Discovering data-driven microbial growth models with symbolic regression

T. Anthony Sun[1], Dovydas Kičiatovas[1], Inga-Katariina Aapalampi[2], Teemu Kuosmanen[1], Teppo Hiltunen[2], and Ville Mustonen[1]

[1]Organismal and Evolutionary Biology Research Programme, Department of Computer Science, University of Helsinki, Helsinki, Finland.

[2]Department of Biology, University of Turku, Turku, Finland.

Abstract

1. Connecting mathematical models with empirically measured microbial growth has remained challenging, as numerous competing models based on different theoretical approaches can fit observations. Therefore, we develop a method to automatically propose growth models from microbial data alone. We validate this approach using an available dataset of *E. coli* grown on known resources, and study sixteen species across various concentrations of a rich medium.

2. The inherently interpretable approach of symbolic regression infers explicit dynamical models directly from growth data. Using symbolic

regression natively, does not favour biologically interpretable models, but we find cumulative population gain to be a more informative machine-learning feature than population size.

3. Random Forest machine learning allows us to relate this finding to the approximation of a constant-rate per capita resource consumption. This suggests that the area under the growth curve (AUC) measured in routine experiments provides information on the effective resource dynamics governing microbial growth. Finally, we use theoretical insights to inform the symbolic regression algorithm and favour biologically interpretable models.

4. Overall, we found that balancing between data fit, parsimony and biological relevance favoured both the simplest, linear approximation, and models based on a Monod dynamics, with either one or two underlying resources. Therefore, our approach to read growth laws off of microbial batch cultures in a data-driven manner provides insights on the models' data-based parsimony.

Keywords

bacterial growth; Monod dynamics; Random Forest regression; resource-consumer model; symbolic regression.

1 Introduction

Microbial growth is a key process in the ecology and evolution of microbes. For instance, understanding it quantitatively is essential for public health in the face of rising antibiotic resistances in human pathogens (Antimicrobial Resistance Collaborators, 2022; Wong, 2024), and for various biotechnology applica-

tions (Lässig et al., 2023; Yeoman et al., 2020). Yet, quantifying how microbial populations convert resources into biomass remains challenging, with most experiments relying on chemically complex media and short culturing times. So far, there has been no simple way to use batch culture data to infer how microbial populations grow, without relying on strong prior modelling assumptions. We know for certain that a growing microbial population uptakes resources from the environment and converts them to new biomass, and previous research has established that growth in batch cultures involve two intrinsic nonlinearities. On the one hand, the typical lag phase observed in growth experiments (Allen and Waclaw, 2018) implies that minimal models must include a time-dependent adjustment function to account for physiological changes occurring in bacterial cells encountering a fresh culture medium (Baranyi and Roberts, 1994; Huang, 2011; Ram et al., 2019a; Borse et al., 2024). On the other hand, the maximum per capita growth rate has a saturating response to an increase in resource concentration, demonstrated by Monod's classical work on a single resource (Monod, 1949).

A wealth of theoretical work has proposed models for multi-resource growth (Zwietering et al., 1990; Baranyi and Roberts, 1994; Huang, 2011; De Jong et al., 2017; Allen and Waclaw, 2018; Letten, 2022; Held et al., 2024; Ghenu et al., 2024). Yet, quantitative analyses of empirical bacterial growth have often circumvented any explicit account of resource dynamics by using traditional models based on logistic growth (Ram et al., 2019a; Asnicar et al., 2023), which might misrepresent microbiological processes at work (Turchin, 2001; Balsa-Canto et al., 2019; Ram et al., 2019b; Picot et al., 2023; Hatton et al., 2024; Ishizawa et al., 2024). Our inability to fully use the biologically more appropri-

ate resource-consumer models does not arise from a lack of theoretical tools (Cui et al., 2024), but from the fact that model complexity often far surpasses experimental evidence. This evidence typically covers observed population data but not resource dynamics, which creates a disconnect between observable phenotypes related to macroscopic growth (*e.g.,* yield, effective growth rate, lag time) and underlying microscopic mechanisms (*e.g.,* maximum growth rate, resource consumption) (Miguel Trabajo et al., 2024).

Indeed, several incompatible multi-resource models may fit experimental data equally well and produce indistinguishable observations (Breiman, 2001; Peleg and Corradini, 2011; Tang and Riley, 2021; Ghenu et al., 2024; Held and Manhart, 2024). We must consider not only parameter identifiability but also model discoverability to be generally insoluble in microbiology, where observations are subjected to a limited resolution and where dynamical chaos is rarely expected (Shumaylov et al., 2025). Thus, how the data could allow us to infer a true, generative model of bacterial growth, is not a well-posed problem. Instead, we ask what kind of null-models for bacterial growth are *allowed* by empirical observations, irrespective of our theoretical priors at first, then considering additional theoretical priors.

So far, an interpretable, yet automated, method to uncover dynamical growth models from experimental data has been missing. One fully interpretable alternative to black-box machine learning is symbolic regression, a method providing explicit mathematical expressions directly from data (Kronberger et al., 2025), which can fruitfully serve as a hypothesis generation approach (Radwan et al., 2025). Symbolic regression has yielded impressive results in recovering and expanding known laws in physics (Course and Nair, 2023) and, in the context

of microbial ecology, in reverse-engineering Lotka-Volterra and logistic-based models (Gaucel et al., 2014; Martin et al., 2018; Regalado, 2021; Hsin et al., 2023). But empirically indistinguishable models (Tang and Riley, 2021; Held and Manhart, 2024) imply that we cannot unambiguously infer any unique, true, generative model based on observations alone. Instead, a data-driven approach should ask which models suffice to reproduce the observed behaviour, so as to propose data-driven null-hypotheses compatible with observations.

The main strength of symbolic regression is the unbiased discovery of relationships between predictors and a dependent variable, done by directly searching the space of mathematical expressions (Kronberger et al., 2025). At its basic usage, the algorithm requires no assumptions or any other inputs than the data itself. While computationally intensive, computing power advancements in the past decades have made this methodology viable and increasingly popular across a multitude of scientific disciplines. However, the true potential of the method is in the ability to specify various constraints for the mathematical expressions by defining allowed mathematical operations and partial functional expressions. The latter constraint is extremely important for utilisation of the known mathematical laws and general domain knowledge, as showcased in this study.

