Quantifying Drift-Selection Balance Using an Agent-Based Biofilm Model of Identical Heterotrophs Under Low Nutrient Conditions

Joseph Earl Weaver*

School of Civil Engineering &Geosciences, Newcastle University, Cassie Building, Newcastle upon Tyne, NE1 7RU, United Kingdom

* Corresponding author(s). E-mail: Joe.Weaver@newcastle.ac.uk

Abstract

Both deterministic and stochastic forces shape biofilm communities, but the balance between those forces is variable. Quantifying the balance is both desirable and challenging. For example, drift-driven drift failure, a stochastic force, can be thought of as an organism experiencing 'bad luck' and manipulating 'luck' as a factor in real world systems is difficult. We used an agent-based model to manipulate luck by controlling seed values governing random number generation. We determined which organism among identical competitors experienced the greatest drift-driven failure, gave it a deterministic growth advantage, and re-ran the simulation with the same seed. This enabled quantifying the growth advantage required to overcome drift, *e.g.,* a 50% chance to thrive may require a 10-20% improved growth rate. Further, we found that crowding intensity affected that balance. At moderate spacings, there were wide ranges where neither drift nor selection dominated. Those ranges shrank at extreme spacings; close and loose crowding respectively favoured drift and selection. We explain how these results may partially illuminate two conundrums: the fact that a stably operating wastewater treatment plant's

microbial community can vary greatly over time and the difference between equivalent and total community size in neutral community assembly models.

Keywords: agent-based model, biofilm, drift, neutral assembly, community assembly, individual based model

1 Introduction

Both stochastic and deterministic assembly processes can shape biofilm communites.[1,2] Those processes, however, rarely act equally and the balance between them is determined by many conditions related to competition intensity. Such conditions include population size,[3,4] available space,[5] and resource availability.[6] Understanding how this balance shifts under differing conditions provides insights into biofilm-associated systems such as environmental bioreactors, healthcare, industrial production, and natural ecosystems.

Here, we attempt to quantify the balance between drift, a pure stochastic process,[1,3] and a more deterministic kinetic advantage. Under this balance, even if losing the 'drift lottery,' an individual's progeny may thrive if their maximum growth rate (μ_{max}) or half saturation constant (K_s) confers a selection advantage over their competitors.

Such quantification is challenging. Drift is an inherently random process and experimental manipulation of a random process, distinct from simply controlling for it, is difficult. Despite that difficulty, there have been some physical experiments in which drift is isolated as an experimental factor,[4,7,8] often requiring subtle statistical analyses or extremely precise experimental work.

An alternative approach, used here, is to perform the experiments *in silico* where drift may be directly manipulated via random number generation. We used an agent-based model (NUFEB)[9,10] to simulate spatially competing bacteria under low nutrient conditions. The bacteria were identical and evenly spaced, differentiated only by random growth directions and biomass allocations during division. Drift was therefore the only selection process and was controlled by the seed value initializing the random number generator.

Our goal was to determine the degree to which a deterministic factor (here, Monod kinetics) must improve to overcome drift-driven failure so subsequent simulations using identical seeds were run. The difference was that the 'biggest loser', the lineage with the lowest relative abundance, was

assigned different kinetics. This approach allowed us to relate quantifiable kinetic changes to the likelihood that the failing lineage would overcome drift-driven failure and thrive. We also determined how the required degree of change varied under differing crowding intensities (*e.g.,* closer spacing and increased initial population size).

We found that under purely stochastic conditions the losing lineage varied unpredictably between runs, showing the expected effects of drift. Further, altering kineticsdid enable losing lineages to overcome drift. For example, for an initial population of 9 cells evenly spaced 10 diameters apart either K_s or μ_{max} had to improve by at least 10-20% for a 50% chance of thriving. Crowding affected both the improvement needed for a 50% chance of thriving and the ranges over which both drift and fitness influenced success. The strong and sometimes non-linear interactions between terms could not be adequately reproduced using simple linear estimators but could be adequately expressed with a generalized additive model.

2 Methods

2.1 Agent Based Model

The agent-based model employed NUFEB (Newcastle University Frontiers in Engineering Biology),[9,10] which is based on the LAMMPS[9] molecular dynamics simulation framework and has successfully been used to model multi-species biofilms,[10] including development and detachment,[7] trade-offs in extracellular polymeric substance production,[11] and phototroph-heterotroph metabolic interactions.[12].

NUFEB is not lattice based, cells were positioned in three dimensions and had individual dynamic sizes. The directions in which cells divided and biomass allocations (40 to 60%) during division were randomly determined using a Park-Miller pseudorandom number generator and were the two factors contributing to drift.

The individually simulated bacterial cells physically interacted using realistic physics and grew according to Monod-style models described by Equation (1) where μ is the substrate-dependent growth rate (1/hr), μ_{max} is the maximum specific growth rate (1/hr), [S] is the concentration of the relevant substrate (kg/m^3), and K_s is the half-saturation constant for the substrate (kg/m^3). Additional descriptions of NUFEBs mechanics are detailed in previous publications.[9,10]

$$\boldsymbol{\mu = \mu_{max} \frac{[S]}{K_s + [S]}} \quad \textbf{(1)}$$

The simulation volume height (2x10^{-4} m) was defined to be in the Z-dimension, the bulk substrate concentration boundary condition at the top of the simulation volume was 1x10^{-4} kg/m^3 and the initial substrate concentration throughout the volume was set to the same value. The X and Y dimensions were equal and varied based on spacing and number of initial cells. Additionally, the X and Y boundaries were periodic, allowing biomass and substrates to wrap from one side of the simulation to the other.

2.1.1 Model Implementation Details

NUFEB simulates bacterial growth, physical interactions, and substrate diffusion and reactions within a cuboid volume. Bacterial growth is given as mass over time and determined by a summation of Monod-style rate equations and the change in mass is used to calculate the diameter of a spherical organism. When an individual grows beyond a user-defined threshold (here 1.36 microns), it divides into two organisms. The first cell receives 40-60% of the biomass (uniformly randomly selected) and the second cell receives the remainder. The three-dimensional direction of division relative to the centre of the initial cell is randomly chosen.Mechanically, the individuals are subjected to contact, adhesion, and fluid forces which are implemented as respective as spring and dashpot, spring, and simple one-way coupling physical models. A mechanical relaxation step is performed to address the mechanical in-equilibrium introduced by organism division. With respect to the crowding explored in this research, the result of mechanical relaxation is that a freshly cell which finds itself

98 ‘overlapping’ with existing biomass will be part of a ‘shoving’ match in which all relevant
99 individuals will be pushed into nearby empty space.

100 In this simulation a generic nutrient substrate is modelled and oxygen is non-limiting. The substrate
101 is modelled within the cuboid by solving a standard advection-diffusion-reaction equation. The
102 equation is discretized across and solved for small voxel subsections of the cuboid with a short
103 timestep.

104 The implementation used here does not differ from previous detailed explanations[10] employing thee
105 the ODD protocol (Overview, Design concepts, Details), which is a standard for agent-based model
106 description. Specifically, the underlying equations regarding growth, transport, and physical
107 interactions have not been modified and the interested reader is guided specifically to the supporting
108 information of reference 10 for an exhaustive, canonical description.

109 ## 2.2 Experimental Approach

110 The base experimental unit was an agent-based simulation initially seeded with identical bacterial
111 cells with starting diameters of 1×10^{-6} m, K_s of 3.5×10^{-5} kg/m^3, μ_{max} of 1 h^{-1}, and yield 0.61 kg
112 biomass per kg substrate consumed. The initial cells (total population 4, 9, or 16) were arranged
113 along evenly spaced (2.5, 5, or 10 cell diameters) *MxM* points at the base of the simulation volume.
114 Bacteria were allowed to grow and compete until 20% of the simulation volume consisted of
115 heterotrophic biomass.

