Mathematical Medicine and Biology Page 1 of 25
doi:10.1093/imammb/dqnxxx

A mathematical framework for modelling 3D cell motility; applications to glioblastoma cell migration

M. SCOTT, K. ŻYCHALUK AND R. N. BEARON
Department of Mathematical Sciences, University of Liverpool, Liverpool, L69 7ZL, UK.

[Received: date / Accepted: date]

Abstract

The collection of 3D cell tracking data from live images of micro-tissues is a recent innovation made possible due to advances in imaging techniques. As such there is increased interest in studying cell motility in 3D *in vitro* model systems, but a lack of rigorous methodology for analysing the resulting data sets. One such instance of the use of these *in vitro* models is in the study of cancerous tumours. Growing multicellular tumour spheroids *in vitro* allows for modelling of the tumour microenvironment and the study of tumour cell behaviours, such as migration, which improves understanding of these cells and in turn could potentially improve cancer treatments.

In this paper we present a workflow for the rigorous analysis of 3D cell tracking data, based on the Persistent Random Walk Model, but adaptable to other biologically-informed mathematical models. We use statistical measures to assess the fit of the model to the motility data and to estimate model parameters, and provide confidence intervals for those parameters, to allow for parametrisation of the model taking correlation in the data into account. We use *in silico* simulations to validate the workflow in 3 dimensions before testing our method on cell tracking data taken from *in vitro* experiments on glioblastoma tumour cells, a brain cancer with a very poor prognosis. The presented approach is intended to be accessible to both modellers and experimentalists alike in that it provides tools for uncovering features of the data set that may suggest amendments to future experiments or modelling attempts.

Keywords: three-dimensional cell migration, mathematical model, *in silico* modelling, persistent random walk, mathematical oncology

1. Introduction and Background

The ability of a cell to migrate is fundamental to its survival. Cells migrate through all manner of different environments, and in order to further our understanding of how systems in the body, both of humans and other organisms, function normally and under the influence of disease, we should endeavour to be able to describe cell motility under different conditions and rigorously test hypotheses about this motion. One way to do this is using mathematical models.

It is becoming increasingly evident that mathematical models can aid discovery in the life sciences, particularly when modelling complex phenomena such as cell migration and systems in which cells and their properties are being studied e.g. in cancer research (Deisboeck *et al.*, 2009; Anderson & Quaranta, 2008; Friedl *et al.*, 2012; Lee, 2018). To be predictive, mathematical models of cell migration should be informed by biology, dictating the relevant terms to be included in a model, the initial and boundary conditions needed to constrain the system and providing model specific values of important cell motil-

ity parameters. In return, these mathematical models can inform biology by analysing experimental data, confirming or rejecting proposed cell motility hypotheses, testing a system's sensitivity to model parameters and being able to make quantitative predictions from numerous *in silico* simulations under different conditions. This can aid biologists in deciding which experiments may be useful for a study without wasting time, money or resources.

Much of the body of work concerning the study of cell motility includes studies which have been conducted in 2D, or on single cells in 3D. Due to the advent of advanced techniques in microscopy and *in vitro* models for studying cell motility in 3D, live 3D tracking of cells in tissues is now becoming increasingly possible (Hoarau-Véchot *et al.*, 2018; Yamada & Cukierman, 2007; Hakkinen *et al.*, 2011; Lee *et al.*, 2014; Paul *et al.*, 2016). This in turn has exposed major differences in the way that cells move in 3D environments compared to 2D (Wu *et al.*, 2018; Yamada & Cukierman, 2007; Antoni *et al.*, 2015), and how cells interact with their environment and each other, highlighting the need for new models of cell migration in 3D. A major difference in 3D cell migration compared to 2D is the way that cells interact with each other and the extracellular matrix (ECM) surrounding them in many different ways. Further complexity arises because individual cells can behave very differently from each other in this environment. Because of the complexity of this 3D system and the potential for cellular heterogeneity, stochastic individual-based models capable of describing cells as individuals may be crucial to reveal the underlying mechanisms of cell motility in 3D.

The recently developed biological methods for studying cell motility produce large datasets in the form of cell tracks, and up to now there is a lack of mathematical tools to rigorously and systematically analyse this data, test proposed cell motility hypotheses and compare this analysis across different models (Driscoll & Danuser, 2015; Friedl *et al.*, 2012). There are few mathematical models of 3D cell motility in existence, much less in number than their 2D counterparts, though numerous biophysical models are found in the literature (Schlüter *et al.*, 2012; Paul *et al.*, 2017; Wu *et al.*, 2018). Rangarajan & Zaman (2008) provide a helpful review of existing mathematical models of 3D cell motility, which can be loosely categorised into force-based models, lattice-based models and stochastic models. Force-based models focus on traction forces in cells due to the ECM and the protrusion of cells into it, as well as drag and adhesion forces that arise as a cell moves. Zaman *et al.* (2005, 2006) make use of such a model, calculating the forces on a cell at each time step in an attempt to describe the cell's motility as a function of time. Lattice-based Monte Carlo methods are based on a 3D lattice and a set of criteria which dictates a cell's movement at each time step (Zaman *et al.*, 2007). Stochastic models are generally based around stochastic differential equations and random walks (Parkhurst & Saltzman, 1992; Wu *et al.*, 2015), the Persistent Random Walk (PRW) model being of particular interest to our current work. Wu *et al.* (2014) investigated the fit of the PRW model to 3D motility data, concluding that the model was incapable of describing motility in 3D, and adding an adjustment to the model in 2D to explain heterogeneity seen in experimental data. In a later work they propose the Anisotropic Persistent Random Walk (APRW) model which they claim better describes motility data in 3D with consideration of anisotropy in motility that the standard PRW model does not take into account (Wu *et al.*, 2015).

The PRW model has long been used to describe cell motility in 2D (Gail & Boone, 1970; Dunn & Brown, 1987; Stokes & Lauffenburger, 1991; Tranquillo & Lauffenburger, 1987; Dimilla *et al.*, 1992), though many have questioned whether the statistical measures defined by the model actually fit experimentally collected data. Most commonly these studies find that the Mean Squared Displacement (MSD) of cells is found to follow a power law rather than being a linear function of time as the PRW predicts (Dieterich *et al.*, 2008; Upadhyaya *et al.*, 2001; Metzner *et al.*, 2015; Loosley *et al.*, 2015; Cherstvy *et al.*, 2018). The Velocity Autocorrelation Function (VACF) is found to be better modelled by a sum of two exponentials rather than a single exponential (Dieterich *et al.*, 2008; Wu *et al.*, 2014) and

non-Gaussian distributions in cell velocities are found in some studies (Dieterich *et al.*, 2008; Metzner *et al.*, 2015). These model properties are discussed in more detail below. Some studies have shown that cells migrating in 3D, particularly cancer cells, display sub- or superdiffusive behaviour (Yurchenko *et al.*, 2019; Luzhansky *et al.*, 2018; Takagi *et al.*, 2008), meaning the PRW model description of the MSD would over- or underestimate this quantity for a population of cells.

Nevertheless, the PRW model is historically one of the most widely used models of cell motility and we use it here to demonstrate the power and usability of our framework. We provide mathematical tools to analyse 3D cell tracking data, using statistical measures to validate the model and provide parameter estimates to allow for parametrisation of the model in specific cases. We believe the framework is adaptable and the description presented in this paper is meant as a starting point to demonstrate a rigorous protocol for such analysis. Whilst our framework is based on the PRW model, we present it as a method for analysing 3D cell tracks, easily adapted to different models and the inclusion of biologically-informed terms in the governing equation of a model. We first carry out *in silico* simulations of the model to build the framework and then test it using experimental data from U87 glioblastoma (GBM) cell tracks *in vitro*, a subset of the data found in Richards *et al.* (2018).

Our workflow is tested using tumour cells from GBM, a particularly fatal brain tumour for which treatment methods inevitably fail due to the highly proliferative and invasive nature of the cells. The recent rise of the field of mathematical oncology (Rockne *et al.*, 2019) has seen many mathematical models attempt to describe many different aspects of cancer. This area of research aims to use mathematical models to assist in the fight against cancer, a disease which is characterised by excessive cell motility, especially invasion of cells into healthy tissue. Improving our understanding of cell motility will thus likely improve mathematical models in this field and eventually lead to better outcomes for patients with fatal brain tumours like GBM. Models of tumours in 3D are becoming increasingly predictive due to data-integration and increased knowledge of the tumour microenvironment that comes with an ability to replicate experimentally the conditions found in this environment.

Many models of tumours in 3D are found in the literature, together describing a range of features of tumours in a 3D environment. Data-integrated continuum models are a popular choice due to the wide range of analytical tools available for investigating these systems. The ability to integrate experimental data into these models makes them suitable for predicting survival times and potential treatment regimens for individuals (Hathout *et al.*, 2016; Swanson *et al.*, 2008; Colombo *et al.*, 2015; Rockne *et al.*, 2015; Agosti *et al.*, 2018; Jackson *et al.*, 2015). However, these continuous models are incapable of modelling individual cells in a tumour, and due to the inherently stochastic nature of cell motility, and cancer in particular, it is evident that discrete, stochastic models will be needed to further this field of study. Stochastic models of tumours and cancer cells broadly fall into one of two categories: agent-based models which can be on- (Gerlee & Nelander, 2012; Hamis *et al.*, 2019; Scianna & Preziosi, 2014) or off-lattice (Lowengrub *et al.*, 2010; Macklin *et al.*, 2010) and those based on stochastic differential equations and random walks (Stein *et al.*, 2007; Antonopoulos & Stamatakos, 2015; Antonopoulos *et al.*, 2019; Wu *et al.*, 2015), both of which attempt to use the properties of individual cells to elucidate the population behaviour under different conditions. We note that cell-based and continuum models of cell motility can be connected using scaling techniques, as described in (Othmer & Xue, 2013), for example.

The rest of this paper is set out as follows. We continue this introduction with an overview of the theory of the PRW model and the statistical measures used to both test the goodness of fit and estimate model parameters. Then follows a description of how we used the model to simulate *in silico* cell trajectories with known parameter values and checked our framework was able to extract good estimates for these parameter values directly from the trajectories. We present and discuss these simulations in 3 dimensions before discussing the parameters and the output of the framework. We finally test

the framework on 3 sets of experimental tracking data taken from *in vitro* tumour spheroid models of glioblastoma U87 cells to see if the PRW model really can describe their migration in 3D.