Here, we make minimal theoretical assumptions to better understand growth data obtained for sixteen bacterial species in R2A medium (figure 1). We demonstrate that symbolic regression can reverse-engineer explicit dynamical models for the microscopic growth process. We find cumulative population gain to be a more informative machine-learning feature than population size, then link this finding to the approximation of a constant-rate per capita resource

consumption. We use theoretical insights to inform the symbolic regression algorithm and favour biologically interpretable models. We validate our approach using available data on *E. coli* with known chemical resources (Held et al., 2024). The algorithm suggests that bacterial growth widely follows a Monod dynamics in our experiments, with either one or two effective underlying resources. This is generally a slightly a better approximation than a linear-consumption model.

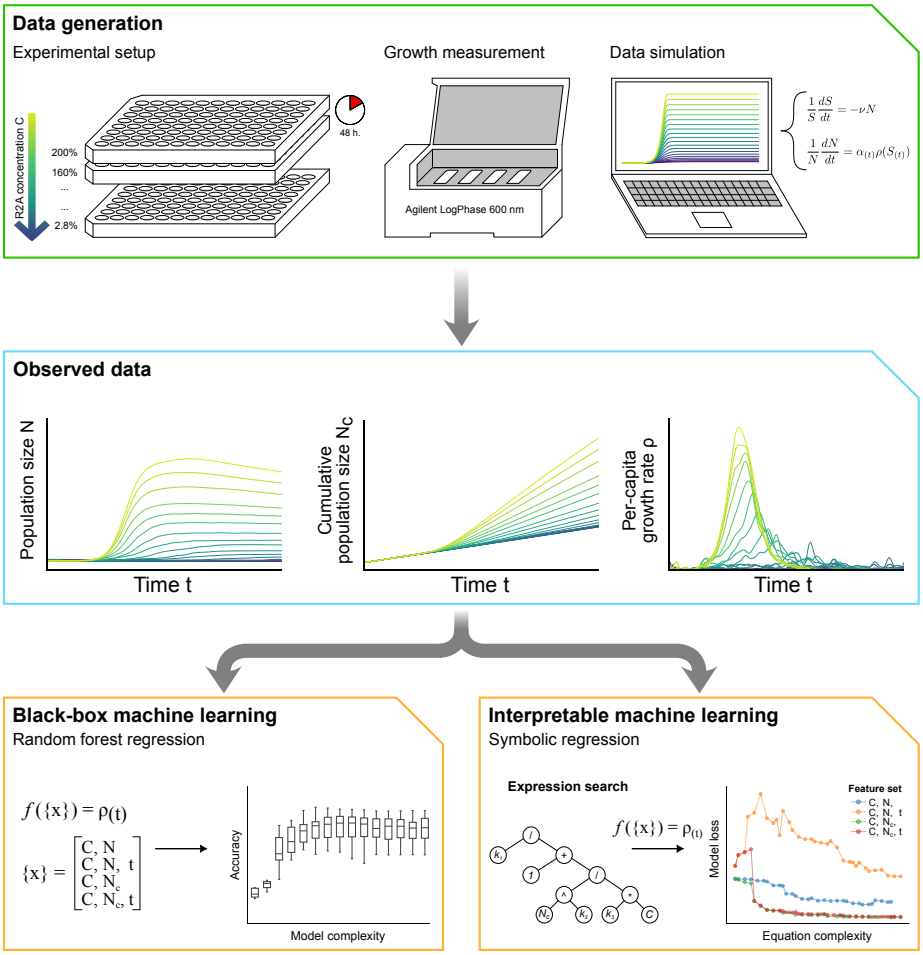

Figure 1: Workflow for the production and analysis of the datasets.

2 Materials and methods

2.1 Experimental methods for bacterial cultures

Our experiments involved 16 species from the HAMBI collection (Helsinki Institute of Life Science HiLIFE, Biodiversity Collections Research Infrastructure HUBCRI, Microbial Domain Biological Resource Centre HAMBI) summarised in Table S1. All were able to grow in Reasoner's 2A medium and the following culture protocol yielded our data, summarised in Figure S1.

Solution M9 contained 11.28 g of M9 minimal medium salts (MP Biomedicals, lot `142662`) in 1.0 L of Milli-Q purified water. The culture medium (100 %) was prepared by dissolving then autoclaving 3.0 g of Reasoner's 2A broth powder (Neogen, lot `UK208078/190`) into 1.0 L of solution M9. We prepared a range of 20 different concentrations C of culture medium, separated by an incremental factor 0.8, from 200 % to 2.88 %, by successive dilutions into autoclaved solution M9. Bacterial pre-cultures were run by inoculating each species into 6.0 mL of culture medium (100 %), then incubating for 48 h in aerobic conditions at 30 °C under agitation at 70 rpm, and finally diluted into autoclaved solution M9 by a factor 10^{-4}.

Monocultures were run in four batch replicates by inoculating 110 µL of each culture medium concentration C with 20 µL of pre-culture onto a 96-well plate, then incubating for 48 h in aerobic conditions at 30 °C under agitation at 800 rpm. An Agilent BioTek LogPhase 600 device automatically recorded time and optical density at 600 nm every 10 min as a proxy to population size. We derived all other observations numerically.

2.2 Theoretical methods and models

Consider a bacterial population of size N growing in a substrate of concentration S. The variables S and N are time-dependent. To account for the lag period observed in microbiological batch cultures, minimal models of bacterial growth assume a time-dependent physiological adjustment α in their population dynamics (Zwietering et al., 1990; Baranyi and Roberts, 1994; Ghenu et al., 2024):

$$\rho_{\text{obs}} \stackrel{\text{def}}{=} \frac{1}{N}\frac{dN}{dt} = \alpha(t) \times \rho(S), \tag{1}$$

where ρ is the time-independent and potentially resource-dependent growth function. How physiological adjustment delays bacterial cells' growth dynamics is usually represented by similar mathematical forms (Baranyi and Roberts, 1994; Huang, 2011; Lo Grasso et al., 2023; Borse et al., 2024; Ghenu et al., 2024). From fits to real data, we know that just two phenomenological parameters suffice to capture the adjustment function (Ram et al., 2019a; Borse et al., 2024), namely the initial physiological state of the cell (population) q and the rate of change m in adjusting its physiology to the present environmental state:

$$\alpha(t) = \frac{q}{q + \mathrm{e}^{-mt}}. \tag{2}$$

Monod (1949) showed that the maximum $\rho_{\max}$ of the empirical per capita growth rate ρ_{obs} was an increasing then saturating function of the resource concentration. The demonstration concerns only the maximum per capita growth rate $\rho_{\max}(S)$ in a bacterial growth experiment (Monod, 1949; Angaroni et al., 2025). However, this type II functional response pattern (Holling, 1959a) is usually modelled using a hyperbolic function (Michaelis and Menten, 1913; Monod, 1949; Holling, 1959b) and integrated as a function of time into a dynamical