116 Each combination of populations sizes and spacings was run 120 times using different seed values to
117 initialize the random number generator and the ‘biggest loser’ from each run was identified (see 2.3).
118 Those simulations were then run again, but with the failed lineage given altered kinetic values (see
119 2.4). The results of the runs were used to determine how the altered kinetics contributed to the
120 probability of transitioning from drift-driven failure to a thriving state (see 2.5) under various
121 crowding intensities.

All combinations of the factor levels listed in **Table 1** (1089 combinations) were simulated for each of the 120 seeds, resulting in a total of 130680 runs. Each run required between 2 to 36 hours to complete, so the simulations were carried out on a high-performance computing cluster (see 2.6).

Table 1: Experimental factors and levels

Factor	Values										
Spacing (cell diameters)	2.5	5	10								
Initial Population Size	4	9	16								
% Change in K_s	-50	-40	-30	-20	-10	0	10	20	30	40	50
% Change in μ_{max}	-50	-40	-30	-20	-10	0	10	20	30	40	50

2.3 Determining Failed Lineages

For a system initialized with N bacterial lineages, the total biomass X_t is the sum of the biomass for each lineage X_i, as expressed by equation(2).

$$\boldsymbol{X_t = \sum_{i}^{N} X_i} \tag{2}$$

In a system where each initial cell is identical, with no competition, and with no random effects, all X_i are expected to be equal, thus the expected relevant abundance of any lineage (X_E) is given as:

$$\boldsymbol{X_E = X_T / N} \tag{3}$$

In the first round of simulations, all initial cells were identical and evenly spaced, but cell division directions and biomass allocations during division were determined randomly. As a result, the distribution biomass for any lineage at any particular time was often not equal to the expected relevant abundance, $X_i \neq X_E$. In practice, there were often one or two lineages which strongly dominated with $X_i \gg X_E$, one or two lineages which became vanishingly small with $X_i \ll X_E$ (the 'biggest losers'), and the rest persisted at some noticeable abundance that was however below X_E. Moreover, the outcomes appeared to be determined early in the simulation, especially for the best and worst performing lineages. (Supporting Information Figure S1, Table S1, and Video SV1). We have defined three classifications of lineage survival based on the difference between X_E and X_i: *languishing* ($X_i < 0.3\ X_E$), *thriving* ($X_i > 0.9 X_E$), and *barely surviving* ($0.3\ X_E \leq X_i \leq 0.9 X_E$). The

threshold for thriving is lower than XE to accommodate situations where single lineage massively dominated (*e.g.,* $Xi > 0.6$) leading to lineages which were clearly otherwise doing well but with low relative abundance.

2.4 Kinetic Alteration for Potential Selective Advantange

The worst-performing bacterial lineages from each of the initial homogenous runs were modified by altering their individual maximum specific growth rate (μ_{max}) and/or their half-saturation constant (K_s) (**Figure 1**), potentially giving them a competitive advantage. The altered values were selected as described in **Table 1**. We acknowledge that not all combinations of μ_{max} and K_s were advantageous and that μ_{max} and K_s are often strongly correlated; here our goal was to thoroughly explore the parameter space.

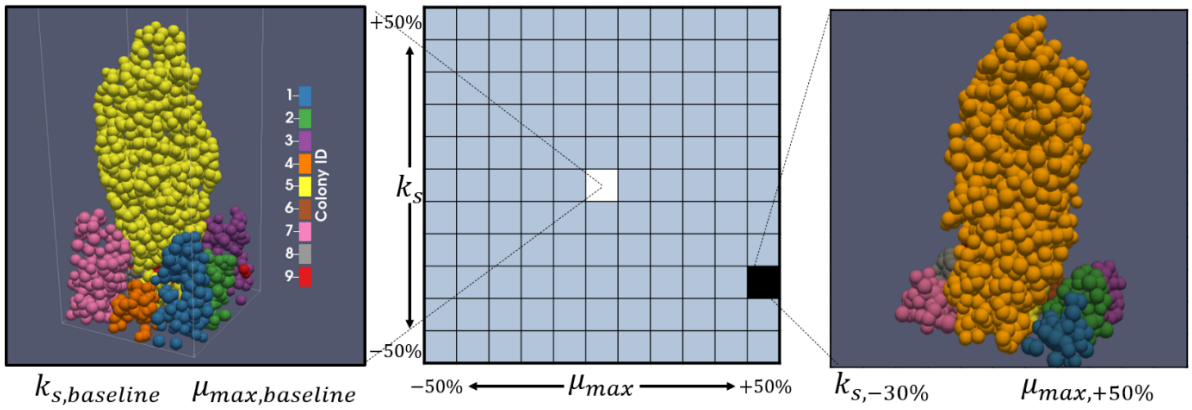

Figure 1: Illustration of a parameter sweep. Under baseline conditions when all bacteria are identical (left hand side), colony 4 was the worst performing lineage. When colony 4 was given a potental selective advantage (right hand side) via reduced K_S and increased μ_{max}, colony 4 transitioned to thriving. This result along with all other parameter combinations across 120 random seeds was used to estimate p_{thrive}, the probability that the worst-performing colony would transition to thriving under given altered kinetics. The trend of upward growth by the bacteria is due to substrate concentration gradients and is characteristic of growth under low-nutrient conditions.[10]

A two-dimensional parameter space was chosen because both μ_{max} and K_s met two desirable criteria. First, they directly associate growth and substrate concentration. Second, they are major parameters used when designing bioreactors, calibrating associated models, and when discussing kinetic control of microbial populations within reactors. A composite ratio of the parameters did not appear usable due to a lack of symmetry in results (*e.g.,* across the upper-left to lower-right diagonals in Figure 4). The disadvantage of such an approach is the large computational cost. For similar work where those

criteria do not apply, a one-dimensional parameter space is suggested. Ideally, this single-parameter would be part of the underlying biological model (such as yield), rather than a generic multiplicative 'selective advantage' variable .

2.5 Probability Map Generation

The kinetic parameter sweeps were used to generate tables for each combination of factors which listed the final relative biomass of each bacterial lineage, that lineage's status as the 'biggest loser', and the lineage's success under each run. Within each combination population size and spacing, the percentage of failing lineages which transitioned to thriving during the parameter sweep was recorded across all seeds. $_{s.}$ Those percentages represent the probabilities that the selective advantage (if any) conferred by altered kinetics would outweigh drift-driven failure under the given conditions.

2.6 Simulation Management

Simulations were run and their results tabulated on the Newcastle University Rocket High Performance Computing environment and managed using Snakemake[13,14] workflows populating a SLURM[15] queue. Each simulation was run on a single core, with multiple hundreds of simulations run in parallel. Job submissions encompassed all kinetic parameter sweeps for each combination of other parameters, *e.g.,* a single batch submission would consist of all combinations of μ_{max} and K_s for 4 bacteria, spaced 5 diameters apart.

2.7 Data Analysis

Simulation results were saved as tabular comma separated value (CSV) text files and aggregated using BASH[16] (v. 4.2) shell and Python[17] (v. 3.8) scripts which included the NumPy[18] and pandas[19] libraries. Further processing of the data was performed off the cluster and used R[20] (v. 4.2) scripts incorporating various Tidyverse[21] and other supporting packages.[22–43]

2.7.1 Parameters Quantifying the Balance Between Drift and Selection

Each probability map was conceptually analogous to a cliffside; a continuous sharp probability threshold gradient separated by two flat regions of either 100% lineage success or failure (**Figure 2** A). We wished to quantify the midpoint and steepness of the gradient along lines of constant K_s for each crowding condition. A cross-section of the probabilities along μ_{max} for any constant K_s produces a sigmoid-shaped profile (**Figure 2** B). The profiles were fit to a logistic function of μ_{max} with a maximum value of 1 given by equation (**4**), where p_{thrive} is the probability of transitioning to a thriving colony, k is a parameter affecting the steepness of the curve, and μ_{50} is the μ_{max} value at which there is a 50% probability of thriving.

$$p_{thrive} = \frac{1}{1 + e^{-k*(\mu_{50} - \mu_{max})}} \tag{4}$$

The relevant k and μ_{50} parameters from each fit were recorded. We also determined the domains of μ_{max} values associated with the p_{thrive} ranges covering either a 2.5-97.5% or 16-84% chance of thriving. These domains, respectively named $spread_{95}$ and $spread_{68}$ quantified the regions over which both drift and selection influenced success.