The Persistent Random Walk Model

The Persistent Random Walk (PRW) model has long been used as a way to describe random cell motility. The model, derived from the stationary, mean-reverting Ornstein-Uhlenbeck (OU) process (Dunn & Brown, 1987), describes a correlated random walk in velocity which sees the correlation between subsequent velocities of the same cell decay over time (see section 1 of the Supplementary Information for more details). A cell's velocity in a subsequent time step is assumed to be conditional on the velocity in the current time step with past velocities having no influence, and tends to be in the same direction. Cells are assumed to be identical, and independent - no interaction between cells is modelled.

In 1D, the probability density $p(v,t)$ of velocity v at time t is assumed to be governed by the OU process and can be described by the Fokker-Planck equation

$$\frac{\partial p}{\partial t} = \frac{\partial(\beta v p)}{\partial v} + \frac{1}{2}\alpha\frac{\partial^2 p}{\partial v^2}. \tag{1.1}$$

Here, α is the diffusion coefficient which represents the magnitude of random movement accelerations, and β is the drift coefficient which represents the velocity decay rate (Stokes & Lauffenburger, 1991). The time evolution of this OU process can be described by a stochastic differential equation (SDE) for cell velocity:

$$dv = -\beta v\,dt + \sqrt{\alpha}\,dW(t), \tag{1.2}$$

where v is the velocity and $W(t)$ is the Wiener process. In 1D, for an initial distribution of velocity taking the value v_0 with probability one, the solution of this equation is the Gaussian distribution with mean $\mu = v_0 e^{-\beta t}$ and variance $\sigma^2 = \frac{\alpha}{2\beta}(1-e^{-2\beta t})$.

In 2D, Stokes & Lauffenburger (1991) express α and β in terms of more intuitive parameters: $P = 1/\beta$, the persistence time of a cell; and $S = \sqrt{\alpha/\beta}$, the root mean squared speed (RMSS) of cells at steady state. More generally, in n-dimensions, Campos *et al.* (2010) express the process in terms of the persistence time $P = 1/\beta$, and D, the spatial diffusion coefficient of the cells in n-dimensional physical space:

$$d\mathbf{v} = -\frac{1}{P}\mathbf{v}\,dt + \frac{\sqrt{2D}}{P}\,d\mathbf{W}(t). \tag{1.3}$$

In n-dimensions, the diffusion coefficient is related to the RMSS of cells at steady state by $D = S^2P/n$, and thus we can relate the original parameters of the OU process α and β to the intuitive parameter S by $S = \sqrt{\alpha n/2\beta}$.

Statistical Measures for Comparison

To decide on whether the PRW model is an appropriate model for a given dataset and if this is the case to estimate what the correct values of S and P are, we must implement statistical measures. Such statistical measures are drawn from the model using equation 1.3. More details on the derivations are provided in sections 2, 3 and 4 of the Supplementary Information.

The first such measure, the Mean Squared Displacement (MSD) is commonly used when looking at cell motility, and for the PRW model in n-dimensions is given by (Campos *et al.*, 2010)

$$\text{MSD(t)} = 2nDP(e^{-\frac{t}{P}} + \frac{t}{P} - 1) = 2S^2P^2(e^{-\frac{t}{P}} + \frac{t}{P} - 1). \tag{1.4}$$

We see the PRW model displays classic diffusion behaviour of MSD tending to a linear function of time, i.e. $\text{MSD(t)} \to 2S^2Pt$ as $t \to \infty$.

Secondly, we use the Velocity Autocorrelation Function (VACF) for the PRW model given in n-dimensions at time t by (Campos *et al.*, 2010)

$$\text{VACF(t)} = \frac{nD}{P}e^{-\frac{t}{P}} = S^2e^{-\frac{t}{P}}. \tag{1.5}$$

The VACF quantifies the correlation between cell velocity at time 0 and at time t. This is calculated at a population level, averaging over all cells for each time. The correlation decays at rate $1/P$, meaning that cells 'forget' their previous velocity over times long compared with P.

We finally consider the stationary speed distribution of the population. At steady state, velocities should follow an n-dimensional Gaussian distribution according to the PRW model. For 3D this implies that the speed, u, follows a Maxwell-Boltzmann distribution with density

$$f(u;S) = \left(\frac{3}{2\pi S^2}\right)^{3/2} 4\pi u^2 e^{-\frac{3u^2}{2S^2}}. \tag{1.6}$$

More detail on this distribution is given in section 4.2 of the Supplementary Information.

2. Using the PRW model to describe cell motility

In silico tests

In order to use the PRW model to describe motility in 3D, we have created a workflow to rigorously assess the fit of the PRW model to cell tracking data by using the dataset to parametrize the model before verifying the fit using the statistical measures outlined above. This framework involves: inputting formatted cell tracking data; estimating S and P; verifying model fit using additional statistical measures. A diagram of the workflow is provided for clarity in Figure 1. Validation of our framework is important to ensure our method extracts the correct parameters S and P; it also allows us to assess if the model is appropriate for the data. In order to validate the workflow, we used *in silico* data generated from the 3D SDE

$$d\mathbf{v} = -\frac{1}{P}\mathbf{v}\,dt + \sqrt{\frac{2S^2}{3P}}\,d\mathbf{W}(t), \tag{2.1}$$

with specified values of S and P. This allowed us to create a data set similar to the experimental set and conduct the validation tests with prior knowledge of the parameters. Refinement of the method was then carried out until the estimates were sufficiently accurate. Cell tracking data entered into the framework must be an array outlining the positions and velocities of each cell at each time point. If only positions within tracks are available, as will be seen in the experimental data later, the velocity of each cell at each time point is estimated from the difference in the current and previous position divided by the time step. For the *in silico* data sets we numerically simulate equation 2.1 along with $d\mathbf{x}/dt =$

v using MATLAB's `simByEuler` function (MATLAB 2017a, Financial Toolbox) to simultaneously obtain both cell positions and velocities in the data set. In addition to S and P, it is necessary to specify the numerical time step, dt, the total time of the simulation, $dt \times$ Nperiods, with Nperiods being the number of simulation periods, the number of cells N, and the initial position and velocity vectors $\mathbf{x}_0$ and $\mathbf{v}_0$ for all cells. Figure 2(a,b) shows 3D sample plots of the tracks generated by the workflow.

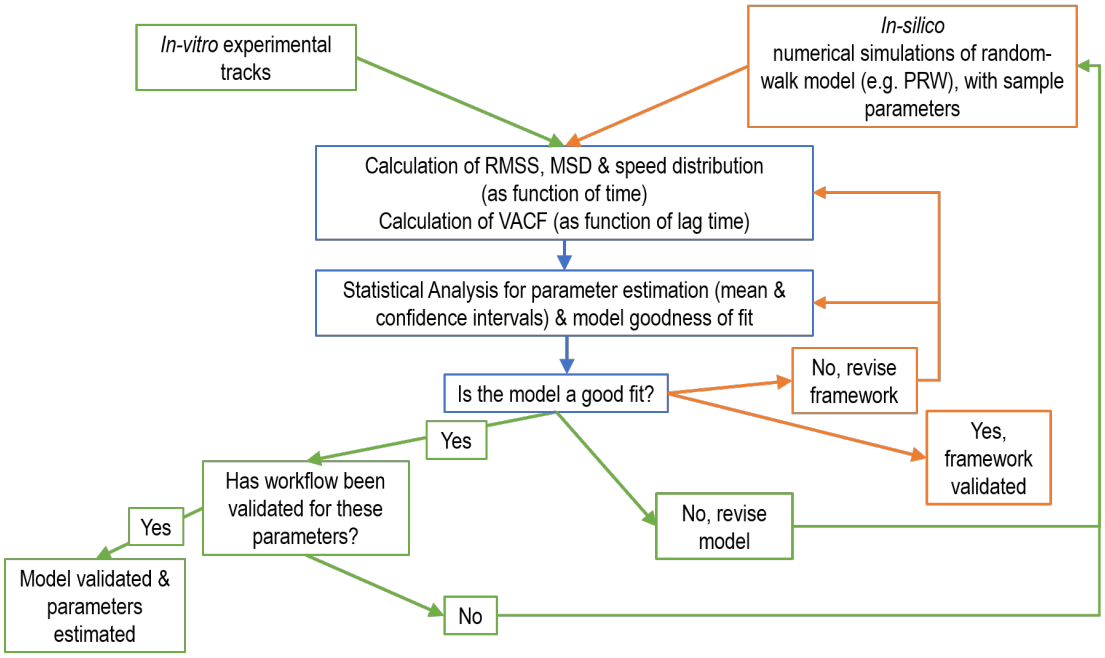

FIG. 1. **Diagram of the workflow.** An overview of the flow of the framework described. Code to carry out this analysis can be found at https://github.com/m-scott22/PRW3DCellMotilityFramework

S estimate

Parameter S is defined as the root mean squared speed of cells once the system reaches steady state. The root mean squared speed (RMSS) at time t across all cells is calculated in 3D as

$$\mathrm{RMSS(t)} = \sqrt{\langle v_x(t)^2 + v_y(t)^2 + v_z(t)^2 \rangle}, \tag{2.2}$$

where the average $<>$ is over all cells, and the 3D components of the velocity at time t are given by $v_x(t)$, $v_y(t)$ and $v_z(t)$. We take the average of RMSS(t) at all times to obtain an overall estimate of S, $\hat{S}$,

$$\hat{S} = \frac{1}{T}\sum_{t=0}^{T} \mathrm{RMSS(t)} = \frac{1}{T}\sum_{t=0}^{T} \sqrt{\langle v_x(t)^2 + v_y(t)^2 + v_z(t)^2 \rangle}, \tag{2.3}$$