model (Miguel Trabajo et al., 2024):

$$\rho(S) = g_{\max}\, g(S) \;\text{ with }\; g(S) \stackrel{\text{def}}{=} \frac{S}{K+S}, \tag{3}$$

where K is the substrate concentration inducing half-maximum response. Without further assumptions, this functional form cannot model multi-resource growth (Monod, 1949), which may be complicated by non-mutually-exclusive resources (Perrin et al., 2020), colimitation (Saito et al., 2008; Held et al., 2024), and sequential preference (Kremling et al., 2018). As it is unclear how several resources $(S_0, \ldots, S_s)$ together impact bacterial growth, multi-resource growth models combine each resource's Monod contribution to the total growth in various ways, notably with multiplicative or additive effects (Saito et al., 2008; Letten, 2022; Held and Manhart, 2024; Held et al., 2024):

$$\rho_M(S_0, \ldots, S_s) = g_{\max} \prod_{x=0}^{s} g_x(S_x), \tag{4}$$

$$\rho_A(S_0, \ldots, S_s) = g_{\max} \sum_{x=0}^{s} g_x(S_x). \tag{5}$$

The multiplicative-effect multi-resource model combines Eq. (1) and Eq. (4). It assumes that growth requires all resources, which are thus biologically not substitutable. The additive-effect multi-resource model combines Eq. (1) and Eq. (5). It assumes that any resource allows for some growth, which implies substitutable resources. A common framework can unify several formulations (Tang and Riley, 2021), but at the cost of a combinatorial explosion of theoretical parameters, as the number of resources increases. Overall, experimental data provides no definitive, general support for either model over the other (Tang and Riley, 2021; Held and Manhart, 2024), so that we have no prior hypothesis in our experimental setting involving several species and a rich medium.

2.3 Numerical methods

Each dataset consisted of a range of bacterial growth curves at different initial medium concentrations C for a given species. We used PYTHON (3.10.14) for analysis, NUMPY (1.26.4) and PANDAS (2.2.2) for numerical protocols, MATPLOTLIB (3.8.4) for visualisations, SCIKIT-LEARN (1.4.2) for Random Forest methods, PySR (1.3.1) for symbolic regression (Cranmer, 2023). Code [will be archived and made available on GITHUB upon acceptance]. Since the experimental batch replicates were extremely similar, we averaged population sizes across the four replicates, for each time point. This has no impact on the study but facilitates computation. The supplementary information details the protocol used for simulated data and expands on the symbolic regression methods.

For symbolic regression, the aim was to find explicit mathematical growth models which we could then interpret, in light of biological knowledge, as revealing latent resources. We split each dataset into a training set (50 %) and a testing set (50 %) comprising every other growth curve along the range of medium concentrations. For a given complexity bound on output expressions, the algorithm optimised data fits based on mean squared-error using the default method (with crossover and mutation) and parameters in reference (Cranmer, 2023). More complex functional forms resulted from randomly combining feature variables and constants using: addition (+), multiplication ($\times$), division (/), and constant power ($\hat{}\,^{k}, k > 0$). We imposed biology-grounded restrictions against negative constants and non-integer powers of the relative medium concentration C. To facilitate numerical computations, we normalised each dataset's cumulative population gain N_c by its maximum value, and we used the exponential negative scaled time $U = \mathrm{e}^{-t}$ and the exponential negative cu-

mulative population gain $F = \mathrm{e}^{-N_c}$ as features. Finally, we used expression templates constraining the search for mathematical expressions as detailed in the results.

For Random Forest regression, the aim was to suggest which observed variables would best predict growth rate and thus better reflected a latent resource dynamics. We split each dataset into a training set (75 %) and a testing set (25 %) comprising every fourth medium concentration along the range of medium concentrations. The predicted response variable was the per capita growth rate ρ_{obs}. Feature variables included: one population size feature, namely N or N_c, the relative initial medium concentration C, the time t. We used 100 estimators, with the coefficient of determination R^2 as a performance score, and different values for maximum tree depth.

3 Results

3.1 Unconstrained symbolic regression predicts growth

Our hypothesis was that growth curves in a range of medium concentrations should allow for the inference of dynamical growth models. We demonstrated it with a symbolic regression algorithm. Symbolic regression consists in optimising a functional form predicting a response variable under constraints on formula complexity, thus providing explicit mathematical models with little theoretical bias (Cranmer, 2023). We used this tool to explore which key observables are more appropriate for a blind model search.

We found that models relying on cumulative population gain N_c systematically outperformed models relying on population size N (figure 2, $\{C, N\}$ and

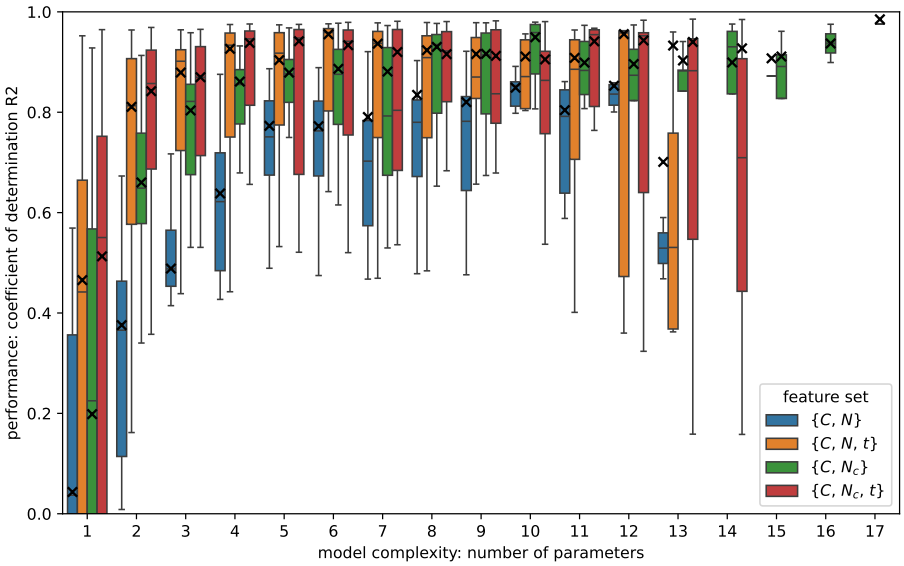

Figure 2: Performance of the unconstrained symbolic regressor in predicting the per capita growth rate from different feature sets in experimental data, given by the coefficient of determination R^2. Crosses show the median train score across species. The cumulative population gain predictor N_c outperforms the population size predictor N as well as its combination with time t.

