migration models quickly reach a point of mathematical intractability even under fairly simple assumptions. Depending on the context, one might opt to exclude the effect of the ECM, chemotaxis and cell-to-cell adhesion/repulsion. The choice of stochastic process modelling the independent propulsion of each cell is of great importance when selecting a model, and whether to include cell division into the model can have an impact as well. In this section, we will discuss the inclusions and omissions made in our model, given the context of our experimental data. The data set, described in greater detail in Section 3, consists of microscopy images of glioblastoma cells migrating through stem cell medium in a well coated with laminin. This implies that the physical and chemical environment surrounding the cells is fairly homogeneous, and hence we assume a complete absence of ECM or chemotaxis. We have access to high resolution data of the spatial evolution of the cells, making an off-lattice agent based model a natural choice. Specifically, the approach will be an interacting system of stochastic differential equations with an interaction potential in the drift term. As the purpose of the

study is to investigate inference methods on parameters in the drift term, the cell proliferation rate will be set to zero in our *in silico* experiments.

2.1 *Interaction potentials*

In mathematical biology, a potential is used to aggregate mechanical effects that can be difficult to disentangle. Essentially, potentials can be either *repulsive*, *attractive* or both (see Oelschläger (1989) for a nice summary). Every agent i, in our case cancer cells, is assigned a potential. If another cell is at a distance r from i where $\nabla U(r) < 0$, the cells repel one another. If $\nabla U(r) > 0$, they attract. The mechanism behind repulsion and attraction can be for any number of reasons, as the potential in itself encode some type of average behaviour, and is not a model of a specific biophysical phenomena. Example of phenomena that results in attraction cell is adhesion, while volume exclusion is a source of repulsion. However, many of the standard potentials resulted in problems during simulation and inference, especially pertaining smoothness of their derivatives and the presence of singular points. As such, we introduce a moderately interacting soft-core potential $U(r)$ governed by a vector of six parameters $\theta = [k_1, \ell_1, \alpha_1, k_2, \ell_2, \alpha_2]$. The potential is designed to be a flexible and infinitely differentiable version of repulsion–attraction potentials such as the Lennard–Jones potential and the Morse potential, and is given as

$$U(r) = \alpha_1 \frac{1}{1 + e^{-k_1(r^2 - \ell_1^2)}} - \alpha_2 \Big(\frac{1}{1 + e^{-k_2(r^2 - \ell_2^2)}} - \frac{1}{1 + e^{-k_1(r^2 - \ell_1^2)}} \Big). \tag{2.1}$$

2.2 *SDE model*

We chose to model our cell population using a system of interacting stochastic differential equations with isotropic diffusion. As we will come to see, isotropic diffusion vastly simplifies some implementation aspects and given the homogeneity of the environment it also serves as a fair assumption. It is however observed that some cells are more motile than others, and as such we choose to give every cell indexed by i its own diffusion coefficient σ_i. At a particular moment in time t, the system evolves according to the following set of equations

$$\mathrm{d}\mathbf{x}_i(t) = -\sum_{j \neq i} \nabla U(r_{ij}(t))\mathrm{d}t + \boldsymbol{\sigma}_i \mathrm{d}W_i(t) \tag{2.2}$$

for $i = 1, \ldots, N$ where N is the number of cells and $r_{ij}(t) = \|\mathbf{x}_i(t) - \mathbf{x}_j(t)\|$.

2.3 *Transition probability*

The quantity of main interest for our inference algorithm will be the N_t-particle density $P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, t) := P(\mathbf{x}, t)$. Let Ω be the domain on which the particle diffusion takes place, $P_0(\mathbf{x})$ be an initial distribution and assume Neumann boundary conditions, as the experiment takes place in an enclosed space. $P(\mathbf{x}, t)$

is then given as the solution to the $2N_t$-dimensional Fokker–Planck equation

$$\partial_t P(\mathbf{x}, t) = \nabla_{\mathbf{x}} \cdot \Big[D\nabla_{\mathbf{x}} P(\mathbf{x}, t) + \nabla_{\mathbf{x}} u(\mathbf{x}; \theta) P(\mathbf{x}, t) \Big], \tag{2.3}$$

$$u(\mathbf{x}; \theta) = \sum_{i=1}^{N_t} \sum_{j=i+1}^{N_t} U(\|\mathbf{x}_i - \mathbf{x}_j\|), \tag{2.4}$$

$$P(\mathbf{x}, 0) = P_0(\mathbf{x}) \tag{2.5}$$

$$\nabla_{\mathbf{x}} P(\mathbf{x}, t) \cdot \mathbf{n} = 0, \qquad \mathbf{x} \in \partial\Omega, \tag{2.6}$$

where $\mathbf{n}$ is a the normal vector the boundary $\partial\Omega$. We note that (2.3) is an unwieldy equation for all but the smallest number of particles, as the spatial dimension is $2N$. Reduction of its dimension can be done by a number of techniques such as a mean field approximation or closure at the two-particle density Bruna *et al.* (2017), but explicit study of this partial differential equation lies outside the scope of this article. Nevertheless, its interpretation lies at the heart of our inference problem, as $P(\mathbf{x}, t)$ is used to construct the likelihood function used in our inference algorithm.

2.4 *Estimating the transition probability through simulation*

Solutions to (2.3) over an interval $[t_k, t_{k+1}]$, $0 = t_0 < t_1 < t_2 \ldots, t_K = T$ correspond to the transition probability from a given state $\mathbf{x}(t_k) = [\mathbf{x}_1(t_k), \mathbf{x}_2(t_k), \ldots, \mathbf{x}_N(t_k)]$ to any future state at the time t_{k+1} in the following fashion Graham *et al.* (2006)

$$\mathbb{P}(\mathbf{x}(t) = \mathbf{x} | \mathbf{x}(t_k)) := P_k(\mathbf{x}, t),$$

where $P_k(\mathbf{x}, t)$ satisfies (2.3)–(2.6) with initial condition

$$P_k(\mathbf{x}, t_k) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_{ik}}(\mathbf{x}),$$

where δ denotes the empirical measure. Assume now that we wish to find the distribution $P_k(\mathbf{x}, t_{k+1})$, which is needed to construct a likelihood function

$$\pi_k(\mathbf{x}_{k+1} | \mathbf{x}_k) = P_k(\mathbf{x}_{k+1}, t_{k+1}). \tag{2.7}$$

Now remember that the cancer cells are observed at equally spaced times t_k, $k = 0, \ldots, K$, but we do not know the state of the cancer cells between these times. Thus, we can view states $\mathbf{x}(\tau)$ for $t_k < \tau < t_{k+1}$ as *hidden states*. The idea is now to construct an approximation of $\pi_k(\mathbf{x}_{k+1} | \mathbf{x}_k)$ using L hidden states. Illustrating the idea using a single hidden state at time $\tau_l \in (t_k, t_{k+1})$, we have

$$\pi_k(\mathbf{x}_{k+1} | \mathbf{x}_k) = \int_{\Omega} \pi_l(\mathbf{x}_{k+1} | \mathbf{x}_l) \pi_k(\mathbf{x}_l | \mathbf{x}_k) \mathrm{d}\mathbf{x}_l. \tag{2.8}$$

(2.8) can now be evaluated using Monte Carlo integration, simulating $\mathbf{x}_l := \mathbf{x}(\tau_l)$ given an observed $\mathbf{x}_k$. With S samples for the MC integration, we can obtain an approximation of the integral (2.8)

$$\pi_k(\mathbf{x}_{k+1}|\mathbf{x}_k) \approx \frac{1}{S}\sum_{s=1}^{S} \pi(\mathbf{x}_{k+1}|\mathbf{x}_{sl}), \tag{2.9}$$

where $\mathbf{x}_{sl}$ is the s:th generated sample of $\mathbf{x}_l$, where $s = 1, \ldots, S$. This method to construct approximate likelihood functions by simulating hidden states using the model is known as a *particle filter* Schön & Lindsten (2015); Chopin (2002). The chief question to answer now is how to sample from the hidden states. The Monte Carlo integration will be constructed using an approximation of the SDE system given by (2.2). There exist an extensive literature on approximations of nonlinear stochastic dynamical systems, see e.g. Kloeden & Platen (1992). These are based on the Itô–Taylor expansion, and the most frequently used method is the Euler–Maruyama scheme. However, since our SDE model (2.2) is driven by *isotropic* diffusion, there exist a higher order-scheme where convergence and stability is vastly improved compared to the Euler scheme. This scheme is studied in detail in Lindwall & Gerlee (2023), but we will give a short description of it here. We subdivide the interval $[t_k, t_{k+1})$ with a finer time grid $\tau_l = t_k + \frac{l}{L}(t_{k+1} - t_k)$ for $l = 0, 1, \ldots, L$ and use the model itself to estimate the unobserved cell distributions $\mathbf{x}_{1:N}(\tau_l)$ between the observations. We do this by generating S proposal paths that $\mathbf{x}_{1:N}$ could have traversed from time t_k to t_{k+1} using the Euler–Maruyama numerical scheme:

$$\hat{\mathbf{x}}_{is}^{\tau_{l+1}} = \hat{\mathbf{x}}_{is}^{\tau_l} + \mathbf{m}_{is}^{\tau_l}(t_{k+1} - t_k) + \sigma_i\sqrt{\frac{t_{k+1} - t_k}{L}}Z, \tag{2.10}$$

$$\mathbf{m}_{is}^{\tau_l} = \sum_{j\neq i} \nabla U(\hat{\mathbf{x}}_{is}^{\tau_l} - \hat{\mathbf{x}}_{js}^{\tau_l}) \tag{2.11}$$

$$\hat{\mathbf{x}}_{is}^{\tau_0} = \mathbf{x}_{ik} \tag{2.12}$$

for $s = 1, \ldots, S$, $l = 0, \ldots, L$ and $i = 1, \ldots, N$. Here Z is standard normally distributed random variable. We can use the $L - 1$:th positions found using the propagation of (2.10)–(2.12) to construct for cell i an approximation $\hat{\pi}_{ik}(\mathbf{x}_{i(k+1)}|\mathbf{x}_k; \theta, \sigma^2)$ of the desired transition density (2.7). It has the distribution $\hat{\pi}_{ik}(\mathbf{x}_{i(k+1)}|\mathbf{x}_k; \theta, \sigma^2) \sim \hat{p}_{ik}(\mathbf{x}_k, \theta, \sigma^2)$ where

$$\hat{p}_{ik}(\mathbf{x}_k, \theta, \sigma^2) = \frac{1}{S}\sum_{s=1}^{S} \hat{p}_{ik}^s(\mathbf{x}_k, \theta, \sigma^2), \tag{2.13}$$

$$\hat{p}_{ik}^s(\mathbf{x}_k, \theta, \sigma^2) = \mathcal{N}(\hat{\mathbf{x}}_{is}^{\tau_{L-1}} + \mathbf{m}_{is}^{\tau_{L-1}}(t_{k+1} - t_k), \sigma_i^2\frac{t_{k+1} - t_k}{L}).$$

More precisely, (2.13) is a Monte Carlo approximation of the *marginal likelihood* for particle i at time t_{k+1}; i.e. an approximation of the distribution

$$p_{ik}(\mathbf{x}_1, t_{k+1}) = \int_\Omega \cdots \int_\Omega P_k(\mathbf{x}, t_{k+1}) \mathrm{d}\mathbf{x}_2 \cdots \mathrm{d}\mathbf{x}_{N_k}. \tag{2.14}$$

Techniques for reconstructing the N_t-particle distribution using the marginals are covered in Section 4.

