Inferring interactions from microbiome data

Román Zapién-Campos[1,2*], Florence Bansept[1,3], and Arne Traulsen[1*]

[1]Max Planck Institute for Evolutionary Biology, Plön, Germany

[2]Centre for Life's Origins and Evolution, University College London, London, UK

[3]Aix Marseille Université, CNRS, Laboratoire de Chimie Bactérienne (UMR7283), IMM, IM2B, Turing Center for Living Systems, Marseille, France

[*]correspondence: zapien@evolbio.mpg.de, traulsen@evolbio.mpg.de

July 13, 2023

Parameter inference of high-dimensional data is challenging and microbiome time series data is no exception. Methods aimed at predicting from point estimates exist, but often even fail to recover the true parameters from simulated data. Computational methods to robustly infer and quantify the uncertainty in model parameters are needed. Here, we propose a computational workflow addressing such challenges – allowing us to compare mechanistic models and identify the values and the certainty of inferred parameters. This approach allows us to infer which kind of interactions occur in the

microbial community. In contrast to point-estimate inference, the distribution for the parameters, our outcome, reflects their uncertainty. To achieve this, we consider as many equations for the statistical moments of the microbiome as parameters. Our inference workflow, which builds upon a mechanistic foundation of microscopic processes, can take into account that commonly metagenomic datasets only provide information on relative abundances and hosts' ensembles. With our framework, we move from qualitative prediction to quantifying the likelihood of certain interaction types in microbiomes.

Keywords: microbiome, interactions, Bayesian inference, parameters, dynamical equations.

Numerous studies have shown how important the microbiome is for their hosts, ranging from development to health [1, 2]. The promise of manipulating the microbiome relies on having understood the ecological and evolutionary processes operating on it [3]. Although metagenomics studies have widely characterized microbiome samples [4], their connection to mathematical models and eco-evolutionary theories lags behind. Part of the gap is explained by an intrinsic difficulty in analysing microbiome data [5], in particular the inverse problem of robustly inferring model parameters – and thus interactions between microbes – from data (Fig.1 A). Pioneering studies have achieved point ("best guess") estimates that allow accurate predictions of the dynamics [6], but fall short in estimating the true parameters from simulations [5]. This apparent contradiction, rooted in the high dimensionality of the parameter space and incomplete nature of data, has been discussed by Cao et al. [5]. Others have suggested to use Bayesian inference – where probabilities are assigned to parameter values – to go beyond point-estimates [7]. Here, we present a computational workflow that starts from defining microscopic transition rates in a mathematical model – describing ecological and evolutionary events (such as birth, migration, mutation or speciation) – and works with macroscopic statistical moments of the microbiome composition. Our Bayesian inference workflow, which naturally bypasses known limitations of point-estimate inference [5], is sufficiently flexible to test mathematical models as diverse as microbiome samples while quantifying the parameter uncertainty stemming from data limitations (Fig.1 C) – including its extrinsic noise. Two classical ecological models are used to illustrate its application on datasets describing absolute or relative abundances of microbes – namely, logistic growth and Lotka-Volterra models. The inference workflow outlined here bridges a gap between microbiome data and theoretical modelling.