$\{C, N_c\}$) in terms of fitting performance. Feature set $\{C, N_c\}$ yielded the best overall performance (figure 2), especially with more complex models. Thus, our naïve symbolic regression algorithm showcases the potential for N_c to inform us on microbial growth resources while providing mathematically explicit models. Yet, it could not yield substantial biological insight beyond the importance of N_c, because the models that it inferred optimised only their fitting performance and did not generally reflect basic theoretical expectations from the literature.

3.2 Cumulative population gain allows for black-box prediction of growth

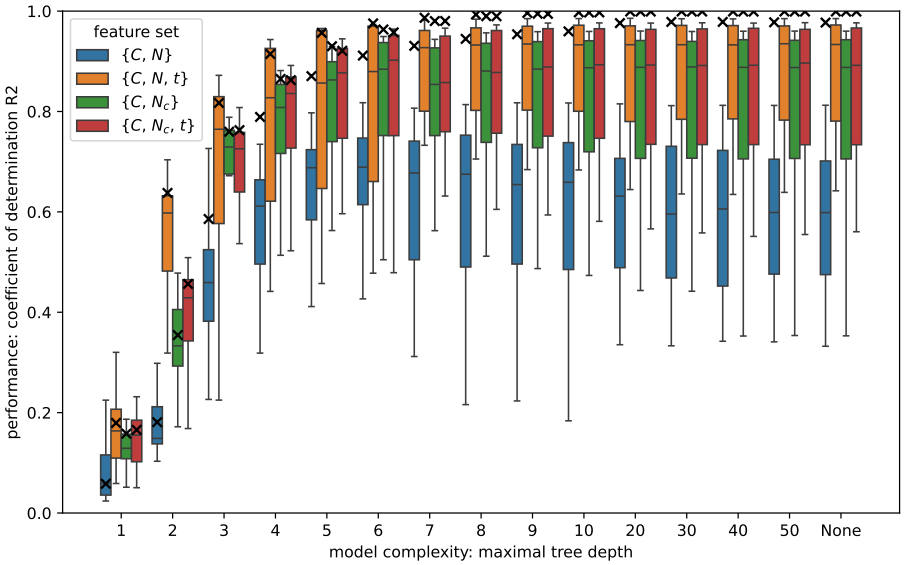

Figure 3: Performance of a black-box Random Forest regressor in predicting the per capita growth rate from different feature sets in experimental data, given by the coefficient of determination R^2. Crosses show the median train score across species. The cumulative population gain predictor N_c outperforms the population size predictor N.

Before addressing the issue of biological relevance, we asked why the cumulative population gain feature N_c yielded better results than population size N alone. We found that the observed behaviour was at least consistent with the theoretical approximation that resource consumption is a constant-per-capita-rate process. Indeed, the law of mass action assumes that a bacterial population of size N growing in a substrate of concentration S uptakes resources at a con-

stant rate ν relative to its size:

$$\frac{1}{S}\frac{dS}{dt} = -\nu N \Rightarrow S = CS^0 \exp\left[-\nu \underbrace{\int_0^t N(u)du}_{N_c}\right], \tag{6}$$

where C is the relative initial concentration of the growth medium compared to a reference, and $S^0 = S_{(C=1,t=0)}$ is the resource density in the reference concentration. This constant-rate model for per capita resource consumption readily extends to several independent resources and naturally gives a paramount role to the cumulative population gain N_c. In laboratory terms, the cumulative population gain N_c corresponds to the area under the (growth) curve (AUC) sometimes reported in microbiology.

This theoretical expectation yielded a testable implication, independent from symbolic regression: models relying on the traditional, instantaneous population size N should perform worse than models relying on the cumulative population gain N_c, but could be supplemented by providing time data t to approximate the integral in Eq. (6). To test that independently from symbolic regression, we compared the performance of a black-box Random Forest regressor using either N or N_c.

Effectively, using N_c instead of N improved the performance of a Random Forest regressor in predicting per capita growth rate ρ_{obs} (figure 3, $\{C, N\}$ and $\{C, N_c\}$). Feature N_c effectively captured the impact of both N and t since feature sets $\{C, N, t\}$, $\{C, N_c\}$, and $\{C, N_c, t\}$ achieved a similar improvement. They also yielded simpler decision trees: while relying on N led to a median maximum tree depth of 35, relying on N_c instead reduced it to 25 (table 1). When capping the maximum depth of all decision trees from the start, all three yielded a systematic improvement of the Random Forest regression compared

to feature set $\{C, N\}$ (figure 3).

Table 1: Unrestricted complexity of a Random Forest regressor using 100 trees to predict the per capita growth rate: median across 16 experimental datasets, each comprising 5780 points.

feature set	train score	test score	median maximum tree depth across species (SD)
$\{C, N\}$	98.1 %	63.7 %	35 (5)
$\{C, N, t\}$	99.9 %	90.8 %	27 (3)
$\{C, N_c\}$	99.9 %	87.2 %	25 (3)
$\{C, N_c, t\}$	99.9 %	87.4 %	26 (3)

When relying on population size N to predict the per capita growth rate ρ_{obs}, the Random Forest regressor showed systematic overfitting, with a gap between train and test scores. Models relying instead on cumulative population gain N_c reduced this overfit. Both this overfit and its reduction support a resource-consumer framework against a logistic-based model (figure S4).

Our data confirmed that, compared to population size N, cumulative population gain N_c improved the performance of a Random Forest black-box, including by reducing the overfitting. This cannot *prove* that resource consumption has a constant per capita rate but showed that this was a reasonable approximation, consistent with the data in our experiments.

3.3 Constrained symbolic regression infers biological growth model

Next, we hypothesised that a symbolic regression algorithm could infer a biologically interpretable growth function, when guided by theory-based expectations.

Thus, we ran three rounds of symbolic regression: one with minimal biological assumptions (unconstrained as previously), and two informed by supplementary theoretical assumptions, using the following expression templates:

$$\rho_{\text{obs}} = \frac{\hat{q}}{\hat{q} + \mathrm{e}^{-\hat{m}t}} \times \hat{\rho}(C, N_c), \tag{7}$$

$$\rho_{\text{obs}} = \frac{\hat{q}}{\hat{q} + \mathrm{e}^{-\hat{m}t}} \times \hat{\rho}(C \times \hat{h}(N_c)), \tag{8}$$

where the algorithm looked for the undefined parameters $\hat{q}$ and $\hat{m}$, and for the undefined functions $\hat{\rho}$ and $\hat{h}$. Template (7) assumed Eq. (1) and Eq. (2) by imposing the search for adjustment parameters q and m. Template (8) inspired by Eq. (6) additionally assumed that the growth's resource-dependency is proportional to the medium's relative concentration C. We validated our approach using available batch culture data where *E. coli* is grown on glucose as a carbon source and ammonium as a nitrogen source (Held et al., 2024). Our algorithms recovered a bidimensional Monod model, and the best-fitting, biologically interpretable model was compatible with the two inferred resources' metabolic non-substitutability (figure S5).

