Inferring characteristics of bacterial swimming in biofilm matrix from time-lapse confocal laser scanning microscopy

Guillaume Ravel, Michel Bergmann, Alain Trubuil, Julien Deschamps, Romain Briandet, Simon Labarthe

Inferring characteristics of bacterial swimming in biofilm matrix from time-lapse confocal laser scanning microscopy

Guillaume Ravel[1,2], Michel Bergmann[3,4,5], Alain Trubuil[6], Julien Deschamps[7], Romain Briandet[7], and Simon Labarthe[*,1,2,6]

[1]Univ. Bordeaux, INRAE, BIOGECO, F-33610 Cestas, France
[2]Inria, INRAE, Pléiade, 33400, Talence, France
[3]Memphis Team, INRIA, F-33400 Talence, France
[4]Univ. Bordeaux, IMB, UMR 5251, F-33400 Talence, France
[5]CNRS, IMB, UMR 5251, F-33400 Talence, France
[6]Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France
[7]Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350 Jouy-en-Josas, France.

January 11, 2022

Abstract

Biofilms are spatially organized microorganism colonies embedded in a self-produced matrix, conferring to the microbial community resistance to environmental stresses. Motile bacteria have been observed swimming in the matrix of pathogenic exogeneous host biofilms. This observation opened new promising routes for deleterious biofilms biocontrol: these bacterial swimmers enhance biofilm vascularization for chemical treatment or could deliver biocontrol agent by microbial hitchhiking or local synthesis. Hence, characterizing swimmer trajectories in the biofilm matrix is of particular interest to understand and optimize its biocontrol.

In this study, a new methodology is developed to analyze time-lapse confocal laser scanning images to describe and compare the swimming trajectories of bacterial swimmers populations and their adaptations to the biofilm structure. The method is based on the inference of a kinetic model of swimmer population including mechanistic interactions with the host biofilm. After validation on synthetic data, the methodology is implemented on images of three different motile *Bacillus* species swimming in a *Staphylococcus aureus* biofilm. The fitted model allows to stratify the swimmer populations by their swimming behavior and provides insights into the mechanisms deployed by the micro-swimmers to adapt their swimming traits to the biofilm matrix.

*Corresponding author: `simon.labarthe@inrae.fr`

1 Introduction

Biofilm is the most abundant mode of life of bacteria and archaea on earth [16, 15]. They are composed of spatially organized communities of microorganisms embedded in a self-produced extracellular polymeric substances (EPS) matrix. EPS are typically forming a gel composed of a heterogenous mixture of water, polysaccharides, proteins and DNA [14]. The biofilm mode of life confers to the inhabitant microbial community strong ecological advantages such as resistance to mechanical or chemical stresses [3] so that conventional antimicrobial treatments remain poorly efficient against biofilms [6]. Different mechanisms were invoked such as molecular diffusion-reaction limitations in the biofilm matrix and the cell type diversification associated with stratified local microenvironments [5]. Biofilms can induce harmful consequences in several industrial applications, such as water [2], or agri-food industry [12], leading to significant economic and health burden [23]. Indeed, it was estimated that the biofilm mode of life is involved in 80% of human infection and usual chemical control leads to serious environmental issues[3]. Hence, finding efficient ways to improve biofilm treatment represents important societal sustainable perspectives.

Motile bacteria have been observed in host biofilms formed by exogenous bacterial species [18, 27, 34, 14]. These bacterial swimmers are able to penetrate the dense population of host bacteria and to find their way in the interlace of EPS. Doing so, they visit the 3D structure of the biofilm, leaving behind them a trace in the biofilm structure, i.e. a zone of extracellular matrix free of host bacteria (Fig. 1a). Hence, bacterial swimmers are digging a network of capillars in the biofilm, enhancing the diffusivity of large molecules [18], allowing the transport of biocide at the heart of the biofilm, reducing islands of living cells. The potentiality of bigger swimmers has also been studied for biofilm biocontrol, including spermatozoa [30], protozoans [11] or metazoans [22]. Recent results suggest a deeper role of bacterial swimmers in biofilm ecology with the concept of microbial hitchhiking: motile bacteria can transport sessile entities such as spores [32], phages [41] or even other bacteria [37], enhancing their dispersion within the biofilm. Hence, characterizing microbial swimming in the very specific environment of the biofilm matrix is of particular interest to decipher biofilm spatial regulations and their biocontrol, but more generally in an ecological perspective of microbial population dynamics in natural ecosystems.

Bacterial swimming is strongly influenced by the micro-topography and bacteria deploy strategies to sense and adapt their motion to their environment [25], with specific implications for biofilm formation and dynamics [9]. Model-based studies were conducted to characterize bacterial active motion in interaction with an heterogeneous environment. An image and model-based analysis showed non-linear self-similar trajectories during chemotactic motion with obstacles [24]. Theoretical studies explored Brownian dynamics of self-propelled particles in interaction with filamentous structures such as EPS [20] or with random obstacles, exhibiting continuous limits and different motion regimes depending on obstacle densities [8, 7]. Image analysis characterized different swimming patterns in polymeric fluids [33], completed by detailed comparisons between a micro-

Species	Batch	# traject.	traj. length	time points	Duration	Δt
B. pumilus	1	122	40 (7.4)	4,590	30	0.134
	2	152	25 (5.7)	3,543	30	0.134
	3	243	38 (6.9)	8,825	30	0.134
B. sphaericus	1	98	40 (7.6)	3,762	30	0.134
	2	91	43 (7.7)	3,771	30	0.134
	3	48	55 (7.9)	2,543	23	0.134
B. cereus	1	105	47 (7.9)	4,766	30	0.069
	2	53	36 (7.7)	1,808	30	0.069
	3	121	43 (7.1)	5,006	30	0.069

Table 1: **Dataset characteristics.** We detailed, for each batch, the number of trajectories, the average number of time points by trajectory (and standard deviation), the total number of time points in the dataset, the total movie duration and the time interval between two snapshots.

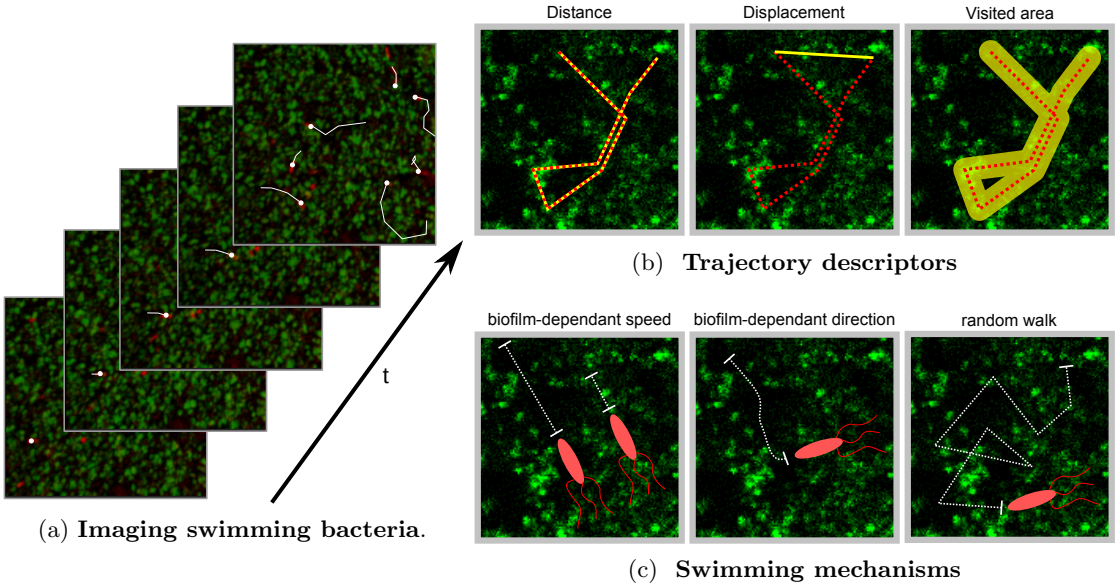
scale model of flagellated bacteria in polymeric fluids and high-throughput images [29]. Models of bacterial swimmers in visco-elastic fluids were also developed to study the force fields encountered during their run [26]. However, to our knowledge, no study tried to characterize swimming patterns in the highly heterogeneous environment presented by an exogenous biofilm matrix.

In this study, we aim at providing a quantitative characterization of the different swimming behaviours in adaptation to the host biofilm matrix observed by microscopy. We focus on identifying potential species-dependent swimming characteristics and quantifying the swimming speed and direction variations induced by the host biofilm structure. To address these goals, three different *Bacillus* species presenting contrasted physiological or swimming characteristics are selected. First, different trajectory descriptors accounting for interactions with the host biofilm are defined, allowing to discriminate the swim of these bacterial strains by differential analysis. Then, a mechanistic random-walk model including swimming adaptations to the host biofilm is introduced. This model is numerically explored to identify the sensitivity of the trajectory descriptors to the model parameters. An inference strategy is designed to fit the model to 2D+T microscopy images. The method is validated on synthetic data and applied to a microscopy dataset to decipher the swimming behaviour of the 3 Bacillus.

2 Results

2.1 Characterizing bacterial swimming in a biofilm matrix through image descriptors

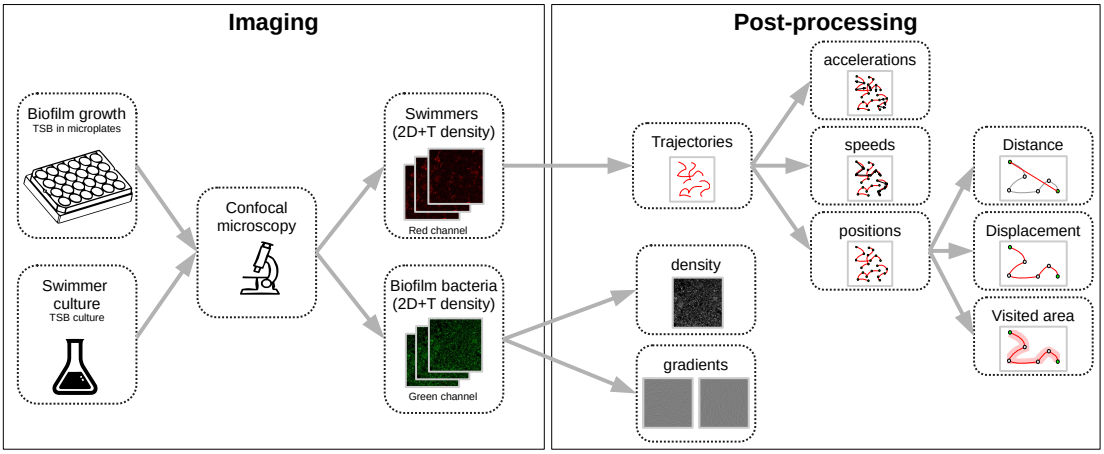
2D+T Confocal Laser Scanning Microscopy (CLSM) images of three bacterial swimmer populations –*Bacillus pumilus* (*B. pumilus*), *Bacillus sphaericus* (*B. sphaericus*) and



(a) **Imaging swimming bacteria.**

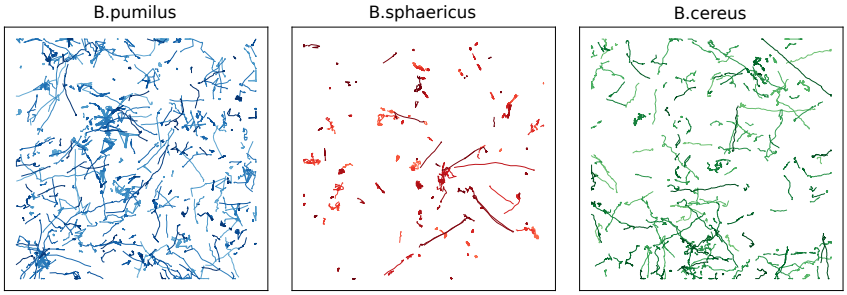
(b) **Trajectory descriptors**

(c) **Swimming mechanisms**

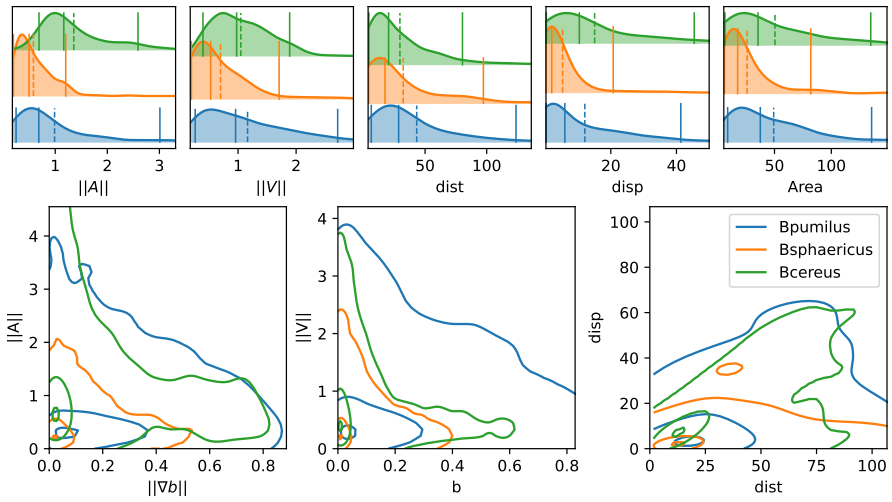


(d) **Image acquisition workflow**

Figure 1: **Microscopy data and model outlines. (a)** Temporal stacks of 2D images are acquired, with different fluorescence colors for host bacteria (*Staphylococcus aureus*, green) and swimmers (*Bacillus pumilus*, *Bacillus sphaericus* or *Bacillus cereus*, red). Bacterial swimmers navigate in a host biofilm and are tracked in the different snapshots. Swimmer trajectories are represented with white lines. High density and low density zones of host cells are visible in the biofilm (green scale). **(b)** Additionally to speed and acceleration distributions, three trajectory descriptors are considered. *Distance* is the distance between the initial and final points of the trajectory. *Displacement* is the total length of the trajectory path. *Visited area* is the total area of the pores left by the swimmer during its path. Hence, when a swimmer retraces its steps, the displacement is incremented but not the visited area. **(c)**Three different mechanisms are considered in the mechanistic model. *Biofilm-dependant speed.* A target speed is defined accordingly to the local density of biofilm and asymptotically reached after a relaxation time. *Biofilm-dependent direction.* Swimming direction is defined accordingly to the local biofilm density gradient. *Random walk.* A Brownian motion is added. **(d)** The image acquisition workflow is composed of a first step at the wet lab where host biofilm and swimmer are plated and imaged in different color channels. Then a post-processing phase recomposes the swimmer trajectories with tracking algorithms. Then, temporal positions, speeds and accelerations are computed. On the biofilm channel, density and density gradient maps are processed at each time step.