3. Experimental data

The raw data used in this study is an image sequence of glioblastoma cells obtained from the Human Glioma Cell Culture (HGCC) resource Xie *et al.* (2015), consisting of $K = 232$ images with a temporal resolution of 20 minutes. The dimensions of the images are 1408×1040 pixels. The glioblastoma cells were suspended in stem cell medium and plated onto well plates coated in laminin. The cells were cultured at 37 $\circ$C and 5% CO_2, and imaged using an IncuCyte microscope. The images were then tracked using The Baxter Algorithms Magnusson (2016). The output of this procedure is a list of identified unique cells, and associated features such as time step first and last observed, position over these time steps, average size and a family tree over mitosis relationships. After cleaning the data by removing nuisance observations, taken as tracks with fewer than 3 observations or objects smaller than 112 pixels, we calculate the *radial distribution function* (RDF) of our data set, using a Matlab script Weeks & Zhang (2023). The RDF describes the local density of cells at a distance r from a reference cell, and the first peak of this function gives information on the typical distance between neighbouring cells. With enumerating the observation times as $k = 1, \dots, K$, the RDF $g_{ik}(r)$ for cell i at time t_k is computed numerically as

$$g_{ik}(r) = \frac{1}{2\pi r \mathrm{d}r} \sum_{j \neq i} \mathcal{I}\big[|\|\mathbf{x}_{ik} - \mathbf{x}_{ik}\| - r| < \mathrm{d}r/2\big] \tag{3.1}$$

$$g(r) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} g_{ik}(r). \tag{3.2}$$

Here, $\mathcal{I}$ is the indicator function and dr is a reasonably small radius 'shell', resulting in (3.1) returning the number density of cells within a distance $r \pm \mathrm{d}r/2$ of cell i at time t_k. We call this an *empirical RDF*. The total RDF given by (3.2) for our *in vitro* dataset is visualized in Fig. 1, displaying a peak at a distance of 32 pixels, corresponding to a a distance of 24 μm. We take the distance to this first peak as the *characteristic length scale* of our data, and normalize so that one such distance is one length unit. While the cells come in a variety of shape and sizes, taking that into account when conducting the inference is unfortunately intractable. We thus make the assumption that the cell nuclei are circular, and essentially we transform every single cell to circular disc with a diameter of 1. This is to ensure numerical stability of the inference algorithm covered in Section 4. The entire process from microscopy images to cleaned and normalized tracks is summarized in Fig. 2.

4. Inference method

With the numerical approximation of the marginal likelihood for each particle given by (2.13), we have the ingredients needed to construct a likelihood-maximizing algorithm. However, straightforward maximization of the likelihood function will be infeasible given the potentially hundreds of diffusion coefficients. Luckily, this hurdle can be bypassed by noting that a conjugate relationship for σ_i exists, drastically simplifying the process of finding the maximum likelihood for these parameters. This conjugate relationship is studied in detail in Lindwall & Gerlee (2023). Thus the inference can be carried out in two blocks, one for the inference on diffusion coefficients conditioned on some set of interaction parameters, and one for inference on interaction parameters given a set of diffusion coefficients. This alternating process is repeated until convergence is reached, and we will specify the

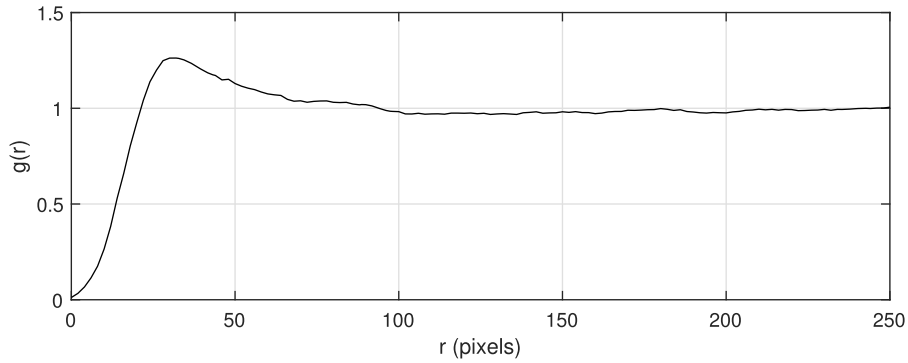

FIG. 1. Radial distribution function of the considered data set with length scale given in pixels. Note that the distribution function have been normalized as to converge to $\lim_{r\to\infty} g(r) = 1$.

details of the procedure throughout this section. The approach mirrors closely the SMC-within-Gibbs approach Wilkinson *et al.* (2011).

4.1 *Inference on diffusion coefficient—conjugate relationship*

The inference algorithm used for the diffusion coefficients is based on a technique covered Lindwall & Gerlee (2023), and we will provide just a brief summary of the method here. For each individual cell i we have a maximum likelihood estimate $\hat{\sigma}_i$ given by

$$\hat{\sigma}_i^2 = \frac{\sum_{k=k_i}^{K_i-1} \left(\mathbf{x}_{i(k+1)} - \mathbf{m}_{ik}\right)^T \mathbf{S}_{ik}^{-1} \left(\mathbf{x}_{i(k+1)} - \mathbf{m}_{ik}\right)}{2K} \tag{4.1}$$

$$\mathbf{m}_{ik} = \mathbf{x}_{ik} + (t_{k+1} - t_k) \sum_{j\neq i} \nabla U(\mathbf{x}_{ik} - \mathbf{x}_{jk}), \tag{4.2}$$

$$\mathbf{S}_{ik} = \mathbf{S}_{1k}^T \mathbf{S}_{1k} + \mathbf{S}_{2k}^T \mathbf{S}_{2k}, \tag{4.3}$$

$$\mathbf{S}_{1k} = \sqrt{t_{k+1} - t_k}\left(\mathbf{I} + \frac{t_{k+1} - t_k}{2} \sum_{j\neq i} \nabla^2 U(\mathbf{x}_{ik} - \mathbf{x}_{jk})\right) \tag{4.4}$$

$$\mathbf{S}_{2k} = \frac{(t_{k+1} - t_k)^{\frac{3}{2}}}{\sqrt{12}} \sum_{j\neq i} \nabla^2 U(\mathbf{x}_{ik} - \mathbf{x}_{jk}), \tag{4.5}$$

where k_i is the index for the first observation of this cell, and K_i is the index of the last observation. This maximum likelihood estimate is based on the approximate transition density given by (2.13), and the fact that it results in a Gamma-conjugacy for isotropic diffusion for σ_i. Again, for more details we refer to Lindwall & Gerlee (2023).

4.2 *Inference on interaction parameters—surrogate likelihood and RDF*

For the inference on θ, we will use a maximum likelihood (ML) approach where the analytically intractable likelihood function is replaced with a Monte Carlo surrogate based on the