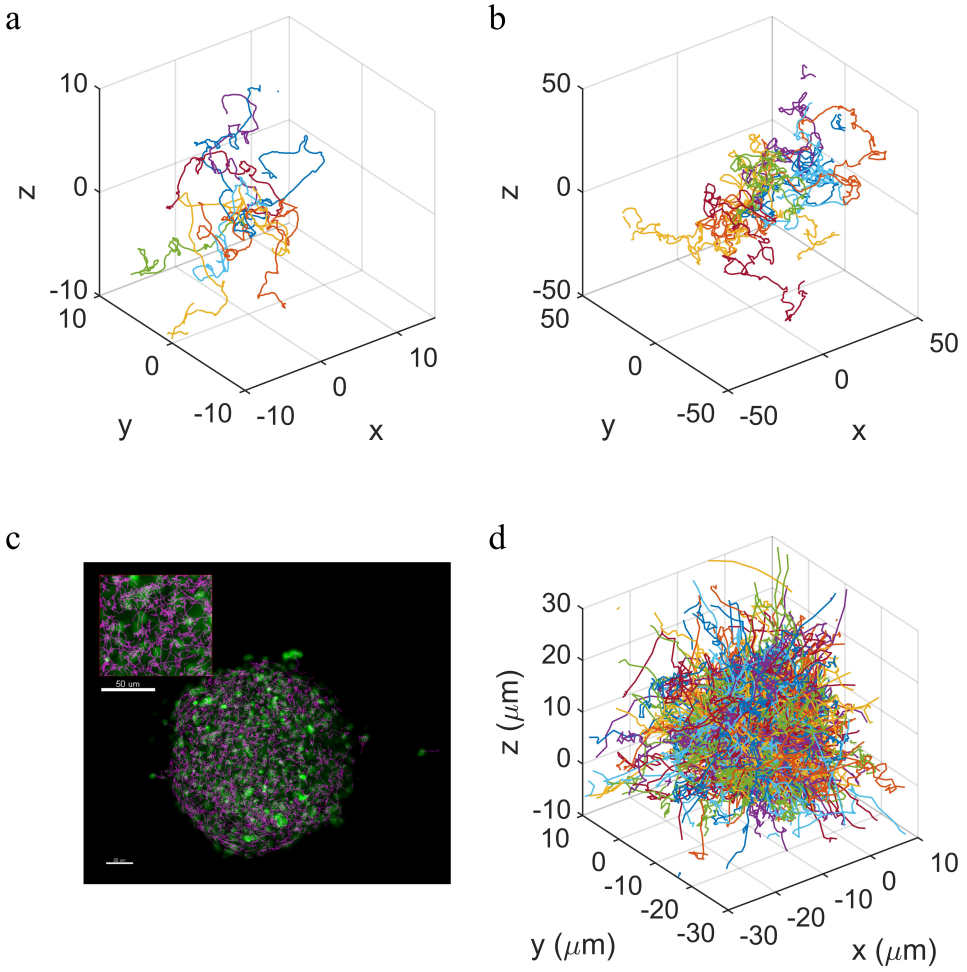

FIG. 2. **Examples of cell trajectories. a)** *in silico* data with parameters $S = 1$, $P = 1$, $dt = 0.05$ and NPeriods $= 1000$. Cells are initialised at the origin, $\mathbf{x}_0 = \mathbf{0}$, with speed S and orientation sampled uniformly from the unit sphere. Plot shows tracks from 10 cells as example trajectories. **b)** *in silico* data with parameters $S = 25$, $P = 0.1$, $dt = 0.05$ and NPeriods $= 480$. Initial positions and velocities as in a). Plot shows tracks from 10 cells as example trajectories. **c)** Experimental *in vitro* images with green indicating location of cell nuclei, and purple the overlay of cell tracks identified using tracking software, from Richards *et al.* (2018). Inset of zoomed in tracks & scalebar. **d)** The corresponding experimental trajectories from c) plotted within the framework. Initial positions and velocities taken from first entries for each track.

as it is assumed that experimental data would initially be at steady state.

The framework outputs a plot of the RMSS time series from which an estimate of S is obtained, and a histogram of the speed distribution at specified time points with the corresponding Maxwell-Boltzmann density function with estimated parameter $\hat{S}$ overlaid. This is depicted in Figure 3 for a simulation with $S = 1$ and $P = 1$. In this simulation, all cells had initial speed of 1, allowing us to obtain the stationary speed distribution more rapidly. Plots c(i)-(iv) demonstrate how the speed distribution of cells settles to the stationary distribution.

We will also look at a confidence interval for the S estimate. The calculated RMSS values at each time point form a time series when taken in sequence, which can be modelled as an autoregressive process of lag 1 (see section 5.1 of the Supplementary Information for more details). This means the current value of the process depends on the past only through the value of the process in the previous time step. We use the fact that the mean of this time series is S to construct a 95% confidence interval for $\hat{S}$.

Due to the serial correlation present in the data, we use an adjustment to the sample size when calculating the confidence interval for a sample mean. The calculation of this adjusted sample size, or effective sample size, is taken from Zwiers & von Storch (1995) and is calculated as

$$n_e = \frac{\text{Nperiods}}{1 + 2\sum_{\tau=1}^{\text{Nperiods}-1}\left(1 - \frac{\tau}{\text{Nperiods}}\right)\rho_1^{\tau}}, \tag{2.4}$$

where Nperiods is the number of observations in the RMSS time series and ρ_1 is the lag-1 correlation coefficient obtained using the MATLAB `autocorr` function (MATLAB R2006a, Econometrics Toolbox) which calculates the sample autocorrelation coefficient for the time series using neighbouring time points.

We use the following formulae to obtain a 95% confidence interval for the estimate of S, $\hat{S}$. For $n_e > 30$ we can assume normality and calculate the interval using

$$\left[\hat{S} \pm Z(0.025)\frac{s}{\sqrt{n_e}}\right], \tag{2.5}$$

where s is the sample standard deviation of the RMSS values, $Z(0.025)$ is the critical value of the Cumulative normal distribution at 0.975 and n_e is the equivalent sample size as above. When $n_e \leqslant 30$ we must use the t-distribution with $n_e - 1$ degrees of freedom, thus the interval here is calculated using

$$\left[\hat{S} \pm t_{n_e-1}(0.025)\frac{s}{\sqrt{n_e}}\right]. \tag{2.6}$$

An example 95% confidence interval for $\hat{S}$ from the *in silico* data in Figure 3 where $S = 1$ and $\hat{S} = 0.9973$ is $[0.9937, 1.0010]$.

P estimate

The VACF is used to estimate P. This is done by first calculating the VACF from 3D data using

$$\text{VACF(t)} = \langle (v_x(0) \cdot v_x(t)) + (v_y(0) \cdot v_y(t)) + (v_z(0) \cdot v_z(t)) \rangle, \tag{2.7}$$

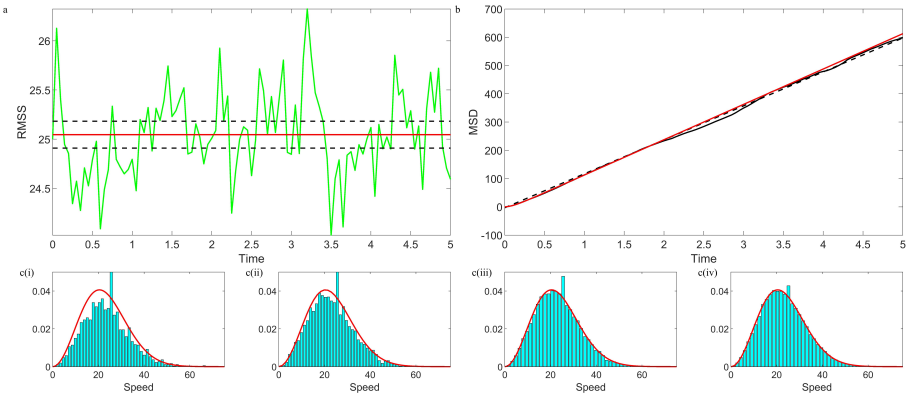

FIG. 3. **Example *in silico* output from the framework in 3D** for 1000 cells with NPeriods $= 1000$, $dt = 0.05$ and $S = P = 1$. Cells are initialised at the origin, $\mathbf{x}_0 = \mathbf{0}$, with speed S and orientation sampled uniformly from the unit sphere. **a)** RMSS over time (grey line, green online) with estimated speed $\hat{S} = 0.9973$ (solid black line, red online) 95% confidence interval $[0.9937, 1.0010]$ ($n_e = 51$) (black dashed lines). **b)** Calculated MSD vs time (solid black line) with model predicted MSD (thin black line, red online) and a straight line fitted to the calculated MSD (black dashed line), enabling S and P estimates to be verified through the gradient of the line being equated to $2S^2P$. The inferred P estimate here is $\hat{P} = 0.9806$, whilst the framework estimated value is $\hat{P} = 0.9951$ with 95% confidence interval [0.9232, 1.0791]. **c)** Histograms of cell speed distributions at $t =$**i** 0.25, **ii** 0.5, **iii** 2.5 and **iv** 25, and Maxwell-Boltzmann distribution (black curve, red online) with estimated parameter $\hat{S} = 0.9973$ overlaid.

where the average $<>$ is over all cells. From equation 1.5, we can estimate $-1/P$ as the gradient of a plot of $\ln(\text{VACF})$ against t. We note that we are using a special case of the O-U process in which correlations in the increments of v_x, v_y and v_z are absent, simplifying the VACF calculation and thus potentially affecting the model's ability to describe any data sets where these correlations may be present.

To obtain an estimate of the gradient of this line, we consider a simple linear regression model fitted to the observed $\ln(\text{VACF})$ values and in doing so directly calculate the estimate for P. Given that our observations are serially correlated, and thus the errors involved in fitting this regression line will also be correlated, we fit this line using feasible generalised least squares (FGLS) instead of the traditional ordinary least squares (OLS) method. To this end we use the MATLAB `fgls` function (MATLAB R2014b, Econometrics Toolbox) to obtain the line of best fit along with estimates for the slope coefficient and its corresponding standard error estimate.(See Supplementary Information Section 5.2.2 for more details).

As VACF tends towards zero there is increasing noise in the estimate of $\ln(\text{VACF})$, and an estimate of P that uses all of this data would be erroneous. Figure 4, particularly plot a(i), shows just how noisy the data can be. To ensure the estimates are not affected by this noise, we only fit our regression model to a subset of data points by implementing a cut-off value. Observations of $\ln(\text{VACF})$ falling below this value are excluded from the dataset.

To determine this new subset, we systematically try a range of cut-off values for $\ln(\text{VACF})$, the line being fitted only to those values above the cut-off, by defining a cut-off test vector with equally spaced entries between the minimum and maximum values of $\ln(\text{VACF})$. We subsequently calculate the mean squared error (MSE) for each fit and choose the cut-off for which the subset includes the most data points such that MSE < 0.5, and proceed as above using the `fgls` function to carry out the rest of the analysis. This choice of MSE cut-off will depend on the simulation parameters, for example number

of cells N, the required accuracy of parameter estimates and MSE obtained from fitted models, but the methodology provides a repeatable and adjustable method for estimating P and the cut-off is one of the parameters that is easily changed. We also restrict the search to subsets with more than 5 data points to allow FGLS to be used, as we are fitting 4 parameters in the regression model (intercept, slope, variance and autocorrelation).