Our experimental datasets were obtained in R2A, a rich culture medium containing yeast extract, thus yielding no expectations regarding resource limitations. The algorithms generated 1068 models compatible with the data. A data-driven approach cannot prove nor disprove any mathematical expression: instead, the interpretation is that proposed models are sufficient to approximate empirical observations. To avoid focusing on well-fitting models with no biological interpretability, we manually annotated models to pick those which were compatible with theoretical expectations. Overall, we found few of the proposed models to be interpretable in light of the literature in microbiology (figure 4).

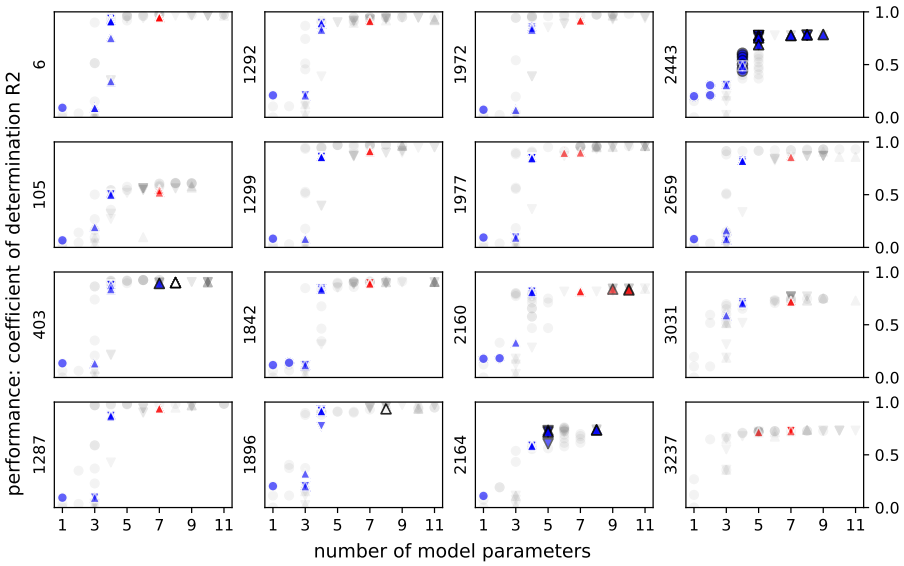

Figure 4: Fitting performance of models proposed by the symbolic regressor for R2A data. Blue: linear-based model; red: Monod-based model; white: both; gray: not biologically interpretable. White edge: unidimensional resource; black edge: multidimensional resource. Circle: no template; triangle pointing down: Template (7); triangle pointing up: Template (8).

Table 2 summarises the best-fitting models that remained biologically interpretable in experiments. In many species, the algorithm initially proposed a linear model, where the resource-dependent growth ρ was a linear function of a unidimensional, latent environmental resource S. This remained the best interpretable model for one of 16 experimental species in R2A medium, suggesting that experimental conditions did not cover the saturating part of Monod's formulation. In all other species indeed, we retrieved a dynamical formulation of the Monod model corresponding to Eq. (1) and Eq. (6), which slightly improved fits. This also implied an effectively unidimensional resource space where the per capita growth rate was a saturating function of a single environ-

mental resource. In four cases, the algorithm further found that the observed growth was better explained by an effectively multidimensional resource, with constant-per-capita-rate or Monod-like components.

Table 2: Formula, interpretation and inferred resource dimensionality of the best-fitting, biologically interpretable growth models for ρ in Eq. (1) and Eq. (3).

model for $\rho(S)$	interpretation	res. dim.	number of experimental species
$g_{\max} S$	linear	1	1 sp.: `2443`
$g_{\max} \frac{S}{K+S}$	Monod	1	11 sp.
$g_{\max} S_1 \times \frac{S_2}{K+S_2}$	linear $\times$ Monod	2	2 sp.: `403`, `1896`
$g_{\max} \frac{S_1}{K_1+S_1} \times \frac{S_2}{K_2+S_2}$	Monod $\times$ Monod	2	1 sp.: `2160`
more complex	linear and Monod	≥ 2	1 sp.: `2164`

However, the fitting improvement brought about by more complex models was often limited (figure 4). This means that linear models are good first approximations in many cases. Modellers may improve fitting performance by considering instead Monod's model, which has a limited cost in terms of parameters and complexity. Whether the fitting improvement is worth the increased complexity will depend on the specific objective at hand, especially when considering multidimensional Monod models.

4 Discussion

4.1 Biological implications

In this study, we used symbolic regression to learn explicit growth models from population size data, without relying on prior theory. While fits to the data

were excellent, an unconstrained use of the method did not ensure biological relevance, thus emphasising the need of biologically bespoke usage of symbolic regression. However, our approach suggested that cumulative population gain N_c provides a powerful and simple way to indirectly observe the latent resource space experienced by growing microbial populations. We confirmed this using Random forest machine learning and linked it with theoretical expectations resulting from the law of mass action expressed by Eq. (6). This implies that the study of microbial growth should systematically report the area under the curve (AUC) computed from growth curves.

Finally, we combined symbolic regression with theoretical insights based on the literature, spelt out here: biological constants are positive; expressions combine through addition, multiplication, division and power; biological mechanisms involve only integer powers of the medium concentration C; the law of mass action of Eq. (6) applies, *i.e.,* N_c is an important state variable; the effective per capita growth rate follows Eq. (1) as the product of an adjustment function and a resource-dependent part. We validated our approach using an available dataset in which resources were known (Held et al., 2024) before applying it to our experimental data in R2A. We found that the dynamical resource-consumer model based on Monod's formulation in Eq. (3), which we never assumed, best trades off among fit to data, symbolic simplicity, and biological interpretability, which extends previous non-dynamical findings (Angaroni et al., 2025). Our results suggests the general approximation, formulated by Eq. (1) and Eq. (6), that bacterial growth follows a dynamical Monod model with latent resources consumed at a constant per capita rate. Overall, they highlight the potential of symbolic regression as a data-driven hypothesis gen-

eration method while emphasising the need to integrate existing theory and biological knowledge when interpreting its outputs.