(a) Trajectories



(b) Descriptor distributions.

Figure 2: **Analysis of swimming characteristics.** (a) The whole set of trajectories of each species is displayed. (b) Trajectory descriptors. Upper panel: acceleration, speed, distance and displacement distributions structured by species are displayed, together with quantile $0.05, 0.5$ and 0.95 (plain lines) and mean (dashed line). T-test pairwise comparison p-values are displayed in A.1. Lower panel: we display the distribution of the instantaneous acceleration norm respectively to the local biofilm density gradient (i.e. $||A_i(t)||$ function of $\nabla b(X_i(t))$) and of the instantaneous velocity norm respectively to the local biofilm density (i.e. $||V_i(t)||$ function of $b(X_i(t))$, structured by population. The point cloud of each species is approximated by a gaussian kernel and gaussian kernel isolines enclosing 5, 50 and 95% of the points centered in the densest zones are displayed to facilitate comparisons between species (see Materials and Methods 4.10).

Bacillus cereus (*B. cereus*) – swimming in a *Staphylococcus aureus* (*S. aureus*) host biofilm were acquired (see Fig.1d). Swimmers and host biofilms were imaged with different fluorescence dies, allowing their acquisition in different color channels, and to recover in the same spatio-temporal referential the swimmer trajectories and the host biofilm density (see Materials and Methods and Fig. 1). Namely, for each species s and individual swimmer i, we recover the initial ($T^s_{0,i}$) and final ($T^s_{end,i}$) observation times (when the swimmer goes in and out the focal plane, see sect. 4.2), and the number T^s_i of time points in the trajectory. We then extract from the 2D+T images the observed position, instantaneous speed and acceleration time-series

$$t \mapsto X^s_i(t), \quad t \mapsto V^s_i(t), \quad t \mapsto A^s_i(t), \qquad \text{for } t \in (T^s_{0,i}, T^s_{end,i}).$$

Noting $b^s(t,x)$ the dynamic biofilm density maps obtained from the biofilm images, we also compute the local biofilm density and density gradient

$$t \mapsto b^s(t, X^s_i(t)), \qquad \text{and } t \mapsto \nabla b^s(t, X^s_i(t)).$$

Different swimming patterns can be deciphered by qualitative observations of the trajectories $X^s_i(t)$ (Fig. 2a). For *B. sphaericus* and to a minor extent *B. pumilus*, the trajectories are divided between back and forth paths around the starting point and long runs. By contrast, *B. cereus* swimmers nearly never get stuck in the same place and describe longer curves in the biofilm.

For quantitative analysis, trajectory descriptors are defined. We first investigate the distribution of the population-wide average acceleration and velocity norms $\frac{1}{T^s_i}\sum_t \|A^s_i(t)\|$ and $\frac{1}{T^s_i}\sum_t \|V^s_i(t)\|$, where $\|\cdot\|$ denotes the Euclidian norm. We also quantify the swimming kinematics by computing the travelled distance $dist^s_i$ along the path and the total displacement $disp^s_i$, i.e. the distance between the initial and final trajectory points, with

$$dist^s_i = \int_{T^s_{0,i}}^{T^s_{end,i}} \|V^s_i(t)\| dt \quad \text{and} \quad disp^s_i = \|X(T^s_{end,i}) - X(T^s_{0,i})\| = \| \int_{T^s_{0,i}}^{T^s_{end,i}} V^s_i(t) dt \|.$$

We finally compute the total biofilm area visited by a swimmer along its path (see Fig. 1b).

The three species present contrasted distributions for these descriptors (Fig. 2b). *B. sphaericus* has the smallest mean and median values of acceleration and speed, while *B. pumilus* has the widest distributions. *B. cereus* for its part shows the highest accelerations, indicating larger changes in swimming velocities, but median and mean speeds comparable to *B. pumilus* (Fig. 2b, $\|A\|$ and $\|V\|$ panels). We also note that *B. sphaericus* and to a lower extent *B. pumilus* trajectories have a significant amount of null or small average speeds, while *B. cereus* trajectories have practically no zero velocity, consistently with the qualitative analysis (Fig. 2b, $\|V\|$ panels). Small velocities episodes of *B. sphaericus* and *B. pumilus* could occur during their back-and-forth trajectories, which produce small displacements and pull the displacement distribution towards lower values than *B. cereus* (Fig. 2b, *Disp* panel). *B. pumilus* displacement is intermediary.

Conversely, back-and-forth trajectories can produce large swimming distances for *B. sphaericus* and *B. pumilus* so that *B. sphaericus* has a distance distribution comparable to *B. cereus* (Fig. 2b, *Dist* panel), but lower than *B. pumilus* which also shows the widest speed distribution. Observing conjointly displacement and distance (Fig. 2b, lower-right panel) provides consistent insights: *B. sphaericus* shows a large variability of small displacement trajectories, from small to large distances, while *B. cereus* trajectory displacement seems to vary almost linearly with the distance at least for the points inside the isoline 50%. *B. pumilus* has again an intermediary distribution, with a large range of displacement-distance couples. The distributions of visited areas of *B. pumilus* and *B. cereus* are almost identical, and higher than *B. sphaericus* one.

All together, this data depict 1) a long-range species, *B. cereus*, which moves efficiently in the biofilm during long, relatively straight, rapid runs, 2) a short-range species, *B. sphaericus*, that moves mainly locally in small areas with lower accelerations and speeds except few exceptions and 3) a medium-range species, *B. pumilus*, with a large diversity of rapid trajectories, from small to large displacement. These kinematics discrepancies for *B. pumilus* and *B. cereus* allow them however to cover identical visited areas.

These global descriptors do not inform however about potential adaptations of the swimmers to the biofilm matrix. We first want to check if swimmer velocities are directly linked to the local biofilm density, and if the swimmers adapt their trajectory according to density gradients by plotting the points $(\|\nabla b(t, X_i^s(t))\|, \|A_i^s(t)\|)$ and $(b(t, X_i^s(t)), \|V_i^s(t)\|)$ (Fig. 2 f, g). Clear differences between the three species can be deciphered. We first observe that the three *Bacillus* do not have the same distribution of visited biofilm density and gradient. *B. pumilus* swimmers visit denser biofilm with higher variations than the other species while *B. sphaericus* and *B. cereus* stay in less dense and smoother areas, the quantile 0.5 area of these species being circumscribed in low gradient and low density values. Next, we see that *B. cereus* has a wider distribution of accelerations, specially for small density gradients, compared to *B. pumilus* and *B. sphaericus*. This could indicate that when the biofilm is smooth, *B. cereus* samples its acceleration in a large distribution of possible values. Finally, we observe that the speed distribution rapidly drops for increasing biofilm densities for *B. sphaericus* and *B. cereus*, while the decrease is much smoother for *B. pumilus*. These observations provide additional insights in the species swimming characteristics: *B. pumilus* swimmers seem to be less inconvenienced by the host biofilm density than the other species, while *B. cereus* and *B. sphaericus* bacteria appear to be particularly impacted by higher densities and to favor low densities where it can efficiently move. Though, *B. sphaericus* has lower motile capabilities than *B. cereus* when the biofiilm is not dense.

2.2 Swimming model

This descriptive analysis does not allow to clearly identify potential mechanisms by which the swimmers adapt their swim to the biofilm structure or to simulate new species-dependant trajectories. We then build a swimming model based on a Langevin-like equation on the acceleration. Once fitted, this model will allow to identify the respective

influence of the deterministic mechanisms it includes but also to generate synthetic data by predicting new swimmer random walks sharing characteristics comparable to the original data.

2.2.1 Governing swimming equation

We consider bacterial swimmers as Lagrangian particles and we model the different forces involved in the update of their velocity $\mathbf{v}$. We assume that the swimmer motion can be modelled by a stochastic process with a deterministic drift (Fig. 1c):

$$\mathrm{d}\mathbf{v} = \underbrace{\gamma(\alpha(b) - \|\mathbf{v}\|)\frac{\mathbf{v}}{\|\mathbf{v}\|}\mathrm{dt}}_{\text{speed selection}} + \underbrace{\beta\frac{\nabla b}{\|\nabla b\|}\mathrm{dt}}_{\text{direction selection}} + \underbrace{\eta\mathrm{dt}}_{\text{random term}} \tag{1}$$

where the right hand side is composed of two deterministic terms in addition to a gaussian noise, each weighted by the parameters γ, β and ϵ.

The first term implements the biological observation (Fig. 2b) that the bacterial swimmers adapt their velocity to the biofilm density. This term can be interpreted as a speed selection term that pulls the instantaneous speed of the swimmer towards a prescribed target velocity $\alpha(b)$ that depends on the host biofilm density b. The weight γ can be interpreted as a penalization coefficient, proportionally inverse to a relaxation time τ, $\gamma \sim \frac{1}{\tau}$. As a first order approximation of the speed drop observed in Fig. 2b for increasing b, the target speed $\alpha(b)$ is modeled as a linear variation between v_0 and v_1, the swimmer characteristic speed in the highest and lowest density regions respectively:

$$\alpha(b) = v_0(1-b) + bv_1 = v_0 + b(v_1 - v_0)$$

The second term updates the velocity direction according to the local gradient of the biofilm density ∇b. The sign of β indicates if the swimmer is inclined to go up (negative β) or down (positive β) the host biofilm gradient, while the weight magnitude indicate the influence of this mechanism in the swimmer kinematics. We note that this term does not depend on the gradient magnitude but only on the gradient direction: this reflects the implicit assumption that the bacteria are able to sense density variations to find favorable directions, but that the biological sensors are not sensitive enough to evaluate the variation magnitudes.

The third term is a stochastic diffusive process that models the dispersion around the deterministic drift modelled by the two first terms. We define

$$\eta \sim \mathcal{N}(0, \epsilon)$$

The term η can also be interpreted as a model of the modelling errors, tuned by the term ϵ. Eq. (1) is supplemented by an initial condition by swimmer. For vanishing $\|v\|$ or $\|\nabla b\|$, the corresponding term in the equation is turned off.

We can define characteristic speed and acceleration V^* and A^* in order to set a dimensionless version of Eq. (1)

$$\mathrm{d}\mathbf{v} = \gamma'(v_0' + b(v_1' - v_0') - \|\mathbf{v}\|)\frac{\mathbf{v}}{\|\mathbf{v}\|}\mathrm{dt} + \beta'\frac{\nabla b}{\|\nabla b\|}\mathrm{dt} + \eta'\mathrm{dt} \tag{2}$$

where $\gamma' = \frac{\gamma V^*}{A^*}$, $v_0' = \frac{v_0}{V^*}$, $v_1' = \frac{v_1}{V^*}$, $\beta' = \frac{\beta}{A^*}$, $\eta' \sim \mathcal{N}(0, \epsilon')$ and $\epsilon' = \frac{\epsilon}{A^{*2}}$.

This dimensionless version will strongly improve the inference process and will allow an analysis of the relative contribution of the different terms in the kinematics.

2.2.2 Numerical exploration and sensitivity analysis

To illustrate the impact of each parameter on the interplay between the host biofilm and the swimmers trajectories, the model 2 was first computed on two mock biofilms. The first one is a square linear density gradient and the second is composed of large pores on a textured background mimicking the dense biofilm zones (Fig. 3a). A basal simulation is computed with $\gamma = \beta = \epsilon = 1$, and this three parameters are alternatively set to zero to assess the resulting trajectories when the speed selection, the direction selection or the random term is shut down. Suppressing speed selection results in rectilinear trajectories ($\gamma = 0$, Fig. 3c), which is rather counter-intuitive since the remaining terms are designed to tune the direction. A discussion of this phenomena is provided in the Annex A.7. When suppressing direction selection ($\gamma = 0$, Fig. 3d), the trajectories are no longer drifted downwards the gradient in the upper panel as in the basal simulation, and no longer follow the pores (lower panel). If the stochastic term is shut down ($\epsilon = 0$, Fig. 3e), the trajectories directly go down the gradients and are trapped in the center of the image in the upper panel. When a pore is found along the run, the swimmer keeps following it without being able to escape the pore any longer unlike the basal situation (lower panel).

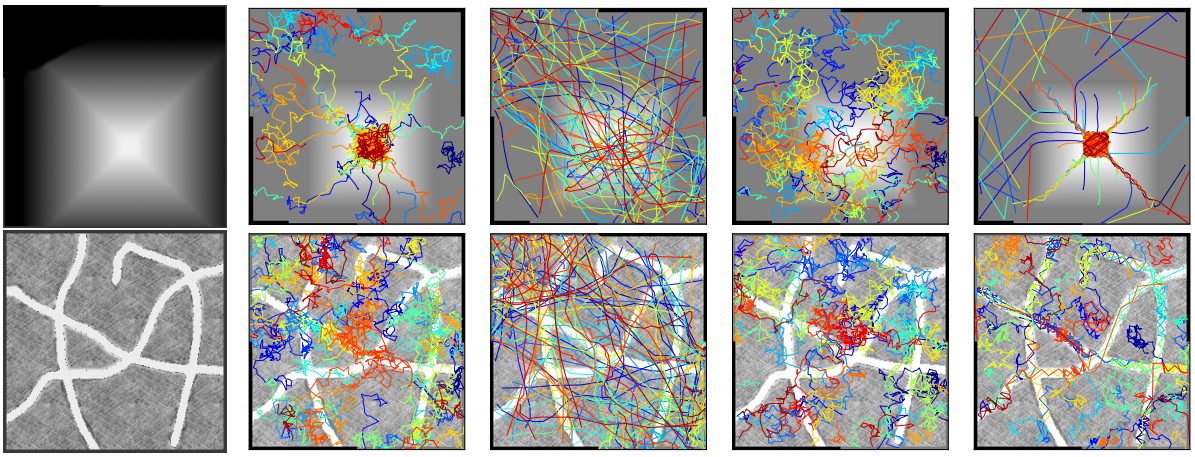
The link between the model parameters and the global trajectory descriptors introduced in Section 2.1 is less intuitive. A global sensitivity analysis of the trajectory descriptors (mean acceleration and speed, distance, displacement and visited areas) with respect to the parameters γ, v_0, v_1, β and ϵ is conducted in A.4 by computing their first order Sobol index (SI) and their pairwise correlation coefficient (PCC). The sensitivity analysis shows that the mean speed is mainly influenced by γ and ϵ with slightly negative and positive impact respectively, while acceleration is rather influenced by β and ϵ with positive impact. The link between the parameters and the other descriptors is more complex, including non linear effects (strong SI and small PCC) and parameter interactions (higher SI residuals, see Sec. A.4 and Fig. Fig. A.3 for detailed analysis).