To obtain an estimate, $\hat{P}$ we apply the `fgls` function to the resulting subset of ln(VACF), choosing to fit to $-\ln(\text{VACF})$ to simplify the algebra and make $P = 1/\hat{\beta}$, where $\hat{\beta}$ is the estimated slope coefficient. We can then obtain a confidence interval for our P estimate by building the 95% confidence interval for slope coefficient $\hat{\beta}$ as

$$\left[\hat{\beta}_L = \hat{\beta} - t_{n-2}(1-\alpha/2)\, SE_\beta,\, \hat{\beta}_U = \hat{\beta} + t_{n-2}(1-\alpha/2)\, SE_\beta\right]$$

where n is the number of data points in the subset, $\alpha = 0.05$, t_{n-2} denotes the t-distribution with $n-2$ degrees of freedom and SE_β is the estimated standard error of $\hat{\beta}$, and transforming this to obtain the 95% interval for $\hat{P}$ as

$$\left[\frac{1}{\hat{\beta}_U}, \frac{1}{\hat{\beta}_L}\right].$$

This formula can also be used to calculate a 99% confidence interval where necessary by setting $\alpha = 0.01$.

The plots in Figure 4 are formed from fitting the regression model to ln(VACF), and show how different ln(VACF) data sets force subsets of this data of different lengths to be used for FGLS fitting and subsequent $\hat{P}$ estimation, according to the MSE cut off algorithm explained above. We expect this regression line to have an intercept, which is also fitted in the model, at $\ln(S^2)$, and so it can be useful to compare this value to the estimated intercept given by the regress function as another way of assessing how well the PRW model can explain a dataset. Figure 4 shows examples of the framework output ln(VACF) plots for $S = 1$, $P = 1$ and 10 and where dt is taken to be 0.01, 0.1 and 1, and the choice of cut-off is determined by the above algorithm. This produces P estimates, along with their 95% confidence intervals, of $\hat{P} = 0.9893$ [0.9473, 1.0352], $\hat{P} = 1.0919$ [0.9999, 1.2505] and $\hat{P} = 1.1497$ [1.0661, 1.2474] for input parameters $S = 1$, $P = 1$ and $\hat{P} = 10.3752$ [9.9281, 10.8644], $\hat{P} = 9.8853$ [9.5095, 10.2919] and $\hat{P} = 10.3866$ [9.8946, 10.9032] for inputs of $S = 1$, $P = 10$ for dt as stated in the above order. Numerical simulations of stochastic differential equations, and associated statistical measures, are strictly valid in the limit as $dt \to 0$. In our simulations, when the persistence time, P, is comparable to dt, we see the reduction in predictive power; for example when $dt = P = 1$, as in figure 3a(iii), the confidence interval doesn't include what we know to be the true value of P. More details of the exact process of estimating P are given in section 5.2 of the Supplementary information.

Mean Squared Displacement

Upon calculating estimates for both S and P, the theoretical MSD from equation (1.4) can be compared with the calculation from the data:

$$\text{MSD(t)} = \left\langle (x(t)-x(0))^2 + (y(t)-y(0))^2 + (z(t)-z(0))^2 \right\rangle, \tag{2.8}$$

where the average $<>$ is over all cells and the position vector at time t is given by $(x(t), y(t), z(t))$. Figure 3b) shows a plot of the calculated MSD vs model MSD for $S = 1, P = 1$ as an example.

We note from equation 1.4 that in the limit as $t \to \infty$, the expression for MSD becomes $\text{MSD(t)} = 2S^2Pt$, the equation of a straight line with a slope of $2S^2P$. Fitting a regression model to the calculated

MSD vs t plot, making use of FGLS since the MSD observations from each time step will depend on previous MSD observations, we can also infer $\hat{P}$ using $\hat{S}$ as

$$\hat{P} = \frac{\text{slope}_{\text{MSD}}}{2\hat{S}^2},$$

with $\text{slope}_{\text{MSD}}$ being the estimated slope coefficient from the FGLS fit to the MSD vs t plot.

In doing so for data shown in Figure 4, for simulated data with parameters $S = 1$, $P = 1$ we obtain $\hat{P} = 0.9141, 0.9533, 0.9845$ for $dt = 0.01, 0.1, 1$, and $\hat{S} = 0.9978, 0.9993, 1.0020$ respectively and for simulated parameters $S = 1$, $P = 10$, we obtain $\hat{P} = 3.6942, 9.4376, 9.7890$ for $dt = 0.01, 0.1, 1$, and $\hat{S} = 1.0097, 0.9943, 1.0007$ respectively. For each of these datasets we took NPeriods to be 1000. We posit that the total simulation time for the dataset in Figure 4b(i) is not large enough compared to P, 10, to use the fact that $\text{MSD(t)} \to 2S^2Pt$ as $t \to \infty$ to justify a linear fit to the data, hence the very poor estimate of P found through this method. In this case a non-linear fit of equation 1.4 should be carried out to infer an estimate for P.

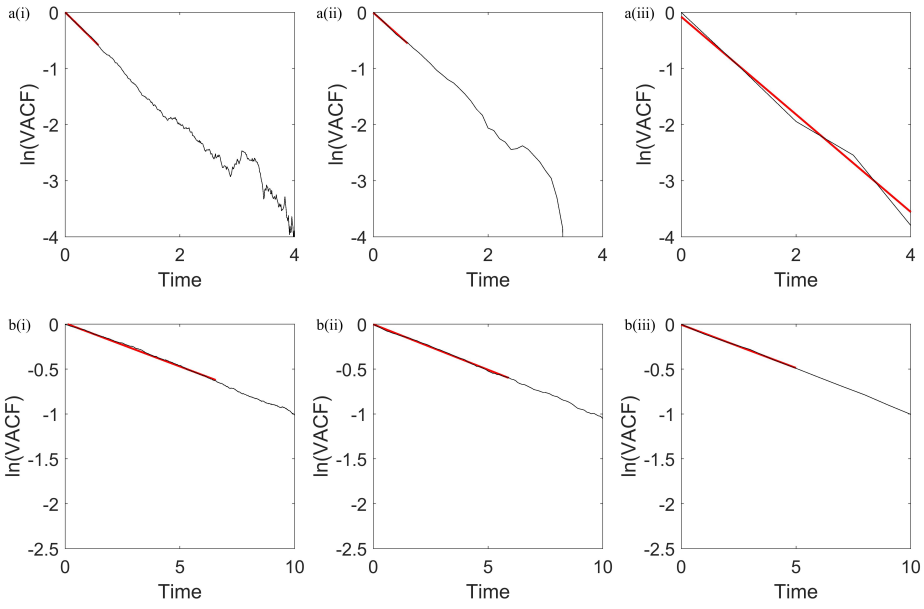

FIG. 4. **Estimation of P using** ln(VACF) **with algorithmic cut-off points for line fits**. Calculated values of ln(VACF) are shown (main black line) with FGLS line fits (straight black line, red online). FGLS line fits differ in length in each panel due to subsets of ln(VACF) data of varying length used in the estimation of $\hat{P}$, according to the MSE cut off algorithm defined in the main text. **a)** $S = 1, P = 1$, Nperiods = 1000, **i** $dt = 0.01$, **ii** $dt = 0.1$ and **iii** $dt = 1$ respectively for 1000 cells. P estimates from left to right along with 95% confidence intervals are $\hat{P} = 0.9893$ [0.9473, 1.0352], $\hat{P} = 1.0919$ [0.9999, 1.2505] and $\hat{P} = 1.1497$ [1.0661, 1.2474]. **b)** $S = 1, P = 10$, Nperiods = 1000, **i** $dt = 0.01$, **ii** $dt = 0.1$ and **iii** $dt = 1$ respectively for 1000 cells. P estimates from left to right along with 95% confidence intervals are $\hat{P} = 10.3752$ [9.9281, 10.8644], $\hat{P} = 9.8853$ [9.5095, 10.2919] and $\hat{P} = 10.3866$ [9.8946, 10.9302].

Discussion of Model Parameters and output

Estimation of parameters from *in silico* data allows us to validate our method and assess the accuracy of our estimates. Having demonstrated our framework can successfully extract these values from 3-dimensional *in silico* simulations, we will go on to estimate S and P from experimental data in the next section, and also check that the workflow is robust for *in silico* data generated from experimental estimates for S and P.

It is clear from Figure 3 that the simulated speeds follow the Maxwell-Boltzmann distribution after enough time has passed for the stationary distribution to be reached. This means we can be reasonably confident that the RMSS will be a good estimate of S in a population that follows the PRW model, and this is confirmed by the narrow confidence intervals calculated for the examples given.

We can also consider the velocity distribution for each of the components of the velocity which we assume to be Gaussian. For consistency, we conduct an Anderson-Darling test using the function (MATLAB `adtest`, Statistics and Machine Learning Toolbox, R2013a) on each of the components of velocity (v_x, v_y, v_z) to check that the assumption is indeed satisfied. If this assumption is violated then we wouldn't expect speeds to follow the Maxwell-Boltzmann distribution which depends on these Gaussian velocities. The Anderson-Darling test was conducted at each time point across all cells with the final time point being taken particularly into consideration. For all *in silico* data sets in the paper, the Anderson-Darling test showed that at the final time point all components of velocity were normally distributed, hence giving further confidence in the S estimate.

When we are looking at estimating P we also need to be careful with the timescale we are simulating over. The simulation interval dt needs to be much smaller than P to be able to see the persistence in velocity over several time periods and subsequent decay of the velocity autocorrelation. We should also ensure that the total simulation time is much larger than P to be able to see the effect of the decay in correlation. We should therefore get a more accurate P estimate with values of dt much smaller than P and a high number of simulation periods.

We also note that the choice of MSE threshold is important here. When testing the framework with *in silico* data, estimates of both S and P were seen to be robust to MSE choice, even when the threshold was as small as 0.05. The MSE should not be too large, but overfitting to the *in silico* data could lead to poor prediction in experimental datasets. The MSE threshold was thus set at 0.5 to be consistent with the chosen threshold for the experimental data sets in our analyses. In practice the MSE threshold should be set based on the dataset being investigated, it being sensitive to sample size. The choice will be dependent on the amount of data once observations have been removed as per the cut-off algorithm, and values of MSE that an investigator deems acceptable in relation to the context of the experimental data itself.

Figure 4 shows the framework output when different values of dt are used and demonstrates how P estimates vary as dt varies between 0.01 and 1. We would expect estimates to become more accurate as dt decreases, and this is seen here when $P = 1$ but not when $P = 10$, possibly due to the way that we choose a subset of data to use when estimating P. In reality the choice of dt, number of cells and the number of simulation periods may be restricted by the data from an experiment, and so consideration of how to amend the framework in these cases may be necessary.