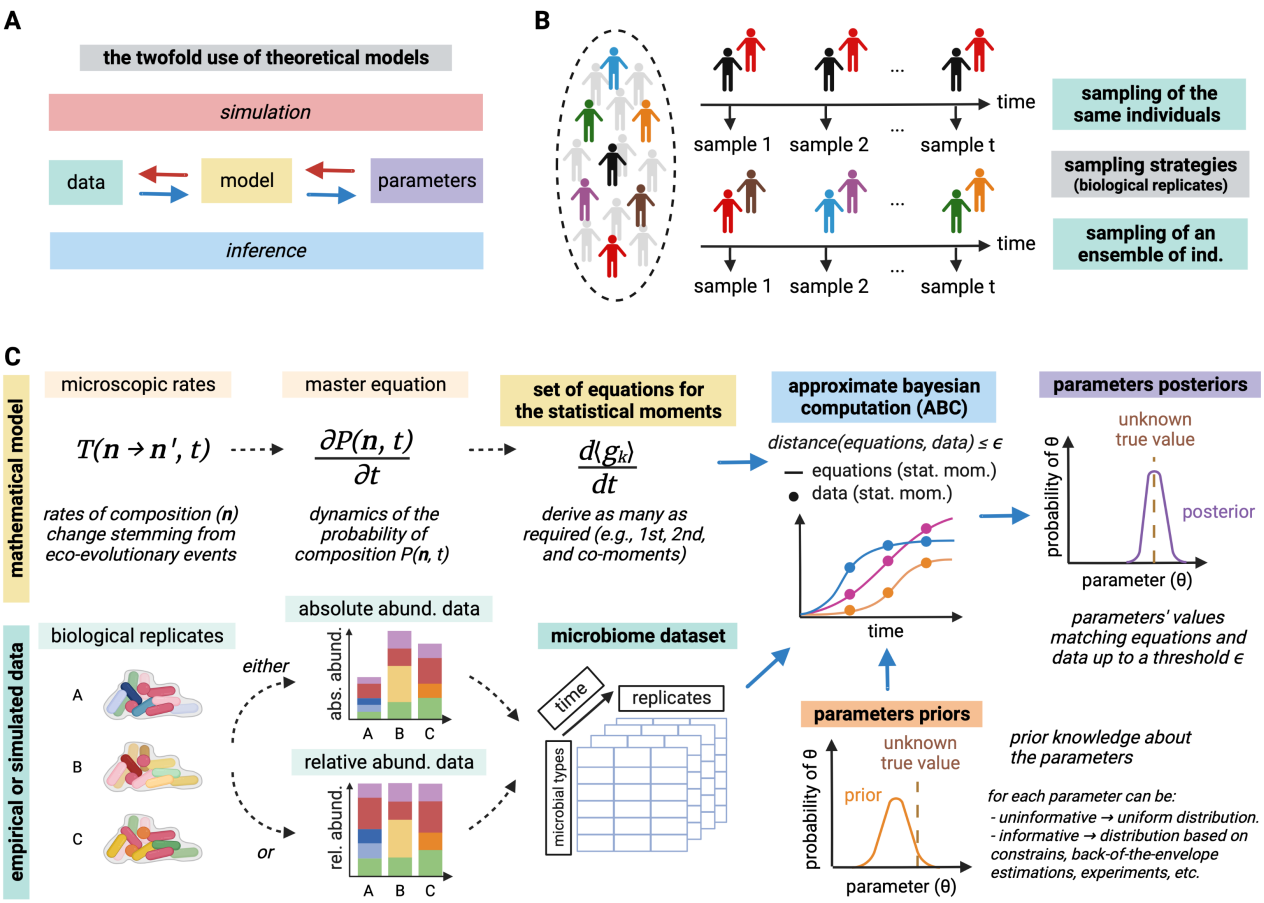

Figure 1: **Parameter inference workflow proposed and microbiome data properties**. (**A**) Mathematical models serve as a link between parameters and data. Either to simulate biological processes or to infer parameters from data. (**B**) Longitudinal sampling of the same hosts or an ensemble of them are used to obtain datasets. (**C**) Workflow from microscopic rates of a model and experimental data to inference of parameters values by Approximate Bayesian Computation (ABC). The microscopic rates describe possible eco-evolutionary events (such as birth, migration, mutation or speciation), leading to macroscopic patterns (statistical moments of the abundance). Datasets describe absolute abundances (counts) or relative abundances (frequencies) of microbes. To quantify the probability of parameter values given a dataset, prior knowledge about the parameters is updated to a posterior distribution based on the agreement of the model with the data.

Results

Developing an inference workflow

We propose a parameter inference workflow grounded on a mechanistic description of the dynamics of absolute abundances in a microbiome (Fig.1 C). First, we write down microscopic transition rates T describing changes in the microbiome composition of one host, the vector $\mathbf{n}$, to other compositions

$(\mathbf{n}')$, given the parameter set $\boldsymbol{\theta}$. Note that these rates can be functions of time t as well,

$$T(\mathbf{n} \to \mathbf{n}', t) = f(\mathbf{n}, \boldsymbol{\theta}, t). \tag{1}$$

Now, instead of tracking the microbiome composition $\mathbf{n}$ in a single host, we can describe how the probability of observing a microbiome composition $\mathbf{n}$ in an ensemble of hosts, $P(\mathbf{n}, t)$, changes with time,

$$\frac{\partial P(\mathbf{n}, t)}{\partial t} = \underbrace{\sum_{\mathbf{n}'} P(\mathbf{n}', t) T(\mathbf{n}' \to \mathbf{n}, t)}_{\text{probability influx to } \mathbf{n}} - \underbrace{\sum_{\mathbf{n}'} P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n}', t)}_{\text{probability outflux from } \mathbf{n}}. \tag{2}$$

In this expression, called the master equation [8], the probability influx and outflux terms indicate an increase or decrease in the probability of composition $\mathbf{n}$. This dynamics depends on the ecological and evolutionary processes contained in the microscopic transition rates.