2.3 Inferring swimming parameters from trajectory data

For each bacterial swimmer population, we now seek to infer with a Bayesian method population-wide model parameters governing the swimming model of a given species from microscope observations.

2.3.1 Inference model setting

Equation (2) is re-written as a state equation on the acceleration for the bacterial strain s and the swimmer i



(a) Mock biofilm (b) Basal (c) $\gamma = 0$ (d) $\beta = 0$ (e) $\epsilon = 0$

Figure 3: **Numerical exploration of the model.** To illustrate the influence of each term of Eq. (2), they are alternatively turned off (Fig. 3c to 3e), and swimmer trajectories are computed on mock biofilms (3a) displaying marked density gradients (upper pannel) or marked pores (lower pannel). Trajectories can be compared to a basal simulation (3b) when all the terms have the same intensity ($\alpha = \beta = \epsilon = 1$).

$$A_i^s(t) = \gamma(v_0^s + b(t, X_i^s(t))(v_1^s - v_0^s) - \|V_i^s(t)\|)\frac{V_i^s(t)}{\|V_i^s(t)\|} + \beta^s \frac{\nabla b(t, X_i^s(t))}{\|\nabla b(t, X_i^s(t))\|} + \eta^{\mathbf{s}} \quad (3)$$

$$:= f_A\left(\theta^s, b(t, X_i^s(t)), V_i^s(t), X_i^s(t)\right) + \eta^{\mathbf{s}} \quad (4)$$

where

$$\theta^s := (\gamma^s, v_0^s, v_1^s, \beta^s)$$

are species-dependant equation parameters. The function f_A can be seen as the deterministic drift of the random walk, gathering all the mechanisms included in the model. The inter-individual variability of the swimmers of a same species comes from the swimmer-dependent initial condition, the resulting biofilm matrix they encounters during their run, and the stochastic term.

Inferring the parameters θ^s can then be stated in a Bayesian framework as solving the non linear regression problem

$$A_i^s(t) \sim \mathcal{N}\left(f_A\left(\theta^s | b(t, X_i^s(t)), V_i^s(t), X_i^s(t)\right), \epsilon^s\right) \quad (5)$$

from the data $b(t, X)$, $X_i^s(t)$, $V_i^s(t)$ and $A_i^s(t)$, with truncated normal prior distributions

$$\theta^s \sim \mathcal{N}(0, 1) \quad (6)$$

$$\epsilon^s \sim \mathcal{N}(0, 1). \quad (7)$$

parameter	ground truth	mean	std	confidence interval [2.5% - 97.5%]	n_{eff}	R_{hat}
γ	1.094	1.08	1.00×10^{-2}	$[1.06 - 1.1]$	3,569	1.0
v_0	0.669	0.66	1.00×10^{-2}	$[0.64 - 0.68]$	3,710	1.0
v_1	0.134	0.13	2.00×10^{-2}	$[0.09 - 0.17]$	3,431	1.0
β	0.146	0.16	6.20×10^{-3}	$[0.15 - 0.17]$	5,050	1.0
ϵ	0.586	0.59	3.00×10^{-3}	$[0.58 - 0.59]$	4,906	1.0

Table 2: **Inference results on synthetic data.** The normalized ground-truth parameter values (*i.e.* ground truth parameter rescaled with A_{ref} and V_{ref}) are compared with the inference outputs on synthetic data: posterior distribution mean and standard deviation are indicated, together with the inferred confidence intervals for the true parameters. Convergence diagnosis indices are also given, with n_{eff} the effective sample size per iteration and R_{hat} the potential scale reduction factors, indicating that convergence occurred for all parameters.

and additional constrains on the parameters

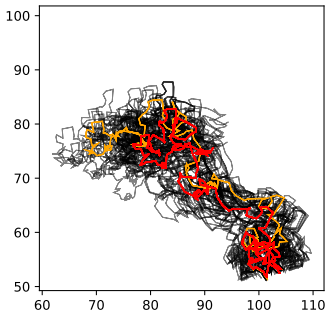
$$\gamma^s \geq 0, \quad v_0^s \geq 0, \quad v_1^s \geq 0, \quad \epsilon^s \geq 0$$

We note that Equation (5) can be seen as a likelihood equation of the parameter θ^s knowing $A_i^s(t), b(t), V_i^s(t)$ and $X_i^s(t)$. The parameter ϵ^s can now be seen as a corrector of both modelling errors in the deterministic drift and observation errors between the observed and the true instantaneous acceleration. Alternative settings where these uncertainties sources are separated and a true state for position and acceleration is inferred can be defined (see Annex A.8). The inference problem is implemented in the Bayesian HMC solver Stan [38] using its *python* interface *pystan* [35].

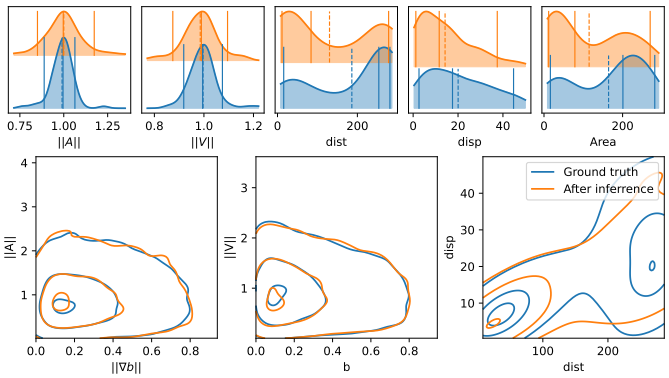
2.3.2 Assessment of the inference with synthetic data

To assess the inference method, synthetic data are built. We arbitrarily fix a parameter vector and solve system (1) from random initial positions, in a host biofilm arbitrarily chosen in the image dataset. We then extract the swimmer positions at given time-steps and recover accelerations and speeds with the same post-processing pipeline as for microscopy images and solve the inverse problem (5)-(7).

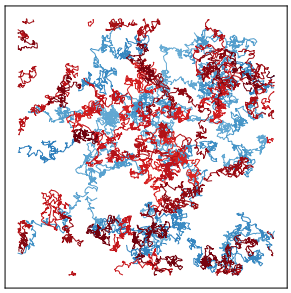
The ground truth parameters are correctly recovered by the inference procedure (Table 2), indicating that the parameters are correctly identifiable and that the inverse problem is well-posed. An error of respectively 1.28, 1.34, 2.98 and 0.68% on the parameters γ, v_0, v_1 and ϵ is observed in this controlled situation, β being inferred with lower accuracy (9.59 %). To assess the impact of parameter inference uncertainties on trajectory computation, the posterior parameter distribution is sampled and new trajectories are computed, replacing the ground-truth parameters by the sampled ones. The swimmer ground truth trajectories are accurately recovered: the sampled trajectories tightly frame the original swimmer path as illustrated on a randomly chosen trajectory



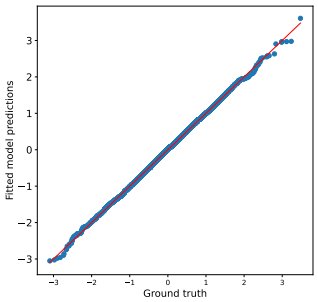
(a) **Trajectory sampling**.



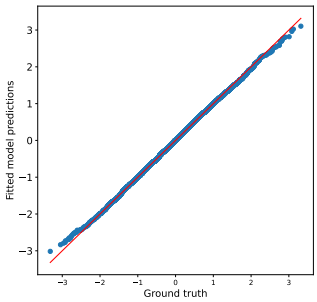
(b) **Trajectory descriptors**



(c) Ground truth (red) vs fitted model (blue) trajectories



(d) A_x qqplot



(e) A_y qqplot

Figure 4: **Inference assessment on synthetic data.**(a) **Predicted vs true trajectories**. Trajectories are recovered by sampling the parameter posterior distribution starting from the same initial condition than in the data. We represent a ground truth trajectory extracted randomly from the original dataset in red, the corresponding sampled trajectories with thin gray lines, and the trajectory obtained with the posterior means in orange. (b) **trajectory descriptors** Trajectories are re-computed replacing the ground-truth parameters by the inferred parameters. The trajectory descriptors introduced in 2.1 are computed on the synthetic data (blue curves) and on the data obtained with the inferred parameters (orange curves). **Qqplot of fitted model output vs ground truth**. After inference, the fitted model is used to re-compute the synthetic dataset (ground truth). We plot the x (left panel) and y (right panel) components of the accelerations in a qqplot: the fitted model output quantiles are plotted against the ground truth quantiles with blue dots, together with the $y = x$ line (red).

(Fig. 4a). We note that an identical random seed has been taken for these simulations, including the ground truth trajectory, in order to turn off the stochastic uncertainties and only focus on the propagation of inference errors during simulations of swimmer trajectories.

Finally, we re-assemble a synthetic dataset by replacing the ground-truth parameters by the inferred ones, *i.e.* the posterior mean. Qqplot of the fitted model accelerations versus the ground truth accelerations give an excellent accuracy (Fig. 4d), with all the points lying on the bisector, except slight divergences on the distribution tails. The fitted model trajectories visually reproduce the qualitative characteristics of the original dataset (Fig. 4c). The trajectory descriptors of section 2.1 are then computed on both datasets (ground truth and inferred) and compared (Fig. 4b). The kinematics descriptors, *i.e.* acceleration and speed distributions, are very accurately recovered with a relative error of 0.1%, 3.2%, 5% for respectively the mean, quantiles 0.05 and 0.95 of the acceleration (resp. 0.9%, 2.5% ,2% for speed). Some small discrepancies can be observed on the distance and displacement distributions, even if the mean and the quantiles 0.05 and 0.95 are close. The interactions between the host biofilm and the acceleration and speed distribution are also recovered with high accuracy. We note that part of the observed discrepancies comes from an additional source of variability of the simulation framework: when a swimmer reaches a domain boundary during a simulation, its trajectory is stopped and a new swimmer is randomly introduced elsewhere in the biofilm (see Materials and Methods for more details). This simulation strategy seems to be responsible of the over-representation of short trajectories in the inferred dataset, compared to the ground truth (Fig. 4b upper panel, distance and displacement distributions).

2.3.3 Analysis of the confocal microscopy dataset

We now solve the inference problem (5)-(7) on the confocal microscopy dataset to identify population-wide swimming model parameters. The inference process is assessed by comparing the descriptors obtained on trajectories predicted by the fitted model (Fig. 5a) with descriptors of real trajectories (Fig. 2). The mean values of acceleration and speeds are accurately predicted for the three species (Figs. 5a, panels $\|A\|$ and $\|V\|$, dashed lines). Relative positions of distance, displacement and visited area mean values are also correctly simulated (Figs. 2 and 5a, upper panel). *B. sphaericus* presents the lowest predicted accelerations and speeds while *B. pumilus* has the widest speed and acceleration distributions and *B. cereus* shows the highest accelerations, consistently with the data. The visited area and the distances are slightly over estimated, but the relative position and the shape of the distributions are conserved. The amount of null velocities for *B. sphaericus* is under estimated by the fitted model and not rendered for *B. pumilus*. The distance distributions of the three species are accurately predicted by the fitted model. When displaying conjointly the distance and the displacement (5a, right lower panel), the distribution of *B. sphaericus* is correctly predicted by the simulations, but *B. cereus* and *B. pumilus* displacements are underestimated. Some qualitative fea-

tures can be recovered, such as the higher distribution of distance-distribution couples for *B. cereus* or higher displacement for *B. cereus* compared to *B. sphaericus*.

Descriptors of swimming adaptations to the host biofilm are also correctly preserved for the main part (Figs. 2 and 5a, lower panel). *B. pumilus* is the species that crosses the highest biofilm densities in the fitted model simulations, showing the highest speeds in this crowded areas, and that visits the most frequently areas with high density gradients, consistently with the data. As in the confocal images, the simulated *B. sphaericus* and *B. cereus* favor smoother zones of the biofilm with lower biofilm densities. The *B. cereus* fitted model correctly render the highest acceleration variance observed in the data for low biofilm gradients, while *B. sphaericus* speed and acceleration variance is the lowest for all ranges of biofilm densities and gradients, both in the data and in the fitted model predictions. The drop of speeds and accelerations for increasing biofilm densities and gradients is well predicted for *B. pumilus*, but is smoother in the simulation compared to the data for *B. sphaericus* and *B. cereus*. In particular, the sharp drop of speeds for $b \simeq 0.25$ observed in the data for *B. cereus* and *B. sphaericus* is underestimated by the fitted model. All together, the model reproduces very accurately the mean values of acceleration, speed and visited area, renders relative positions and the main characteristics of distributions for distance, displacement and interactions with the host biofilm matrix, but produces less variable outputs than observed in the data.

To further inform the fitted model accuracy, the coefficient of determination R^2_{det} of the deterministic components $f_A(\theta^s, b(t), V_i^s, X_i^s(t))$ of eq. (4) is computed (Table 4), in order to quantify the goodness of fit of the friction and gradient terms of eq. (2) that represent interactions with the biofilm. These results highlight that *B. cereus* bacteria do present an important stochastic part in the accelerations, while the *B. pumilus* species is the best represented by our deterministic modelling.