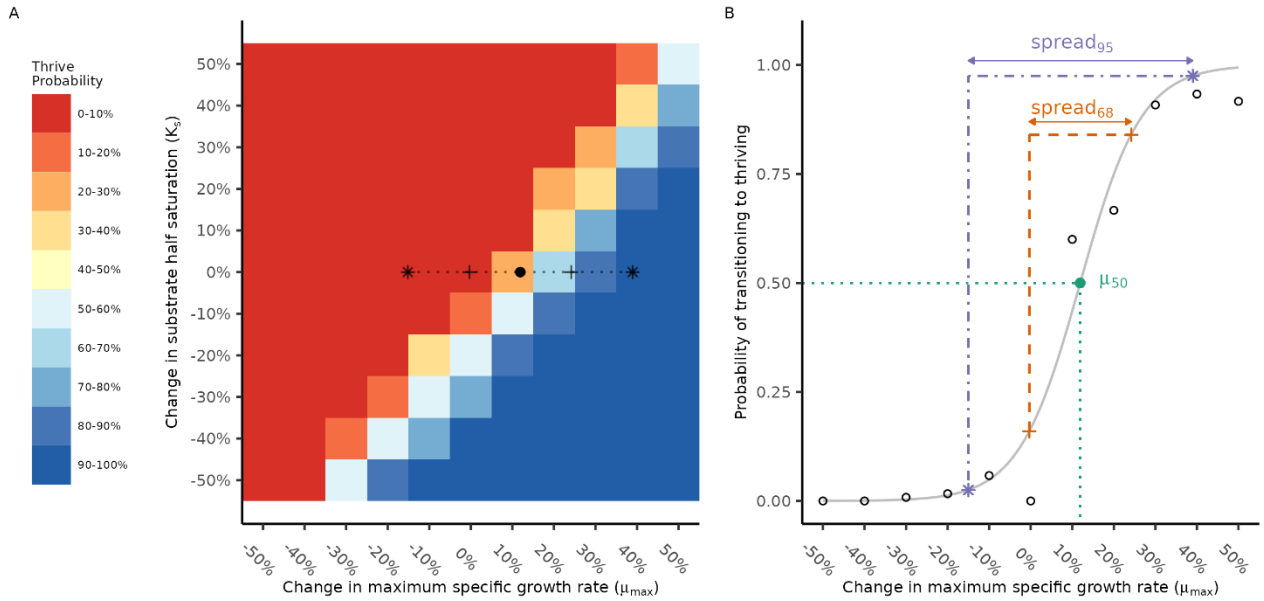

Figure 2: Illustration of how the μ_{50} and *spread* parameters were calculated. In this example, the probability map corresponding to 4 initial organisms placed 5 diameters apart is shown (A), and the dashed line is drawn along a line of constant K_s. The full length of the line denotes the $spread_{95}$ region, the portion between crosses denotes $spread_{68}$, and the solid point represents the μ_{50} mark. When the p_{thrive} values are plotted as a function of μ_{max} along the line of constant K_s, (B) it is apparent that a logistic function (grey solid line)

may be fitted to the points (black rings). The fitted function was used to estimate both the value of μ corresponding to μ_{50} and the widths of the *spread* regions. This analysis was repeated for all crowding conditions along all lines of constant K_s.

The results of all sigmoid fits are shown in Supporting Information Figures S2-S10.

2.7.2 Analysing Balance Parameters

Within each crowding scenario, the extracted parameters were analysed using simple linear regression models of the parameters as functions of K_s. The effect of crowding pressure (spacing and total population) was then analysed by comparing the results of the fits between scenarios.

We note that although the linear fits for a 2nd order polynomial on μ_{50} generally resulted in marginally improved R^2 scores and removed parabolic patterns from the residuals, the simple linear regressions were still excellent and more interpretable; care should be taken if extending this work to larger ranges of kinetic values.

2.7.3 Modelling the Effect of Competitive Pressure and Altered Kinetics

We wished to determine if a model based on the simulation results could accurately reproduce the transition probabilities for each crowding scenario. The ultimate goal of these models was not prediction, but to provide a descriptive framework[44] showing which factors, interactions, and potential non-linearities were important. Variations on both multiple linear regression models (MLR) and Generalized Additive Models (GAMs)[45] were fitted to either the log-likelihood of p_{thrive} (for MLRs) or directly to p_{thrive} (GAMs).

In both cases, backward step selection from factorial models incorporating up to three-way interactions was performed to select the final model. Non-significant ($p > 0.05$) terms were iteratively removed from the model starting with the highest order interactions. Main effects were retained even if non-significant when they were part of a significant interaction term.

The final models were selected based on R^2 and Akaike Information Criterion (AIC) values as well as interpretability. The potential models and the associated fit criteria are included in Supporting Information Tables S2-S5.

3 Results

3.1 Drift Occurred When All Cells Were Identical

A foundational assumption of this approach is that even in a system with equally spaced, identical microbes, random growth will lead to drift. We tested this assumption for crowding scenarios where all microbes had identical base K_s and μ_{max} parameters by determining the number of times each lineage was the 'biggest loser' over 120 simulations (**Figure 3**) and, similar to testing m dice for fairness, applied a Chi-Square test (α=0.05/m) where m is a Bonferroni correction for multiple testing (m=9 at 3x3 initial spacings and population sizes). Each initial site was statistically as likely as any other to be the biggest loser (Supporting Information Table S6).

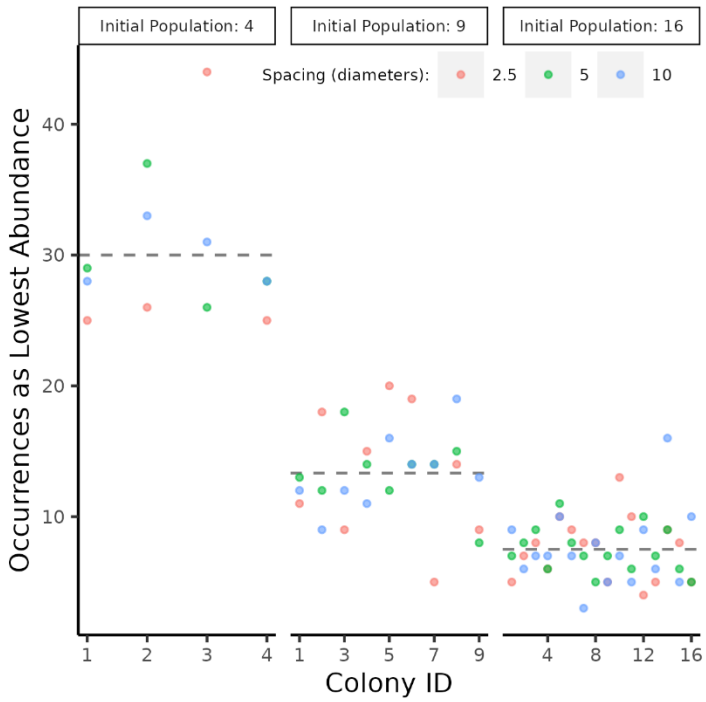

Figure 3: The number of times each colony was the least successful performer during all 120 runs of the baseline simulation where all bacteria were identical. Dashed grey lines indicate the expected value. Points are colored based on spacings between initial sites. For each set of initial populations, no colony appeared biased away from the expected number of failures.