Experimental Data

After validating the method using pre-determined parameter values for *in silico* data, we can now reliably use it to extract parameter values from a 3D experimental dataset. The cell tracking data used here was obtained from *in vitro* tumour spheroids consisting of glioblastoma cells. These spheroids were

grown and imaged with a Light Sheet Fluorescent Microscope, as described in Richards *et al.* (2018), and of importance here is the fact that images were collected every 3 minutes over a 24 hour period, meaning there are 480 periods of 0.05 hours in the dataset. Though the spheroids were in some instances treated with drugs, the 3 datasets we use are all controls.

The data is in the form of individual cell tracks, there being a velocity at each time step for each cell as required. Plots of the tracks from one of these control spheroids are shown in Figure 2 d) compared to the experimental image in 2 c). There were 3780, 3861 and 3808 cells in each of the three experimental data sets though only cell tracks that are recorded as starting at time 0 are included in the analysis, thus we analyse the 549, 929 and 1054 cells with such tracks for 149, 93 and 78 periods of 0.05 hours in control spheroid datasets 1, 2 and 3 respectively, meaning we look at time periods of 7.5, 4.7 and 3.9 hours.

Parameter estimation using framework applied to experimental data

Upon running the data through our framework we were able to obtain estimates for parameters S and P along with 95% confidence intervals as $\hat{S}_1 = 27.3137\,\mu$m/h [25.2892, 29.3382], $\hat{P}_1 = 0.0863$h (5.18 min) [0.0697, 0.1130], $\hat{S}_2 = 26.9272\,\mu$m/h [25.9613, 27.8930], $\hat{P}_2 = 0.0789$h (4.73 min) [0.0677, 0.0946] and $\hat{S}_3 = 28.0600\,\mu$m/h [27.3979, 28.7222], $\hat{P}_3 = 0.0976$h (5.86 min) [0.0804, 0.1241]. In the calculations of the confidence intervals for $\hat{S}$ we found effective sample sizes of $n_e = 16.6, 19.5$ and 29.5, resulting from sample autocorrelations of 0.8064, 0.6635, and 0.4617 at lag 1. Output plots from the framework can be seen in Figure 5 for each of the three spheroids.

Our speed estimates agree well with the estimate of 27μm/h obtained from the same dataset for cells located inside the spheroid boundary in Richards *et al.* (2018). In terms of our estimates for P, there are very few sources in the literature which predict persistence time for any type of cell, less so for GBM cells, but we note Stein *et al.* (2007) carried out similar analysis to ours studying GBM U87 cells from 2D projections of 3D images and obtained a value of $\beta = 9.3/$h, corresponding to $P = 1/\beta = 0.1075$h which is similar to the values we find.

Our estimates for both S and P are additionally very consistent across the controls, making them fairly reliable for this experiment. We can also again infer P from the MSD calculations for comparison, using $\hat{S}_1, \hat{S}_2$ and $\hat{S}_3$ as specified above and obtaining corresponding values for $\hat{P}$ of 0.0940h, 0.1289h, 0.1017h, all of which are reasonably consistent with the estimates taken from the VACF.

Looking closer at the regression line fitted to $\ln(\text{VACF})$ we can gain a further two estimates for S, those coming from studying the intercept of the regression line and the actual experimental value of the autocorrelation function at time 0, which the model says are equal to $\ln(S^2)$ (from equation 1.5). For control spheroid 1 the experimental value is 7.2049 leading to an S estimate of $\hat{S}_1 = 36.6880\mu$m/h and the value predicted by the regression is 7.4978 giving an S estimate of $\hat{S}_1 = 42.4743\mu$m/h. Similarly for spheroid 2 we get $\ln(\hat{S}_2^2) = 6.7969$ giving $\hat{S}_2 = 29.9177\mu$m/h and a regression intercept of 7.0927 giving $\hat{S}_2 = 34.6865\mu$m/h. Finally for spheroid 3 we obtain $\ln(\hat{S}_3^2) = 6.8386$ giving $\hat{S}_3 = 30.5480\mu$m/h and regression intercept 7.1186 giving $\hat{S}_3 = 35.1386\mu$m/h. Compared to the estimates from the model framework obtained through RMSS, independently of P, the predictions from the regression overestimate in each case, though the experimental values are also above the values that the framework estimates. We suggest that the most reliable method of estimating S is still the one using the RMSS as this encompasses the definition of parameter S and provides the estimates closest to those found by experimentalists.

As further validation that our framework should be able to correctly extract parameters from the data, Figure 6 shows the framework output for a realistic set of parameter values as informed by running the experimental data through the framework ($S = 25$, $P = 0.1$, $dt = 0.05$, NPeriods = 100, 550 cells).

This shows that our framework is still capable of estimating S and P accurately when the experimental parameter values are used, even with the restricted dt value. The estimated value of P is 0.0996 with 95% confidence interval [0.0978, 0.1015], which includes the true value of $P = 0.1$. This is a good sign we can be reasonably confident in our intervals for $\hat{P}$ that come from the experimental data P estimates.

By conducting this analysis we are able to explore 'realistic' parameters in the framework and see how well it is capable of estimating parameters of this magnitude. This enables us to see if we can indeed make accurate predictions about the experimental data using the framework, but also to test the robustness of it when parameters take values similar to these. For example we have estimated P to be around 0.1, and since we need $dt << P$ to see persistence over several time intervals, we should determine whether the framework can handle values of P which are quite close to dt, as in the experimental case where dt is 0.05. We see from the output of this *in silico* simulation with experimental parameters that the framework is capable of handling such parameters, and thus can go on to make conclusions about the experimental data knowing that any discrepancies arising are not down to the framework's estimation capabilities, but to experimental errors or biological phenomena.

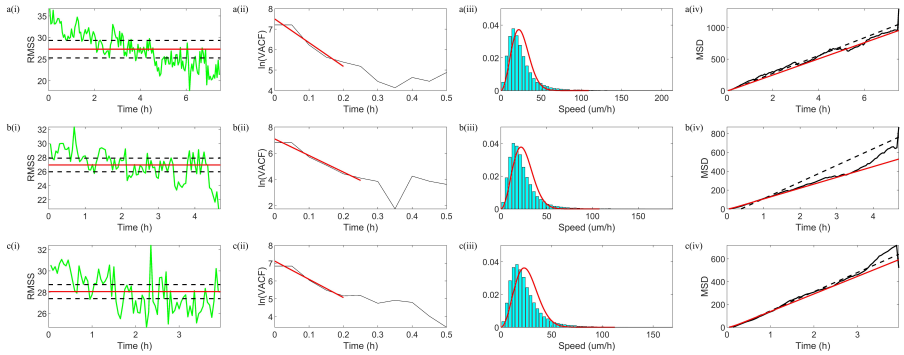

FIG. 5. **Framework outputs for experimental spheroids**, data from Richards *et al.* (2018). (i) RMSS vs time (grey line, green online) with estimated average speed (solid black line, red online) and 95% confidence intervals (black dashed lines). (ii) Velocity autocorrelation, calculated ln(VACF) (main black line) and FGLS fit (straight black line, red online). (iii) Histogram of speeds with Maxwell-Boltzmann density with parameter S overlaid (black curve, red online). (iv) Calculated MSD vs time plot (black line) and model predicted MSD (thin black line, red online) with line fit (black dashed line). Each row corresponds to an independent control spheroid.

Probing model assumptions using framework applied to experimental data

We shall now highlight some fundamental differences depicted in Figure 5 between model predictions and experimental observations, and where possible propose rigorous statistical tests to examine whether the PRW model should be rejected. By comparing figure 6, which shows a very good fit of model to *in silico* data, to figure 5 we have confidence that differences between model predictions and observations are not due to sample size or parameter values in this case, but instead that perhaps the PRW model is not sufficient to describe this data.

To undertake statistical tests to determine whether the PRW model should be rejected, we choose to consider in more detail a subset of cell tracks which last the full length of the experiment. Firstly, we

see that the Maxwell-Boltzmann distribution appears unable to completely explain the speed distribution data. If we consider only the final cell speeds in each of these tracks (Figure 7a,b,c(iii)), we have a set of independent speeds which should follow the Maxwell-Boltzmann distribution, as we are looking at a fairly large number of cells (76, 71 and 56) over a long time (149, 93 and 78 periods respectively). We can first conduct the Anderson-Darling test on the velocities in the experimental data sets and upon doing so, even if we restrict the test to just the full length tracks in each data set, we are still led to reject the null hypothesis in all cases. This suggests that the velocities are not normally distributed and so consequently we shouldn't expect the speeds to follow the Maxwell-Boltzmann distribution.

Further, carrying out a Kolmogorov-Smirnov test (MATLAB `kstest`, Statistics and Machine Learning Toolbox, R2006a) on the final cell speeds of full length tracks for each control spheroid with parameter $\hat{S}$ as estimated from the data through our framework, we see that in all cases this test instructs us to reject the null hypothesis that the data follows the Maxwell-Boltzmann distribution. Furthermore, we see from Figure 7a,b,c(iv) that the mean speeds of each cell with a full length track are not clustered around the mean of the expected Maxwell-Boltzmann distribution based on the estimated speed parameter S. This leads us to believe that each cell monitored over the full experiment isn't itself displaying speeds following the Maxwell-Boltzmann distribution with this parameter S.

All of this suggests that the cell speeds are not what we would expect if the cells behaved as per the model, and so there are some cells travelling quite a bit faster and some cells quite a lot slower than the estimated mean speed (estimated mean speed $\pm$ standard deviation, spheroid 1: 27.3137 ± 0.0027 μm/h, spheroid 2: 26.9272 ± 0.0028 μm/h, spheroid 3: 28.0600 ± 0.0026 μm/h). This provokes interesting biological questions about why some cells are able to travel at higher speeds than their counterparts and perhaps looking at where these cells lie in the spheroid would provide some insight into this difference and differences in motility mechanisms across cells. Upon plotting individual cell speeds across the experiment we see that there are indeed some cells with abnormally high speeds at certain times, and the peaks in speed are coming from the same cells, generally those with higher mean speeds overall, though their speed is not consistently higher than we would expect. These plots can be seen in Figure 1 in section 6 of the Supplementary information. We could probe this more by looking more in detail at how the speed distributions vary over time, monitoring when this shift in the peak of the distribution happens and when the high-speed outliers become so, to determine whether these mean speeds are so high due to extreme values at later times, or are simply down to chance.