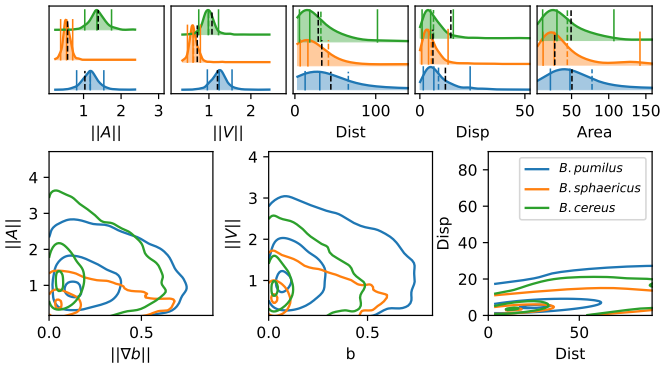
The three species present very different inferred parameter values (Fig. 5b and table 3), showing that the model inference captures contrasted swimming characteristics of this *Bacillus*. Due to the mechanistic terms introduced in Eq. (1), these differences can be interpreted in term of speed and direction adaptations to the host biofilm. First, *B. pumilus* shows the highest v_0 value, and the highest amplitude between v_0 and v_1, inducing a higher ability for *B. pumilus* to swim fast in low density biofilm zones. In comparison, *B. sphaericus* presents quasi no difference between v_0 and v_1 showing a poor adaptation to biofilm density. *B. cereus* has the highest γ value, showing a reduced relaxation time toward the density dependant speed: in other words, *B. cereus* is able to adapt its swimming speed more rapidly than the other species when the biofilm density varies. *B. cereus* swimmers are also better able to change their swimming direction in function of the biofilm variations they encounter along their way, their β distribution being markedly higher than the other species which have very low β. Finally, the stochastic parameter ϵ is also contrasted, from a low distribution for *B. sphaericus* to high values for *B. cereus*. All together, the inference complete the observations made in Fig. 2b: *B. pumilus* poorly adapts its swimming direction to the host biofilm (low β) but has a wide range of possible speeds when the biofilm density varies (high v_0, low v_1), that it can reaches quite rapidly (intermediary γ) with intermediary stochastic

species	param	mean	std	confidence interval [2.5% - 97.5%]	n_{eff}	R_{hat}
B. pumilus	γ	0.77	3.95×10^{-3}	[0.77−0.77]	4,507	1.0
	v_0	0.14	8.67×10^{-3}	[0.12−0.16]	3,879	1.0
	v_1	1.69×10^{-3}	1.69×10^{-3}	$[5.18 \times 10^{-5} - 6.26 \times 10^{-3}]$	4,821	1.0
	β	9.84×10^{-3}	5.07×10^{-3}	$[1.45 \times 10^{-5} - 2.07 \times 10^{-2}]$	5,223	1.0
	ϵ	0.62	2.48×10^{-3}	[0.61−0.62]	5,307	1.0
B. sphaericus	γ	0.61	4.53×10^{-3}	[0.60−0.62]	4,965	1.0
	v_0	2.75×10^{-4}	2.75×10^{-4}	$[4.91 \times 10^{-6} - 1.01 \times 10^{-3}]$	4,019	1.0
	v_1	4.84×10^{-3}	4.77×10^{-3}	$[9.39 \times 10^{-5} - 1.45 \times 10^{-2}]$	5,001	1.0
	β	4.25×10^{-3}	3.33×10^{-3}	$[-2.18 \times 10^{-3} - 1.15 \times 10^{-2}]$	4,668	1.0
	ϵ	0.32	1.55×10^{-3}	[0.31−0.32]	5,943	1.0
B. cereus	γ	0.83	1.11×10^{-2}	[0.80−0.86]	2,700	1.0
	v_0	6.44×10^{-2}	1.07×10^{-2}	$[3.22 \times 10^{-2} - 9.66 \times 10^{-2}]$	2,510	1.0
	v_1	6.65×10^{-3}	6.33×10^{-3}	$[1.50 \times 10^{-4} - 2.15 \times 10^{-2}]$	4,061	1.0
	β	2.78×10^{-2}	9.04×10^{-3}	$[1.39 \times 10^{-2} - 5.56 \times 10^{-2}]$	4,230	1.0
	ϵ	0.90	4.17×10^{-3}	[0.89−0.92]	4,852	1.0

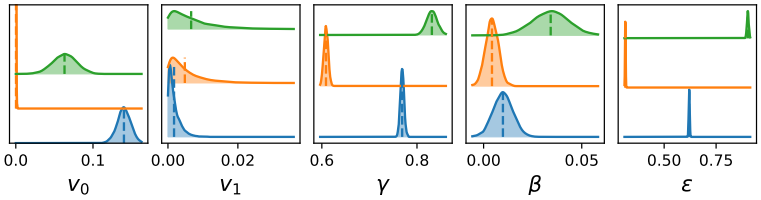
Table 3: **Inference outputs for the three species.** The posterior mean, standard deviation and inferred confidence interval are indicated for each parameter and each specie. Convergence diagnosis index n_{eff} and R_{hat} are provided.

correction (ϵ). In contrast, *B. cereus* reaches lower speed values (intermediary v_0, low v_1) but is more agile to adapt its swimming to its environment by changing rapidly its speed when the biofilm density is more favorable and adapting it swimming direction to biofilm variations, with higher stochastic variability (large ϵ). Finally, *B. sphaericus* is the less flexible of the three bacteria: less fast (small v_0 and v_1), they are also less responsive to biofilm variations (small γ and β) with low random perturbations (small ϵ).

Finally, after inference, the impact of each term in the overall acceleration data can be quantified and analyzed by displaying its relative contribution in a ternary plot (Fig. 6). The direction selection is the less influential mechanism for the three species, with a slightly higher impact for *B. cereus* (50 and 95 % isolines slightly shifted towards $A(\nabla b)$ in Fig. 6 b). When zooming in, the three *Bacillus* show differences in the balance between speed selection and the random term: while *B. pumilus* is slightly more influenced by the friction term than by stochasticity, these mechanisms are perfectly balanced in *B. sphaericus* accelerations, while *B. cereus* is more influenced by the random term.

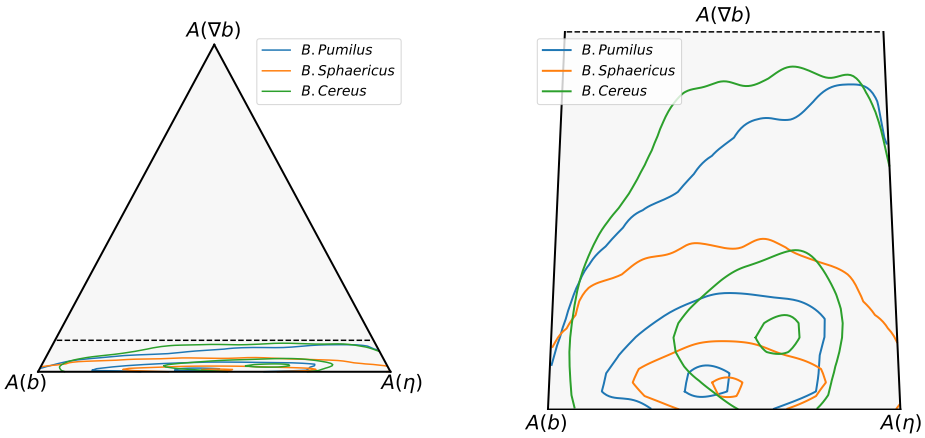


(a) **Trajectory descriptors.**



(b) **Posterior distribution of the parameters**.

Figure 5: **Inference result on the experimental images.** (a) To validate the inference process, a synthetic dataset is assembled by computing eq. (1) with the inferred parameters and the trajectory descriptors introduced in section 2.1 are computed and can be compared to the data descriptors in Fig. 2. Acceleration, speed, distance and displacement distributions are displayed in the upper panel, with quantiles 0.05, 0.5 and 0.95 (plain lines) and mean (dashed line). The mean values observed in the image data are also displayed for comparison (black dashed line). Interactions between the host biofilm and, respectively, acceleration and speed distributions are displayed in the lower panel with isolines enclosing 5, 50 and 95% of the points, centered in the densest zones. (b) Inferred parameter posterior distributions after analysis of the confocal swimmer images, and posterior mean (dashed line).



(a) **Entire ternary plot** (b) **Zoom in**

Figure 6: **Respective influence of stochastic effects, speed or direction adaptation to the host biofilm.** We plot in a ternary plot the respective influence of the speed selection (V), the direction selection (D) and the random term (ϵ) of Eq.(1) in the acceleration distribution of each species. Each squared instantaneous acceleration is mapped in the ternary plot coordinates through the relative contribution of V^2, D^2 and ϵ^2, and this point cloud is approximated in the ternary plot coordinates with a gaussian kernel to display the point distributions. The 0.05, 0.5 and 0.95 quantile isovalues of these distributions are plotted. (a) The entire ternary plot is displayed. The dashed line represents the zoom box represented in Fig. (b), where the same isolines are displayed, but with a zoom in in the y direction to highlight differences between species.

data	N	A_{ref}	V_{ref}	$\sigma(A)$	$R^2_{det}[\%]$	ϵ^2
B. pumilus	33,916	81.08	7.89	0.87	58.80	0.36
B. sphaericus	20,152	44.93	4.74	0.58	48.50	0.30
B. cereus	23,160	108.92	7.03	0.63	32.72	0.42

Table 4: **Reference acceleration and speed, and acceleration variance decomposition between stochastic and deterministic terms.** The number N of acceleration times points is indicated for each specie. Then, reference values for acceleration A_{ref} and speed V_{ref} used for adimensionalization are computed by averaging the corresponding values by specie. Descriptive statistics of acceleration variance decomposition are then computed in order to illustrate the contribution of the deterministic terms in the observed acceleration distribution, and the part of the residual mechanisms that are not included in the model. We indicate for each species the acceleration variance $\sigma(A)$, the part of the variance explained by the deterministic terms R^2_{det} (see 4.9) and the variance of the stochastic term ϵ^2. We note that in order to compare species at vizualisation step, they are re-normalized with the average of the species reference values : $A_{ref} = 78.31$ and $V_{ref} = 6.55$

2.3.4 Ultrastuctural bacterial morphology

Both kinematic descriptors and swimming parameters can then be reinterpreted through the insights provided by the morphology of each bacteria species (Fig. 7). First, *B. sphaericus* bacteria are much longer than the other two species, which may explain why this species is the less motile in terms of acceleration and kinematics: its length may be a drawback for navigating in crowded areas. Besides, the three *Bacillus* do not have the same type of flagella: while both *B. pumilus* and *B. sphaericus* species present several long flagella distributed over the whole surface of the membrane, *B. cereus* shows a unique brush-like group of very thin flagella, at the tail of the bacteria. The kind, size and disposition of the flagella may helps *B. cereus* swimmers to adapt their runs to their environment by changing directions to follow lower density areas (higher impact of direction selection term of the three *Bacillus* in Fig. 6) or to adapt rapidly when biofilm density varies (largest γ). *B. cereus* being the bacteria with the strongest stochastic part (highest ϵ, density shifted towards $A(\epsilon)$ in Fig. 6), this morphology could also help the swimmer to go through the biofilm by random navigation. *B. pumilus*, which has the highest number of flagella, is also the bacteria that reaches the highest speeds specially in low-density areas with rather fast changes for varying biofilm densities (intermediary γ value), indicating that this characteristic may be an advantage for swimming fast in the extracellular matrix.

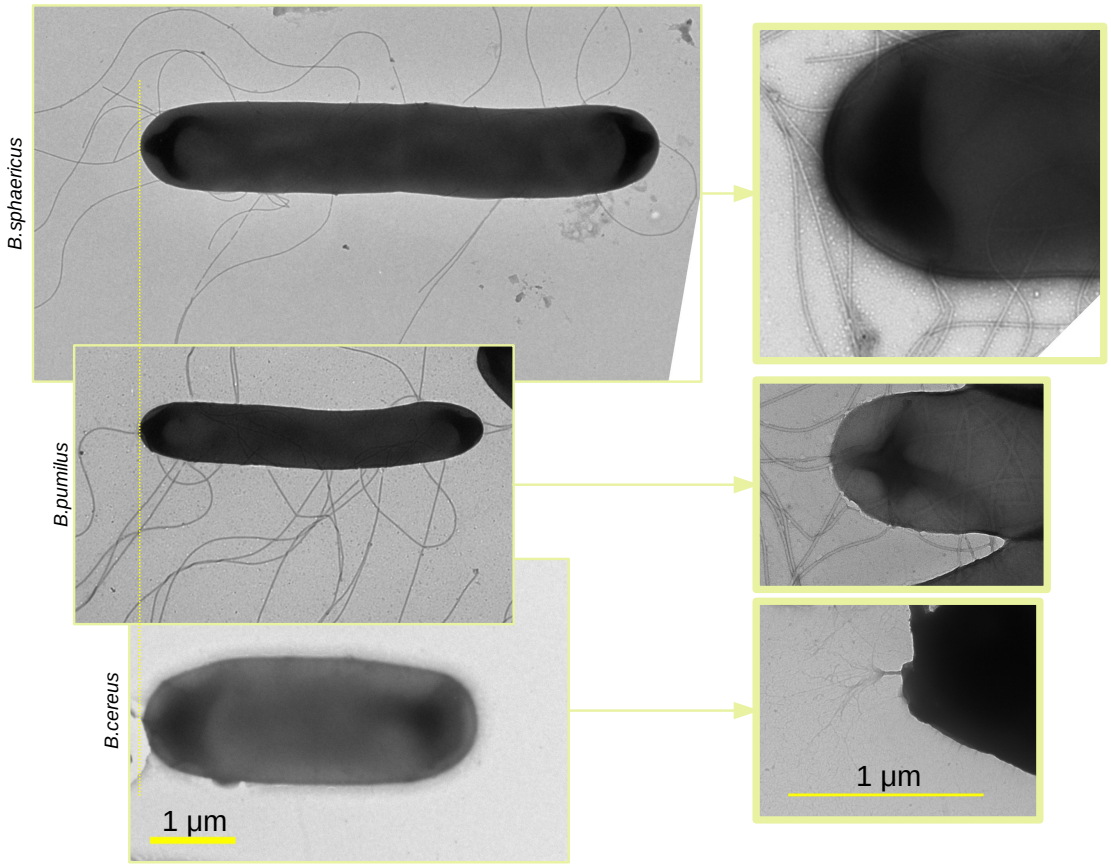


Figure 7: **TEM images of the three *Bacillus*.** TEM images of the three *Bacillus* are acquired, scaled in the same dimension and aligned (left panel). Images at lower scale are made with a zoom in on the flagella insertion (right panel).