Additionally, the relative proportion of lineages which languished, survived, or thrived for each set of crowding conditions was determined. Simulations, on average, had between one and two thriving lineages, with the rest languishing (65-75% for 4 initial sites, 80-88% others), and a few (0-5%) which did not thrive but grew to non-negligible abundance (Supporting Information Table S1). When

4 organisms were initially present, only languishing and thriving lineages existed, there was otherwise no clear trend between these ratios and either the number or spacing of initial bacteria.

3.2 The Least Successful Lineages Could Overcome Drift with Altered Kinetics

As expected, altering the kinetics of an organism could give it a chance to overcome drift-driven failure (**Figure 4**).

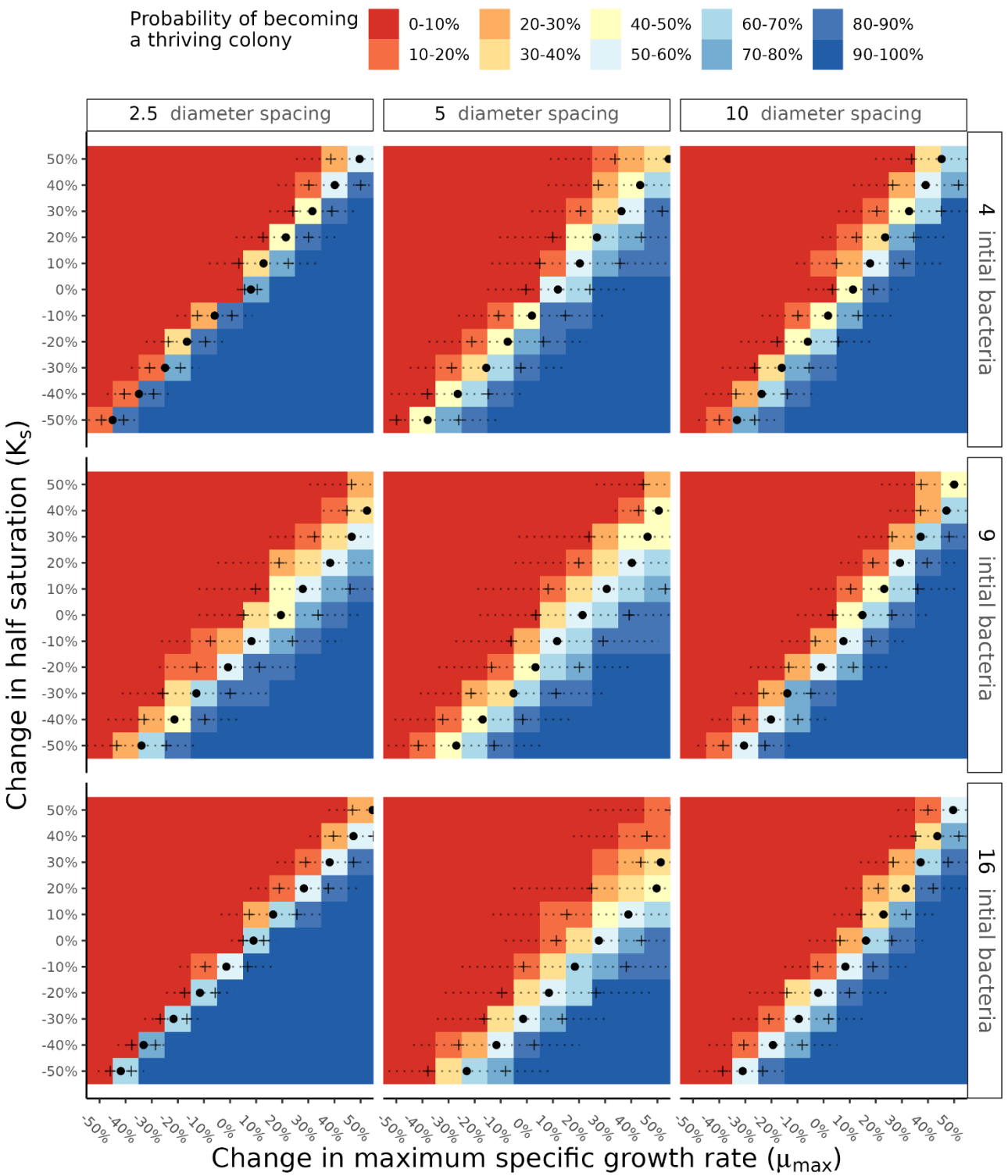

Figure 4: Changing the μ_{max} and K_s of the least successful lineage was associated with a probability of transitioning to a thriving status. Solid dots represent μ_{50}, the percent change in μ_{max} at a given K_s associated with 50-50 odds of thriving. Dashed lines show the range of μ_{max} corresponding to a p_{thrive} of 2.5 to 97.5 (*i.e.*, $spread_{95}$). Crosses indicate the analagous $spread_{68}$ region.

The increases in μ_{max} corresponding to the least successful lineage having a 50% chance to become thriving, which we denote as μ_{50}, are represented by the dark circles in **Figure 4**. At the baseline K_s a typical μ_{50} is in the range of 10-30%, with the exact value affected by initial spacing and population

size (*i.e.,* crowding). Decreasing $K_{s,}$ as expected, reduces μ_{50} – even to the point where so long as substrate uptake affinities are 'good enough', the initially failing organism may have excellent odds despite having a μ_{max} notably lower than its peers. The overall effect, for a given crowding condition, is a semi-linear 'cliff 'of μ_{50} values where μ_{50} changes inversely with K_s. Qualitatively speaking, the location of that 'cliff' was shifted to the right (higher μ_{50}) when crowding was increased via initial population size or when comparing between the extremes of spacing.

Areas where the probability of thriving is neither 0 nor 1, are, by definition, areas where drift and selection both influence success. The widths of these areas are denoted as *spread* and are indicated by the dotted horizontal lines and crosses in **Figure 4**. The full length of the line denotes the $spread_{95}$ area, which is the range of μ_{max} for a given K_S which corresponds to a 2.5% to 97.5% chance of thriving. The crosses represent a similar range, $spread_{68}$, which corresponds to a 16% to 85% chance of thriving.

Because the μ_{50} values are also the centre point of the *spread* regions, spread shifted in the same manner as μ_{50}. However, the actual magnitudes of *spread* did not necessarily follow the same patterns. First, there was no guaranteed symmetry about $K_{s.}$ For example, for 9 initial organisms separated by 5 diameters, the $spread_{95}$ for K_s of -30% and 30% are visibly different (**Figure 4**, row 2 column 2). Though the asymmetry varied between crowding conditions, it generally manifested as spread widening with increasing K_s. Second, there was no clear monotonic trend with spread values corresponding to crowding. A spacing of 5 diameters appeared to produce the widest spreads, *ceteris paribus*. Further, there was no clear rule determining which of the two spacing extremes would have a larger *spread.* For example, with 4 initial bacteria a spacing of 10 diameters resulted in larger spreads than in 2.5 diameters, but the opposite occurred with 16 initial bacteria.

3.3 Quantitative Effect of Crowding on μ_{50} and *spread*

The qualitative effects of crowding described in the previous section were quantified via simple linear regression as described in section 2.7.2.