Our approach also provides an access to the effective resource dimensionality experienced by a growing microbial population. The effective resource dimensionality represents the minimal number of effective, independent resources theoretically required to reproduce the observed population dynamics. In most cases, the growth dynamics can be approximated as relying on a unidimensional resource (Table 2) (Monod, 1949). Among multidimensional resource models, multiplicative effects indicate non-substitutable colimiting resources (Held and Manhart, 2024), whereas additive effects suggest that the resources are substitutable. Modelled effective resources are defined purely from their impact on the growth dynamics. Departing from a chemical view allows us to translate population dynamics into effective ecological niches, which can be generally disconnected from chemical resources (Meszéna et al., 2006). Thus, our approach connects microbiology experiments with a rich literature on multidimensional niches (Hutchinson, 1961; MacArthur and Levins, 1967; Posfai et al., 2017) and their overlap, which determine species coexistence in communities (Chesson, 2018; Badali and Zilman, 2020; Pastore et al., 2021; Rozas Garcia et al., 2023). Future research could modulate the concentration of single chemical resources in a culture medium, to allow for a high-definition view of the growth process in a complex environment, and discuss the interpretation of symbolic regression outputs in light of more chemical evidence.

4.2 Methodological impact of symbolic regression in microbial ecology

Despite its success in physics (Course and Nair, 2023), symbolic regression has mainly been constrained to search logistic-based models in microbiology (Gaucel et al., 2014; Martin et al., 2018; Regalado, 2021; Hsin et al., 2023), thus overlooking two limitations. First, such models fail (Turchin, 2001; Balsa-Canto et al., 2019; Ram et al., 2019b; Picot et al., 2023; Hatton et al., 2024; Ishizawa et al., 2024) to account for key resource dynamics mediating the interactions (Meroz et al., 2024; Miguel Trabajo et al., 2024) that explain how bacterial communities assemble (Coyte et al., 2021; Lee et al., 2023; Ho et al., 2024), grow, organise and change (Pacheco et al., 2021), through competition and resource utilisation traits (Litchman et al., 2015; Ho et al., 2022). Secondly, assuming a functional form *a priori* leaves the plethora (Zwietering et al., 1990; Baranyi and Roberts, 1994; Huang, 2011; De Jong et al., 2017; Allen and Waclaw, 2018; Letten, 2022; Held et al., 2024; Ghenu et al., 2024) of growth functions proposed in the literature unexploited. Rather than developing more and more complex theory, it is essential to better connect empirical data on microbial growth with available models offering easy computation and best understanding.

We demonstrated how to combine pre-existing knowledge with the potential for symbolic regression to automatically infer an effective growth function from data, with little to no prior constraints on its functional form. Since symbolic regression does not provide automatic scientific interpretation, it is useful in generating, but not testing, hypotheses (Radwan et al., 2025). Clearly, it should not be confused for a tool to recover arbitrary generative models: instead, it constructs models that sufficiently reproduce observations, which is independent

from generative processes by virtue of well-known parametric identifiability and model discoverability caveats (Ram et al., 2019b; Tang and Riley, 2021; Held and Manhart, 2024; Shumaylov et al., 2025). Notably, our proposed models involve exclusively observable variables. Only scientific interpretation in light of pre-existing knowledge expressed by Eq. (6) allows for a translation into a latent resource space. Indeed, symbolic regression outcomes cannot reject alternative hypotheses, which may still receive other empirical support, *e.g.,* through knowledge about specific metabolites involved. Thus, future research should further investigate the link between effective resources inferred by symbolic regression and specific chemicals.

5 Conclusions

Our study demonstrates that symbolic regression offers a promising, empirically viable, and scalable way to analyse routine experiments in microbiology to infer null-models for growth and effective resource dimensionalities. Compared to the traditional method of fitting a posited family of models to data, our approach does not predefine candidate models. Instead, we build a cohort of candidate models based on the data first, then assess their theoretical and biological relevance. We found that it is reasonable to assume a constant-rate per capita resource consumption and, based on this assumption, we expect the dynamical Monod model to capture the growth behaviour of most species in batch monocultures. Microbiologists can now assess this assumption using experimental data reporting the area under growth curves, and knowledge of resource consumption dynamics could inform our predictions and control of microbial

community interactions.

Acknowledgements

The authors wish to thank Frédéric Guillaume for comments on an earlier version of the manuscript, as well as three anonymous reviewers, and CSC — IT Center for Science, Finland, for computational resources. This work for was in part supported by Research Council of Finland (`345829`, `346128`, and `364234`) to VM and TH.

References

R. J. Allen and B. Waclaw. Bacterial growth: a statistical physicist's guide. *Reports on Progress in Physics*, 82(1):016601, 2018. doi: 10.1088/1361-6633/aae546.

F. Angaroni, A. Peruzzi, E. Z. Alvarenga, and F. Pinheiro. Translating microbial kinetics into quantitative responses and testable hypotheses using Kinbiont. *Nature Communications*, 16(1), 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-61592-6.

Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, 2022. doi: 10.1016/s0140-6736(21)02724-0.

F. Asnicar, A. M. Thomas, A. Passerini, L. Waldron, and N. Segata. Machine learning for microbiologists. *Nature Reviews Microbiology*, 22(4):191–205, 2023. ISSN 1740-1534. doi: 10.1038/s41579-023-00984-1.

M. Badali and A. Zilman. Effects of niche overlap on coexistence, fixation and invasion in a population of two interacting species. *Royal Society Open Science*, 7(2):192181, 2020. doi: 10.1098/rsos.192181.

E. Balsa-Canto, J. Alonso-del-Real, and A. Querol. Mixed growth curve data do not suffice to fully characterize the dynamics of mixed cultures. *Proceedings of the National Academy of Sciences*, 117(2):811–813, 2019. doi: 10.1073/pnas.1916774117.

J. Baranyi and T. A. Roberts. A dynamic approach to predicting bacterial growth in food. *International Journal of Food Microbiology*, 23(3-4):277–294, 1994. doi: 10.1016/0168-1605(94)90157-0.

F. Borse, D. Kičiatovas, T. Kuosmanen, M. Vidal, G. Cabrera-Vives, J. Cairns, J. Warringer, and V. Mustonen. Quantifying massively parallel microbial growth with spatially mediated interactions. *PLOS Computational Biology*, 20(7):e1011585, 2024. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011585.

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199–215, 2001. doi: 10.1214/ss/1009213725.

P. Chesson. Updates on mechanisms of maintenance of species diversity. *Journal of Ecology*, 106(5):1773–1794, 2018. doi: 10.1111/1365-2745.13035.