Using the master equation, we can derive equations for the statistical moments of the microbiome composition in an ensemble of hosts – namely, the product of the master equation by a variable of interest (g_k, where k is an identifying index) summed over all possible microbiome compositions $\mathbf{n}$,

$$\frac{d\langle g_k \rangle}{dt} = \sum_{\mathbf{n}} g_k \frac{\partial P(\mathbf{n}, t)}{\partial t}. \tag{3}$$

For example, computing the average abundance of microbial type i implies setting $g_k = n_i$, whereas for the co-moment of microbial types i and j, $g_k = n_i n_j$. Equations like these describe the expected macroscopic dynamics of the microbiome. To obtain a system of equations of the moments that fully determines the parameters, we have to derive at least as many equations for them as parameters in the model. For example, in a Lotka-Volterra model with S microbial types, there are S growth rates and S^2 intra- and inter-specific interactions, amounting to $S + S^2$ parameters. Thus, we would need $S + S^2$ equations. We can use the S equations for the first moments $\langle n_k \rangle$, S for the second moments $\langle n_k^2 \rangle$, and $S(S-1)$ for the co-moments $\langle n_k n_l \rangle$ and covariances $\langle n_k, n_l \rangle$ (see Sup. Methods A). Note that each equation can depend on the vector of other moments, i.e., $\langle g_k \rangle = f(\langle \mathbf{g} \rangle, \boldsymbol{\theta}, t)$. A step by step derivation from microscopic rates up to second order moments for a logistic growth and the Lotka-Volterra models can be found in the Sup. Methods A. These models include conventional ecological events, such as growth, death, immigration, and direct and indirect interactions.

We now have the elements to infer the parameters $\boldsymbol{\theta}$ from microbiome data. The idea of Approximate Bayesian Computation (ABC) is to identify feasible parameters values by comparing the data to model predictions [7]. Specifically, for a given set of parameters values $\boldsymbol{\theta}$, a distance metric

between the numerical solution of the equations for the moments, $\langle g_k \rangle$, and the equivalent moments from data, $\bar{g}_k$, is estimated, e.g.,

$$\sum_k \sum_i |\underbrace{\langle g_k \rangle(\langle \mathbf{g} \rangle, \boldsymbol{\theta}, t_i)}_{\text{model solution}} - \underbrace{\bar{g}_k(t_i)}_{\text{data}}|, \tag{4}$$

for the Euclidean distance. If this distance is smaller than a threshold ε, the set is considered valid. By testing sets of parameters sampled according to an expectation – prior distribution – and recording those below the threshold ε, a posterior distribution of the parameters reflecting the uncertainty of the inference can be obtained (Fig. 1 C). With a smaller threshold ε this posterior can become the new prior and the process can be iterated to improve it. This is called Approximate Bayesian Computation - Sequential Monte Carlo (ABC-SMC). We show how to choose prior distributions of the parameters in the Sup. Methods C.

Properties of microbiome data

Given a microbiome dataset, all statistical moments $\bar{g}_k$ can be estimated from raw data. Concretely, it can be done by averaging the variable of interest g_k, over all replicates in each specific time point (Fig. 1 C). For example, for $g_k = n_i$, the replicates of n_i are simply summed over and divided by the number of replicates, while for $g_k = n_i n_j$, the products of n_i and n_j for each replicate are computed, then summed over and finally divided by the number of replicates.

Microbiome data is nowadays typically produced by metagenome sequencing. Conventionally, for technical reasons metagenomics only quantifies the *relative abundance* of each microbial type in a sample (Fig. 1 C) [9]. More recently, some studies have measured absolute numbers of culturable [10] and non-culturable microbes in samples [11]. We call these counts *absolute abundances*.

Our former equations only track moments of absolute abundance, $\langle g_k \rangle$. As Gloor et al. [9] show, inferring parameters from relative abundance (x_k) data using them would lead to spurious correlations (Fig. 2 A-B). To find equivalent expressions for the statistical moments of relative abundance, first, we propose $n_\Sigma \equiv \sum_j n_j$ to be defined as a scaling factor and a dynamical equation for its first moment, $\langle n_\Sigma \rangle$. Then, a transformation to moments of relative abundances, $\langle \gamma_k \rangle$, is given by,

$$\frac{d\langle \gamma_k \rangle}{dt} = f(\langle \boldsymbol{\gamma} \rangle, \langle n_\Sigma \rangle, \boldsymbol{\theta}, t). \tag{5}$$

Because relative abundance datasets lack information about the scaling factor, its initial condition, $\bar{n}_\Sigma(0)$, must be inferred as a free parameter, one parameter more than for absolute abundance