3 Discussion

3.1 Modelling and analysis of swimming trajectories

When analyzing microbial swimming trajectories, two general strategies can be found in the literature. The first one aims at designing statistical tests quantifying similarities with or deviations from typical motion of interest such as diffusion [33]. Another strategy consists in providing a generative model of the data, analyzing it [8, 7] and comparing model outputs with real data [24, 20], possibly after inference. The model that is studied in this paper belong to the second category: the model includes deterministic mechanisms describing interactions with the host biofilm, together with a random correction counterbalancing the modelling errors. The parameter inference allows to interpret the data variance relatively to speed or direction adaptations to the host biofilm versus residual effects gathered in the stochastic term. Furthermore, the fitted model allows to simulate typical swimming trajectories of a given species.

3.2 Population-wide swimming characteristics vs true-state inference.

In this study, we do not aim to recover 'true' swimmer trajectories (a.e. the blue trajectory in Fig. A.6), i.e. identifying through smoothing techniques an approximation of the specific realization of the stochastic modeling and observation errors that lead to a given 'observed' trajectory. Rather, the goal is to identify common characteristics shared by a population of trajectories by inferring the 'population-wide' parameters (the parameters α, β, v_0, v_1, γ and ϵ) that best explain the whole set of observed accelerations in a same population of swimmers. For this reason, we did not introduced swimmer-specific terms nor individual noise: they would have increased the model accuracy, but to the price of a blurrier characterization of the species specificities.

This choice determined our inference framework. Despite several alternative options for recovering hidden states, in particular SSM (space state models) which are common in spatial ecology [1], the Bayesian method we opted for is a simpler non-linear regression problem that proved to be sufficient to recover macroscopic swimmer trajectories and species stratification. We discuss in section A.8 the different options that were tested and present in Sec. 4.7 the method for noise model selection. Among other interesting features, the Bayesian method provides confidence intervals on the final parameter estimation, and on the resulting trajectories as in Fig. 4a.

3.3 Predictive capabilities of the model

The deterministic terms of the model explain only half of the variance (Table 4). A major part of the underlying mechanisms is not correctly described by our model which is a common feature since it is a phenomenological model which only considers interactions with the underlying biofilm at a macroscopic level, without taking into account nanoscale physical mechanisms. A more detailed description of the underlying physics could have been designed as in [29], but it would have made more complex the analysis of the interactions between the host biofilm and the swimmer trajectories and the

extraction of species-specific patterns. However, we note that our model correctly renders observations made through macroscopic trajectory descriptors, even though the inference process has not been made based on these observables. Furthermore, several repetitions of the same models with different samples of the stochastic terms give very similar values for the trajectory descriptors (see Fig. A.7 and section A.7), showing that these descriptors are robust to stochastic perturbations. Hence, the model (2) can be used to produce synthetic data sharing the same global characteristics than the original ones specifically taking into accounts interactions between the swimmers and the host biofilm. Furthermore, these predictions also reproduce the species stratification observed in the original data using the global descriptors.

3.4 Biological interpretation of the fitted models

The direction selection term of the equation driven by β has little impact in the swimmer model fitted on real data. The parameter β can however have a sensible impact on the kinematics as shown in the sensitivity analysis, and on the trajectories as shown in mock biofilms (Fig. 3d). This could indicate that direction selection based on biofilm gradients is marginally effective in real-life swimming trajectories in a biofilm matrix. On the contrary, the speed selection term is more effective for the three *Bacillus*, showing that these micro-swimmer are able to adapt their swimming velocity to the biofilm density faced during their run. This term acts as an inertial term which enhances the stochastic term to provide direction and velocity changes.

The model has been used to decipher different adaptation strategies to the host biofilm of the three species during their swim. It indicates that *B. sphaericus* are the less motile bacteria, whereas *B. pumilus* can adapt their speed to the biofilm density they encounter and *B. cereus* are more driven by stochastic effects with slight capabilities to adapt their direction to density gradients. This characterization methodology could be used to drive species selection for improved biofilm control. Furthermore, the model can be used to predict new trajectories and the resulting biofilm vascularization, in a similar framework as in [18]. Coupled with a model of biocide diffusion, these simulations could be used to test numerically the efficiency of mono- or multi-species swimmer pretreatment to improve the removal of the host biofilm.

4 Materials and Methods

4.1 Infiltration of host biofilms by bacilli swimmers

Infiltration of *S. aureus* biofilms by bacilli swimmers were prepared in 96-well microplates. Submerged biofilms were grown on the surface of polystyrene 96-well microtiter plates with a μclear® base (Greiner Bio-one, France) enabling high-resolution fluorescence imaging [4]. 200 μL of an overnight *S. aureus* RN4220 pALC2084 expressing GFP [28] cultured in TSB (adjusted to an OD 600 nm of 0.02) were added in each well. The microtiter plate was then incubated at 30 °C for 60 min to allow the bacteria

to adhere to the bottom of the wells. Wells were then rinsed with TSB to eliminate non-adherent bacteria and refilled with 200 μL of sterile TSB prior incubation at 30 celsius for 24 h. In parallel, *B. sphaericus 9A12*, *B. pumilus 3F3* and *B. cereus 10B3* were cultivated overnight planktonically in TSB at 30°C. Overnight cultures were diluted 10 times and labelled in red with 5 μM of SYTO 61 (Molecular probes, France). After 5 minutes of contact, 50 μL of labelled fluorescent swimmers suspension were added immediately on the top of the *S. aureus* biofilm. All microscopic observations were collected within the following 30 minutes to avoid interference of the dyes with bacterial motility. Three replicates were conducted.

4.2 Confocal Laser Scanning Microscopy (CLSM)

The 96 well microtiter plate containing 24h *S. aureus* biofilm and recently added *bacilli* swimmers were mounted on the motorized stage of a Leica SP8 AOBS inverter confocal laser scanning microscope (CLSM, LEICA Microsystems, Germany) at the MIMA2 platform (`https://www6.jouy.inra.fr/mima2_eng/`). Temperature was maintained at 30 celsius during all experiments. 2D+T acquisitions were performed with the following parameters: images of 147.62 x 147.62 μm were acquired at 8000 Hz using a 63×/1.2 N.A. To detect GFP, an argon laser at 488 nm set at 10% of the maximal intensity was used, and the emitted fluorescence was collected in the range 495 to 550 nm using hybrid detectors (HyD LEICA Microsystems, Germany). To detect the red fluorescence of SYTO61, a 633 nm helium-neon laser set at 25% and 2% of the maximal intensity was used, and fluorescence was collected in the range 650 to 750 nm using hybrid detectors. Images were collected during 30 s (see 1 for sampling period).

Bacterial swimmers navigate within a three-dimensional biofilm matrix and confocal microscope refreshment time is not small enough to allow 3D+T images. To limit 3D trajectories, a focal plane near the well bordure has been selected, where the well wall physically constrains the swimmer trajectories in one direction, which select longer trajectories in the 2D plane that can be tracked in time. Therefore, experimental data are composed of two-dimensional trajectories captured between the swimmer arrival and departure times in the focal plane, and the associated 2D+T biofilm density images that change over time due to swimmer action.

4.3 Transmitted Electron Microscopy

Materials were directly adsorbed onto a carbon film membrane on a 300-mesh copper grid, stained with 1% uranyl acetate, dissolved in distilled water, and dried at room temperature. Grids were examined with Hitachi HT7700 electron microscope operated at 80 kV (Elexience – France), and images were acquired with a charge-coupled device camera (AMT).

4.4 Post-processing of image data

See Fig. 1 for a sketch of the datastream from microscope raw images to model inputs and Fig. A.1 for data visualization at each step of the post-processing pipeline.

Swimmer tracking has been applied on the red channel of the raw temporal stacks with *IMARIS* software (Oxford Instruments) using the tracking function after automated spots detection to get position time-series for each swimmer. Time-series with less than 8 time steps were filtered out.

Then, swimmer speed and acceleration time-series were computed from their position by finite-difference approximations and trajectory descriptors were extracted. The *RGB* biofilm density temporal images were converted into grayscale and rescalled between 0 and 1 (linear scalling). Post-processed data are available at https://forgemia.inra.fr/bioswimmers/swim-infer/SwimmerData.

Trajectory descriptors are defined as follows. The mean acceleration and speed values, distance and displacement are computed with $\|A\|_i^s = \frac{1}{T_i^s}\sum_t \|A_i^s(t)\|$, $\|V\|_i^s = \frac{1}{T_i^s}\sum_t \|V_i^s(t)\|$, $dist_i^s = \Delta t \sum_{T_{0,i}^s}^{T_{end,i}^s} \|V_i^s(t)\|$ and $disp_i^s = \|X(T_{end,i}^s) - X(T_{0,i}^s)\|$. To compute the visited area, each trajectory piece was subsampled by computing $X_i^s(t_k) = \frac{k}{n_s}X_i^s(t) + (1-\frac{k}{n_s})X_i^s(t+1)$ for $k = 1, n_s$, with $n_s = 10$ and the pixels included in the ball $B(X_i^s(t_k), r)$ with radius $r = 2$ where labelled. The total area of the labelled pixels is defined as the visited area of the swimmer i of species s.

4.5 Computation of the forward swimming model

Time integration of equations (2) has been solved with an explicit Euler scheme regarding positions $\mathbf{x}_{i,t}^s$ and velocities $\mathbf{v}_{i,t}^s$ of the swimmer i of species s at time t:

$$\mathbf{x}_{i,t+1}^s = \mathbf{x}_{i,t}^s + \mathbf{v}_{i,t}^s \mathrm{dt} \tag{8}$$

$$\mathbf{v}_{i,t+1}^s = \mathbf{v}_{i,t}^s + \mathbf{dv}_{i,t}^s \tag{9}$$

where $\mathbf{dv}_{i,t}^s$ is given by eq. (2), and depend on θ^s, $V_{i,t}^s$, $x_{i,t}^s$, $b(t, x_{i,t}^s)$ and $\nabla b(t, x_{i,t}^s)$. In practice, the biofilm density and gradient maps b and ∇b are discretized with a Cartesian grid corresponding to the image voxels.

During random walks, swimmer may exit the biofilm domain. When the swimmer reaches the domain boundary, a new swimmer is introduced with a velocity oriented towards the interior of the domain while the original trajectory is stopped at the boundary.

4.6 Sensitivity analysis

A local sensitivity analysis (Fig. 3) is performed by comparing basal simulation obtained with $\gamma = \beta = \epsilon = 1$ (v_0 and v_1 where taken as in Table 2) with 3 simulations where γ, β and ϵ are alternatively set to 0, resulting in 3 alternative models where the speed or the direction selection or the random term is turned off. The interaction between the speed selection term (set by γ) and the random term is illustrated in Fig. A.2 where 5

repetitions of the same trajectory of a simplified Langevin equation (10) are displayed with or without friction ($\gamma = 1$ or $\gamma = 0$), but with the same random seed for the stochastic term so that the stochastic part is strictly identical.

To analyze the impacts of the non-dimensionalized swimming parameters γ, v_0, v_1, β, ϵ on the locomotion behaviour, a global sensitivity analysis has been performed. The parameter space $[0,1]^5$ was uniformly sampled with $n = 1,000$ points using the Fourier Amplitude Sensitivity Test (FAST) sampler of the *SALib* library *i.e.* the function *SALib.sample.fast_sampler.sample* [10, 36]. We note that the interval $[0,1]$ covers a large parameter domain for some parameters, in particular β which remains small after inference. For this parameter, the sensitivity analysis will show potential impact on the output, that may be ineffective in the parameter range of the inferred model.

For each point in the parameter space, a forward simulation is conducted on a population of swimmers on a representative biofilm extracted from the dataset (first batch of the *B. pumilus* dataset). Trajectory descriptors are then extracted and taken as observable of the sensitivity anaylsis that requires both the parameters sampling and the associated descriptors. Sobol indices of first order are then returned and pairwise partial correlations matrix has been calculated. Convergence of the Sobol indices has been checked by taking sub-samples containing less than $1,000$ points.

4.7 Inference

Numerical implementation The inverse problem (4)-(7) has been implemented using a Hamiltonian Monte Carlo (HMC) method to solve this Bayesian inference problem.

The three replicates for each swimmer species are pooled (trajectories and biofilm density maps) and the input data required for the inference procedure (velocity $\mathbf{yV}$ and acceleration $\mathbf{yA}$ times series for the whole batch of swimmers, biofilm densities yb and gradient $\mathbf{yGb}$ extracted at swimmer positions) were assembled in a customed data structured. Normal standard prior distributions were set for all swimming parameters $\theta = (\gamma, v_0, v_1, \beta, \epsilon)$. Additional positivity constrained were imposed for all parameters but β. Therefore, the implemented model can be summarized as:

$$\theta \sim \mathcal{N}(0,1), \quad \gamma \geq 0, \quad v_0 \geq 0, \quad v_1 \geq 0, \quad \epsilon \geq 0$$

$$\mathbf{yA} \sim \mathcal{N}(f_A\left(\gamma, v_0, v_1, \beta | yb, \mathbf{yV}, yb, \mathbf{yGb}, dt\right), \epsilon)$$

A *warmup* of 1,000 runs is followed by the Markov chains construction (4,000 iterations for 4 Markov chains). Markov chain convergence is assessed by direct visualization (Fig. A.4) by checking for biaised covariance structures in pair-plots (Fig. A.5). Standard convergence index were additionnaly computed: effective sample size per iteration (n_{eff}) and potential scale reduction factor (R_{hat}).

Noise model selection Different noise models have been evaluated for the regression model (5) to take into account batch or individual effects. Namely, we decomposed the noise in Eq. (5) by replacing $\eta^{\mathbf{s}}$ by ${\eta^{\mathbf{s}}}_i$ and/or $\eta^{\mathbf{s,b}}$ for individual i and experimental batch b. Model selection has been conducted by computing the WAIC for the different

noise models. A huge degradation of the WAIC has been observed for individual or batch dependant noises, indicating that the enhancement of the inference accuracy provided by the additional parameters can be considered as over-fitting and discarded.