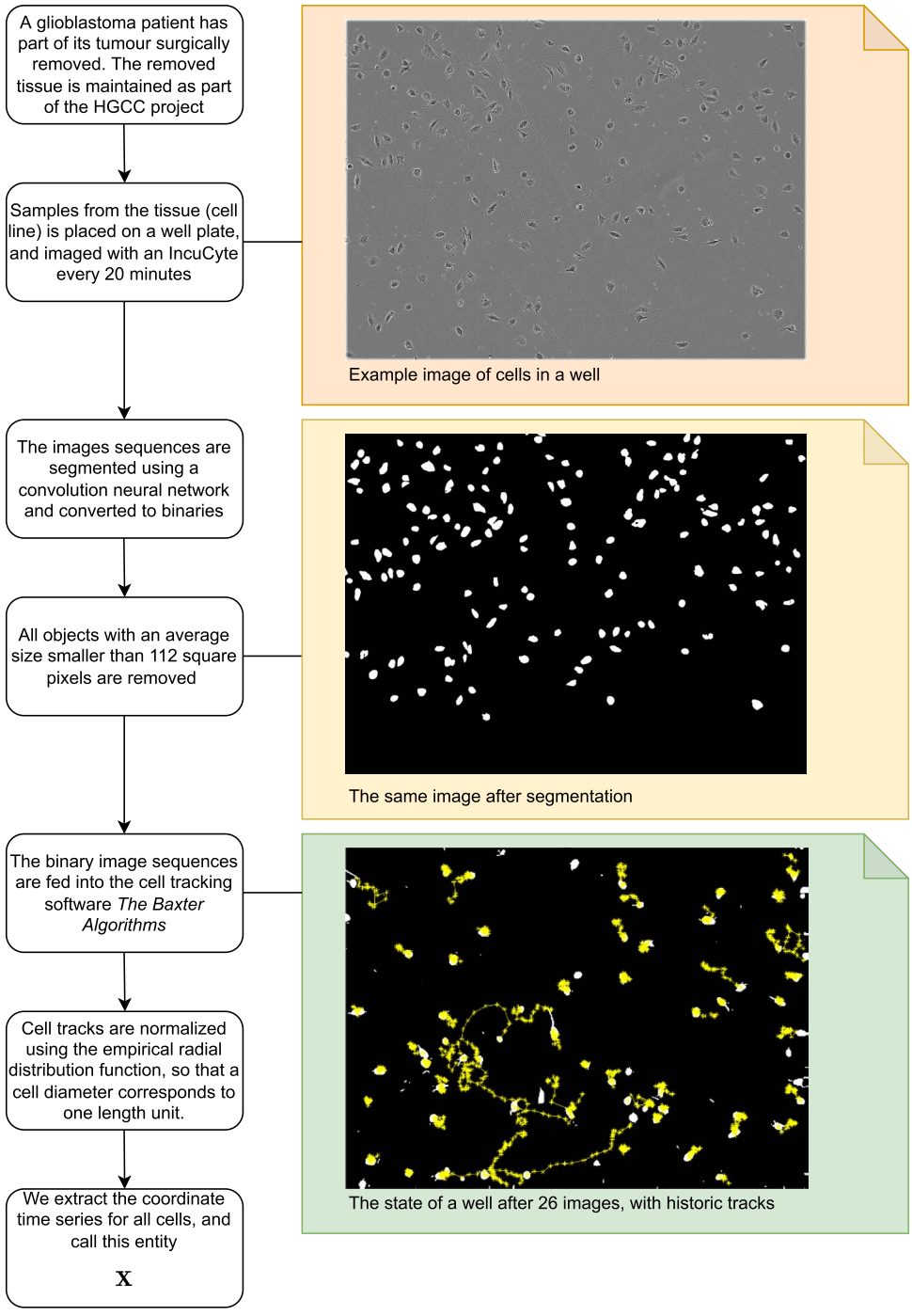

FIG. 2. Pipeline detailed in Section 3, showing the process from laboratory to annotated track data useful for our inference algorithm.

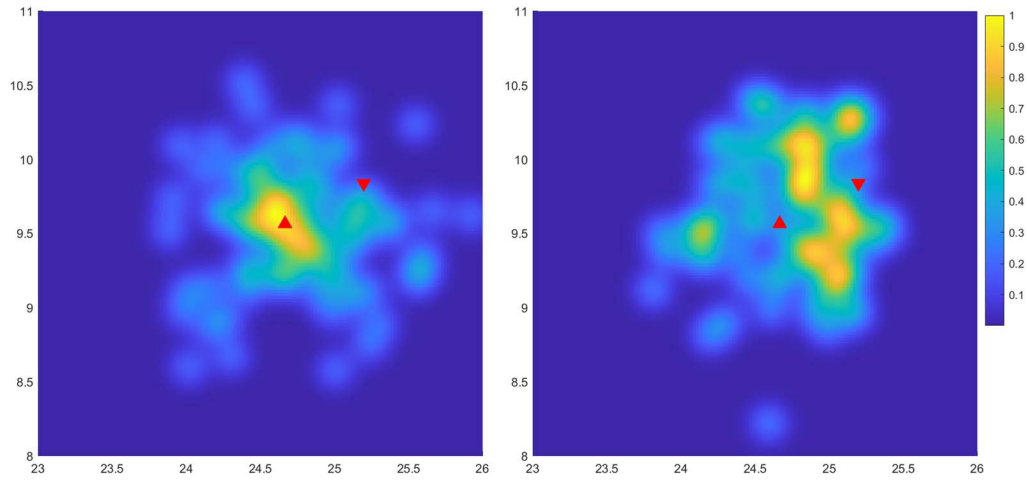

FIG. 3. A snapshot of the particle filter predicting the location of an *in silico* cell at time t_{k+1} (marked with a Δ) given its location at time t_k (marked with a ∇). Colourbar shows normalized log likelihood calculated using (4.6), where 1 the maximum. There are 45 minutes between the observations. In the picture on the left, the particle filter uses the same θ and σ that were used simulate the data set. Here the prediction performs well, indicated by the fact that the colour around the cell's location at time t_{k+1} (marked by a Δ) is warm. In the picture on the right, a randomly perturbed θ was used. This resulted in a bad overlap between the location of the cell at time t_{k+1} and the log-likelihood given by (4.6). For both images, particle filter hyperparameters of $L = 45$ and $S = 80$ have been used.

$\hat{pi}_{ik}(\mathbf{x}_{i(k+1)}; \mathbf{x}_k, \theta, \sigma^2)$ computed using (2.13). We then use this simulation to approximate the log likelihood of a parameter set θ. Here however, we have to tackle the problem that (2.13) only approximates the marginal likelihoods, and we must find a way to express the full N_t-particle density using the marginals. We suggest to simply assume independence of the particle distributions;

$$P(\mathbf{x}, t) = \prod_{i=1}^{N_t} p_i(\mathbf{x}_i, t)$$

leading to the mean field approximation (MFA) of the N_t-particle distribution. This gives us a log-likelihood given by

$$\hat{\ell}_{\text{MFA}}(\theta, \sigma) = \sum_{k=0}^{K-1} \sum_{i=1}^{N_k} \log(\hat{\pi}_{ik}(\mathbf{x}_{i(k+1)} | \mathbf{x}_k; \theta, \sigma^2)), \tag{4.6}$$

where N_k is the number of observed particles at time t_k. It is possible to include higher order approximations in the particle distributions, see e.g. Bruna *et al.* (2017), but given the complexity of the inference algorithm we here opt for a simple approach to the problem of calculating the N_t-particle distribution from the marginals. We see a visual representation of (4.6) in Fig. 3, where we also indicate how it is used for inference purposes. We refrain to the Figure text for further details on the interpretation.

In the event of mitosis taking place between observation k and $k + 1$ when applying the algorithm to *in vitro* data, our simulation scheme ignores the presence of the new born cell until the next set of forward simulation, from observation $k + 1$ to $k + 2$. The logic behind this decision is illustrated in Fig. 4. The exact birth time of a cell is outside of what we can possibly know, so we simply assume

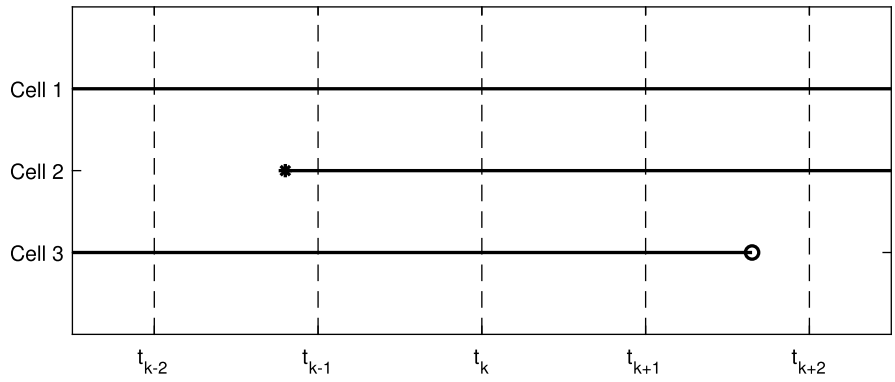

FIG. 4. An illustration of how cell lifelength is dealt with in our inference algorithm. The solid lines represent at what times the cell is actually alive, and the dashed lines represent observation times. Cell 1 is alive for the entire duration of this snapshot over five observed times $t_{k-2}, \ldots, t_{k+2}$. Cell 2 is born after observed time t_{k-2}, but before t_{k-1}. We thus assign Cell 2 a birth time of t_{k-1}. Cell 3 dies after time t_{k+1}, but its death is first noted at time t_{k+2}. We assign that Cell 3 died at the time it was last observed; at t_{k+1}.

that it was born the moment we first detected it, and concede that for some time between its 'actual' birth time and the time we first detect it, it has had some influence on the dynamics of the entire system that is not taken into account. The same applies for cell death—the time of death of a cell is set to be the last time it was observed alive. The bias these assumptions introduce is mitigated by increasing the observation frequency. We also want to make sure that not only is the predictive power of a proposed parameter configuration (θ, σ) resulting in a maximized likelihood, but also that the *spatial structure* of our synthetic cell population mimic that of the underlying data. The spatial structure is encoded in the *radial distribution function*, previously discussed in Section 3, and as such we will also consider a penalty for widely diverging radial distribution functions for our forward simulations when compared to the ground truth RDF, namely the *radial distribution function deviation* (RDF-deviation)

$$\hat{R}(\theta, \sigma) = \frac{1}{S} \sum_{k=1}^{K} \left(\int_0^{r_{\max}} (g_k(r) - \hat{g}_{sk}(r; \theta, \sigma))^2 \mathrm{d}r \right)^{1/2}. \tag{4.7}$$

Here, $g(r)$ is the RDF from our dataset (visualized in Fig. 1) and $\hat{g}_{sk}(r)$ is an RDF calculated using (3.1) from the forward simulation-generated positions $\hat{\mathbf{x}}_{is}^{\tau_L}$ given (θ, σ); see equations 2.102.12. The integral is evaluated numerically by simply considering the Euclidean norm of the discretized radial distribution functions, which are represented as vectors in Matlab.

4.3 *Full inference algorithm*

The inference algorithm is a variation of the classic particle swarm optimization method introduced in Kennedy & Eberhart (1995). Particle swarm optimization is a stochastic optimization method running over T generations, and is useful when optimizing objective functions that one can assume satisfy some smoothness conditions, but whose derivatives are exceedingly hard to evaluate. The objective function itself however should be readily available for evaluation. It is also applicable when we suspect that the objective exhibits many local maxima. The set up is as follows; a set of A agents enumerated $(a) = 1, \ldots, A$ are initiated in the parameter space Θ, i.e. the space of possible values of θ. Their locations,

dubbed $\theta_{(a)}$, are then used to calculate a set of diffusion coefficients $\sigma_{(a)}$ given the data $\mathbf{x}_k$, $k = 1, \ldots, K$ along with (4.1)–(4.5). $(\theta_{(a)}, \sigma_{(a)})$ are then used to forward simulate from one observation to the next, and we evaluate the surrogate likelihood for the parameters using (4.6). This is the object of greatest interest for the optimization algorithm. However, we note by inspection of (2.10)–(2.12) that the number of operations required for evaluating the particle filter used to estimate (4.6) is $\mathcal{O}(N^2KSL)$, where we remind ourselves that N is the number of observed cells, K is the number of observations, S is the number of particles used in the filter and L is the time resolution used in the particle filter.