For any given crowding condition μ_{50}, the relative change of μ_{max} at which the worst performing lineage had a 50% chance to transition towards thriving, was essentially linear with respect to K_S and the correlation coefficient was uniformly high (**Figure 5**). The slopes of these relationships indicate the change in μ_{50} required to compensate for a change in K_S. At the tightest spacing, μ_{50} had to change the most, with a ratio of essentially 1:1 and a slight monotonic increase corresponding to initial population size. As initial spacings widened, the ratio almost always decreased for any initial population size. Across initial population sizes, the ratio for 5 and 10 diameter spacings appeared to follow a general trend of increasing, but this was not monotonic.

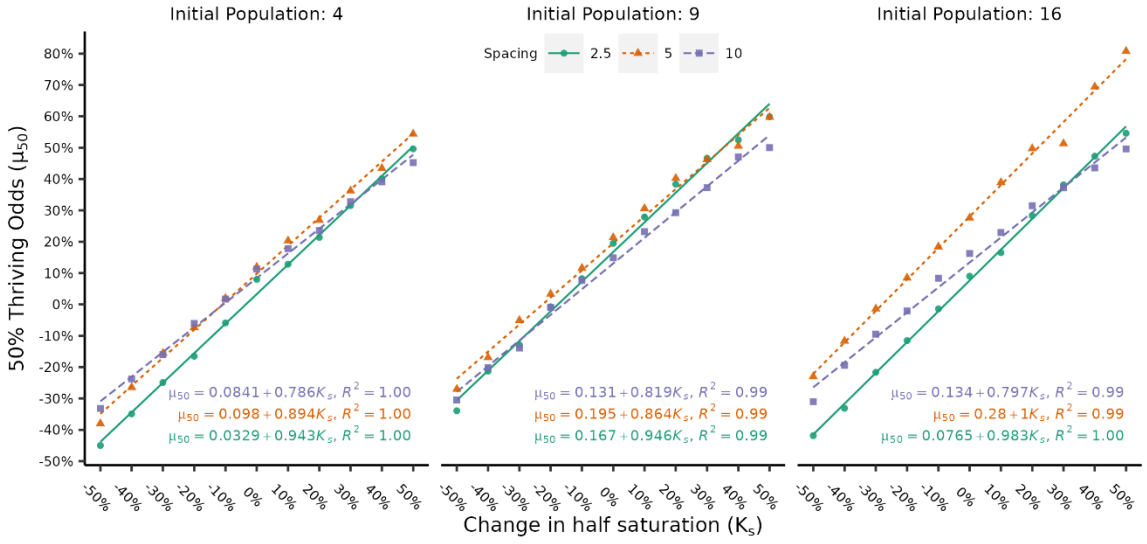

Figure 5: Under each crowding condition, μ_{50} changed linearly with K_S. Large initial population sizes increased the differences between spacings, moderate spacings generally required the largest absolute μ_{50}, but the tightest spacings required the largest change μ_{50} in per unit change in K_S.

The absolute value of μ_{50} was strongly affected by differences between the fitted intercepts. For example, a 2.5 diameter spacing under an initial population size of 16 had a high slope (0.983) but also the lowest required μ_{50} of all spacings under the same conditions until a 30% change in K_S. The practical difference between spacing was largest at high initial population size, indicating a potential interaction between these factors.

Unlike μ_{50}, the range over which both drift and selection effects influenced success, $spread_{95}$ did not have a simple linear relationship with K_S, with many poor R^2 values, residual patterns, and high leverage datapoints (**Figure 6**). There was also no clear, consistent relationship applicable across factors. In general, linear fits became worse with increasing population size which appeared to produce higher variance and generated more high-leverage points, especially at separation distances of 5 diameters. These issues were largely the same when the analysis was repeated for $spread_{68}$ (Supporting Information Figure S13). There is little to concretely say except that the *spread* was most often widest at moderate spacings, generally increased with K_s, and had a noisy, complicated relationship with initial population size and spacing.

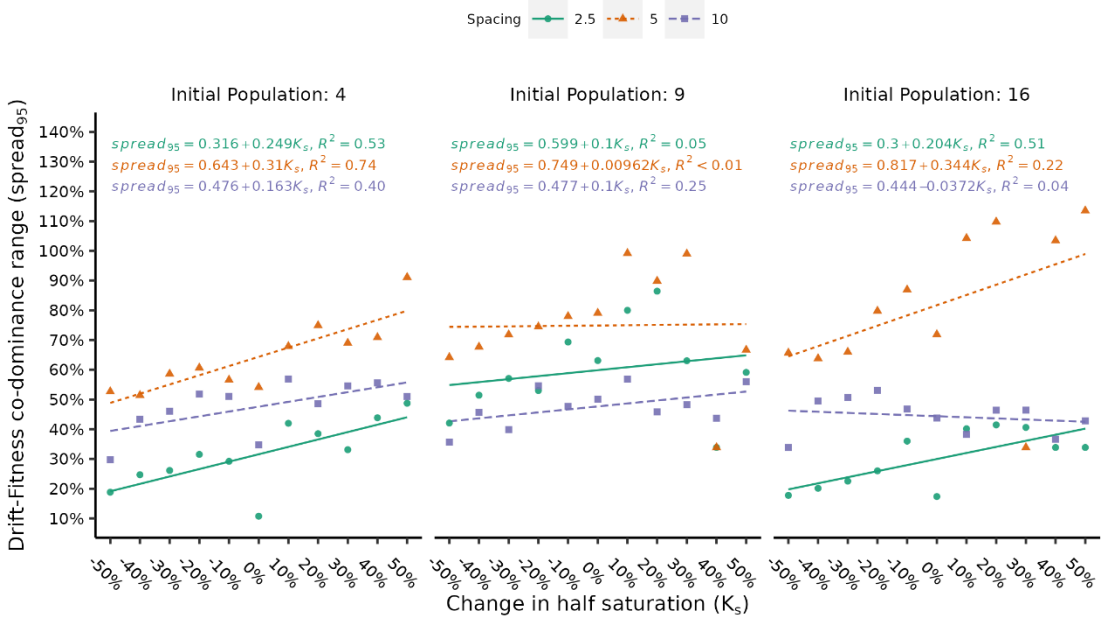

Figure 6: Under each crowding condition, $spread_{95}$ changed with K_S. Insofar as trends were present, moderate spacing produced the widest $spread_{95}$ and the differences between spacings increased with population size.

3.4 Description via Multiple Linear Regression and Generalized Additive Models

The simulation results were modelled using both multiple linear regression (MLR) and a generalized additive model(GAM) respectively described by equations (**5**) and (6) where: p_{thrive} is the probability of transitioning to a thriving status, μ_p and K_p are the respective percent changes from the baseline μ_{max} and K_s, N_0 is the initial population size, s_i is the initial spacing (in diameters) between organisms, and ε is a small pseudo-probability ($1x10^{-6}$) added to avoid division by 0 and issues with

log transformation. For linear terms in equations (**5**) and (6), β_i denotes the fitted coefficient for term i with i=0 representing the intercept. Terms to which GAM smoothing was applied are represented by s(...) in equation (6) with interactions between a smoothed variable x and linear variable y denoted as s*(x,* by *y)*. Significant terms ($p < 0.05$) are highlighted in bold. The associated coefficients, significance values, and other relevant fitting information are included in Supporting Information Tables S2-S5.

$$\log\left(\frac{\boldsymbol{p_{thrive}}}{\boldsymbol{1-p_{thrive}+\varepsilon}}+\boldsymbol{\varepsilon}\right)=\boldsymbol{\beta_0+\beta_1\mu_p+\beta_2K_p+\beta_3N_0+\beta_4s_i+\beta_5\mu_ps_i+\beta_6K_ps_i} \quad (5)$$

$$p_{thrive}=\beta_0+s(\mu_p)+s(K_p)+\boldsymbol{s}(N_0)+s(\boldsymbol{s_i})+s(\boldsymbol{\mu_pK_p})+s(\boldsymbol{\mu_ps_i})+s(\boldsymbol{K_ps_i})+s(\boldsymbol{N_0},\text{by }\boldsymbol{s_i})+s(\boldsymbol{\mu_pK_pN_0})+s(\boldsymbol{\mu_pK_ps_i}) \quad (6)$$

The MLR model captured the general behaviour of the shift in the boundary between low and high thriving probabilities but did not adequately reproduce changes in *spread* (**Figure 7** A *vs.* C). The overall root-mean-squared error (RMSE) of the model was 0.125. While most predicted probabilities differed from the simulation by no more than ±0.1, some predictions were subject to large error (**Figure 7** A, D, F and Supporting Information Figures S11 and S14-S15). The largest errors unsurprisingly appear closest to the boundary between low and high p_{thrive} regions with the MLR model over-optimistic at the extremes of spacing and lower initial population size. Conversely, the model tended towards overly pessimistic at moderate spacing.