K. Course and P. B. Nair. State estimation of a physical system with unknown governing equations. *Nature*, 622(7982):261–267, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06574-8.

K. Z. Coyte, C. Rao, S. Rakoff-Nahoum, and K. R. Foster. Ecological rules

for the assembly of microbiome communities. *PLOS Biology*, 19(2):e3001116, 2021. doi: 10.1371/journal.pbio.3001116.

M. Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv*, 2023. doi: 10.48550/arXiv.2305.01582.

W. Cui, R. Marsland, and P. Mehta. Les Houches lectures on community ecology: From niche theory to statistical mechanics. *arXiv*, 2024. doi: 10.48550/arXiv.2403.05497.

H. De Jong, S. Casagranda, N. Giordano, E. Cinquemani, D. Ropers, J. Geiselmann, and J.-L. Gouzé. Mathematical modelling of microbes: metabolism, gene expression and growth. *Journal of The Royal Society Interface*, 14(136): 20170502, 2017. ISSN 1742-5662. doi: 10.1098/rsif.2017.0502.

S. Gaucel, M. Keijzer, E. Lutton, and A. Tonda. Learning dynamical systems using standard symbolic regression. In *Genetic Programming*, page 25–36. Springer Berlin Heidelberg, 2014. ISBN 9783662443033. doi: 10.1007/978-3-662-44303-3{_}3.

A.-H. Ghenu, L. Marrec, and C. Bank. Challenges and pitfalls of inferring microbial growth rates from lab cultures. *Frontiers in Ecology and Evolution*, 11, 2024. ISSN 2296-701X. doi: 10.3389/fevo.2023.1313500.

I. A. Hatton, O. Mazzarisi, A. Altieri, and M. Smerlak. Diversity begets stability: Sublinear growth and competitive coexistence across ecosystems. *Science*, 383(6688), 2024. ISSN 1095-9203. doi: 10.1126/science.adg8488.

N. A. Held and M. Manhart. Are microbes colimited by multiple resources?

Current Opinion in Microbiology, 80:102509, 2024. ISSN 1369-5274. doi: 10.1016/j.mib.2024.102509.

N. A. Held, A. Krishna, D. Crippa, R. R. Battaje, A. J. Devaux, A. Dragan, and M. Manhart. Nutrient colimitation is a quantitative, dynamic property of microbial populations. *Proceedings of the National Academy of Sciences*, 121(52), 2024. ISSN 1091-6490. doi: 10.1073/pnas.2400304121.

P.-Y. Ho, B. H. Good, and K. C. Huang. Competition for fluctuating resources reproduces statistics of species abundance over time across wide-ranging microbiotas. *eLife*, 11, 2022. doi: 10.7554/elife.75168.

P.-Y. Ho, T. H. Nguyen, J. M. Sanchez, B. C. DeFelice, and K. C. Huang. Resource competition predicts assembly of gut bacterial communities in vitro. *Nature Microbiology*, 9(4):1036–1048, 2024. ISSN 2058-5276. doi: 10.1038/s41564-024-01625-w.

C. S. Holling. The components of predation as revealed by a study of small-mammal predation of the European pine sawfly. *The Canadian Entomologist*, 91(5):293–320, 1959a. ISSN 1918-3240. doi: 10.4039/ent91293-5.

C. S. Holling. Some characteristics of simple types of predation and parasitism. *The Canadian Entomologist*, 91(7):385–398, 1959b. ISSN 1918-3240. doi: 10.4039/ent91385-7.

J. Hsin, S. Agarwal, A. Thorpe, L. Sentis, and D. Fridovich-Keil. Symbolic regression on sparse and noisy data with gaussian processes. *arXiv*, 2023. doi: 10.48550/arXiv.2309.11076.

L. Huang. A new mechanistic growth model for simultaneous determination

of lag phase duration and exponential growth rate and a new Bělehdrádek-type model for evaluating the effect of temperature on growth rate. *Food Microbiology*, 28(4):770–776, 2011. ISSN 0740-0020. doi: 10.1016/j.fm.2010.05.019.

G. E. Hutchinson. The paradox of the plankton. *The American Naturalist*, 95 (882):137–145, 1961. doi: 10.1086/282171.

H. Ishizawa, Y. Tashiro, D. Inoue, M. Ike, and H. Futamata. Learning beyond-pairwise interactions enables the bottom–up prediction of microbial community structure. *Proceedings of the National Academy of Sciences*, 121(7), 2024. ISSN 1091-6490. doi: 10.1073/pnas.2312396121.

A. Kremling, J. Geiselmann, D. Ropers, and H. de Jong. An ensemble of mathematical models showing diauxic growth behaviour. *BMC Systems Biology*, 12(1), 2018. ISSN 1752-0509. doi: 10.1186/s12918-018-0604-8.

G. Kronberger, B. Burlacu, M. Kommenda, S. M. Winkler, and M. Affenzeller. *Symbolic Regression*. Chapman and Hall, CRC Press, Boca Raton, 2025. ISBN 9781138054813. doi: 10.1201/9781315166407. URL `http://dx.doi.org/10.1201/9781315166407`.

M. Lässig, V. Mustonen, and A. Nourmohammad. Steering and controlling evolution — from bioengineering to fighting pathogens. *Nature Reviews Genetics*, 24(12):851–867, 2023. ISSN 1471-0064. doi: 10.1038/s41576-023-00623-8.

H. Lee, B. Bloxham, and J. Gore. Resource competition can explain simplicity in microbial community assembly. *Proceedings of the National Academy of Sciences*, 120(35), 2023. ISSN 1091-6490. doi: 10.1073/pnas.2212113120.

A. D. Letten. rescomp: An R package for defining, simulating and visualizing ODE models of consumer-resource interactions. *BioRxiv*, 2022. doi: 10.1101/2022.01.11.475574.

E. Litchman, K. F. Edwards, and C. A. Klausmeier. Microbial resource utilization traits and trade-offs: implications for community structure, functioning, and biogeochemical impacts at present and in the future. *Frontiers in Microbiology*, 06, 2015. doi: 10.3389/fmicb.2015.00254.

A. Lo Grasso, A. Fort, F. F. Mahdizadeh, A. Magnani, and C. Mocenni. Generalized logistic model of bacterial growth. *Mathematical and Computer Modelling of Dynamical Systems*, 29(1):169–185, 2023. ISSN 1744-5051. doi: 10.1080/13873954.2023.2236681.

R. MacArthur and R. Levins. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, 101(921):377–385, 1967. doi: 10.1086/282505.