4.3.1 *Penalty formulation of optimization problem.* To bypass this ballooning complexity, we chose to approximate the cell-cell interactions by, for each cell, only consider a limited set of neighbouring cells. This list is updated for each of the K observations. Thus, between the observations, when we propagate from a state given by $\mathbf{x}_k$ using the particle filter, we will only evaluate a fraction of all possible interactions. This is not a concession when forward-simulating, since the interaction potential is short-range by design. However, without taking long-range interactions into account when constructing the surrogate likelihood, the fact that the interaction must remain short-range must be enforced in some other way. We must also enforce a number of intrinsic properties of the interaction potential, which we will formulate as constraints of the maximum likelihood optimization. The optimization problem of maximizing the likelihood and minimizing the RDF-deviation given the constraints on the interaction function is given by

$$\begin{aligned} \max_{\theta,\sigma} \quad & \hat{\ell}_{\mathrm{MFA}}(\theta,\sigma) - \hat{R}(\theta,\sigma) & \\ \text{s.t.} \quad & |U(R;\theta)| - \varepsilon|U(1;\theta)| & \leq 0 \\ & U(0;\theta) & \geq 0 \\ & \tfrac{\partial}{\partial r}U(0;\theta) & \leq 0 \\ & U(1;\theta) & \leq 0 \\ & \tfrac{\partial}{\partial r}U(1;\theta) & = 0 \\ & \tfrac{\partial^2}{\partial r^2}U(1;\theta) & \geq 0. \end{aligned} \tag{4.8}$$

Here, R and ε are parameters that determine the range of the potential. The parameter R determines the distance at which the potential has decreased by a factor ε compared to the equilibrium, which is always present at a distance of 1 l.e. In Andolfi *et al.* (2014) it was shown that glioblastoma cells interact mechanically up to a distance of two cell diameters and we therefore set $R = 3$. The value of ε was set to 10^{-2}, which signifies a substantial drop in intra-cellular repulsion. The remaining constraints are to make sure that U retains the general desired shape of the interaction potential. In practice, we enforce these constraints using an *exterior penalty method*. Let $g_c(\theta)$ be the c:th constraint, and express the constraint $\frac{\partial}{\partial r}U(1;\theta) = 0$ as two inequality constraints of opposite signs. In generation t of our optimization algorithm, our objective is to maximize the unconstrained problem

$$\max_{\theta,\sigma} \quad \psi(\theta,\sigma) := \hat{\ell}_{\mathrm{MFA}}(\theta,\sigma) - \hat{R}(\theta,\sigma) - \lambda_t\Big(\sum_{c=1}^{7} \max(0, g_c(\theta))^2\Big), \tag{4.9}$$

where λ_t is the *penalty term* that increases every generation; we chose the penalty term $\lambda_t = 2^{\sqrt{t}}$. We will refer to $\psi(\theta,\sigma)$ as the *fitness value* for these parameters.

4.3.2 *Outline of optimization algorithm.* With the objective function clearly in our minds, we now move on to a brief description of the optimization algorithm. We begin by specifying the attributes of each agent.

TABLE 2 *Attributes of particle swarm agents*

Attribute	Notation	Initial value
Diffusion coefficients	$\sigma_{(a)}$	`NaN`
Interaction parameters	$\theta_{(a)}$	Uniformly drawn from Θ
Velocity	$v_{(a)}$	$10^{-3} \cdot \theta_{(a)}$
Fitness value	$\psi_{(a)}$	`NaN`
Best ever location	$\theta^*_{(a)}$	$\theta_{(a)}$
Best ever fitness value	$\Psi_{(a)}$	$-\infty$

Note that it does not matter what we initialize $\sigma_{(a)}$ as—it will be calculated using our initial $\theta_{(a)}$ regardless. At initialization, we start with deciding for how many generations T we wish to run the algorithm, set the hyperparameters S and L for the particle filter and R, ε for the constraints. We then generate A agents with the attributes specified in Table 2. These attributes are then used to evaluate (4.9) and a fitness value $\psi_{(a)}$ is assigned. If this fitness value is higher than the current best fitness $\Psi_{(a)}$, it replaces $\Psi_{(a)}$, and $\theta_{(a)}$ replaces $\theta^*_{(a)}$. The fitness-comparison aspect of a standard particle swarm generation is complicated by the presence of the penalty method. In order to bypass this, we 're-penalize' the previously held best belief about θ, to even out the playing field. This is vital, as otherwise beliefs held in earlier generations are given an unfair advantage when compared to more heavily penalized beliefs. At the end of every generation, the agents are compared to one another, and a global highest fitness Ψ with corresponding $\theta^*_{(a)\max}$ is crowned. If these would trump the current historical best values Ψ^* and θ^*, they are replaced. To account for situations where an unduly high score was given to a bad choice of θ^* in an early generation, we once again 're-penalize' the best ever choice θ^* in every generation to make sure that the algorithm does not run astray. The agents then propagate using simulated annealing to update the velocity, with annealing functions

$$f_1(t) = \frac{1}{3}\left(1 + e^{2-\frac{3t}{T}}\right)^{-1},$$
$$f_2(t) = \frac{1}{3}\left(1 + e^{4-\frac{6t}{T}}\right)^{-1},$$

where $f_1(t)$ governs annealing for propagation towards $\theta^*_{(a)}$ and $f_2(t)$ is for the collective best θ^*. The hyperparameters governing the annealing functions were chosen so that for the first half of the iterations, particles would favor exploring towards $\theta^*_{(a)}$, i.e. expressing more individual behaviour. After that, $f_2(t) > f_1(t)$, meaning that θ^* will be favoured. We then update the time index t and start over. This algorithm is summarized in Fig. 5. One important aside is that the propagation takes place in a *logarithmic* parameter space, in order to counteract the varying magnitude of the different parameters.

5. Results

5.1 *Results on in silico data*

To benchmark the algorithm laid out in section 5, we ran an *in silico* experiment generated using the model as stated in (2.2) and simulated using the same scheme as in (2.10)–(2.12). The parameters used in these experiments are specified in Table 3, along with the initial population size N_0. The choice of σ

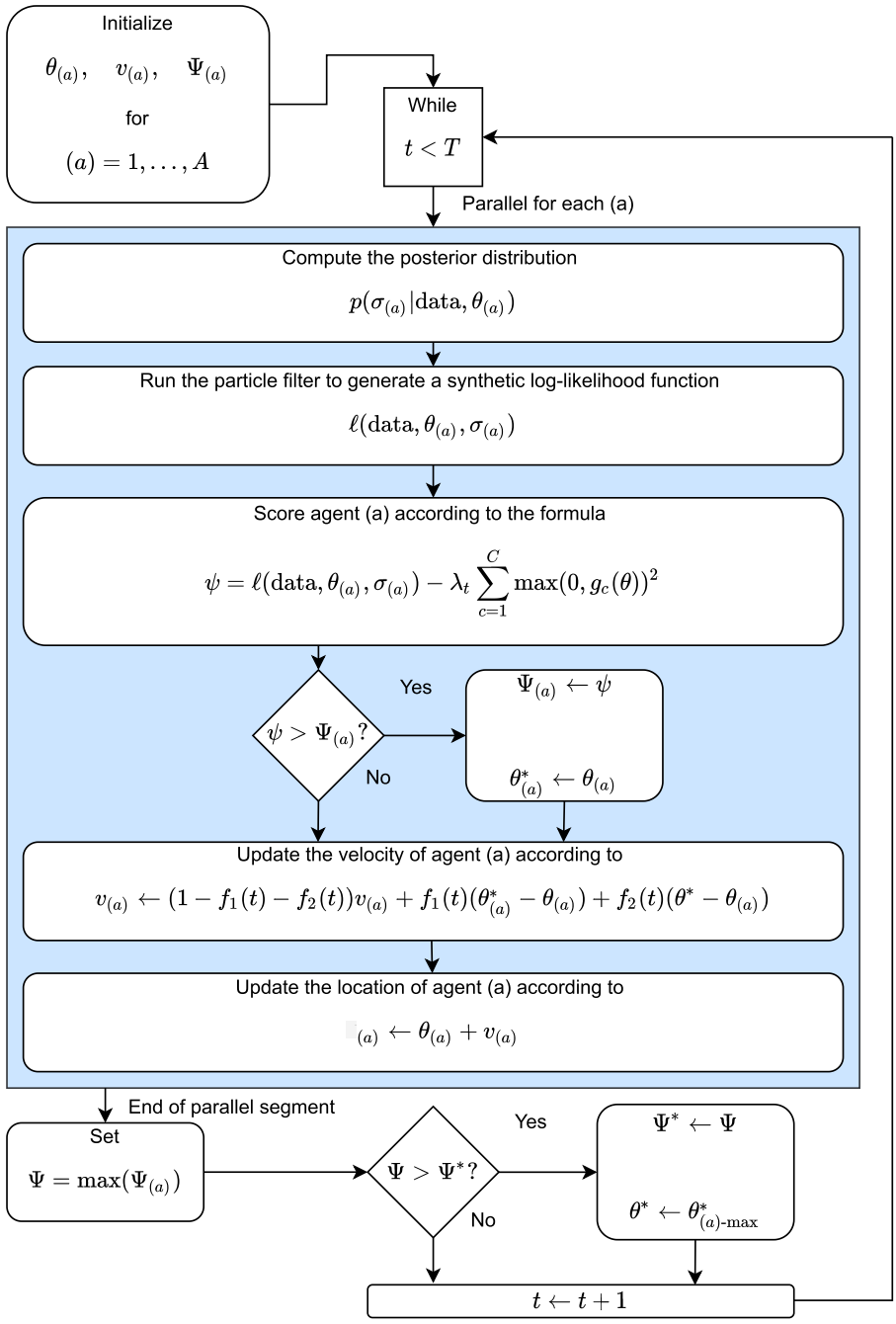

FIG. 5. Outline of the stochastic optimization algorithm. Note that the re-penalization step is not written out explicitly, but rather implied in all comparison steps.