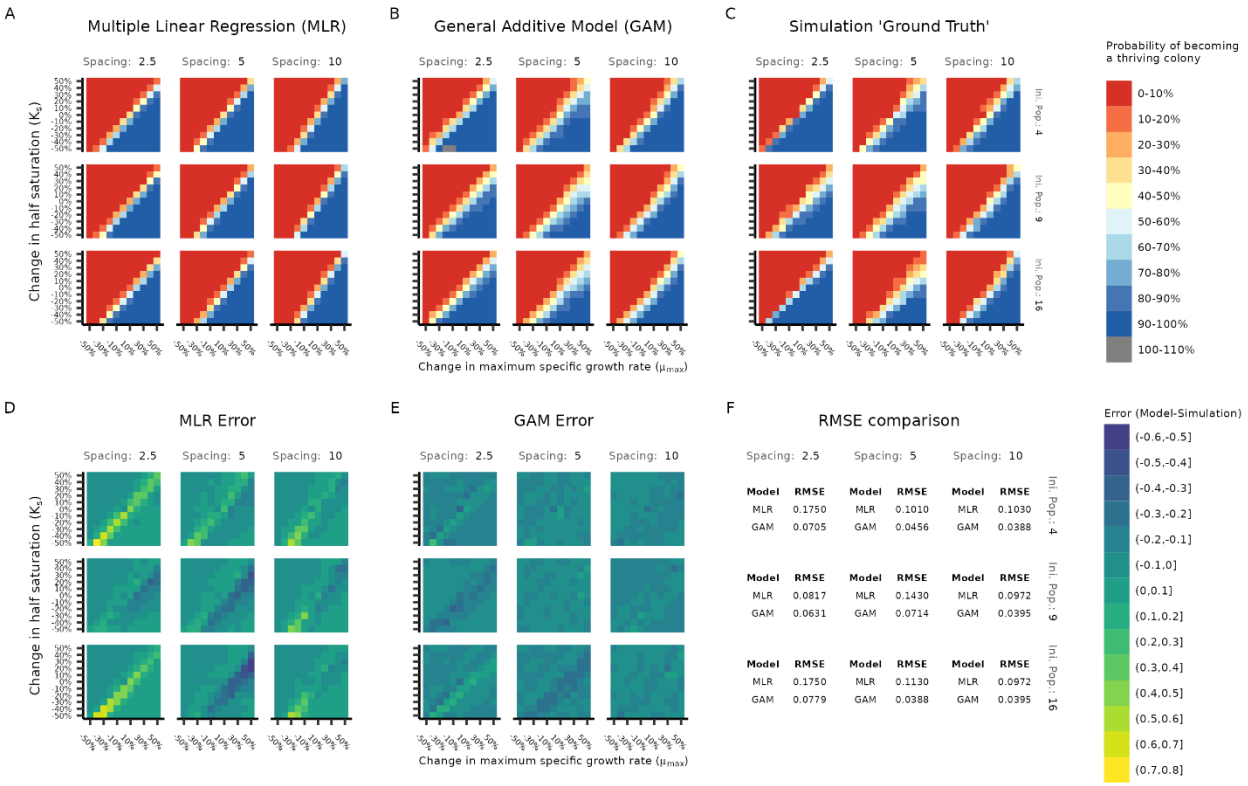

Figure 7: Predictions of MLR model (A) and GAM (B). Simulation results in (C) are presented for ease of comparison. The model errors for the MLR (D) and GAM (E) are presented visually as well as quantified per-crowding condition in (F). The GAM outperformed the MLR, which particularly failed to capture *spread*, was overly optimistic at spacing extremes, and pessimistic at moderate spacing. The small region of greater than 100% odds occured because the GAM was not constrained to predicting values in the range of [0,1]. Larger individual plots of panels A, B, D, and E are available in Supporting Information figures S14-S17.

In comparison to the MLR model, the GAM not only captured the general boundary shift but also the changes in *spread* (**Figure 7** B *vs.* C in contrast to A *vs.* C). The overall RMSE of the GAM was 0.0563, or somewhat better than half the RMSE of the MLR model. As with the MLR model, most predicted probabilities differed from the simulation by no more than ±0.1. Unlike the MLR model, there were fewer exceptionally large errors and those which did occur were of smaller magnitude (**Figure 7** B, E, F and Figures S12 and S16-S17). The GAM followed the same trends in over- and under-prediction as the MLR.

4 Discussion

4.1 Crowding Affects the Balance Between Drift and Selection

The two parameters describing the balance between drift and selection, μ_{50} and *spread*, were both affected as crowding became more intense due to either decreased initial spacing or increased initial

population size. It was originally expected that as crowding intensity increased, greater selective advantages would be required (μ_{50}) along with a decrease in the range of values over which both drift and selection influenced success (*spread*). That was not the case.

Instead, the largest *spread* values predominately occurred at moderate (5 diameter) initial spacing. We suggest the cause is physical competition for space, specifically the practical significance of single 'bad' random choices in division direction and biomass allocation. When bunched tightly together, competition for space is intense and even a few poor random events can consign a lineage to languishing despite a moderate growth advantage. At the other extreme, spatial competition is lessened sufficiently that a few missteps do not guarantee ruin, allowing a lineage to take the full benefit of any growth advantage. Meanwhile, at moderate spacing, immediate neighbours are close enough so that poor random events are harmful but not necessarily disastrous and, at the same time, growth advantages are somewhat hindered, but still helpful. Remembering that spread quantifies the region where both fitness and drift influenced success, it then makes sense that we observed the largest spread values at moderate spacing.

The 50-50 odds point, μ_{50}, was also slightly larger at moderate spacings, although not consistently and the effect size was not practically different except at large population sizes. The underlying basis for why is not entirely clear, numerically it was due to the consistently larger intercept (Figure 5). The trend of the slopes is, however, more easily explained and we attribute it to competition for substrate. For any initial population size, smaller spacings resulted in higher slopes. In other words, to maintain the 50-50 odds when K_s was poor, μ_{50} had to change more at closer spacing. This makes intuitive sense – closer spacings result in lower local substrate concentrations, and any deficit to K_s is more deleterious to selection.

Increased initial population sizes had more straightforward, secondary, effects on μ_{50} and K_s. As the initial population size increased, the differences between spacings became more pronounced, but the

general trends remained unchanged. In other words, more competitors are problematic, especially as it relates to diffusible substrate, but the major influence on success is competition for space between immediate neighbours.