B. T. Martin, S. B. Munch, and A. M. Hein. Reverse-engineering ecological theory from data. *Proceedings of the Royal Society B: Biological Sciences*, 285(1878):20180422, 2018. ISSN 1471-2954. doi: 10.1098/rspb.2018.0422.

N. Meroz, T. Livny, and J. Friedman. Quantifying microbial interactions: concepts, caveats, and applications. *Current Opinion in Microbiology*, 80:102511, 2024. ISSN 1369-5274. doi: 10.1016/j.mib.2024.102511.

G. Meszéna, M. Gyllenberg, L. Pásztor, and J. A. J. Metz. Competitive exclusion and limiting similarity: A unified theory. *Theoretical Population Biology*, 69(1):68–87, 2006. doi: 10.1016/j.tpb.2005.07.001.

L. Michaelis and M. L. Menten. Die Kinetik der Invertinwirkung. *Biochemische Zeitschrift*, 49:333–369, 1913.

T. Miguel Trabajo, I. Guex, M. Dubey, E. Sarton-Lohéac, H. Todorov, X. Richard, C. Mazza, and J. R. van der Meer. Inferring bacterial interspecific interactions from microcolony growth expansion. *microLife*, 5, 2024. ISSN 2633-6693. doi: 10.1093/femsml/uqae020.

J. Monod. The growth of bacterial cultures. *Annual Review of Microbiology*, 3 (1):371–394, 1949. doi: 10.1146/annurev.mi.03.100149.002103.

A. R. Pacheco, M. L. Osborne, and D. Segrè. Non-additive microbial community responses to environmental complexity. *Nature Communications*, 12(1), 2021. doi: 10.1038/s41467-021-22426-3.

A. I. Pastore, G. Barabás, M. D. Bimler, M. M. Mayfield, and T. E. Miller. The evolution of niche overlap and competitive differences. *Nature Ecology & Evolution*, 5(3):330–337, 2021. doi: 10.1038/s41559-020-01383-y.

M. Peleg and M. G. Corradini. Microbial growth curves: What the models tell us and what they cannot. *Critical Reviews in Food Science and Nutrition*, 51 (10):917–945, 2011. ISSN 1549-7852. doi: 10.1080/10408398.2011.570463.

E. Perrin, V. Ghini, M. Giovannini, F. Di Patti, B. Cardazzo, L. Carraro, C. Fagorzi, P. Turano, R. Fani, and M. Fondi. Diauxie and co-utilization of carbon sources can coexist during bacterial growth in nutritionally complex environments. *Nature Communications*, 11(1), 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16872-8.

A. Picot, S. Shibasaki, O. J. Meacock, and S. Mitri. Microbial interactions

in theory and practice: when are measurements compatible with models? *Current Opinion in Microbiology*, 75:102354, 2023. ISSN 1369-5274. doi: 10.1016/j.mib.2023.102354.

A. Posfai, T. Taillefumier, and N. S. Wingreen. Metabolic trade-offs promote diversity in a model ecosystem. *Physical Review Letters*, 118(2), 2017. doi: 10.1103/physrevlett.118.028103.

Y. A. Radwan, G. Kronberger, and S. Winkler. A comparison of recent algorithms for symbolic regression to genetic programming. In A. Quesada-Arencibia, M. Affenzeller, and R. Moreno-Díaz, editors, *Computer Aided Systems Theory – EUROCAST 2024*, number 15172 in Lecture Notes in Computer Science, pages 157–171. Springer Nature Switzerland, Cham, 2025. ISBN 9783031829499. doi: 10.1007/978-3-031-82949-9_15.

Y. Ram, E. Dellus-Gur, M. Bibi, K. Karkare, U. Obolski, M. W. Feldman, T. F. Cooper, J. Berman, and L. Hadany. Predicting microbial growth in a mixed culture from growth curve data. *Proceedings of the National Academy of Sciences*, 116(29):14698–14707, 2019a. doi: 10.1073/pnas.1902217116.

Y. Ram, U. Obolski, M. W. Feldman, J. Berman, and L. Hadany. Reply to Balsa-Canto et al.: Growth models are applicable to growth data, not to stationary-phase data. *Proceedings of the National Academy of Sciences*, 117 (2):814–815, 2019b. doi: 10.1073/pnas.1917758117.

D. Regalado. Approximate analytic solution to the three species Lotka-Volterra differential equation model. *Recoletos Multidisciplinary Research Journal*, 9 (2):123–128, 2021. ISSN 2408-3755. doi: 10.32871/rmrj.2109.02.09.

E. Rozas Garcia, M. J. Crumpton, and T. Galla. Niche overlap and Hopfield-like interactions in generalized random Lotka-Volterra systems. *Physical Review E*, 108(3), 2023. doi: 10.1103/physreve.108.034120.

M. A. Saito, T. J. Goepfert, and J. T. Ritt. Some thoughts on the concept of colimitation: Three definitions and the importance of bioavailability. *Limnology and Oceanography*, 53(1):276–290, 2008. ISSN 1939-5590. doi: 10.4319/lo.2008.53.1.0276.

Z. Shumaylov, P. Zaika, P. Scholl, G. Kutyniok, L. Horesh, and C.-B. Schönlieb. When is a system discoverable from data? discovery requires chaos. *arXiv*, 2025. doi: 10.48550/ARXIV.2511.08860.

J. Tang and W. J. Riley. Finding Liebig's law of the minimum. *Ecological Applications*, 31(8), 2021. ISSN 1939-5582. doi: 10.1002/eap.2458.

P. Turchin. Does population ecology have general laws? *Oikos*, 94(1):17–26, 2001. ISSN 1600-0706. doi: 10.1034/j.1600-0706.2001.11310.x.

C. Wong. Antibiotic resistance is a growing threat: is climate change making it worse? *Nature*, 2024. ISSN 1476-4687. doi: 10.1038/d41586-023-04077-0.

K. Yeoman, B. Fahnert, D. Lea-Smith, and T. Clarke. Microbial biotechnology and agriculture. In K. Yeoman, B. Fahnert, D. Lea-Smith, and T. Clarke, editors, *Microbial Biotechnology*. Oxford University Press, Oxford, 2020. ISBN 9780198822813. doi: 10.1093/hesc/9780198822813.003.0008.

M. H. Zwietering, I. Jongenburger, F. M. Rombouts, and K. van 't Riet. Modeling of the bacterial growth curve. *Applied and Environmental Microbiology*, 56(6):1875–1881, 1990. doi: 10.1128/aem.56.6.1875-1881.1990.