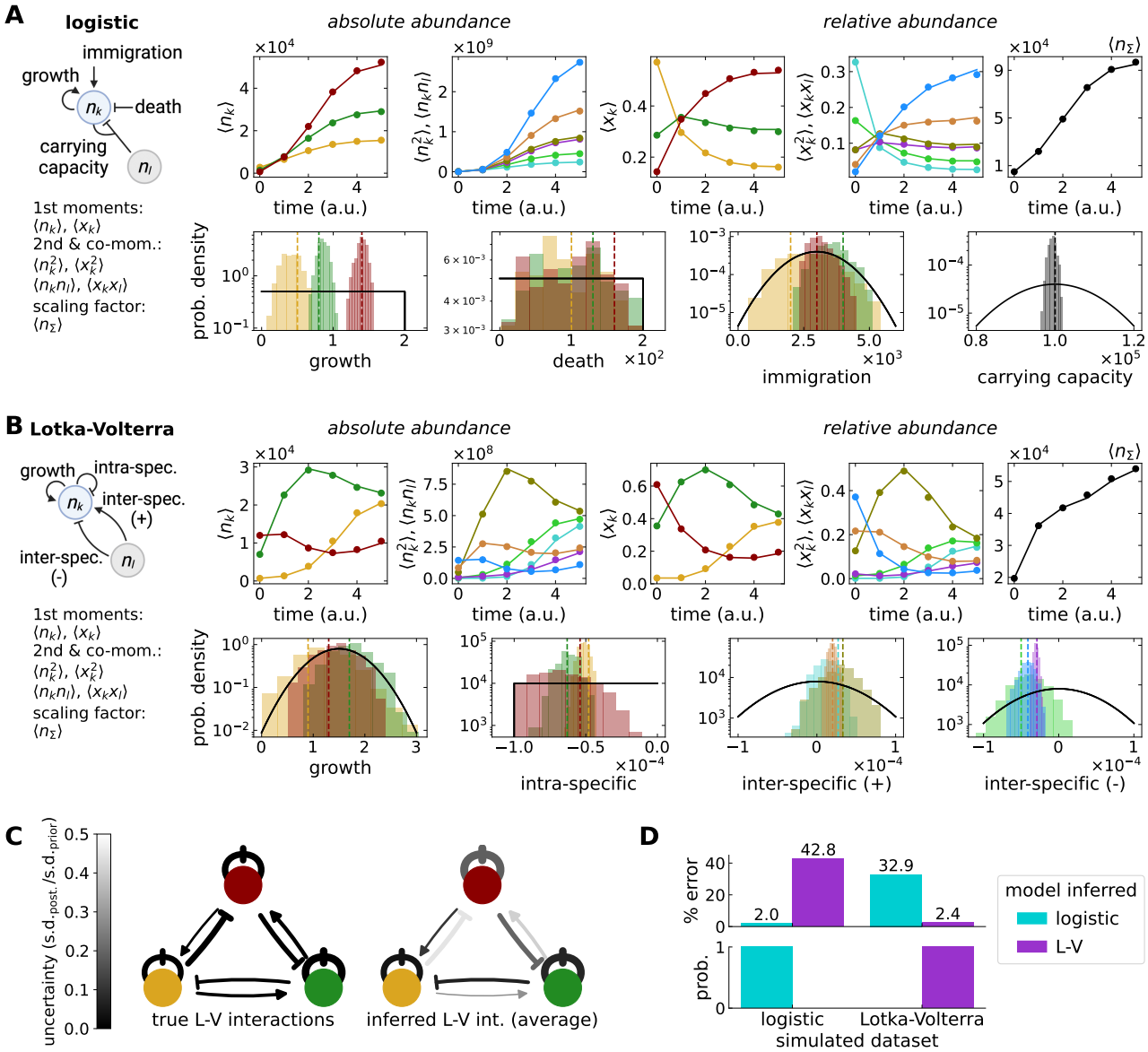

Figure 2: **Inferring true parameters from simulated data**. (**A-B**) Time series comparison between simulations (dots, derived from only 4 replicates) and equations for the statistical moments (lines) of absolute (n_k) and relative abundance (x_k) sharing parameters. Two models with three microbial types ($S = 3$) were tested, (**A**) logistic growth with immigration and death ($3S + 1 = 9$ parameters) and (**B**) Lotka-Volterra ($S + S^2 = 12$ parameters). Inferred parameter posteriors from the relative abundance are compared to true parameters (dashed lines, Sup. Methods B) and priors (black distributions). All microbial types shared the same priors (Sup. Methods C). (**C**) The inferred interactions for the Lotka-Volterra model resembled the true interactions, qualitatively (arrow heads) and quantitatively (arrow thickness), with various certainties (grayscale, defined by the ratio of s.d. of posterior to prior). (**D**) For both datasets, the most probable model was identified correctly (a.u. = time units are determined by the rates, see Sup. Methods B).

data. Note that because $\sum_k x_k = 1$, the number of equations for the microbial types decreases but the number of parameters per type remains. A detailed derivation of transformations to relative abundance for a logistic growth and the Lotka-Volterra models is shown in the Sup. Methods A.