4.2 Interactions Between Factors Incorporating Non-Linear Effects are Important

In the MLR a main-effects only model (RMSE 0.125, R^2 of 0.820) performed essentially identically to the MLR model with interactions (RMSE 0.127, and R^2 of 0.820), however neither adequately reproduced simulation results. Both were especially poor at representing the regions where fitness and drift influenced success. A GAM which incorporated only main effects using non-linear smoothing quantitatively performed slightly worse than either MLR main-effects model (RMSE 0.197 and R^2 of 78.1), but drastically and uniformly overpredicted spread. Only when both interactions and smoothing were incorporated did a model adequately reproduce the simulation results (**Figure 7** and Supporting Information Figure S17). It is visually apparent in the simulation results and quantified in the fitting results (Supporting Information Table S4-5) that interactions are important, particularly those involving spacing. Further, the non-linearity of the interactions (measured as the departure of the term's extended degrees of freedom from a value of 1), is particularly high for any interaction incorporating both μ_p and K_p and less so but still notably for interactions incorporating spacing (Supporting Information Table S5).

4.3 Limitations and Extensions

The simulated conditions were deliberately chosen to isolate the effect of drift. While this made the work tractable, a system wherein every organism is completely identical, starts growing at the same time, and is initially evenly spaced on a grid does not frequently occur in nature. Although we believe the general themes uncovered translate to real ecological systems, the exact quantification does not and is not mean to apply to all situations. Future work should focus on stochastically placed (in time and space) populations with natural variability in Monod parameters.

Extending the work so that the simulated community reflects a more natural distribution would also enable validation of the model, as, despite promising advances,[46] it is currently infeasible to exactly place essentially identical bacteria at the resolution required.

Additional parameters affecting drift and selection should also be evaluated – especially the influence of nutrient-rich conditions[47] and how a change to yield, rather than growth rate, alters success.[48] Adding these factors requires however overcoming the curse of dimensionality, the current simulations took over 1 year of real-world time and 175 years' worth of CPU time. Given the large areas where 'nothing interesting' happens, designing further experiments to incorporate adaptive sampling[49] is a promising solution. Further, adaptive sampling would enable, at the same computational cost, exploring a larger range of μ_{max} and K_S variation (which may vary by orders of magnitude in real-world conditions[50]) and at a greater degree of resolution than 10% changes in the region where the probabilities rapidly change.

5 Conclusion And Relevance to Real World Systems

It is apparent that during biofilm formation in low nutrient conditions, drift strongly determines which organisms thrive and which organisms fail, so long as they have similar growth rates and substrate affinities. Even when those parameters differ between individuals by ±50%, there are still large regions where a selective advantage does not guarantee overcoming drift-driven failure.

In fact, we observed the lineage fates were determined very early in the simulations and for these systems 'well-begun is half done'. We speculate that this may be a piece to the puzzle explaining the apparent contradiction between actual and effective community size in neutral modelling[4] – the bacteria are not in competition with the full steady-state community but only the immediate smaller, community near the beginning of biofilm growth. However, the conditions studied here violate the steady state assumption of that work, so a more careful analysis is warranted.

The conditions we have described are not dissimilar from those within an aerated portion of a wastewater treatment plant, where tightly packed bacterial aggregates are suspended in a bulk liquid and where substrate concentrations are often quite low, especially during operation as a completely mixed stirred reactor (albeit somewhat higher than simulated here). Further, these bacteria are recirculated through the system and relatively well-adapted to domestic wastewater, thus already selected for similarity. Based on the results presented here, we would expect to see a system in which there is a high degree of random turnover in organism identity, but relatively stable functional and biological activity, which is exactly what has been observed in wastewater treatment plants.[51,52]

6 Acknowledgements

We wish to acknowledge the United States National Science Foundation Directorate for Biological Sciences for funding via the Postdoctoral Research Fellowships in Biology (NSF PRFB Award # 2007151) and the Newcastle University Rocket HPC computing cluster. We also thank Tom Curtis for his valuable feedback and encouragement, Denis Taniguchi and Bowen Li for their help in understanding the NUFEB software, and countless helpful comments from many interested parties during poster sessions and talks.

7 Competing Interests

The author has no competing interests.

8 Data Availability Statement

The data analysis code, data from the simulations, and exact NUFEB variant are respectively located in the following repositories:

Analysis: https://github.com/joeweaver/agent_based_biofilm_drift/
Data: https://osf.io/fch3z/
NUFEB variant: https://github.com/nufeb/NUFEB-dev/tree/compute_vol_group

9 References

(1) Vellend, M. Conceptual Synthesis in Community Ecology. *The Quarterly Review of Biology* **2010**, *85* (2), 183–206. https://doi.org/10.1086/652373.

(2) Battin, T. J.; Sloan, W. T.; Kjelleberg, S.; Daims, H.; Head, I. M.; Curtis, T. P.; Eberl, L. Microbial Landscapes: New Paths to Biofilm Research. *Nat Rev Microbiol* **2007**, *5* (1), 76–81. https://doi.org/10.1038/nrmicro1556.

(3) Nemergut, D. R.; Schmidt, S. K.; Fukami, T.; O'Neill, S. P.; Bilinski, T. M.; Stanish, L. F.; Knelman, J. E.; Darcy, J. L.; Lynch, R. C.; Wickey, P.; Ferrenberg, S. Patterns and Processes of Microbial Community Assembly. *Microbiol Mol Biol Rev* **2013**, *77* (3), 342–356. https://doi.org/10.1128/MMBR.00051-12.

(4) Sloan, W. T.; Nnaji, C. F.; Lunn, M.; Curtis, T. P.; Colloms, S. D.; Couto, J. M.; Pinto, A. J.; Connelly, S.; Rosser, S. J. Drift Dynamics in Microbial Communities and the Effective Community Size. *ENVIRONMENTAL MICROBIOLOGY* **2021**, *23* (5), 2473–2483. https://doi.org/10.1111/1462-2920.15453.

(5) Zhou, J.; Liu, W.; Deng, Y.; Jiang, Y.-H.; Xue, K.; He, Z.; Van Nostrand, J. D.; Wu, L.; Yang, Y.; Wang, A. Stochastic Assembly Leads to Alternative Communities with Distinct Functions in a Bioreactor Microbial Community. *MBIO* **2013**, *4* (2). https://doi.org/10.1128/mBio.00584-12.

(6) Dini-Andreote, F.; Stegen, J. C.; van Elsas, J. D.; Salles, J. F. Disentangling Mechanisms That Mediate the Balance between Stochastic and Deterministic Processes in Microbial Succession. *Proceedings of the National Academy of Sciences* **2015**, *112* (11), E1326–E1332. https://doi.org/10.1073/pnas.1414261112.

(7) Cira, N. J.; Pearce, M. T.; Quake, S. R. Neutral and Selective Dynamics in a Synthetic Microbial Community. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115* (42). https://doi.org/10.1073/pnas.1808118115.

(8) Fodelianakis, S.; Valenzuela-Cuevas, A.; Barozzi, A.; Daffonchio, D. Direct Quantification of Ecological Drift at the Population Level in Synthetic Bacterial Communities. *ISME JOURNAL* **2021**, *15* (1), 55–66. https://doi.org/10.1038/s41396-020-00754-4.

(9) Li, B.; Taniguchi, D.; Gedara, J. P.; Gogulancea, V.; Gonzalez-Cabaleiro, R.; Chen, J.; McGough, A. S.; Ofiteru, I. D.; Curtis, T. P.; Zuliani, P. NuFeb: A Massively Parallel Simulator for Individual-Based Modelling of Microbial Communities. *PLoS Computational Biology* **2019**, *15* (12), e1007125. https://doi.org/10.1371/journal.pcbi.1007125.

(10) Jayathilake, P. G.; Gupta, P.; Li, B.; Madsen, C.; Oyebamiji, O.; González-Cabaleiro, R.; Rushton, S.; Bridgens, B.; Swailes, D.; Allen, B.; McGough, A. S.; Zuliani, P.; Ofiteru, I. D.; Wilkinson, D.; Chen, J.; Curtis, T. A Mechanistic Individual-Based Model of Microbial Communities. *PLoS ONE* **2017**, *12* (8), e0181965. https://doi.org/10.1371/journal.pone.0181965.