Secondly, in Figure 5, we see that the RMSS appears to be a function of time, with the data suggesting a linear decrease, in conflict with the underlying assumptions of the PRW model. We questioned whether this trend for decreasing RMSS over time is due to the decreasing number of tracks involved in the calculation as time goes on, due to initial filtering of the data according to the start time of a track. However RMSS plots created with only the full length tracks as used in Figure 7 show a similar downward trend (data not shown) and thus more data is needed to investigate this changing of speed with time. We are also assuming here that the system is already in steady state due to the cells being grown for 3 days before the tracking started and time 0h is 3 hours after the spheroid has been placed in the microscope chamber. This assumption could be wrong and could explain the decrease in speed over the time interval we are considering. Nevertheless, it is clear that the RMSS time series plot is one of the first indicators from the framework of whether a dataset has a constant average speed, and thus one of the first ways to assess the suitability of the PRW model to describe a dataset.

Thirdly, in Figure 5, we see that for all of the control spheroids, the model MSD underpredicts the calculated MSD, leading us to take care with the P estimates inferred from the MSD calculations. This underprediction agrees with the previously observed superdiffusive nature of cells in 3D (Yurchenko *et al.*, 2019; Luzhansky *et al.*, 2018; Takagi *et al.*, 2008).

Finally, in Figure 5, the plots of ln(VACF) against time, for which the PRW model predicts to be linear with slope $-1/P$, show a reasonable fit. Moreover the noise in the data leads the method for estimating P to only use a small subset of data points when fitting the straight lines, possibly affecting the reliability and accuracy of our estimates. The nonlinear nature of the ln(VACF) plots suggests there may be some biological factors which stop the model from being able to accurately estimate P from these plots. In this framework we are assuming that all cells are identical and independent, which is intuitively unrealistic, and accounting for possible differences in persistence time between cells could enable us to better estimate what these parameter values may be. It has been suggested in the literature (Wu *et al.*, 2014; Yurchenko *et al.*, 2019; Takagi *et al.*, 2008) that populations of cells may have several subpopulations with different persistence times. We don't explore this idea here, but one could change the framework accordingly to account for this by using a different governing model that allows for heterogeneity in individual values of population parameters. The statistical measures calculated in the framework currently based on the PRW model could not be adjusted sufficiently to account for significant heterogeneity in the population, the model itself assuming that there is one S and one P value for the entire population.

To illustrate this we ran a heterogeneous *in silico* data set through the framework which consisted of just 2 possible S values. This data set consisted of 100 cells and was run for 1000 periods. We gave half of the population $S = 1$ and the rest $S = 3$, whilst keeping $P = 1$ for all cells. We would expect to retrieve $\hat{S} = 2, \hat{P} = 1$ from the framework in this case. The framework gave $\hat{S} = 2.2638\,[2.2081, 2.318]$, $\hat{P} = 1.5272\,[0.9811, 3.4445]$. We see from this small experiment alone that introducing even a small amount of heterogeneity has meant that the estimate of S is poor and the confidence interval doesn't contain the true value. We also see that the P estimate is poor and has a much wider confidence interval than we have seen with homogeneous data. Thus it is our suggestion that if significant heterogeneity in the data is present then a model different from the PRW model should be used in the framework to ensure that estimates are not biased in this way.

Discussion and conclusions

We present an example of a rigorous combined mathematical and statistical approach for analysis of 3D cell tracking data using stochastic models. The framework we have developed provides tools for calculating various statistical measures for testing goodness of fit and for parametrising the given model, here demonstrated using the Persistent Random Walk model. This model has been chosen in the knowledge that it is perhaps not complex enough to fully capture the motility seen in GBM spheroids, but is one of the most popular stochastic models used in cell motility. The ill-fitting nature of the model though allows us to exploit the framework and show its potential in uncovering features of a data set that may be missed by less rigorous analysis or a more well-fitting, but not optimal model. We also make clear the distinction between the PRW model in all 3 physical dimensions, and how the governing equation changes based on the dimension-dependent diffusion coefficient, something which has not previously been stated as clearly.

Our framework outputs parameter estimates along with confidence intervals and uses statistical measures to provide them, all of which take into account serial correlation in the data. This has been lacking in the literature in this context until now, to the best of our knowledge. We believe the approach we present is adaptable to other models and data sets by simulating *in silico* data sets using the model of choice and using statistical measures appropriate for the data being studied in the same way we have

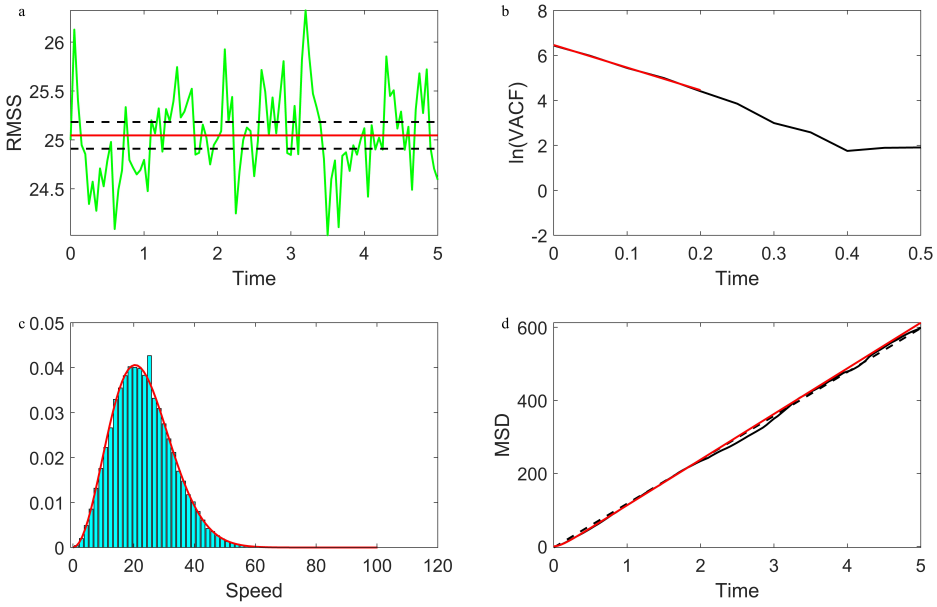

FIG. 6. **Testing *in silico* output based on experimental parameters** for 550 cells over 100 simulation periods with $dt = 0.05$ and $S = 25, P = 0.1$. Cells are initialised at the origin, $\mathbf{x}_0 = \mathbf{0}$, with speed S and orientation sampled uniformly from the unit sphere. **a)** RMSS over time is shown (grey line, green online) with estimated average speed $\hat{S} = 25.0458$ (solid black line, red online) and 95% confidence interval $[24.9091, 25.1825]$ ($n_e = 40$) (black dashed lines). **b)** Calculated ln(VACF) vs time (main black line) with FGLS line fit (straight black line, red online) giving $\hat{P} = 0.0996$ with 95% confidence interval [0.0978, 0.1015]. **c)** Histogram of speeds with Maxwell-Boltzmann density with parameter S overlaid (black curve, red online). **d)** Calculated MSD vs time (black line) with model predicted MSD (thin black line, red online) and a straight line fitted to the calculated MSD (black dashed line). The inferred P estimate from the MSD calculations is $\hat{P} = 0.0962$.

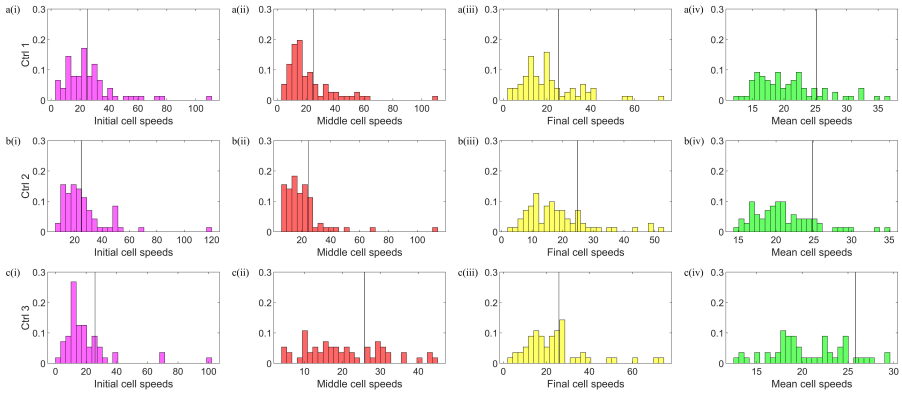

FIG. 7. **Speed distribution for cells with tracks lasting the full length of the experiment** (μm/h) : (i) initial speed distribution; (ii) intermediate speed distribution; (iii) final speed distribution; (iv) mean cell speed distribution. Vertical black lines display the estimate for $2S\sqrt{\frac{2}{3\pi}}$, the mean value of the theoretical Maxwell-Boltzmann distribution. Each row corresponds to an independent spheroid.

demonstrated to compare the model with the data. It is the consistency and thoroughness of the approach which allows for elucidation of possible reasons for mismatch between a model and a data set, but also suggest routes for further exploration of 3D cell tracking data sets such as the ones we explore here.

The framework as a package is useful for experimentalists looking to analyse tracking data without necessarily having the mathematical or statistical background required to carry out such rigorous analysis, but there is also great benefit for modellers alike in being able to test potential models for a data set with the same consistent, thorough method of testing, allowing for direct comparison of population level statistics between models. There are also benefits to those looking at initial analysis of a data set before moving on to more complex considerations by way of the plots that the framework outputs, as well as quantitative descriptions of population and individual track characteristics such as speeds and correlations in velocity.

The framework has been tested on *in silico* datasets in 3 dimensions through the use of statistical measures MSD, VACF and speed histograms, before being applied to experimental 3D cell tracking data collected from GBM tumour spheroids. Results show that the PRW model may not be complex enough to describe these particular data sets well, as shown by the experimental data having different speed distributions and MSDs than predicted by the model. Though others have reached this conclusion before for other cells types (Wu *et al.*, 2014; Dieterich *et al.*, 2008; Metzner *et al.*, 2015; Upadhyaya *et al.*, 2001; Cherstvy *et al.*, 2018; Loosley *et al.*, 2015), we have done so with statistically significant proof and through following the same rigorous procedure for each data set. These findings allow us to question what about the model itself and the biology needs further investigation, likely taking into account the proliferative and heterogeneous nature of cancer cells. For example our investigation of the cell speed distribution allows us to see that although the bulk of the cells are travelling as we expect, there are some outliers moving particularly fast and there is surely an interesting biological reason behind this.

In this analysis, we were grateful to have a large dataset to work with, though we excluded any tracks

that did not begin at the start of the experiment, greatly reducing our available data. We were however still able to reach the conclusions above and reject the PRW model for the data, with strong evidence to back this up. This demonstrates that it is not the amount of data that is vital here, but experimental parameters such as the frequency with which measurements are taken, and the length over which cells are studied. This is just one example of how iterating between mathematical models and experiments will elucidate new directions for modelling and study of biological systems.