TABLE 3 *Parameters used for in silico experiments*

k_1	ℓ_1	α_1	k_2	ℓ_2	α_2	σ	N_0
10.00	0.55	$4 \cdot 10^{-3}$	4.00	1.20	$6 \cdot 10^{-5}$	$e^{-9/2}$	256

is informed by Swanson *et al.* Swanson *et al.* (2000), where *in vivo* invasion of glioma was measured in both gray and white brain matter. The findings in Swansons work is that the average diffusion in gray matter were 0.0013 cm^2/day, which translates to $\sigma = 0.0053$ in our our unit of [cell diameter]^2/second. This is assuming an average cell diameter of 24 μm, corresponding to the peak in Fig. 1. The parameters θ and σ were chosen so that by visual inspection, the simulated system mimics the observed data detailed in Section 3. We simulate the cell set using a time step in the numerical scheme corresponding to $(\tau_k - \tau_{k-1}) = \Delta\tau = 1$ second over 24 hours. Although the algorithm allows for populations of varying size, these are not considered in this *in silico* experiment. One purpose of running the *in silico* experiment is to study how well the algorithm converges for different initial conditions. In order to study this, we initiate the particle swarm randomly at the surface of a 6-dimensional hyper-sphere centered around the logarithm of the ground truth parameter values as presented in 3, with radii $r_\theta = 0.25, 0.5, 0.75$ and 1. For these experiments, we use an observation frequency of $\Delta t = 20$ minutes. The results of these experiments are summarized in Fig. 6 along with Table 4. Another purpose is to investigate to what degree the frequency of observations affect the accuracy of our parameter estimations. In order to study this, we run inference on three variations of a single realization of the experiment, using observations every $\Delta t = 10, 20$ and 40 minutes. The 10 minutes between observations is a much higher frequency compared to the *in vitro* data, while the 20 minutes interval is comparable and the 40 minutes interval data is more infrequent. The results of these experiments are summarized in Fig. 7 along with Table 5. The algorithm runs for $T = 180$ generations for each experiment on the *in silico* data set, and we use $S = 12$ for the number of particles in the particle filter. L is set as $\Delta t/12$. We see that while the parameters inferred as summarized in Tables 4–5 can differ quite a lot from Table 3, though the winning parameter sets generates a potential quite close to the underlying target, measured in the bottom panel of Figs 6–7. The error in these panels is computed as where r_m is the closest distance two cells were ever detected at

$$\int_{r_m}^{5} |U(r;\theta_{(a)}) - U(r;\theta)|^2 \mathrm{d}r. \tag{5.1}$$

Finally, we present inference on the diffusion coefficient for the different time resolutions as well. The $\Delta t = 40$ minutes data underestimates the diffusion coefficient, but this is to be expected; the attractive portion of the potential used to generate the data set imposes a subdiffusive quality onto the system. Since the less frequent observations settled for a 'flatter' interaction potential, the numerical scheme given by (4.1)–(4.5) took less of the interaction into account, reporting the 'actual' motility, and not the underlying. This result is visualized by box plots over the posterior modes in 8.

5.2 *Results on in vitro data*

For the *in vitro* experiment, we considered the data set appended within the supplementary materials of this paper, the radial distribution function of which is given in Fig. 1. In order to explore a large number of parameter configurations, we ran 64 instances of the inference algorithm summarized in Fig. 5 utilizing twelve agents. Using the parameters in Table 3 as a reference point, each run was initiated around a

TABLE 4 *Parameters inferred from the in silico radius experiment—ground truth is found in Table 3*

r_θ	k_1	ℓ_1	α_1	k_2	ℓ_2	α_2
0.25	10.56	0.57	$5.08 \cdot 10^{-3}$	4.14	1.18	$5.90 \cdot 10^{-5}$
0.5	7.59	0.48	$3.03 \cdot 10^{-3}$	3.21	1.44	$4.46 \cdot 10^{-5}$
0.75	10.84	0.68	$2.75 \cdot 10^{-3}$	7.34	1.25	$5.99 \cdot 10^{-5}$
1	10.89	0.50	$9.14 \cdot 10^{-3}$	1.81	1.53	$9.49 \cdot 10^{-5}$

TABLE 5 *Parameters inferred from the in silico time resolution experiment—ground truth is found in Table 3*

Δt	k_1	ℓ_1	α_1	k_2	ℓ_2	α_2
10 min	6.00	0.32	$3.99 \cdot 10^{-4}$	4.04	1.21	$6.23 \cdot 10^{-5}$
20 min	10.37	0.12	$2.61 \cdot 10^{-2}$	4.08	0.56	$1.74 \cdot 10^{-4}$
40 min	9.79	0.62	$1.59 \cdot 10^{-3}$	5.20	0.95	$7.96 \cdot 10^{-5}$

TABLE 6 *5 best parameter sets inferred from the in vitro experiment. Note the difference in range for fitness values Ψ in the initial and final run of the algorithm—these are results of the increased resolution used in the final run*

Final Ψ	Initial Ψ	k_1	ℓ_1	α_1	k_2	ℓ_2	α_2
1727.7	−969.54	47.437	0.3672	$7.4052 \cdot 10^{-3}$	17.239	0.3377	$8.8214 \cdot 10^{-5}$
1717.6	−968.94	19.528	0.0096	$1.1518 \cdot 10^{-3}$	13.450	0.0087	$1.5979 \cdot 10^{-5}$
1639.2	−968.76	20.039	0.1742	$1.3114 \cdot 10^{-3}$	10.429	0.3002	$9.7072 \cdot 10^{-6}$
1610.0	−968.74	18.320	0.3136	$3.4666 \cdot 10^{-4}$	9.7028	0.4892	$3.5659 \cdot 10^{-6}$
1571.2	−969.42	63.925	0.3101	$1.8444 \cdot 10^{-1}$	4.0104	0.5602	$5.0649 \cdot 10^{-6}$

corner of a 6-dimensional cube with side length 2 centered around the the logarithm of the parameters given in Table 3. The 12 agents were then initiated on the surface of a 6-dimensional sphere of radius $r_\theta = 0.5$ centered at those corners. We let the algorithm run for 100 generations, with the same S and L as for the *in silico* experiments. After this, we identify the 20 strongest potentials discovered by this wide sweep, with fitness values ranging from $\Psi = -967.37$ for the highest scoring agent to $\Psi = -970.58$ for the 20:th best fit. In order to further hone in on what potential fit the data set best, we score the top 20 parameters once again, but this time using $S = 100$ and $L = \Delta t/4$, giving us a much more accurate result from the particle filter (2.10)–(2.12). The five best potentials discovered after refined scoring are presented in Table 6 and are visualized in Fig. 9. We refer to the b:th best performing parameter set as θ_b^*, for $b = 1, \ldots, 5$. Broadly, two local maxima has been found by the algorithm. One case, which includes the highest scoring one, features a equilibrium distance of $r \approx 0.6$. The other case is a purely repulsive potential. For the diffusion coefficients, we find that they vary quite a bit across the population, but the high-scoring potentials agree on both the mean and standard deviation of how σ is distributed across the glioblastoma cell culture. Box plots featuring the distribution for σ^* for the top five parameter configurations are visualized in Fig. 10.

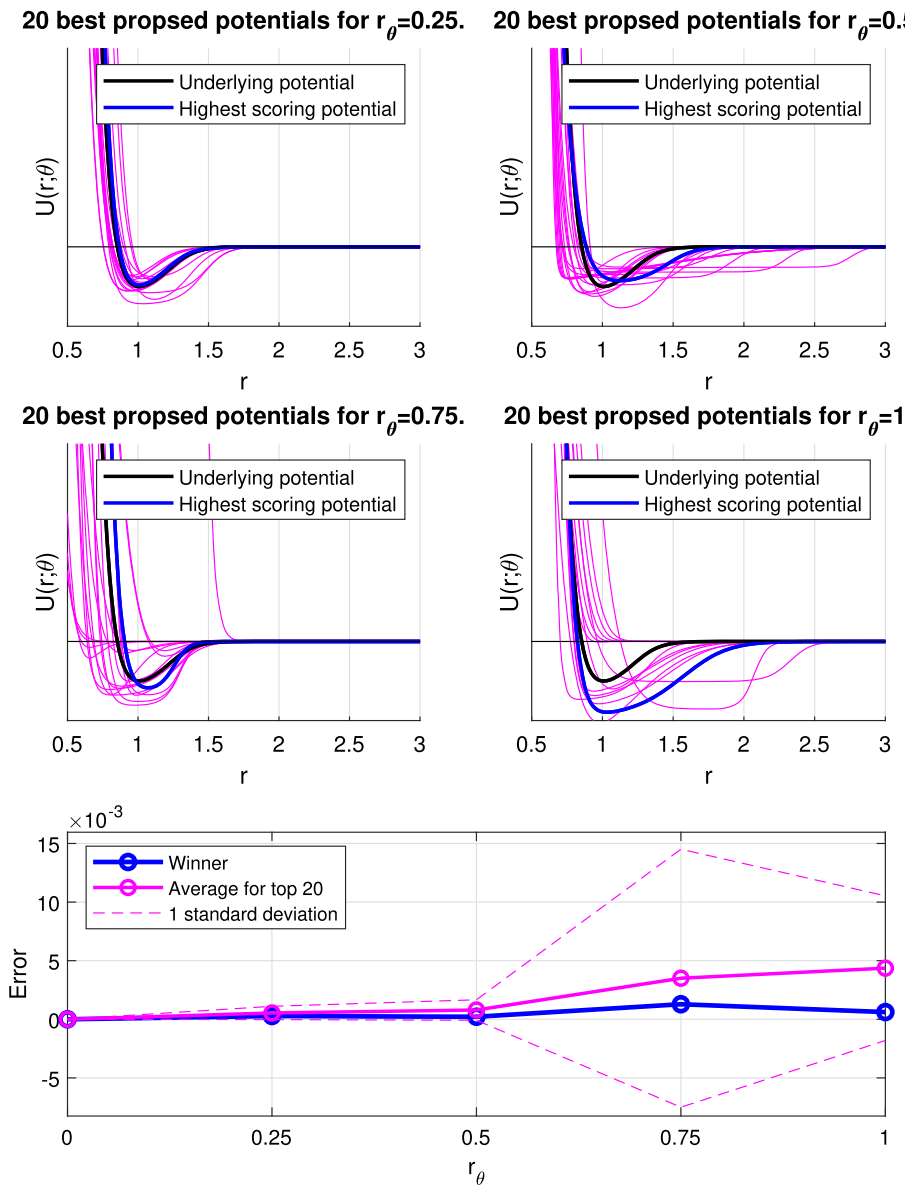

FIG. 6. 20 best potentials generated by the particle swarm from the *in silico* experiment, compared to the underlying potential. Note that the potential marked in red is the one with the best score, and not necessarily the one closest to the underlying potential. For all of these, $\Delta t = 20$ minutes. The error in the bottom panel is given by equation (5.1).