Apart from tracking the microbiome of the same hosts over time, as in animal gut studies, commonly, hosts sampled at different time points are pulled together to produce a single time series (Fig. 1 B). This is the case when hosts are sacrificed while sampling as in experimental studies of *D. melanogaster*, *C. elegans*, and *Hydra vulgaris* [10]. In contrast to deterministic models, our workflow can deal with hosts pulled together thanks to its account of the stochastic demography. More concretely – akin to the concept of biological replicates – if the parameter values and initial conditions are the same in each host sampled, we can account for their emerging demographic differences, i.e., differences in microbiome composition resulting from stochasticity.

Inference from simulated and empirical data

We tested our inference workflow in two ways. Firstly, we recovered the correct parameters from simulated relative abundance data (Fig. 2). Our approach proved successful in cases with and without inter-specific interactions, namely, data simulated from logistic growth and Lotka-Volterra models. In both cases, our approach identified the correct model while in addition estimating their parameters values and certainty.

Secondly, we inferred parameters of an empirical time series of absolute abundance in a reduced mice microbiome – OMM^{12} [12] (Fig. 3). The posteriors suggested the growth rates of *Akkermansia muciniphila*, *Bacteroides caecimuris*, *Bifidobacterium longum*, and *Muribaculum intestinale* to be most certain, with average doubling times ranging from hours to days. Meanwhile, except from *B. caecimuris*, the average death and immigration rates were less certain, ranging from $\approx 4 \cdot 10^5$ to $1.4 \cdot 10^6$ cells per day. Most of the certainties obtained from empirical data (Fig. 3) are smaller than those from simulations (Fig. 2), highlighting the limits of the model tested and inference from, noisy, empirical data. However, in each case we obtained a set of parameters – capturing interactions between microbes – with some level of certainty. To exemplify the utility of our outcome, here, our results point to selection as the ecological driver of the OMM^{12} dynamics, despite a possible compatibility of this data with a neutral hypothesis once it has reached steady state [13]. In this case, neutrality would imply that the parameter posteriors overlap between microbial types, which is not the case (Fig. 3).