4.8 Inference validation on synthetic data

Ground truth data construction Ground truth synthetic data (see section Assessment of the inference with synthetic data) were computed by solving eq. (9) with $\gamma = 10$ s^{-1}, $v_0 = 5$ $m.s^{-1}$, $v_1 = 1$ $m.s^{-1}$, $\beta = 10$ $m.s^{-2}$, $\epsilon = 40$ $m.s^{-2}$ and biofilm maps taken from the first batch of the *B. pumilus* dataset. The number of swimmers was fixed to $N = 50$ and the number of time steps was taken identical to the experimental data *i.e.* $N_t = 224$. Resulting mean speeds and accelerations were $A_{ref} = 68.29$ $m.s^{-2}$, $V_{ref} = 7.47$ $m.s^{-1}$ and were used to rescale the data before inference together with the ground truth parameters (Table 2). In total, the acceleration dataset contains 9,523 samples for each spatial direction.

Comparing ground truth data with the fitted model After inference, a new dataset is obtained by solving eq. (9) with the fitted parameters. The same initial conditions for speeds and positions as the ground truth data are taken. Trajectories are stopped after the same number of time step as in the corresponding trajectory of the ground truth dataset. To discard spurious stochastic uncertainties, the same random seed as the ground truth simulations was taken, so that the unique uncertainty source was inference errors.

4.9 Inference validation on experimental data

Comparing microscopy data with the fitted model The same procedure is repeated on the microscopy data: after inference, a new dataset is obtained by solving eq. (9) with the fitted parameter, taking the same initial conditions for speeds and positions. Trajectories are stopped after the same number of time step as in the corresponding trajectory of the ground truth experimental dataset.

Measuring the deterministic reconstruction The deterministic coefficient of determination R^2_{det} was computed to measure how much the dataset is explained by the deterministic part of the model. Setting $A_i^{s,det} = f_A\left(\gamma, v_0, v_1, \beta | yb, \mathbf{yV}, yb, \mathbf{yGb}, dt\right)$:

$$R^{2,s}_{det} = 1 - \frac{\sum_i (yA_i^s - A_i^{s,det})^2}{\sum_i (yA_i^s - y\bar{A}^s)^2}$$

where $y\bar{A}^s$ is the acceleration mean. $R^{2,s}_{det}$ is expected to tend towards 1 when the stochastic term $\eta = \mathcal{N}(0, \epsilon)$ becomes negligible with respect to A^{det}.

4.10 Plots and statistics

To allow inter-species comparisons in plots, the data and model outputs are re-normalized with common reference values A_{ref} and V_{ref} defined as the average of the species reference values (see Table 4 for values). Uni-dimensional distributions (figures 2 upper panel, 4a upper panel, 5b and 5a upper panel) were obtained with the *gaussian_kde* function of scipy.stats. T tests for mean comparison were performed using scipy.stats *ttest_ind*.

Two-dimensional distribution plots (figures 2, 4a, 5a lower panels) were obtained by first plotting the two-dimensional point cloud and approximating the point distribution with a gaussian KDE using scipy.stats *gaussian_kde* function. Then, the gaussian kde is evaluated at each point of the point cloud and quantiles 0.05, 0.5 and 0.95 of the resulting values are computed. Finally, quantile isovalues are plotted and the point cloud and the KDE are removed (see Fig. A.8 and Sec. A.9 for details): this procedure ensure to enclose 5, 50 and 95 % of the original points, centered in the densest zones of the initial point cloud.

Ternary plots (Fig. 6) were obtained by first computing the contribution of each term of equation 4 to acceleration estimate. Namely, note

$$s(b)_i^s = \|\gamma(v_0^s + b(t, X_i^s(t))(v_1^s - v_0^s) - \|V_i^s(t)\|)\frac{V_i^s(t)}{\|V_i^s(t)\|}\|,$$

$$s(\nabla b)_i^s = \|\beta^s \frac{\nabla b(t, X_i^s(t))}{\|\nabla b(t, X_i^s(t))\|}\|, \quad \text{and} \quad s(\eta)_i^s = \|\eta^{\mathbf{s}}\|.$$

We compute the proportions $A(k)_i^s$ for $k \in \{b, \nabla b, \eta\}$

$$A(k)_i^s = \frac{s(k)_i^s}{s(b)_i^s + s(\nabla b)_i^s + s(\eta)_i}^s.$$

Points $(A(b)_i^s, A(\nabla b)_i^s, A(\eta)_i^s)$ are then plotted in ternary plots using the Ternary python package [13] and approximated by gaussian KDE. Isolines are finally plotted as previously described.

4.11 Code availability

All the image pre- and post-processing, calculations and statistics have been performed with custom scripts using the standard python libraries numpy [17], scipy [40], imageio [21] and pandas [31]. The forward swimming problem computation is computed using customed scripts built upon numpy [17] and H5py (https://www.h5py.org). Sensitivity analysis has been conducted with the function *SALib.analyze.fast.analyze* of the *SALib* library [10, 36] (Sobol index), *pcorr* method of the *pingouin* library [39] (PCC). The Bayesian inference has been conducted using the *STAN* library [38] through its python interface *pystan* [35]. All plots have been made with the matplotlib python library [19].

The whole *python* code has been made available and accessible at the following git repository https://forgemia.inra.fr/bioswimmers/swim-infer.

Acknowledgements

This work has benefited from the facilities and expertise of MIMA2 MET – GABI, INRAE, AgroParistech, 78352 Jouy-en-Josas, France. C. Péchoux is warmly acknowledged for TEM observations. Financial support was provided by the French National Research Agency ANR-12-ALID-0006.

Competing interests

The authors declare no competing interests.

References

[1] Marie Auger-Méthé, Ken Newman, Diana Cole, Fanny Empacher, Rowenna Gryba, Aaron A. King, Vianey Leos-Barajas, Joanna Mills Flemming, Anders Nielsen, Giovanni Petris, and Len Thomas. A guide to state–space modeling of ecological time series. *Ecological Monographs*, n/a(n/a):e01470.

[2] Iwona B Beech and Jan Sunner. Biocorrosion: towards understanding interactions between biofilms and metals. *Current opinion in Biotechnology*, 15(3):181–186, 2004.

[3] Arnaud Bridier, Romain Briandet, V Thomas, and Florence Dubois-Brissonnet. Resistance of bacterial biofilms to disinfectants: a review. *Biofouling*, 27(9):1017–1032, 2011.

[4] Arnaud Bridier, Florence Dubois-Brissonnet, A Boubetra, V Thomas, and Romain Briandet. The biofilm architecture of sixty opportunistic pathogens deciphered using a high throughput clsm method. *Journal of microbiological methods*, 82(1):64–70, 2010.

[5] Arnaud Bridier, Jean-Christophe Piard, Caroline Pandin, Simon Labarthe, Florence Dubois-Brissonnet, and Romain Briandet. Spatial organization plasticity as an adaptive driver of surface microbial communities. *Frontiers in microbiology*, 8:1364, 2017.

[6] Arnaud Bridier, Pilar Sanchez-Vizuete, Morgan Guilbaud, J-C Piard, Murielle Naitali, and Romain Briandet. Biofilm-associated persistence of food-borne pathogens. *Food microbiology*, 45:167–178, 2015.

[7] Oleksandr Chepizhko, Eduardo G Altmann, and Fernando Peruani. Optimal noise maximizes collective motion in heterogeneous media. *Physical review letters*, 110(23):238101, 2013.

[8] Oleksandr Chepizhko and Fernando Peruani. Diffusion, subdiffusion, and trapping of active particles in heterogeneous media. *Physical review letters*, 111(16):160604, 2013.

[9] Jacinta C. Conrad and Ryan Poling-Skutvik. Confined flow: Consequences and implications for bacteria and biofilms. *Annual Review of Chemical and Biomolecular Engineering*, 9(1):175–200, 2018. PMID: 29561646.

[10] Fortuin C.M. Shuler K.E. Petschek A.G. Schaibly J.H. Cukier, R.I. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. *Journal of Chemical Physics*, 59:3873–3878, 1973.

[11] Nicolas Derlon, Maryna Peter-Varbanets, Andreas Scheidegger, Wouter Pronk, and Eberhard Morgenroth. Predation influences the structure of biofilm developed on ultrafiltration membranes. *Water research*, 46(10):3323–3333, 2012.

[12] Agapi I Doulgeraki, Pierluigi Di Ciccio, Adriana Ianieri, and George-John E Nychas. Methicillin-resistant food-related staphylococcus aureus: a review of current knowledge and biofilm formation for future studies and applications. *Research in microbiology*, 168(1):1–15, 2017.

[13] Marc Harper et al. python-ternary: Ternary plots in python. *Zenodo 10.5281/zenodo.594435*.

[14] Hans-Curt Flemming, Thomas R Neu, and Jost Wingender. *The perfect slime: microbial extracellular polymeric substances (EPS)*. IWA publishing, 2016.

[15] Hans-Curt Flemming, Jost Wingender, Ulrich Szewzyk, Peter Steinberg, Scott A Rice, and Staffan Kjelleberg. Biofilms: an emergent form of bacterial life. *Nature Reviews Microbiology*, 14(9):563–575, 2016.

[16] Hans-Curt Flemming and Stefan Wuertz. Bacteria and archaea on earth and their abundance in biofilms. *Nature Reviews Microbiology*, 17(4):247–260, 2019.

[17] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020.

[18] Ali Houry, Michel Gohar, Julien Deschamps, Ekaterina Tischenko, Stéphane Aymerich, Alexandra Gruss, and Romain Briandet. Bacterial swimmers that infiltrate and take over the biofilm matrix. *Proceedings of the National Academy of Sciences*, 109(32):13088–13093, 2012.

[19] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.

[20] Mehdi Jabbarzadeh, YunKyong Hyon, and Henry C Fu. Swimming fluctuations of micro-organisms due to heterogeneous microstructure. *Physical Review E*, 90(4):043021, 2014.

[21] Almar Klein, Sebastian Wallkötter, Steven Silvester, Anthony Tanbakuchi, Paul Müller, Juan Nunez-Iglesias, Mark Harfouche, actions user, Antony Lee, Matt McCormick, OrganicIrradiation, Arash Rai, Ariel Ladegaard, Tim D. Smith, Ghislain Antony Vaillant, jackwalker64, Joel Nises, Miloš Komarčević, rreilink, lschr, Hugo van Kemenade, Maximilian Schambach, Chris Dusold, DavidKorczynski, Felix Kohlgrüber, Ge Yang, Graham Inggs, Joe Singleton, Michael, and Niklas Rosenstein. imageio/imageio: v2.13.1, December 2021.

[22] Theresa Klein, David Zihlmann, Nicolas Derlon, Carl Isaacson, Ilona Szivak, David G Weissbrodt, and Wouter Pronk. Biological control of biofilms on membranes by metazoans. *Water research*, 88:20–29, 2016.

[23] Robin Köck, Karsten Becker, B Cookson, JE van Gemert-Pijnen, S Harbarth, JAJW Kluytmans, Martin Mielke, G Peters, RL Skov, MJ Struelens, et al. Methicillin-resistant staphylococcus aureus (mrsa): burden of disease and control challenges in europe. *Eurosurveillance*, 15(41):19688, 2010.

[24] Hana Koorehdavoudi, Paul Bogdan, Guopeng Wei, Radu Marculescu, Jiang Zhuang, Rika Wright Carlsen, and Metin Sitti. Multi-fractal characterization of bacterial swimming dynamics: a case study on real and simulated serratia marcescens. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2203):20170154, 2017.

[25] Sang Won Lee, K. Scott Phillips, Huan Gu, Mehdi Kazemzadeh-Narbat, and Dacheng Ren. How microbes read the map: Effects of implant topography on bacterial adhesion and biofilm formation. *Biomaterials*, 268:120595, 2021.

[26] Gaojin Li and Arezoo M Ardekani. Collective motion of microorganisms in a viscoelastic fluid. *Physical review letters*, 117(11):118001, 2016.

[27] Yingbo Li, Romain Briandet, and Alain Trubuil. Tracking swimmers bacteria and pores within a biofilm. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 302–305. IEEE, 2014.

[28] Cheryl L Malone, Blaise R Boles, Katherine J Lauderdale, Matthew Thoendel, Jeffrey S Kavanaugh, and Alexander R Horswill. Fluorescent reporters for staphylococcus aureus. *Journal of microbiological methods*, 77(3):251–260, 2009.

[29] Vincent A Martinez, Jana Schwarz-Linek, Mathias Reufer, Laurence G Wilson, Alexander N Morozov, and Wilson CK Poon. Flagellated bacterial motility in

polymer solutions. *Proceedings of the National Academy of Sciences*, 111(50):17771–17776, 2014.

[30] Carmen C Mayorga-Martinez, Jaroslav Zelenka, Jan Grmela, Hana Michalkova, Tomáš Ruml, Jan Mareš, and Martin Pumera. Swarming aqua sperm micromotors for active bacterial biofilms removal in confined spaces. *Advanced Science*, page 2101301, 2021.

[31] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[32] Alise R Muok, Dennis Claessen, and Ariane Briegel. Microbial hitchhiking: how streptomyces spores are transported by motile soil bacteria. *The ISME Journal*, pages 1–10, 2021.

[33] AE Patteson, Arvind Gopinath, M Goulian, and PE Arratia. Running and tumbling with e. coli in polymeric solutions. *Scientific reports*, 5(1):1–11, 2015.

[34] JC Piard, SY Kim, J Deschamps, Y Li, C Dorel, A Gruss, A Trubuil, and R Briandet. Travelling through slime–bacterial movements in the eps matrix. *The Perfect Slime: Microbial Extracellular Polymeric Substances (EPS)*, page 179, 2016.

[35] Allen Riddell, Ari Hartikainen, and Matthew Carter. pystan (3.0.0). PyPI, March 2021.

[36] Andrea Saltelli, Stefano Tarantola, and KP-S Chan. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56, 1999.

[37] Tahoura Samad, Nicole Billings, Alona Birjiniuk, Thomas Crouzier, Patrick S Doyle, and Katharina Ribbeck. Swimming bacteria promote dispersal of non-motile staphylococcal species. *The ISME journal*, 11(8):1933–1937, 2017.

[38] Stan Development Team. The Stan Core Library, 2018. Version 2.18.0.

[39] Raphael Vallat. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026, 2018.