6. Discussion

The model as well as the inference algorithm designed for this this problem contains a diverse array of methods from different fields of applied mathematics. Some immediate mechanistic aspects left out in the SDE model will now be discussed, followed by ideas for improvement of the inference algorithm.

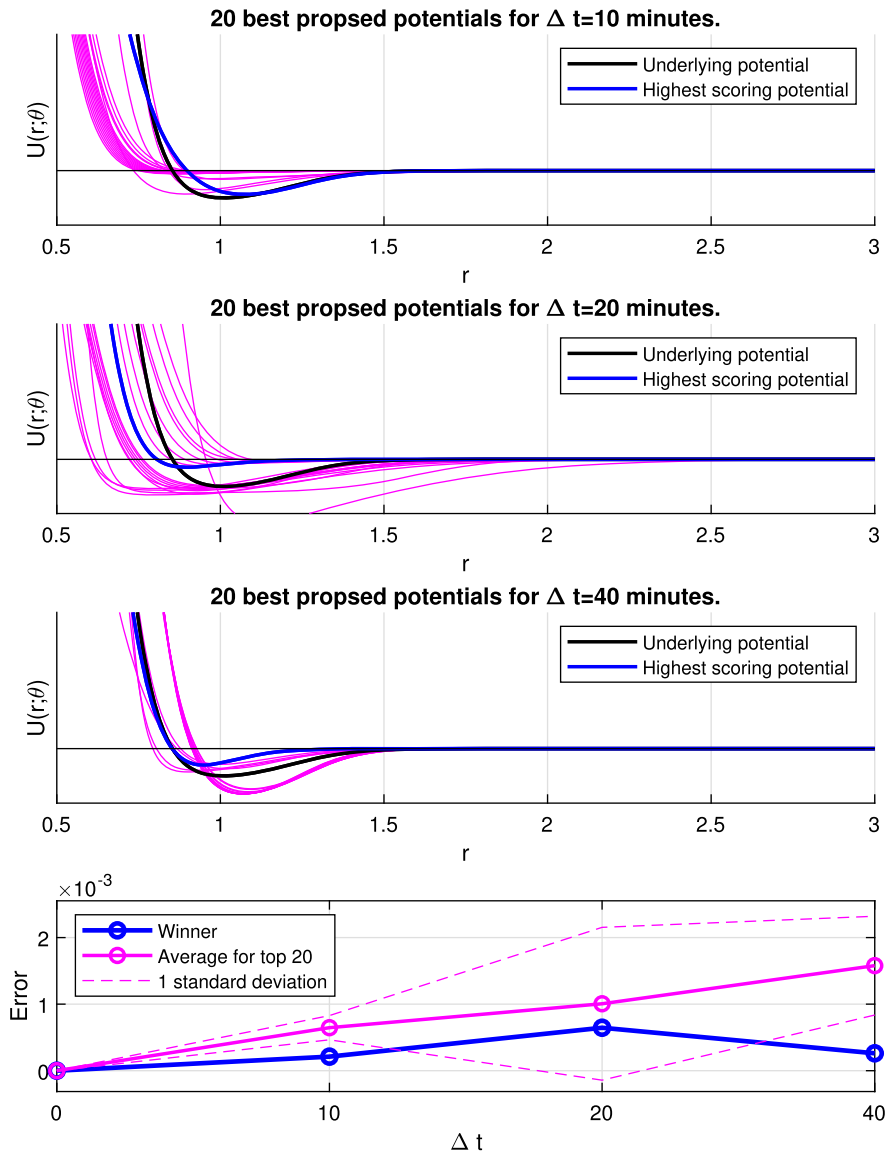

FIG. 7. 20 best potentials generated by the particle swarm from the *in silico* experiment, compared to the underlying potential. Note that the potential marked in red is the one with the best score, and not necessarily the one closest to the underlying potential. For all of these, $r_\theta = 0.75$. The error in the bottom panel is given by equation (5.1).

6.1 *Model development proposals for the agent based model*

Birth and death of cells. While the inference algorithm is flexible in that it can handle a varying number of cells between images, cell birth and death has not been considered in the model. There has been extensive research into the field of branching Brownian motion Bramson (1978), and this stochastic

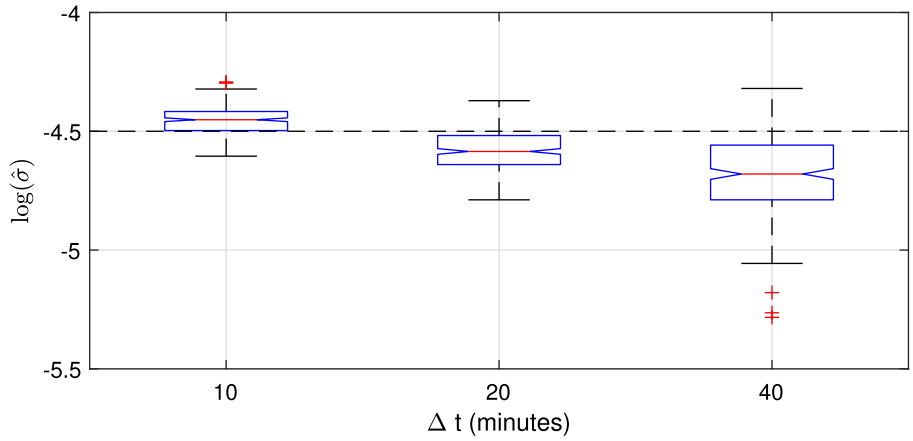

FIG. 8. Distribution of modes for the σ_i found by the winning agents in the 10, 20 and 40 minute cases of the *in silico* experiment, compared to the underlying σ as a black dashed line.

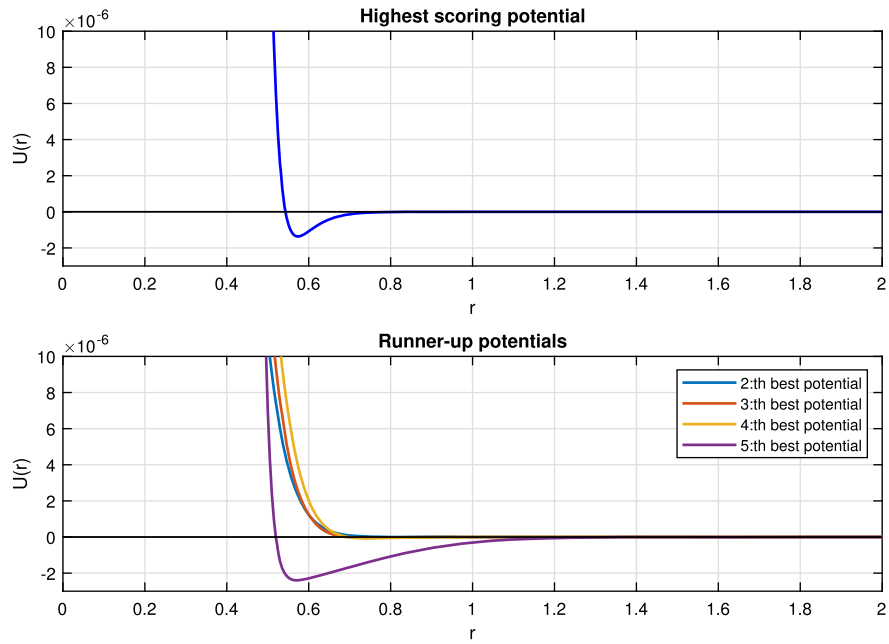

FIG. 9. The interaction potentials found by the winning agents on the *in vitro* experiment. Parameters θ_b^* for potential b given by the b:th row of Table 6. Note that the distance r is given in terms of of the equilibrium distance derived from Fig. 1, so that $r = 1$ denotes the peak of the radial distribution function.

process has a straightforward connection to Fisher-type equations through the use of renewal-reward theory Oelschläger (1989); Smith (1958). Since the model covered here is already Brownian motion based, a natural next step is to add branching properties to the paths $\mathbf{x}_i(t)$, representing mitosis. This could then be modified further to study other effects, such as the Allee effect Neufeld *et al.* (2017).

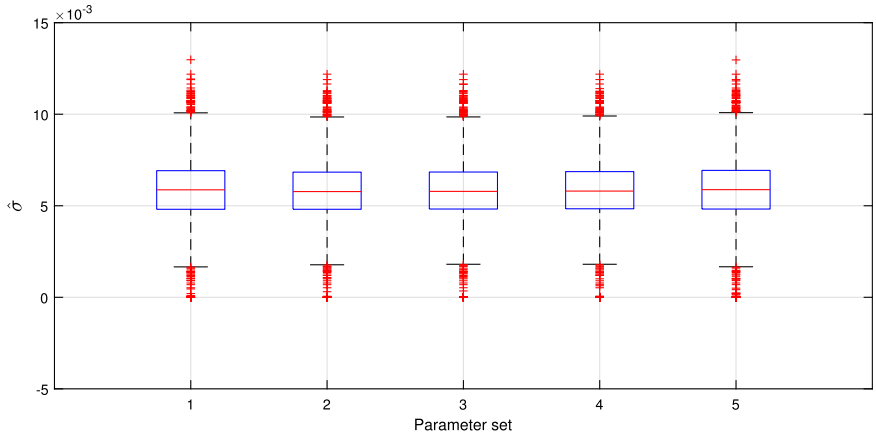

FIG. 10. Box plot over σ_b^* found by using the winning parameter sets θ_b^* found in Table 6. θ_1^* results in an average diffusion coefficient of $\overline{\sigma^*} = 5.83 \cdot 10^{-3}$, and a standard deviation in diffusion coefficient of $\sqrt{\sum_{i=1}^{N}(\sigma_{i1}^* - \overline{\sigma^*})^2/(N-1)} = 1.71 \cdot 10^{-3}$.