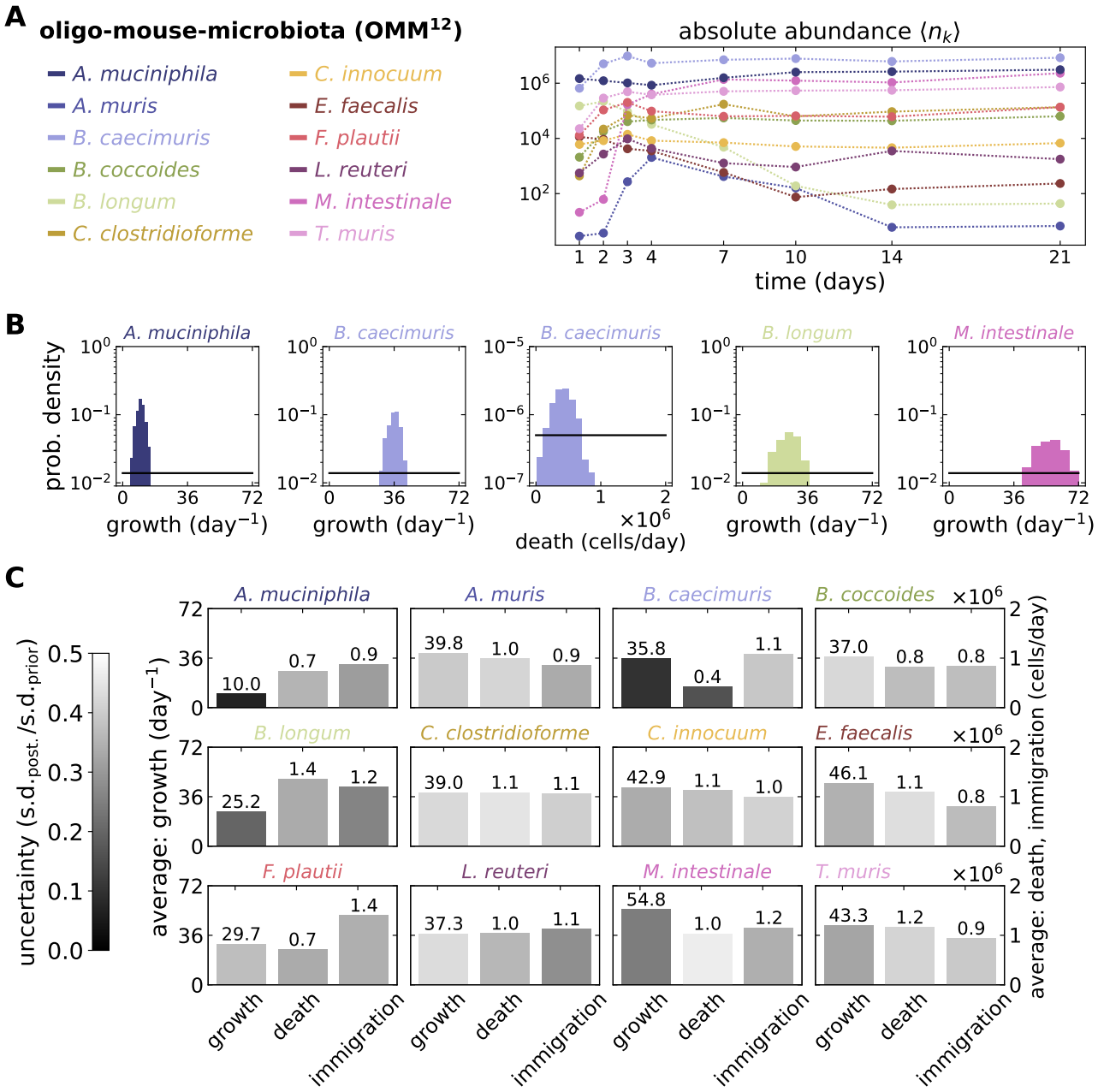

Figure 3: **Inferring parameters from empirical data**. The parameters of a logistic growth with immigration and death model were inferred from a mouse dataset. The Oligo-Mouse-Microbiota (OMM^{12}) dataset [12] tracks a 12-species defined mice microbiome ($S = 12$), where the absolute abundances of the same individuals were sampled 11 times over 99 days. (**A**) We analysed the first 21 days where 4 replicates are available throughout and proposed a logistic growth model with parameters for growth, death and immigration. We illustrate 12 of the $S + S^2 = 156$ moments used for the inference. (**B**) Of the $3S + 1 = 37$ parameters inferred, we show only the posteriors of the five most certain ones (defined by the ratio of s.d. of posterior to prior). All microbial types shared the same uniform priors (black distributions, Sup. Methods C). (**C**) The parameters inferred for each species varied widely with various certainties. For the shared carrying capacity we found an average $N \approx 1.45 \cdot 10^7$ bacterial cells, $\pm 3.49 \cdot 10^5$ cells and uncertainty of 0.0582. A system of 156 equations was solved ($S = 12$ 1st moments and $S^2 = 144$ 2nd moments and co-moments).

Discussion

The work presented here is motivated by the goal to understand how the microbes in a microbiome interact and the need to quantify the uncertainty of parameter estimation from microbiome data. Although inference methods for point estimation have been used [6], several issues limit their quantitative application, restricting them to recreate qualitative patterns of data [5]. A major issue is the indeterminacy of models with more parameters than equations [6, 5, 14]. We suggest a solution to this issue by deriving equations for the statistical moments of the microbiome composition – at least as many as parameters. In fact, our approach is driven by a mechanistic spirit, where microscopic rates must be written down first. As opposed to approaches where analytic solutions – or expensive stochastic simulations – are needed, here, a numerical solution is sufficient to quantify the distance to data [15]. This allows our workflow to handle very diverse models, where model comparison [15] in a Bayesian sense is possible.

The workflow is not limited by the properties of microbiome data [9]. As we have shown, analyzing datasets describing the relative abundance of microbial types – even if the total absolute abundance is dynamic [14] – is possible. Such is the nature of metagenomic sequencing data – the most common method to characterize microbiomes [5]. In addition, by tracking statistical moments of the microbiome, our approach naturally accounts for the diverse types of experimental samplings, such as those where ensembles of hosts are used to obtain a single time series. Concretely, compared to other methods, we track the demographic variation between hosts explicitly, and assign the remaining variation to external noise.

Although by design, our approach deals with longitudinal (time series) data, analyzing single time points (snapshot data) is possible. For example, if data is assumed to be at steady state, the inference method's aim is to find parameters making the dynamical equations for the moments equal to zero. This does not mean that the moments are zero, but that their rate of change is. Nevertheless, as our results illustrate, given the various sources of uncertainty, time series data leads to better parameter inference, in particular, those time intervals of "high activity" where many changes occur [5]. As Cao et al. [5] proposed, several of these intervals could be analysed simultaneously to improve the inference.

Bayesian inference can suffer the curse-of-dimensionality in large and diverse systems [15]. By tracking numerous statistical moments readily solved numerically, we believe our approach combined

with data of sufficiently high quality can overcome this to some extent – exploring the parameter space thoroughly in a reasonable time. We implemented an Approximate Bayesian Computation with Sequential Monte Carlo in our workflow using tools from the Python package pyABC [16] (Sup. Methods C), on which further optimizations could greatly improve its wider application [7]. As proof-of-principle, we applied our workflow to two simulated relative abundance datasets and recovered the true parameter values. We also applied it to a reduced microbiome in mice [12], where we estimated values and certainties of parameters describing logistic growth.