(11) Jayathilake, P. G.; Jana, S.; Rushton, S.; Swailes, D.; Bridgens, B.; Curtis, T.; Chen, J. Extracellular Polymeric Substance Production and Aggregated Bacteria Colonization Influence the Competition of Microbes in Biofilms. *FRONTIERS IN MICROBIOLOGY* **2017**, *8*. https://doi.org/10.3389/fmicb.2017.01865.

(12) Sakkos, J. K.; Santos-Merino, M.; Kokarakis, E. J.; Li, B.; Fuentes-Cabrera, M.; Zuliani, P.; Ducat, D. C. Predicting Partner Fitness Based on Spatial Structuring in a Light-Driven Microbial Community. bioRxiv September 30, 2022, p 2022.09.28.510001. https://doi.org/10.1101/2022.09.28.510001.

(13) Mölder, F.; Jablonski, K. P.; Letcher, B.; Hall, M. B.; Tomkins-Tinch, C. H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S. O.; Kanitz, A.; Wilm, A.; Holtgrewe, M.; Rahmann, S.; Nahnsen, S.; Köster, J. Sustainable Data Analysis with Snakemake. F1000Research April 19, 2021. https://doi.org/10.12688/f1000research.29032.2.

(14) Köster, J.; Rahmann, S. Snakemake—a Scalable Bioinformatics Workflow Engine. *Bioinformatics* **2012**, *28* (19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480.
(15) Jette, M.; Yoo, A.; Grondona, M. SLURM: Simple Linux Utility for Resource Management; 2003. https://doi.org/10.1007/10968987_3.
(16) Ramey, C. Bash, the Bourne- Again Shell. In *Proceedings of The Romanian Open Systems Conference & Exhibition (ROSE 1994), The Romanian UNIX User's Group (GURU)*; 1994; pp 3–5.
(17) Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.
(18) Harris, C. R.; Millman, K. J.; Walt, S. J. van der; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; Kerkwijk, M. H. van; Brett, M.; Haldane, A.; Río, J. F. del; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.
(19) McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*; Walt, S. van der, Millman, J., Eds.; 2010; pp 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.
(20) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
(21) Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L. D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Pedersen, T. L.; Miller, E.; Bache, S. M.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D. P.; Spinu, V.; Takahashi, K.; Vaughan, D.; Wilke, C.; Woo, K.; Yutani, H. Welcome to the Tidyverse. *Journal of Open Source Software* **2019**, *4* (43), 1686. https://doi.org/10.21105/joss.01686.
(22) Aphalo, P. J. *Ggpmisc: Miscellaneous Extensions to "Ggplot2";* 2022.
(23) Aphalo, P. J. *Ggpp: Grammar Extensions to "Ggplot2";* 2022.
(24) Brand, T. van den. *Ggh4x: Hacks for "Ggplot2";* 2022.
(25) Coretta, S. *Tidymv: Tidy Model Visualisation for Generalised Additive Models*; 2022.
(26) Daróczi, G. *Logger: A Lightweight, Modern and Flexible Logging Utility*; 2021.
(27) Garnier; Simon; Ross; Noam; Rudis; Robert; Camargo; Pedro, A.; Sciaini; Marco; Scherer; Cédric. *Viridis - Colorblind-Friendly Color Maps for R*; 2021. https://doi.org/10.5281/zenodo.4679424.
(28) Henry, L.; Wickham, H. *Purrr: Functional Programming Tools*; 2020.
(29) Iannone, R.; Cheng, J.; Schloerke, B. *Gt: Easily Create Presentation-Ready Display Tables*; 2022.
(30) Kassambara, A. *Ggpubr: "ggplot2" Based Publication Ready Plots*; 2020.
(31) Meschiari, S. *Latex2exp: Use LaTeX Expressions in Plots*; 2022.
(32) Müller, K. *Here: A Simpler Way to Find Your Files*; 2020.
(33) Neuwirth, E. *RColorBrewer: ColorBrewer Palettes*; 2022.
(34) Pedersen, T. L. *Patchwork: The Composer of Plots*; 2020.
(35) Rij, J. van; Wieling, M.; Baayen, R. H.; Rijn, H. van. Itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs, 2022.
(36) Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer-Verlag New York, 2016.
(37) Wilke, C. O. *Cowplot: Streamlined Plot Theme and Plot Annotations for "Ggplot2";* 2020.
(38) Wilke, C. O. *Ggtext: Improved Text Rendering Support for "Ggplot2";* 2020.
(39) Wolen, A. R.; Hartgerink, C. H. J.; Hafen, R.; Richards, B. G.; Soderberg, C. K.; York, T. P. Osfr: An R Interface to the Open Science Framework. *Journal of Open Source Software* **2020**, *5* (46), 2071. https://doi.org/10.21105/joss.02071.

(40) Wood, S. N. Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2011**, *73* (1), 3–36.
(41) Pedersen, T. L.; Robinson, D. *Gganimate: A Grammar of Animated Graphics*; 2022.
(42) Pedersen, T. L. *Transformr: Polygon and Path Transformations*; 2022.
(43) Ooms, J. *Gifski: Highest Quality GIF Encoder*; 2022.
(44) Caswell, H. The Validation Problem. *Systems analysis and simulation in ecology* **1976**, *4*, 313–325.
(45) Hastie, T.; Tibshirani, R. Generalized Additive Models: Some Applications. *Journal of the American Statistical Association* **1987**, *82* (398), 371–386. https://doi.org/10.1080/01621459.1987.10478440.
(46) Krishna Kumar, R.; Meiller-Legrand, T. A.; Alcinesio, A.; Gonzalez, D.; Mavridou, D. A. I.; Meacock, O. J.; Smith, W. P. J.; Zhou, L.; Kim, W.; Pulcu, G. S.; Bayley, H.; Foster, K. R. Droplet Printing Reveals the Importance of Micron-Scale Structure for Bacterial Ecology. *Nat Commun* **2021**, *12* (1), 857. https://doi.org/10.1038/s41467-021-20996-w.
(47) Nadell, C. D.; Foster, K. R.; Xavier, J. B. Emergence of Spatial Structure in Cell Groups and the Evolution of Cooperation. *PLOS Computational Biology* **2010**, *6* (3), e1000716. https://doi.org/10.1371/journal.pcbi.1000716.
(48) Kreft, J.-U.; Bonhoeffer, S. 2005. The Evolution of Groups of Cooperating Bacteria and the Growth Rate versus Yield Trade-Off. *Microbiology 151* (3), 637–641. https://doi.org/10.1099/mic.0.27415-0.
(49) Westermann, P. W.; Evins, R. Adaptive Sampling For Building Simulation Surrogate Model Derivation Using The LOLA-Voronoi Algorithm; Rome, Italy; pp 1559–1563. https://doi.org/10.26868/25222708.2019.211232.
(50) Cox, C. D. Statistical Distributions of Uncertainty and Variability in Activated Sludge Model Parameters. *Water Environ Res* **2004**, *76* (7), 2672–2685. https://doi.org/10.1002/j.1554-7531.2004.tb00229.x.
(51) Weaver, J. E. *From Floc to Reactor Scales: A Multi-Faceted Investigation Integrating Microbial Ecological Experiments and Computational Modeling to Understand Aerobic Wastewater Systems*; North Carolina State University, 2021.
(52) Wang, X.; Wen, X.; Criddle, C.; Yan, H.; Zhang, Y.; Ding, K. Bacterial Community Dynamics in Two Full-Scale Wastewater Treatment Systems with Functional Stability. *Journal of Applied Microbiology* **2010**, *109* (4), 1218–1226. https://doi.org/10.1111/j.1365-2672.2010.04742.x.