[40] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

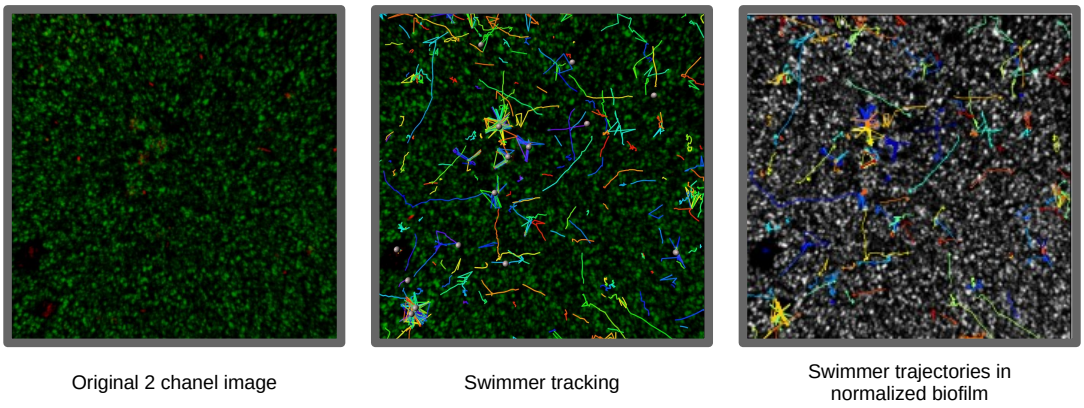


Figure A.1: **Illustration of image data along the post-processing process.** Raw data (2 chanel images) are first displayed. Then, trajectory tracking are obtained. Finally, the biofilm density map is rescalled, and mapped to grayscale.

[41] Zhuodong Yu, Cory Schwarz, Liang Zhu, Linlin Chen, Yun Shen, and Pingfeng Yu. Hitchhiking behavior in bacteriophages facilitates phage infection and enhances carrier bacteria colonization. *Environmental Science & Technology*, 55(4):2462–2472, 2020.

A Appendix

A.1 Illustration of the datastream

Illustrations of the image data at different steps of the data stream are displayed in Fig. A.1, from raw microscopy data to rescalled biofilm density map with trajectories. Trajectories are mapped into the biofilm density map at initial condition of the first *B. pumilus* batch.

A.2 Statistical tests

T-tests were performed to compare mean differences between 1D distribution of Figure 2b. Resulting p-values are displayed in A.1.

	$\|A\|$		$\|V\|$		*disp*		*dist*		*Area*	
	B. cereus	*B. sphaericus*	*B. cereus*	*B. sphaericus*	*B. cereus*	*B. sphaericus*	*B. cereus*	*B. sphaericus*	*B. cereus*	*B. sphaericus*
B. pumilus	$5.e-9$	$1.e-10$	$4.e-2$	$1.e-13$	$1.e-5$	$2.e-3$	$8.e-3$	$7.e-10$	**$7.e-1$**	$4.e-14$
B. cereus		$5.e-51$		$2.e-13$		**$3e-1$**		$5e-15$		$1e-12$

Table A.1: **P-values of pairwise comparison between distributions**. Pairwise comparison were performed between 1D distributions displayed in Figure 2b using T-test and p-values are displayed. Non-significant values are indicated in bold.

A.3 Friction and random term in Langevin equations.

To illustrate the interplay between the friction and the random term during a random walks, we solve the problem

$$\mathrm{d}\mathbf{v} = -\underbrace{\gamma \mathbf{v}\mathrm{dt}}_{friction} + \underbrace{\eta \mathrm{dt}}_{\text{random term}} \tag{10}$$

$$\mathbf{v}(0) = (0,0) \tag{11}$$

$$\mathbf{X}(0) = (0,0) \tag{12}$$

in an unconstrained domain, with η a 2 dimensional white noise with unitary variance. The friction parameter γ is alternatively set to 1 (Fig A.2, upper panel) or 0 (Fig A.2, lower panel). We note that the random seed is the same for the simulations with or without the friction term, so that the stochastic contribution is completely identical in the upper and lower panels. The trajectories produced without the friction term are much more regular and rectilinear that those produced with the friction term, that are much chaotic.

The reason of that behaviour may come from the null mean of the white noise. Roughly speaking, in average, the acceleration shows small variations around zero which leads after temporal integration to regular speeds and rectilinear-like trajectories. By contrast, the friction term reduces the particle inertia, enhancing the impact of the stochastic term, which produces much more chaotic trajectories.

A.4 Model sensitivity analysis

The link between the model parameters and the global trajectory descriptors introduced in Section 2.1 is not intuitive. A global sensitivity analysis of the trajectory descriptors

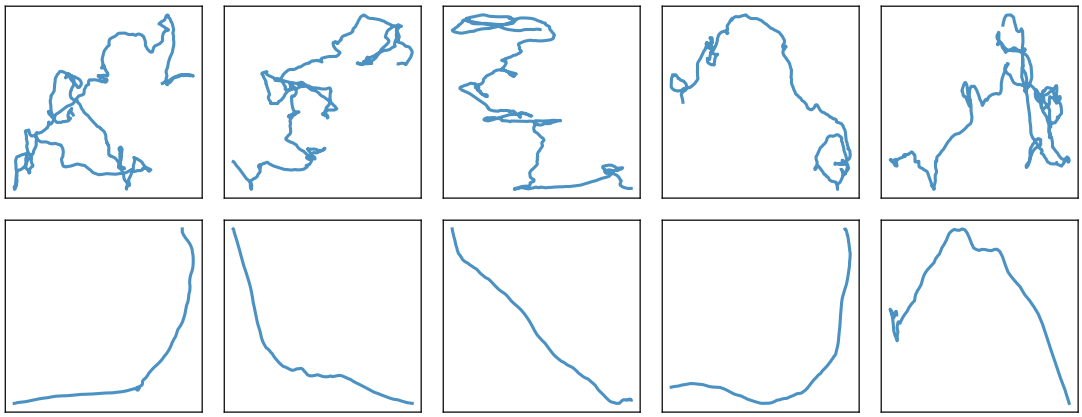


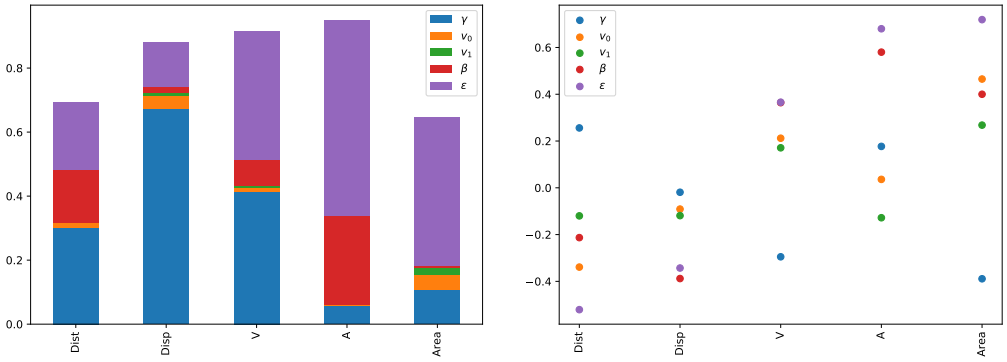
Figure A.2: **Illustration of the interplay between friction and stochastic terms in Langevin equation.** Trajectories produced by different repetitions of eq. 10 are displayed with $\gamma = 1$ (upper panel) and $\gamma = 0$ (lower pannel). We note that the same random seed has been taken for the simulations of the same column with or without the friction term, meaning that the stochastic term is strictly identical in both simulations.

(mean acceleration and speed, distance, displacement and visited areas) with respect to the parameters γ, v_0, v_1, β and ϵ is conducted by computing their first order Sobol index (SI) and their pairwise correlation coefficient (PCC).

The residual variance is small for the median speed and acceleration but slightly larger for the distance, displacement and visited area indicating larger effects of parameter interactions for these outputs, i.e. output variations induced by joint shifts of the parameters (Fig. A.3a). The SI of the parameters v_0 and v_1 are negligible, except for the displacement and the visited area. The parameters γ, β and ϵ, *i.e.* the three weights associated to each component of the state equation 2, are more influential. Distance and speed have several main drivers. The distance is impacted nearly equally by γ, β and ϵ and the PCC of this parameters is quite small, indicating that this parameters may induce indistinctly negative or positive variations of the travelled distance, except for ϵ which is slightly negatively correlated. The median speed is mainly impacted by ϵ (slightly positively) and γ (slightly negatively), with relatively small PCC (Fig. A.3b). The mean acceleration, the displacement and the visited area are preponderantly impacted by a main driver: the mean acceleration and the visited area are particularly impacted by ϵ, the stochastic term weight, with positive influence. The displacement is mainly influenced by γ with no preponderant variation direction (null PCC, Fig. A.3b).

A.5 Markov chains convergence and correlation

Markov chain (Fig. A.4) and markov chain pairplots (Fig. A.5) are displayed. Direct visualization of the posterior sampling allows to detect convergence failure (strong autocorrelation or stationnary markov chain). Markov chain pairplot informs on potential



(a) **Sobol index**. (b) **Pairwise correlation coefficients**.

Figure A.3: **Sensitivity analysis of state observable respectively to state-equation parameters.** The sensitivity of different state observables to parameter shifts is systematically studied through sensitivity analysis methods. a) Sobol indices are displayed for each output through barplots indicating the part of variance explained by a given parameter. The bars do not reach the value 1, indicating a residual variance reflecting interactions between parameters. b) Pairwise correlation coefficient (PCC) of the observable respectively to the input parameters are displayed. A negative PCC indicates a negative impact on the output, and conversely. We note that the red dot indicating the PCC of β for V is confounded with the purple one indicating the PCC of ϵ.

correlation between different parameters posterior samples, showing an interaction between parameter and an identification issue. In Fig. A.4, the markov chains correctly converged for all the parameter. No strong correlation can be observed in Fig. A.5.

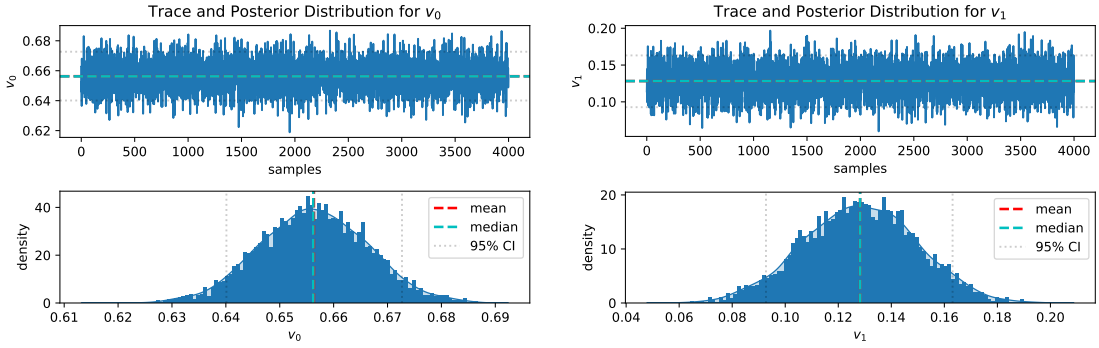
A.6 Impact of the stochastic term

We illustrate the impact of the random walk term on the overall swimmer trajectory with Fig. A.6. In this figure, we display two trajectories computed from model (2) with identical parameters (α, β, v_0, v_1, γ and ϵ), initial condition, host biofilm and time length. Different random samplings of the stochastic term of Equation (2) lead to these very different trajectories. This example illustrate the difference between identifying population-wide characteristics and inferring true trajectories : while the later try to detect the differences between the two trajectories (*i.e.* in this example, identifying and smoothing the different stochastic samples leading to this trajectories), the former focuses on the common features between these apparently different trajectories.

A.7 Influence of inference and stochastic terms on the trajectory descriptors

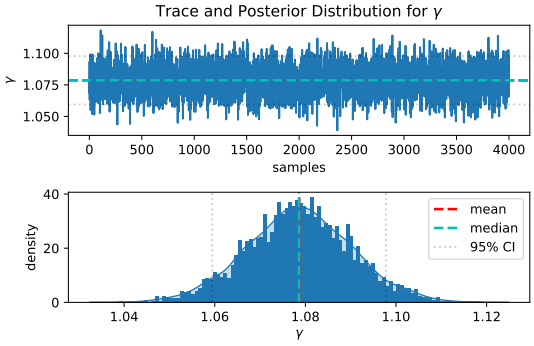
We wonder if the uncertainty sources involved in the inference process and in the stochastic term of the random walk have a decisive impact on the trajectory descriptors. To address this question, a first dataset is assembled by integrating in time Eq.(2) for given parameters (see table 2), initial conditions and host biofilm. Then, this dataset is used as inputs of the inference method to infer the initial parameters (ground truth). Another dataset is produced by replacing the initial parameters by the inferred parameters. We note that we take the same seed for the random number generator than for the initial dataset, so that the only uncertainty that has been introduced until this step comes from the inference procedure. Finally, we produce a last dataset by solving the model with the same inferred parameters as in the second dataset, but changing the seed of the random number generator. Hence, this last dataset involves uncertainties coming from the stochastic terms and from the inference process. This variation results in modifying the sampling of the stochastic terms and leads to strong modifications of the trajectories, like in Fig. A.6.

At end, the trajectory descriptors are computed and plotted in Fig. A.7. We can see that the trajectory descriptor distributions are very similar across the different dataset, except for the total distance and the displacement where discrepancies can be noted. However, these differences are relatively small compared to the mean and the width of the distributions. We can also observe that the interactions with the underlying biofilm is very well conserved, even when the sampling of the stochastic term is very different. This observation grounds the initial guess that these trajectory descriptors captures common global features of the different trajectories rather than specificities of given trajectories.

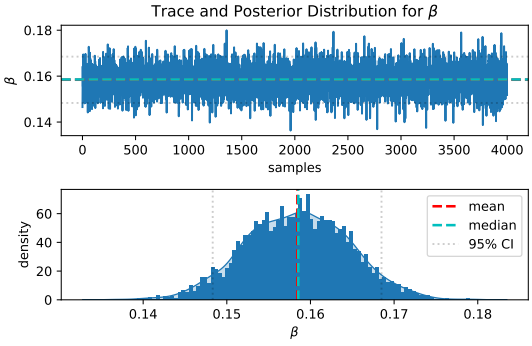


(a) v_0

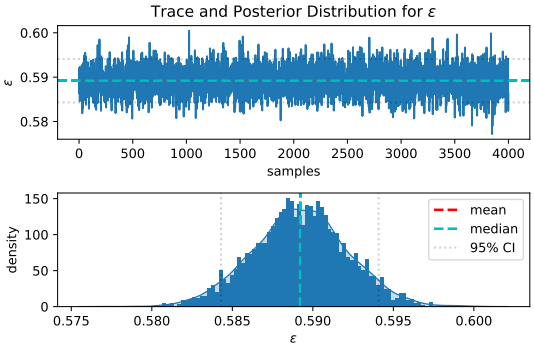
(b) v_1



(c) γ



(d) β



(e) ϵ

Figure A.4: **Inference convergence validation**. The markov chain (upper panel) and the posterior distribution (lower panel) of each parameter is displayed, showing good convergence of the stochastic sampling of the posteriors.