Inference on birth and death rates using single cell tracking has been carried out previously Johnston *et al.* (2014), and as such this is a promising future line of work. **Phenotype switching.** Phenotype switching can be considered, as in Gerlee & Nelander (2012). This would result in an increased number of model parameters, and a model framework for this is summarized in Oelschläger (1989), along with how to treat birth and death rates. However, inference on phenotype switching given the current methodology is a subtle problem, giving rise to queries of whether a phenotype switch has truly taken place, or if a cell slows down due to some other external stimuli. Related to phenotype switching is the concept of population heterogeneity, which amounts to the presence of distinct subpopulations within the cancer cell population. These subpopulations might differ with respect to the model parameters, and this is one possible explanation for the highly variant nature of the histogram of estimated diffusion coefficients in Fig. 10. However, variation between these subpopulations with respect to mechanical properties is not accounted for in the model, and the performance of the algorithm might improve if multiple subpopulations are included. **Persistent random walk.** Brownian motion is by its very nature a physical impossibility, as its trajectory is nowhere differentiable. One can for the purpose of rigour choose to model the random motion of cells using a stochastic processes such as (1.3)–(1.4) instead. This approach brings with it new and exciting modelling opportunities, as we in the continuum perspective move away from pure diffusion type equation to Boltzmann-like equations over time, space and velocity Othmer *et al.* (1988). The trade-off is that the problem doubles in complexity, and useful properties such as the Markovian nature of the fluctuations in space are lost. **Critical re-evaluation of the interaction potential**. The model (2.2) and especially the interaction potential (2.1) were designed with the explicit goal to to create a flexible, multipurpose way of expressing many kinds of particle kinematics situation. Under appropriate choices of θ, (2.1) can approximate everything from hard-sphere interactions to very long-range, soft-core potentials. However, the flexibility comes at a cost, as it introduces many degrees of freedom when solving the inverse problem of finding the parameters. We note this in our results, as both a purely repulsive and an attraction-repulsion potential was proposed as fair solutions. Compared to some classical potentials, the six parameters of (2.1) dwarfs the three parameters of the *Morse potential*, and the two parameters of the *Lennard–Jones potential*. Also, (2.1) is purely motivated by mathematical convenience and lack any derivation from fundamental theory. This is not entirely the case for the

Lennard–Jones and Morse potentials, however. The functional form of the Lennard–Jones potential is given as

$$U_{\mathrm{LJ}}(r) = D\big(\big(\frac{\rho}{r}\big)^{12} - \big(\frac{\rho}{r}\big)^{6}\big)$$

and this expression is based on the fact that the intra-molecular London dispersion force asymptotically decays at a rate of proportional to r^{-6}. The power of r^{12} in the repulsive part of this potential is chosen for mathematical convenience. In the case of the Morse potential, it was introduced as an empirical construct, but recently, justification for its functional form has been found based on first-order principles in quantum mechanics Costa Filho *et al.* (2013). The Morse potential is given by

$$U_{\mathrm{Morse}}(r) = D\Big(1 - e^{-a(r-r_0)}\Big)^2.$$

While biological systems are magnitude orders away from quantum mechanics based arguments, a Lennard–Jones style derivation based on perturbation theory could be motivated by biological principles, and is an area of future research. As it stands, the potential employed in the current study works well for simulation, but is far too complex for accurate inference given the current algorithm and realistic data. Perhaps with the previously discussed improvements this could be mitigated somewhat, and perhaps the flexible potential given by (2.1) holds some merit in future applications. Mathematical modelling must always operate at the intersection of *capturing enough reality to be useful* and *leaving out enough detail to be tractable*. What we were interested in with this model is capturing intra-cellular adhesion-repulsion behaviour and under some type of noise influence. With the above considerations, we can hopefully do this to an even higher degree of precision in future research.

6.2 *Inference algorithm development*

The inference algorithm is essentially based on the idea of a Gibbs sampler, albeit forgoing the extra information that a Bayesian approach gives us in favor of point estimates. More precisely, we use a pure Gibbs sampler when appropriate, i.e. for the diffusion coefficients, and then use a particle filter to approximate the likelihood function of the interaction parameters. This idea is usually referred to as SMC-within-Gibbs Schön & Lindsten (2015) and while intuitively easy to understand, it comes with an array of issues regarding variance of simulated distributions and mixing speed of the underlying Markov chains. Both of these issues can be mitigated by designing a *particle Gibbs sampler* Chopin & Singh (2015) for this problem. A way to reduce the variance in the surrogate distribution (2.13) is to reconsider what underlying stochastic process is used when propagating from observation k to observation $k+1$. As it stands, we do not use the $k+1$:th observation for variance reduction. In computational finance, the method we employ here is known as the Pedersen sampler. Various improvements on this scheme has been proposed, see e.g. Durham & Gallant (2002). One could also implement *guided proposals* Meulen & Schauer (2017) for similar purposes. All of the suggested improvements condition the interim samples in the particle filter (2.10)–(2.12) on a future observation in some way. The presence of multiple local maxima when the algorithm was applied to *in vitro* data indicate that variance reduction might be needed for more accurate results. Another issue is that in the scoring algorithm for the particle swarm optimization model, we make use of the mean-field approximation of $P(\mathbf{x}, t)$ (see Equation (4.6)). The mean-field assumption is a poor approximation, especially when close-range interactions play an important role Bruna *et al.* (2017). An improvement that can be considered is to use closure at the

2-particle density function (2PD). We acquire this by first making the assumption that the full density can be factorized as

$$P(\mathbf{x}, t) = \prod_{i=1}^{N_t} \prod_{j=i+1}^{N_t} P_2(\mathbf{x}_i, \mathbf{x}_j, t).$$

This assumption is not a completely unreasonable one to make, as the interactions that drive the particle system are evaluated pairwise. The tricky part now is to approximate $P_2(\mathbf{x}_i, \mathbf{x}_j, t)$ in a suitable manner. In Bruna *et al.* (2017), a number of different methods are evaluated, among them is the *composite expansion* as covered in Hinch (1991), which is given by

$$P_2(\mathbf{x}_i, \mathbf{x}_j, t) \propto p(\mathbf{x}_i, t)p(\mathbf{x}_j, t)e^{-U(\|\mathbf{x}_i - \mathbf{x}_j\|)}.$$

The log likelihood is now, up to an additive constant,

$$\hat{\ell}_{\text{2PD}}(\theta) = \sum_{k=1}^{K} \Big[\sum_{i=1}^{N_k} \log(\hat{p}_{ik}(\mathbf{x}_k^i; \mathbf{x}_{k-1}, \theta, \sigma^2)) - (t_k - t_{k-1})u(\mathbf{x}_{k-1}; \theta) \Big],$$

where u is the expression from (2.4). The additional term punishes configurations that lead to high intracellular tension in the observed system, which should improve the scoring algorithm further. Better yet, this penalty term is completely derived from the underlying mechanics, giving it further justification, compared to the current penalty that is based on the algorithm designer's subjective choice.

6.3 *Analysis of results on the in silico data*

For the high-frequency observations, the algorithm performed well on the considered synthetic data set. Importantly, the attractive and equilibrium portions of the potential were accurately inferred, whereas the repulsive component was more difficult to identify correctly. However, the amplitude of the repulsion is still large enough to effectively model volume exclusion. When analysing Table 5, one can note that the exact parameter values does not converge towards the ground truth presented in Table 3 for smaller Δt. We remind ourselves that the fitness (4.9) can be high even for parameter configurations that are quite different from the underlying truth. This implies a parameter identifiability issue for our model; we see that multiple widely different parameter configurations can lead to similar-looking potentials. The lower frequency data resulted in the particle swarm settling for flatter interaction potential (i.e. less adhesion). Note however that a completely flat interaction landscape is actually a local maxima when optimizing (4.9), meaning that the algorithm settled for a 'safe' solution. It is a known issue with particle swarm optimization that it can settle for non-global maxima, and that seems to be the case here. Reasonably, the large variance stemming from the Monte Carlo-scheme (2.10)–(2.12) when forward-propagating the particle filter can make evaluations of the objective (4.9) difficult. Alas, in a future implementation, more care will be put into reducing variance in this step. Bench marking using the *in silico* experiments suggests that the diffusion coefficients are accurately predicted by our algorithm, which could indicate one of two things. Either, the propagation of our SDE system (2.2) is so heavily dominated by the isotropic diffusion term that the interaction potential barely affects the inference, or the algorithm carefully accounts for discrepancies generated by (2.1), providing adequate inference on

diffusion coefficient under many different circumstances. Nevertheless, the results in this study further validates the method introduced in Lindwall & Gerlee (2023).

6.4 *Analysis of results on the in vitro data*

The local maxima discovered by the algorithm for the *in vitro* data set (see Table 6 and Fig. 10) gives some merit to the model, while also elucidating how complicated accurate modelling of glioblastoma multiforme is. First we can note that even after rescaling the problem so that the peak of the radial distribution function (see Fig. 1) equals to $r = 1$ in our length scale, the shape of glioblastoma is irregular enough in size to allow about 60% of that distance to be equilibrium, as seen in potential 1 and 5 in Fig. 9. Despite two local maxima existing in the parameter space, the inference on the diffusivity of population settled nicely in both cases. The reason for this can perhaps be explained by the following two arguments. Either the population was sparse enough that the exclusion and attraction modelled by the potential did not influence their diffusivity by much, or the diffusivity was accurately accounted for the θ found in Table 6. Notably, parameter sets explored during the run of the algorithm resulted in quite a variety of distributions over the diffusion coefficients, indicating that θ did have an impact. The best five all agreeing on the diffusion coefficients indicate that we can be confident in the inference on the diffusivity. One can note that the mean of $(\sigma^*)^2/2$ given the σ^* presented in Fig. 10 is $1.70 \cdot 10^{-5}$[cell diameter]^2/second. In the unit of [cm]^2/day, this is three orders of magnitude lower than the values presented by Swanson *et al.* Swanson *et al.* (2000). However, Swanson *et al.* estimated the diffusion coefficient for a solid tumour, making the situations not one-to-one comparable.

7. Conclusions

In this paper, we have introduced an interacting particle system model for glioblastoma migration, and constructed a maximum likelihood algorithm to conduct inference on said model. The strengths and weaknesses of the model has been assessed through analysis and simulation, and a somewhat promising result has been achieved when applying the model and inference algorithm to *in vitro* data of glioblastoma migration. Improvements on both the model and the inference algorithm has been suggested as future venues of research.

Acknowledgements

This research is funded by SSF grant SB16-0066.

References

Urbańska, K., Sokołowska, J., Szmidt, M. & Sysa, P. (2014) Glioblastoma multiforme–an overview. *Contemporary oncology.*, **18**, 307.