We also note that it would be possible to switch the frequentist statistical analysis conducted in the framework currently with its Bayesian equivalents, which comes with its own set of advantages and disadvantages. We have replicated the analysis in this work using Bayesian methodology and obtain comparable estimates with the frequentist approach (Scott, 2021), thus leading us to recommend that a potential user of the framework chooses the methodology that suits them.

Cancer is a complex condition in which cells interact with each other and with many other molecules within a tumour microenvironment. Cells can also vary between themselves, and are capable of changing their own behaviour in response to certain stimuli. This presents a problem with creating models simple enough to test certain motility hypotheses for a dataset such as the one we have been working with, given the wide range of conditions that would need to be taken into account. This challenge only increases when drugs are brought into the system and so there is plenty of scope for the model to be adapted to incorporate any or a range of these complications. We do however see the potential of a framework such as ours to be able to estimate motility parameters under different conditions, particularly when spheroids are treated with drugs.

In future work we would endeavour to consider problems outlined throughout the paper such as ensuring the suitability of experimental data for this framework and how to make use of all available data when carrying out the analysis. We could further consider adapting the model so that the alternative ideas about MSD, VACF and velocity distributions may be studied rigorously (Yurchenko *et al.*, 2019; Luzhansky *et al.*, 2018; Takagi *et al.*, 2008). We would also hope to be able to add alternative terms into the model to better describe how the cells are moving in response to chemical stimuli in addition to random motion. Cells are known to have a 3-step migration cycle (Lauffenburger & Horwitz, 1996; Mitchison & Cramer, 1996) consisting of protrusion of the cell's leading edge, adhesion of this region to the underlying substrate and then contraction of the cell body causing detachment of the rear of the cell, which has been incorporated into some cell migration models, though cells in 3D change the mode of their migration depending on the geometry of their environment (Wolf *et al.*, 2013; Wu *et al.*, 2018; Mierke, 2015). A lot of the differences between 2 and 3D cell migration are as a result of the Extracellular Matrix (ECM) which surrounds cells in 3D. For example, the availability of space for cells to move through (Wolf *et al.*, 2013; Tozluoğlu *et al.*, 2013), resistance cells face from the ECM, the viscosity and stiffness of the matrix (Zaman *et al.*, 2005, 2006; Wang *et al.*, 2014), and the presence or absence of matrix proteins (Fraley *et al.*, 2015; Wu *et al.*, 2018) can all affect the migration of a cell in 3D. Thus, following suit of others in the field, incorporating terms that describe the influence of the ECM would no doubt improve the model fit, as well as considering other phenomena such as chemotaxis, haptotaxis and gradients in nutrients, and more specific to cancer, angiogenesis, hypoxia and necrosis.

It would also be informative to add cell-cell interactions into the model, as this may be one of the reasons for the mismatch between the PRW model and the experimental data. In order to study this further one could look to models of 3D cell motility that include interaction terms such as the Vicsek model (Vicsek *et al.*, 1995; Cziròk *et al.*, 1999; Liu, 2010), that of Sepúlveda *et al.* (2013) or of Matsiaka *et al.* (2019). Generating *in silico* data from any of these models and running this data through the framework would reveal whether interactions do need to be included in the model, evidenced by further

mismatch. This would allow rigorous study of how interactions affect the statistical measures and parameter estimates and potentially suggest sensible avenues of exploration for alternative models to place within the framework.

In order to consider the issues discussed here, it would be beneficial to obtain data collected from the entire spheroid. A drawback of the imaging data collected is that there was difficulty penetrating into the centre of the spheroid, and thus only around half of the movements of the cells within the spheroid were able to be tracked (Richards, 2016). Information on the behaviour of trajectories throughout the spheroid would facilitate investigation into correlations between trajectory location and cell behaviour, elucidating spatial effects and how crowding may impact migration and proliferation.

For now, we present this framework as a data-driven, rigorous methodology for testing whether a cell tracking dataset could reasonably be described by a given model. It provides statistical measures for assessing how realistic the model is for a data set, and tests whether we can obtain estimates of population level parameters using individual cell properties, paving the way for future interrogation of cell tracking data and investigation of cell motility in 3D.

Code used for the framework can be found at https://github.com/m-scott22/PRW3DCellMotilityFramework

Acknowledgements

Author contributions: M.S. and R.N.B designed research, developed the numerical simulations and analysed the experimental data. M.S. and K.Z. designed and implemented the statistical tests, and M.S., K.Z., and R.N.B. interpreted the results and wrote the paper.

This work is supported by EPSRC EP/N014499/1 'EPSRC Centre for New Mathematical Sciences Capabilities for Healthcare Technologies'

The authors would like to thank Rosalie Richards, Dave Mason and Violaine See for allowing the use of their experimental tracking data, and for helpful discussions on the biological aspects and implications of the framework output.

References

AGOSTI, A., GIVERSO, C., FAGGIANO, E., STAMM, A. & CIARLETTA, P. 2018 A personalized mathematical tool for neuro-oncology: A clinical case study. *International Journal of Non-Linear Mechanics* **107**, 170–181.

ANDERSON, A. R. A. & QUARANTA, V. 2008 Integrative Mathematical Oncology. *Nature Reviews Cancer* **8**, 227–234.

ANTONI, D., BURCKEL, H., JOSSET, E. & NOEL, G. 2015 Three-Dimensional Cell Culture: A Breakthrough *in Vivo*. *International Journal of Molecular Sciences* **16**, 5517–5527.

ANTONOPOULOS, M., DIONYSIOU, D., STAMATAKOS, G. & UZUNOGLU, N. 2019 Three-dimensional tumor growth in time-varying chemical fields: a modeling framework and theoretical study. *BMC Bioinformatics* **20** (1), 442.

ANTONOPOULOS, M. & STAMATAKOS, G. 2015 In Silico Neuro-Oncology: Brownian Motion-Based Mathematical Treatment as a Potential Platform for Modeling the Infiltration of Glioma Cells into Normal Brain Tissue. *Cancer Informatics* **14** (Supplement 4), 33–40.

CAMPOS, D., MÉNDEZ, V. & LLOPIS, I. 2010 Persistent random motion: Uncovering cell migration dynamics. *Journal of Theoretical Biology* **267** (4), 526–534.

CHERSTVY, A. G., NAGEL, O., BETA, C. & METZLER, R. 2018 Non-Gaussianity, population heterogeneity, and transient superdiffusion in the spreading dynamics of amoeboid cells. *Physical Chemistry Chemical Physics* **20** (35), 23034–23054.

COLOMBO, M. C., GIVERSO, C., FAGGIANO, E., BOFFANO, C., ACERBI, F. & CIARLETTA, P. 2015 Towards the Personalized Treatment of Glioblastoma: Integrating Patient-Specific Clinical Data in a Continuous Mechanical Model. *PLoS ONE* **10** (7), e0132887.

CZIRÒK, A., VICSEK, M. & VICSEK, T. 1999 Collective motion of organisms in three dimensions. *Physica A: Statistical Mechanics and its Applications* **264**, 299–304.

DEISBOECK, T. S., ZHANG, L., YOON, J. & COSTA, J. 2009 In silico cancer modeling: Is it ready for prime time? *Natue Clinical Practice Oncology* **6**, 34–42.

DIETERICH, P., KLAGES, R., PREUSS, R. & SCHWAB, A. 2008 Anomalous dynamics of cell migration. *PNAS* **105** (2), 459–463.

DIMILLA, P. A., QUINN, J. A., ALBELDA, S. M. & LAUFFENBURGER, D. A. 1992 Measurement of individual cell migration parameters for human tissue cells. *AIChE Journal* **38** (7), 1902–1104.

DRISCOLL, M. K. & DANUSER, G. 2015 Quantifying modes of 3D cell migration. *Trends in Cell Biology* **25** (12), 749–759.

DUNN, G. A. & BROWN, A. F. 1987 A Unified Approach to Analysing Cell Motility. *Journal of Cell Science. Supplement* **8**, 81–102.

FRALEY, S. I., WU, P., HE, L., FENG, Y., KRISNAMURTHY, R., LONGMORE, G. D. & WIRTZ, D. 2015 Three-dimensional matrix fiber alignment modulates cell migration and MT1-MMP utility by spatially and temporally directing protrusions. *Scientific Reports* **5**, 14580.

FRIEDL, P., SAHAI, E., WEISS, S. & YAMADA, K. M. 2012 New dimensions in cell migration. *Nature Reviews Molecular Cell Biology* **13**, 743–747.

GAIL, M. & BOONE, C. 1970 The Locomotion of Mouse Fibroblasts in Tissue Culture. *Biophysical Journal* **10**, 980–993.

GERLEE, P. & NELANDER, S. 2012 The Impact of Phenotypic Switching on Glioblastoma Growth and Invasion. *PLoS Computational Biology* **8** (6), e1002556.

HAKKINEN, K. M., HARUNAGA, J. S., DOYLE, A. D. & YAMADA, K. M. 2011 Direct comparisons of the morphology, migration, cell adhesions, and actin cytoskeleton of fibroblasts in four different three-dimensional extracellular matrices. *Tissue Engineering: Part A* **17** (5-6), 713–724.

HAMIS, S., POWATHIL, G. G. & CHAPLAIN, M. A. J. 2019 Blackboard to Bedside: A Mathematical Modeling Bottom-Up Approach Toward Personalized Cancer Treatments. *JCO Clinical Cancer Informatics* **3**, 1–11.

HATHOUT, L., PATEL, V. & WEN, P. 2016 A 3-dimensional DTI MRI-based model of GBM growth and response to radiation therapy. *International Journal of Oncology* **49** (3), 1081–1087.

HOARAU-VÉCHOT, J., RAFII, A., TOUBOUL, C. & PASQUIER, J. 2018 Halfway between 2D and animal models: Are 3D cultures the ideal tool to study cancer-microenvironment interactions? *International Journal of Molecular Sciences* **19** (181).

JACKSON, P. R., JULIANO, J., HAWKINS-DAARUD, A., ROCKNE, R. C. & SWANSON, K. R. 2015 Patient-Specific Mathematical Neuro-Oncology: Using a Simple Proliferation and Invasion Tumor Model to Inform Clinical Practice. *Bulletin of Mathematical Biology* **77** (5), 846–856.

LAUFFENBURGER, D. A. & HORWITZ, A. F. 1996 Cell Migration: A Physically Integrated Molecular Process. *Cell* **84** (3), 359–369.