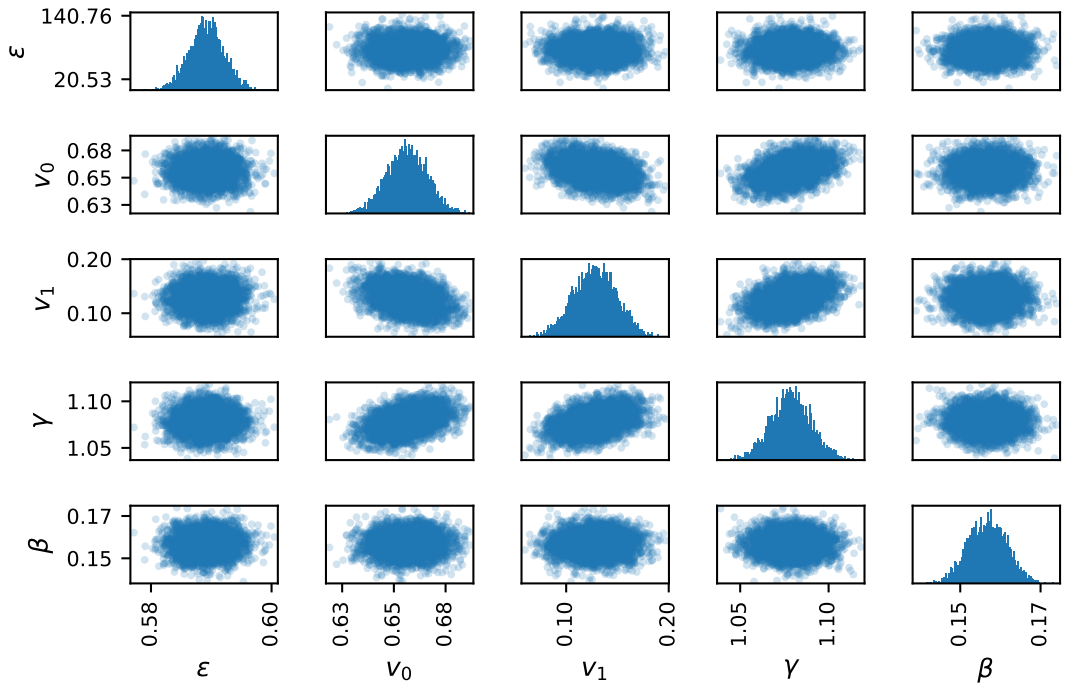


Figure A.5: **Pairplot of parameter markov chains**. No strong covariance effect can be observed, showing that the model can not be reduced by analytical dependence between parameters. Slight correlation is observed between the parameters v_0, v_1 and γ: this feature is not surprising since γ, v_0 and v_1 are in the same term of equation (2). The correlation is however too low to expect a model reduction.

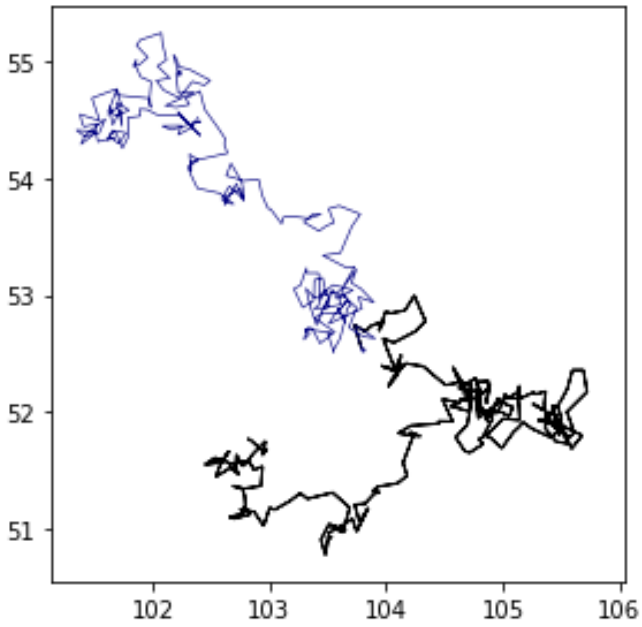


Figure A.6: **Influence of the stochastic process on swimmer trajectories**. We plot two different trajectories computed with the model (2), including the same parameters α, β, v_0, v_1, γ and ϵ, the same initial condition and identical host biofilm. The only uncertainty source comes from the different random samplings of the stochastic term. In this simulation, the ground truth (with default random seed) is plotted in blue.

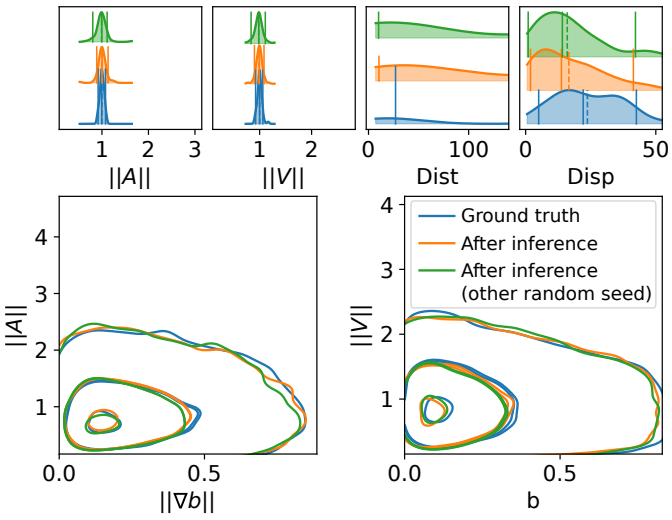


Figure A.7: **Low influence of the stochastic term on the trajectory descriptors.** To assess the influence of the random term on the population-wide trajectory descriptors and overall prediction accuracy, we repeated the experiment displayed in Fig 4b. A synthetic database was first assembled (ground truth) and prediction were performed with a fitted model (After inference). Then, a second repetition of the prediction of the fitted model was computed with another seed for the random number generator, resulting in modifying the sampling of the stochastic terms and strong modifications of individual trajectories, like in Fig. A.6. The population-wide trajectory descriptors are however slightly impacted by this random effect, indicating that the main characteristics of the trajectory populations marginally depend on the stochastic term.

A.8 Various inference models

Different inference models were designed and tested from the dimensionless state equation (2).

A.8.1 SSM model

The inference model can be stated as a space-state model (SSM) which is a framework commonly used in spatial ecology to infer a true state, i.e. true positions and trajectories, and population-wide random walk parameters from time-serie data [1]. The SSM inference model is a generalization of Hidden Markov Models (HMM).

Note $z_i^s(t)$ the true (hidden) position of the individual i of the species s at time t. The state model on acceleration (4) can be rewritten as

$$\frac{d\mathbf{v}_i^s(t)}{dt} = \gamma(v_0^s + b(t, z_i^s(t))(v_1^s - v_0^s) - \|\nu_i^s(t)\|)\frac{\mathbf{v}_i^s(t)}{\|\mathbf{v}_i^s(t)\|} + \beta^s \frac{\nabla b(t, z_i^s(t))}{\|\nabla b(t, z_i^s(t))\|} + \boldsymbol{\eta}^{\mathbf{s}}_{\mathbf{mod}} \quad (13)$$

$$\frac{z_i^s(t)}{dt} = \mathbf{v}_i^s(t) \quad (14)$$

In this equation, $\mathbf{v}_i^s$ is the true hidden swimmer velocity. Starting from observed initial conditions $z_i^s(0)$, $\mathbf{v}_i^s(0)$, equations (14) can be integrated in time to recover hidden $z_i^s(t)$, $\mathbf{v}_i^s(t)$ for all times t.

Then, a likelihood equation can be written to compare the true hidden state to the observations.

$$X_i^s(t) \sim z_i^s(t) + \boldsymbol{\eta}^{\mathbf{s}}_{\mathbf{obs}} \quad (15)$$

We note a link between η_{mod} and η_{obs} in Eqs. (14)-(15) and the random state η in Eq. (4). Namely, noting σ_{mod} and σ_{obs} the standard deviation of the gaussian noises η_{mod} and η_{obs}, direct finite-difference of $A_i^s(t)$ from the true state gives an estimate of the noise variance on the acceleration of the non-linear regression model

$$\epsilon = \sqrt{\left(\frac{\sigma_{mod}}{\Delta t}\right)^2 + \left(\sqrt{6}\frac{\sigma_{obs}}{\Delta t^2}\right)^2}.$$

Compared to problem (5), the main advantages are that the likelihood is written on the original data, i.e. the observed position, and not a post-processed observed acceleration, subject to finite-difference errors. Furthermore, the true trajectories are recovered and modelling errors $\boldsymbol{\eta}^{\mathbf{s}}_{\mathbf{mod}}$ and observation errors $\boldsymbol{\eta}^{\mathbf{s}}_{\mathbf{obs}}$ are separated. The main drawback of this methodology is that the state space is very large since it includes all the positions and speeds at every time for every swimmers, which leads to intractable computations.

A.8.2 Mixing SSM and non-linear inference models

An intermediary strategy has been designed by selecting swimmer trajectories that we want to infer by SSM, the remaining trajectories being kept to compute an acceleration dataset $A_i^s(t)$. Namely, note D_{ssm} the set of swimmer index kept for SSM, and D_A the set of swimmer index kept for non-linear regression. We set, for $i \in D_{ssm}$

$$\frac{d\mathbf{v}_i^s(t)}{dt} = \gamma(v_0^s + b(t, z_i^s(t))(v_1^s - v_0^s) - \|\nu_i^s(t)\|)\frac{\mathbf{v}_i^s(t)}{\|\mathbf{v}_i^s(t)\|} + \beta^s \frac{\nabla b(t, z_i^s(t))}{\|\nabla b(t, z_i^s(t))\|} + \eta^{\mathbf{s}}_{\mathbf{mod}} \quad (16)$$

$$\frac{z_i^s(t)}{dt} = \mathbf{v}_i^s(t) \quad (17)$$

for given initial conditions $z_i^s(0)$, $\mathbf{v}_i^s(0)$, and for $j \in D_A$

$$A_j^s(t) = \gamma(v_0^s + b(t, X_j^s(t))(v_1^s - v_0^s) - \|\nu_i^s(t)\|)\frac{V_j^s(t)}{\|V_j^s(t)\|} + \beta^s \frac{\nabla b(t, X_j^s(t))}{\|\nabla b(t, X_j^s(t))\|} + \eta^{\mathbf{s}} \quad (18)$$

where $X_j^s(t)$, $V_j^s(t)$ and $A_j^s(t)$ are observed positions, speeds and accelerations. This model is completed by a likelihood equation

$$X_i^s(t) \sim z_i^s(t) + \eta^{\mathbf{s}}_{\mathbf{obs}}, \text{ for } i \in D_{SSM} \quad (19)$$

$$A_i^s(t) \sim f_A(\theta^s | b, X_j^s(t), V_j^s(t), A_j^s(t)) + \eta^{\mathbf{s}} \quad (20)$$

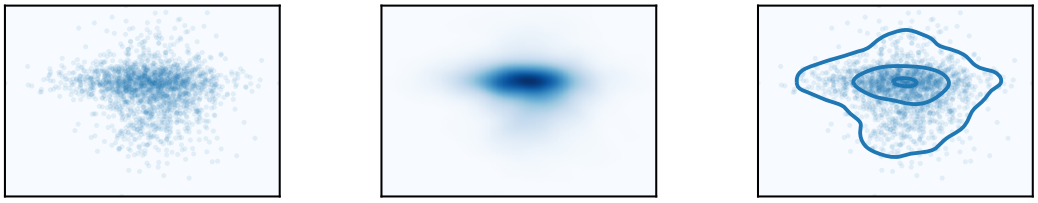
where f_A is defined in equation (4).

This setting kept some advantages of the SSM, like inferring some true hidden trajectories or separating the estimate of modeling and observation errors, while limiting the computational load if D_{SSM} is not too large.

We finally kept the regression model for several reasons. First, we are interested in recovering population wide parameters to characterize strain-specific swims, and not identifying true trajectories. Second, we can consider that the observation error with confocal microscopy is several order of magnitudes under the spatial characteristic lengths involved in equation (2), so that observation errors can be neglected. Hence, the objective of separating the uncertainty sources between model and observation errors, which is a main advantage of the SSM or mixed inference settings, becomes secondary. Furthermore, enhancing the state space dimension provided additional uncertainties, worsening the inference precision on synthetic data. We then opted for the simple regression model that provided sufficient parameter identifiability for limited computational load.

A.9 KDE computation

We illustrate the process of visualization of multiple point distributions in the same graph using KDE and isolines enclosing specific proportions of the data in Fig. A.8. A point cloud is first approximated with a Gaussian KDE. Then, the value of the gaussian



(a) Original point cloud (b) Gaussian KDE (c) Isolines

Figure A.8: **Illustration of the Gaussian KDE isovalues computation**. Starting from a random 2D point distribution, a gaussian KDE is computed using the scipy.stats function. Then, the gaussian KDE is evaluated at the original point positions, and quantiles of the resulting values are computed (quantiles 0.05, 0.5 and 0.95). Gaussian KDE isolines corresponding to this quantiles are finally computed. This isolines enclose respectively 5, 50 and 95 % of the points of the original distribution, centered in the densest area of the initial point cloud. This procedure gives a good representation of the shape of the data, but allows to display several distributions in the same graph, enabling comparison while avoiding superimposition issues.

KDE is evaluated in each point of the original point cloud, which allows to map the 2D map into a 1D set where order relation can be defined. Specific quantiles of the resulting values are computed (namely quantile 0.05, 0.5 and 0.95). By definition, the quantile 0.05 separate 5% of the points of the original dataset (the 5% lowest Gaussian KDE values) from the remainder of the data set. The isoline corresponding to the quantile 0.05 then also separates in the 2D map the 5% lowest Gaussian KDE values from the others.