In summary, we presented a Bayesian inference workflow bridging microbiome data to theoretical modelling. By inferring datasets of microbial absolute and relative abundances, we showed its robustness – identifying likely interactions and certainty of parameters in simulated and empirical data. Because mechanistic rates serve as stepping stones of the workflow, similar microscopic models could replace the two classical ecological models that we illustrated – including experimentally informed models.

Acknowledgements

We thank the *Theoretical Biology Department* in the MPI Plön and the *Collaborative Research Centre 1182: Origins and Functions of Metaorganisms* for the fruitful discussions. Finally, we thank the Max Planck Society and the CRC 1182 for the funding provided.

Author contributions

RZC and AT developed the concept. RZC and FB worked on the methods. RZC wrote the first draft; all authors reviewed and approved the final version.

Availability of data and software

The data generated and software used for the analyses are available in GitHub (https://github.com/romanzapien/microbiome-inference.git).

References

[1] Daniela Esser, Janina Lange, Georgios Marinos, Michael Sieber, Lena Best, Daniela Prasse, Jay Bathia, Malte C Rühlemann, Kathrin Boersch, Cornelia Jaspers, et al. Functions of the microbiota for the physiology of animal metaorganisms. *Journal of innate immunity*, 11(5): 393–404, 2019.

[2] Jack A Gilbert, Martin J Blaser, J Gregory Caporaso, Janet K Jansson, Susan V Lynch, and Rob Knight. Current understanding of the human microbiome. *Nature medicine*, 24(4):392–400, 2018.

[3] Michael A Fischbach. Microbiome: focus on causation and mechanism. *Cell*, 174(4):785–790, 2018.

[4] Luke R Thompson, Jon G Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J Locey, Robert J Prill, Anupriya Tripathi, Sean M Gibbons, Gail Ackermann, et al. A communal catalogue reveals earth's multiscale microbial diversity. *Nature*, 551(7681):457–463, 2017.

[5] Hong-Tai Cao, Travis E Gibson, Amir Bashan, and Yang-Yu Liu. Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons. *BioEssays*, 39(2):1600188, 2017.

[6] Richard R Stein, Vanni Bucci, Nora C Toussaint, Charlie G Buffie, Gunnar Rätsch, Eric G Pamer, Chris Sander, and Joao B Xavier. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS computational biology*, 9 (12):e1003388, 2013.

[7] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

[8] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer, NY, Berlin, third edition, 2004.

[9] Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

[10] Tanita Wein, Tal Dagan, Sebastian Fraune, Thomas CG Bosch, Thorsten BH Reusch, and Nils F Hülter. Carrying capacity and colonization dynamics of curvibacter in the hydra host habitat. *Frontiers in microbiology*, 9:443, 2018.

[11] Gianluca Galazzo, Niels Van Best, Birke J Benedikter, Kevin Janssen, Liene Bervoets, Christel Driessen, Melissa Oomen, Mayk Lucchesi, Pascalle H van Eijck, Heike EF Becker, et al. How to count our microbes? the effect of different quantitative microbiome profiling approaches. *Frontiers in cellular and infection microbiology*, 10:403, 2020.

[12] Claudia Eberl, Diana Ring, Philipp C Münch, Markus Beutler, Marijana Basic, Emma Caroline Slack, Martin Schwarzer, Dagmar Srutkova, Anna Lange, Julia S Frick, et al. Reproducible colonization of germ-free mice with the oligo-mouse-microbiota in different animal facilities. *Frontiers in microbiology*, 10:2999, 2020.

[13] M. Sieber, L. Pita, N. Weiland-Brauer, P. Dirksen, J. Wang, B. Mortzfeld, S. Franzenburg, R. A. Schmitz, J.F. Baines, S. Fraune, U. Hentschel, H. Schulenburg, T. C. G. Bosch, and A. Traulsen. Neutrality in the metaorganism. *PLoS Biology*, 17(6):e3000298, 2019.

[14] Tyler A Joseph, Liat Shenhav, Joao B Xavier, Eran Halperin, and Itsik Pe'er. Compositional lotka-volterra describes microbial dynamics in the simplex. *PLoS computational biology*, 16(5): e1007917, 2020.

[15] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate bayesian computation. *PLoS computational biology*, 9 (1):e1002803, 2013.

[16] Yannik Schälte, Emmanuel Klinger, Emad Alamoudi, and Jan Hasenauer. pyabc: Efficient and robust easy-to-use approximate bayesian computation. *Journal of Open Source Software*, 7 (74):4304, 2022.