LEE, B., ZHOU, X., RICHING, K., ELICEIRI, K. W., KEELY, P. J., GUELCHER, S. A., WEAVER A. M. & JIANG, Y. 2014 A Three-Dimensional Computational Model of Collagen Network Mechanics. *PLoS One* **9** (11), e111896.

LEE, J. 2018 Insights into cell motility provided by the iterative use of mathematical modeling and experimentation. *AIMS Biophysics* **5**, 97–124.

LIU, Z. 2010 Consensus of the 3-dimensional Vicsek model. In *Proceedings of the 29th Chinese Control Conference*, pp. 4635–4640.

LOOSLEY, A. J., O'BRIEN, X. M., REICHNER, J. S. & TANG, J. X. 2015 Describing Directional Cell Migration with a Characteristic Directionality Time. *PLoS ONE* **10** (5), e0127425.

LOWENGRUB, J. S., FRIEBOES, H. B., JIN, F., CHUANG, Y. L., LI, X., MACKLIN, P., WISE, S. M. & CRISTINI, V. 2010 Nonlinear modelling of cancer: bridging the gap between cells and tumours. *Nonlinearity* **23** (1), R1–R91.

LUZHANSKY, I. D., SCHWARTZ, A. D., COHEN, J. D., MACMUNN, J. P., BARNEY, L. E., JANSEN, L. E. & PEYTON, S. R. 2018 Anomalously diffusing and persistently migrating cells in 2D and 3D culture environments. *APL Bioengineering* **2** (2), 026112.

MACKLIN, P., EDGERTON, M. E., LOWENGRUB, J. S. & CRISTINI, V. 2010 Discrete cell modelling. In *Multiscale modeling of cancer: an integrated experimental and mathematical modeling approach* (ed. V. Cristini & J. Lowengrub), pp. 88–122. Cambridge University Press.

MATSIAKA, O. M., BAKER, R. E., SHAH, E. T. & SIMPSON, M. J. 2019 Mechanistic and experimental models of cell migration reveal the importance of cell-to-cell pushing in cell invasion. *Biomedical Physics and Engineering Express* **5**, 045009.

METZNER, C., MARK, C., STEINWACHS, J., LAUTSCHAM, L., STADLER, F. & FABRY, B. 2015 Superstatistical analysis and modelling of heterogeneous random walks. *Nature Communications* **6**, 7516.

MIERKE, C. T. 2015 Physical view on migration modes. *Cell Adhesion and Migration* **9** (5), 367–379.

MITCHISON, T. J. & CRAMER, L. P. 1996 Actin-Based Cell Motility and Cell Locomotion. *Cell* **84** (3), 371–379.

OTHMER, H. G. & XUE, CHUAN 2013 The mathematical analysis of biological aggregation and dispersal: progress, problems and perspectives. In *Dispersal, Individual Movement and Spatial Ecology* (ed. Maini P. Lewis M. & Petrovskii S.), *Lecture Notes in Mathematics*, vol. 2071, pp. 79–127. Springer, Berlin, Heidelberg.

PARKHURST, M. R. & SALTZMAN, W. M. 1992 Quantification of human neutrophil motility in three-dimensional collagen gels - Effect of collagen concentration. *Biophysical Journal* **61**, 306–315.

PAUL, C. D., HUNG, W., WIRTZ, D. & KONSTANTOPOULOS, K. 2016 Engineered Models of Confined Cell Migration. *Annual Review of Biomedical Engineering* **18**, 159–180.

PAUL, C. D., MISTRIOTIS, P. & KONSTANTOPOULOS, K. 2017 Cancer cell motility: lessons from migration in confined spaces. *Nature Reviews Cancer* **17** (2), 131–140.

RANGARAJAN, R. & ZAMAN, M. H. 2008 Modeling cell migration in 3D: Status and challenges. *Cell Adhesion and Migration* **2** (2), 106–109.

RICHARDS, R. 2016 Understanding the Role of the Solid Tumour Microenvironment in Brain Tumour Progression. PhD thesis, University of Liverpool.

RICHARDS, R., MASON, D., LEVY, R., BEARON, R. & SEE, V. 2018 4D imaging and analysis of multicellular tumour spheroid cell migration and invasion. *bioRxiv Preprint* .

ROCKNE, R. C., HAWKINS-DAARUD, A., SWANSON, K. R., SLUKA, J. P., GLAZIER, J. A., MACKLIN, P., HORMUTH, D. A. & JARRETT, A. M. ET AL 2019 The 2019 mathematical oncology roadmap. *Physical Biology* **16**, 041005.

ROCKNE, R. C., TRISTER, A. D., JACOBS, J., HAWKINS-DAARUD, A. J., NEAL, M. L., HENDRICKSON, K. MRUGALA, M. M., ROCKHILL, J. K. & KINAHAN, P. ET AL 2015 A patient-specific computational model of hypoxia-modulated radiation resistance in glioblastoma using ^{18}F-FMISO-PET. *Journal of the Royal Society Interface* **12** (103), 20141174.

SCHLÜTER, D. K., RAMIS-CONDE, I. & CHAPLAIN, M. A. J. 2012 Computational Modeling of Single-Cell Migration: The Leading Role of Extracellular Matrix Fibers. *Biophysical Journal* **103**, 1141–1151.

SCIANNA, M. & PREZIOSI, L. 2014 A cellular Potts model for the MMP-dependent and -independent cancer cell migration in matrix microtracks of different dimensions. *Computational Mechanics* **53** (3), 485–497.

SCOTT, M. 2021 In Preparation. PhD thesis, University of Liverpool.

SEPÚLVEDA, N., PETITJEAN, L., COCHET, O., GRASLAND-MONGRAIN, E., SILBERZAN, P. & HAKIM, V. 2013 Collective Cell Motion in an Epithelial Sheet Can Be Quantitatively Described by a Stochastic Interacting Particle Model. *PLoS Computational Biology* **9**, e1002944.

STEIN, A. M., VADER, D. A., SANDER, L. M. & WEITZ, D. A. 2007 A Stochastic Model of Glioblastoma Invasion. In *Mathematical Modeling of Biological Systems, Vol I: Cellular Biophysics, Regulatory Networks, Development, Biomedicine, and Data Analysis* (ed. A. Deutsch, L. Brusch, H. Byrne, G. DeVries & H. Herzel), pp. 217–224.

STOKES, C. L. & LAUFFENBURGER, D. A. 1991 Migration of individual microvessel endothelial cells: stochastic model and parameter measurement. *Journal of Cell Science* **99** (Part 2), 419–430.

SWANSON, K. R., ROSTOMILY, R. C. & ALVORD, E. C. 2008 A mathematical modelling tool for predicting survival of individual patients following resection of glioblastoma: a proof of principle. *British Journal of Cancer* **98** (1), 113–119.

TAKAGI, H., SATO, M. J., YANAGIDA, T. & UEDA, M. 2008 Functional Analysis of Spontaneous Cell Movement under Different Physiological Conditions. *PLoS ONE* **3** (7), e2468.

TOZLUOĞLU, M., TOURNIER, A. L., JENKINS, R. P., HOOPER, S., BATES, P. A. & SAHAI, E. 2013 Matrix geometry determines optimal cancer cell migration strategy and modulates response to interventions. *Nature Cell Biology* **15**, 751–762.

TRANQUILLO, R. T. & LAUFFENBURGER, D. A. 1987 Stochastic model of leukocyte chemosensory movement. *Journal of Mathematical Biology* **25** (3), 229–262.

UPADHYAYA, A., RIEU, J. P., GLAZIER, J. A. & SAWADA, Y. 2001 Anomalous diffusion and non-Gaussian velocity distribution of Hydra cells in cellular aggregates. *Physica A-Statistical Mechanics and its Applications* **293** (3-4), 549–558.

VICSEK, T., CZIRÒK, A., BEN-JACOB, E., COHEN, I. & SHOCHET, O. 1995 Novel Type of Phase Transition in a System of Self-Driven Particles. *Physical Review Letters* **75** (1226).

WANG, C., TONG, X. & YANG, F. 2014 Bioengineered 3D Brain Tumor Model To Elucidate the Effects of Matrix Stiffness on Glioblastoma Cell Behavior Using PEG-based Hydrogels. *Molecular Pharmaceutics* **11** (7), 2115–2125.

WOLF, K., LINDERT, M., KRAUSE, M., ALEXANDER, S., RIET, J., WILLIS, A. L., HOFFMAN, R. M., FIGDOR, C. G., WEISS, S. J. & FRIEDL, P. 2013 Physical limits of cell migration: Control by ECM space and nuclear deformation and tuning by proteolysis and traction force. *Journal of Cell Biology* **201**, 1069–1084.

WU, P., GILKES, D. M. & WIRTZ, D. 2018 The Biophysics of 3D Cell Migration. *Annual Reviews of Biophysics* **47**, 549–567.

WU, P., GIRI, A., SUN, S. X. & WIRTZ, D. 2014 Three-dimensional cell migration does not follow a random walk. *PNAS* **111** (11), 3949–3954.

WU, P., GIRI, A. & WIRTZ, D. 2015 Statistical analysis of cell migration in 3D using the anisotropic persistent random walk model. *Nature Protocols* **10** (3), 517–527.

YAMADA, K. M. & CUKIERMAN, E. 2007 Modeling Tissue Morphogenesis and Cancer in 3D. *Cell* **130**, 601–610.

YURCHENKO, I., VENSI BASSO, J. M., SYROTENKO, V. S. & STAII, C. 2019 Anomalous diffusion for neuronal growth on surfaces with controlled geometries. *PLoS ONE* **14** (5), e0216181.

ZAMAN, M. H., KAMM, R. D., MATSUDAIRA, P. & LAUFFENBURGER, D. A. 2005 Computational Model for Cell Migration in Three-Dimensional Matrices. *Biophysical Journal* **89**, 1389–1397.

ZAMAN, M. H., MATSUDAIRA, P. & LAUFFENBURGER, D. A. 2007 Understanding Effects of Matrix Protease and Matrix Organization on Directional Persistence and Translational Speed in Three-Dimensional Cell Migration. *Annals of Biomedical Engineering* **35** (1), 91–100.

ZAMAN, M. H., TRAPANI, L. M., SIEMINSKI, A., MACKELLAR, D., GONG, H., KAMM, R. D., WELLS, A., LAUFFENBURGER, D. A. & MATSUDAIRA, P. 2006 Migration of tumor cells in 3D matrices is governed by matrix stiffness along with cell-matrix adhesion and proteolysis. *PNAS* **103** (29), 10889–10894.

ZWIERS, F. W. & VON STORCH, H. 1995 Taking Serial Correlation into Account in Tests of the Mean. *Journal of Climate* **8** (2), 336–351.