Anderson, A. R. & Maini, P. K. (2018) Mathematical oncology. *Bulletin of mathematical biology.*, **80**, 945–953.

Hamis, S., Powathil, G. G. & Chaplain, M. A. (2019) Blackboard to bedside: a mathematical modeling bottom-up approach toward personalized cancer treatments. *JCO clinical cancer informatics.*, **3**, 1–11.

Skog, J., Würdinger, T., Van Rijn, S., Meijer, D. H., Gainche, L., Curry, W. T., et al. (2008) Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. *Nature cell biology.*, **10**, 1470–1476.

Tracqui, P., Cruywagen, G., Woodward, D., Bartoo, G., Murray, J. & AlvordJr., E. (1995) A mathematical model of glioma growth: the effect of chemotherapy on spatio-temporal growth. *Cell proliferation.*, **28**, 17–31.

Gerlee, P. & Nelander, S. (2016) Travelling wave analysis of a mathematical model of glioblastoma growth. *Mathematical biosciences.*, **276**, 75–81.

Othmer, H. G. & Hillen, T. (2000) The diffusion limit of transport equations derived from velocity-jump processes. *SIAM Journal on Applied Mathematics.*, **61**, 751–775.

Painter, K. & Hillen, T. (2013) Mathematical modelling of glioma growth: the use of diffusion tensor imaging (DTI) data to predict the anisotropic pathways of cancer invasion. *Journal of theoretical biology.*, **323**, 25–39.

Oelschläger, K. (1989) On the derivation of reaction-diffusion equations as limit dynamics of systems of moderately interacting stochastic processes. *Probability Theory and Related Fields.*, **82**, 565–586.

Johnston, S. T., Baker, R. E., McElwain, D. S. & Simpson, M. J. (2017) Co-operation, competition and crowding: a discrete framework linking Allee kinetics, nonlinear diffusion, shocks and sharp-fronted travelling waves. *Scientific reports.*, **7**, 1–19.

Davies, K., Green, J., Bean, N., Binder, B. & Ross, J. (2014) On the derivation of approximations to cellular automata models and the assumption of independence. *Mathematical biosciences.*, **253**, 63–71.

Turelli, M. (1977) Random environments and stochastic calculus. *Theoretical population biology.*, **12**, 140–178.

Lemons, D. S. & Gythiel, A. (1997) Paul langevin's 1908 paper "on the theory of brownian motion"["sur la théorie du mouvement brownien," cr acad. sci.(paris) 146, 530–533 (1908)]. *American Journal of Physics.*, **65**, 1079–1081.

Einstein, A. (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der physik.*, **4**.

Black, F. & Scholes, M. (2019) The pricing of options and corporate liabilities. *World Scientific Reference on Contingent Claims Analysis in Corporate Finance: Volume 1: Foundations of CCA and Equity Valuation*. World Scientific, pp. 3–21.

Itô, K. (1944) 109. stochastic integral. *Proceedings of the Imperial Academy.*, **20**, 519–524.

Klebaner, F. C. (2012) *Introduction to stochastic calculus with applications*. World Scientific Publishing Company.

Arulampalam, M. S., Maskell, S., Gordon, N. & Clapp, T. (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on signal processing.*, **50**, 174–188.

Swanson, K. R., AlvordJr., E. C. & Murray, J. (2000) A quantitative model for differential motility of gliomas in grey and white matter. *Cell proliferation.*, **33**, 317–329.

Tremel, A., Cai, A., Tirtaatmadja, N., Hughes, B. D., Stevens, G. W., Landman, K. A., et al. (2009) Cell migration and proliferation during monolayer formation and wound healing. *Chemical Engineering Science.*, **64**, 247–253.

Johnston, S. T., Shah, E. T., Chopin, L. K., DS, M. E. & Simpson, M. J. (2015) Estimating cell diffusivity and cell proliferation rate by interpreting IncuCyte *ZOO*^{TM} assay data using the Fisher-Kolmogorov model. *BMC systems biology*, **9**, 1–13.

Lagergren, J. H., Nardini, J. T., Baker, R. E., Simpson, M. J. & Flores, K. B. (2020) Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. *PLoS computational biology.*, **16**, e1008462.

Everett, R., B., Flores, K., Henscheid, N., Lagergren, J., Larripa, K., Li, D., et al. (2020) A tutorial review of mathematical techniques for quantifying tumor heterogeneity. *Mathematical Biosciences and Engineering*, **17**.

Johnston, S. T., Simpson, M. J., McElwain, D. S., Binder, B. J. & Ross, J. V. (2014) Interpreting scratch assays using pair density dynamics and approximate Bayesian computation. *Open biology.*, **4**, 140097.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics.*, **145**, 505–518.

Browning, A. P., McCue, S. W., Binny, R. N., Plank, M. J., Shah, E. T. & Simpson, M. J. (2018) Inferring parameters for a lattice-free model of cell migration and proliferation using experimental data. *Journal of Theoretical Biology.*, **437**, 251–260.

Browning, A. P., Warne, D. J., Burrage, K., Baker, R. E. & Simpson, M. J. (2020) Identifiability analysis for stochastic differential equation models in systems biology. *Journal of the Royal Society Interface.*, **17**, 20200652.

Brückner, D. B., Ronceray, P. & Broedersz, C. P. (2020) Inferring the dynamics of underdamped stochastic systems. *Physical review letters.*, **125**, 058103.

Schnoerr, D., Grima, R. & Sanguinetti, G. (2016) Cox process representation and inference for stochastic reaction–diffusion processes. *Nature communications.*, **7**, 1–11.

Schön T, Lindsten F. Learning of dynamical systems–Particle filters and Markov chain methods. Draft available. 2015

Hillen, T. & Painter, K. J. (2009) A user's guide to PDE models for chemotaxis. *Journal of mathematical biology.*, **58**, 183–217.

Chauviere, A., Hillen, T. & Preziosi, L. (2007) Modeling the motion of a cell population in the extracellular matrix. *Discrete Contin Dyn Syst*, 250–259.

Gerlee, P. & Nelander, S. (2012) The impact of phenotypic switching on glioblastoma growth and invasion. *PLoS Comput Biol.*, **8**, e1002556.

Malik, A. A. & Gerlee, P. (2019) Mathematical modelling of cell migration: stiffness dependent jump rates result in durotaxis. *Journal of mathematical biology.*, **78**, 2289–2315.

Bodor, D. L., Pönisch, W., Endres, R. G. & Paluch, E. K. (2020) Of cell shapes and motion: the physical basis of animal cell migration. *Developmental cell.*, **52**, 550–562.

Scott, M., Żychaluk, K. & Bearon, R. (2021) A mathematical framework for modelling 3D cell motility: applications to glioblastoma cell migration. *Mathematical Medicine and Biology: A Journal of the IMA.*, **38**, 333–354.

Bruna, M., Chapman, S. J. & Robinson, M. (2017) Diffusion of particles with short-range interactions. *SIAM Journal on Applied Mathematics.*, **77**, 2294–2316.

Graham, C., Kurtz, T. G., Méléard, S., Protter, P. & Pulvirenti, M. (2006) *Probabilistic Models for Nonlinear Partial Differential Equations: Lectures Given at the 1st Session of the Centro Internazionale Matematico Estivo (CIME) Held in Montecatini Terme, Italy, May 22-30, 1995*. Springer.

Chopin, N. (2002) A sequential particle filter method for static models. *Biometrika.*, **89**, 539–552.

Kloeden, P. E. & Platen, E. (1992) *Numerical Solution of Stochastic Differential Equations*. Springer.

Lindwall, G. & Gerlee, P. (2023) Fast and precise inference on diffusivity in interacting particle systems. *Journal of Mathematical Biology.*, **86**, 1–19.

Xie, Y., Bergström, T., Jiang, Y., Johansson, P., Marinescu, V. D., Lindberg, N., et al. (2015) The human glioblastoma cell culture resource: validated cell models representing all molecular subtypes. *EBioMedicine.*, **2**, 1351–1363.

Magnusson, K. E. (2016) *Segmentation and tracking of cells and particles in time-lapse microscopy [Ph.D. thesis]*. KTH Royal Institute of Technology, p. 3.

Eric Weeks. Rui Zhang, editor. *Pair Distribution Function*. 2023. https://www.mathworks.com/matlabcentral/fileexchange/70378-pair-distribution-function.

Wilkinson, R. D., Steiper, M. E., Soligo, C., Martin, R. D., Yang, Z. & Tavaré, S. (2011) Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Systematic biology.*, **60**, 16–31.

Kennedy, J. & Eberhart, R. (1995) Particle swarm optimization. *Proceedings of ICNN'95-international conference on neural networks. vol. 4*. IEEE, pp. 1942–1948.

Andolfi, L., Bourkoula, E., Migliorini, E., Palma, A., Pucer, A., Skrap, M., et al. (2014) Investigation of adhesion and mechanical properties of human glioma cells by single cell force spectroscopy and atomic force microscopy. *PLoS one.*, **9**, e112582.

Bramson, M. D. (1978) Maximal displacement of branching Brownian motion. *Communications on Pure and Applied Mathematics.*, **31**, 531–581.

Smith, W. L. (1958) Renewal theory and its ramifications. *Journal of the Royal Statistical Society: Series B (Methodological).*, **20**, 243–284.

Neufeld, Z., von Witt, W., Lakatos, D., Wang, J., Hegedus, B. & Czirok, A. (2017) The role of Allee effect in modelling post resection recurrence of glioblastoma. *PLoS computational biology.*, **13**, e1005818.

Othmer, H. G., Dunbar, S. R. & Alt, W. (1988) Models of dispersal in biological systems. *Journal of mathematical biology.*, **26**, 263–298.

Costa Filho, R. N., Alencar, G., Skagerstam, B. S. & AndradeJr., J. S. (2013) Morse potential derived from first principles. EPL (Europhysics Letters). **101**, 10009.

Chopin, N., Singh, S. S., et al. (2015) On particle Gibbs sampling. *Bernoulli.*, **21**, 1855–1883.

Durham, G. B. & Gallant, A. R. (2002) Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics.*, **20**, 297–338.

van der Meulen, F., Schauer, M., et al. (2017) Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electronic Journal of Statistics.*, **11**, 2358–2396.

Hinch, E. J. (1991) Perturbation Methods. *Cambridge Texts in Applied Mathematics*. Cambridge University Press.