[17] Omri Allouche and Ronen Kadmon. A general framework for neutral models of community dynamics. *Ecology letters*, 12(12):1287–1297, 2009.

[18] Christian Kuehn. Moment closure – a brief review. *Control of self-organizing nonlinear systems*, pages 253–271, 2016.

[19] Beth Gibson, Daniel J Wilson, Edward Feil, and Adam Eyre-Walker. The distribution of bacterial doubling times in the wild. *Proceedings of the Royal Society B: Biological Sciences*, 285(1880): 20180789, 2018.

Appendix

A. Derivation of dynamical equations for the microbiome moments

To track the statistical moments of a model, e.g. average, variances, and co-variances, we have to account for the stochasticity of events. Thus, describing the probability of microbiome compositions is needed. The change in probability of each microbiome composition is described by the master equation,

$$\begin{aligned}\frac{\partial P(\mathbf{n},t)}{\partial t} =& \underbrace{\sum_i P(\mathbf{n}+\mathbf{e}_i,t)T(\mathbf{n}+\mathbf{e}_i\to\mathbf{n})}_{\text{influx from states } n_i+1} + \underbrace{\sum_i P(\mathbf{n}-\mathbf{e}_i,t)T(\mathbf{n}-\mathbf{e}_i\to\mathbf{n})}_{\text{influx from states } n_i-1} \\ & - \underbrace{\sum_i P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_i)}_{\text{outflux to states } n_i+1} - \underbrace{\sum_i P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_i)}_{\text{outflux to states } n_i-1},\end{aligned} \tag{6}$$

where $\mathbf{n}$ is the vector of absolute microbial abundances, and $\mathbf{e}_i$ is the amount of change, a vector with one in the *i-th* entry and zero otherwise.

Dynamical equations for the statistical moments can be obtained from the master equation by multiplication and subsequent summation. E.g., for the first moment $\langle n_k \rangle$, equivalent to the average, we have

$$\frac{d\langle n_k\rangle}{dt} = \sum_{\mathbf{n}} n_k \frac{\partial P(\mathbf{n},t)}{\partial t} = \ldots \sum_{n_k=0}^{\infty} \ldots n_k \frac{\partial P(\mathbf{n},t)}{\partial t}, \tag{7}$$

where for convenience, we make summations more explicit. For the second moment $\langle n_k^2 \rangle$, we have

$$\frac{d\langle n_k^2\rangle}{dt} = \sum_{\mathbf{n}} n_k^2 \frac{\partial P(\mathbf{n},t)}{\partial t} = \ldots \sum_{n_k=0}^{\infty} \ldots n_k^2 \frac{\partial P(\mathbf{n},t)}{\partial t}, \tag{8}$$

and for the co-moments $\langle n_k n_l \rangle$,

$$\frac{d\langle n_k n_l\rangle}{dt} = \sum_{\mathbf{n}} n_k n_l \frac{\partial P(\mathbf{n},t)}{\partial t} = \ldots \sum_{n_k=0}^{\infty} \sum_{n_l=0}^{\infty} \ldots n_k n_l \frac{\partial P(\mathbf{n},t)}{\partial t}. \tag{9}$$

For models with a finite carrying capacity, the upper sum limit is changed to a finite number.

A.1. Logistic growth with immigration and death

Let us exemplify the former steps with a logistic growth model. Similarly to Allouche and Kadmon [17], let us define the microscopic transition rates for one microbial population i,

$$T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_i) = (f_i n_i + m_i)\frac{N - \sum_j n_j}{N} \tag{10a}$$

$$T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_i) = \phi_i \frac{n_i}{N}, \tag{10b}$$

where N is the shared carrying capacity, f_i is the maximum growth rate, and ϕ_i and m_i are the death and immigration rates for each type i.

Now, we illustrate how to derive dynamical equations for the moments. Let us start with the first moment,

$$\begin{aligned}
\frac{d\langle n_k \rangle}{dt} = &\ldots \sum_{n_k=0}^{N-1} \ldots n_k P(\mathbf{n} + \mathbf{e}_k, t) T(\mathbf{n} + \mathbf{e}_k \to \mathbf{n}) \\
&+ \ldots \sum_{n_k=1}^{N} \ldots n_k P(\mathbf{n} - \mathbf{e}_k, t) T(\mathbf{n} - \mathbf{e}_k \to \mathbf{n}) \\
&- \ldots \sum_{n_k=0}^{N} \ldots n_k P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_k) \\
&- \ldots \sum_{n_k=0}^{N} \ldots n_k P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_k) \\
&+ \ldots \sum_{n_k=0}^{N} \sum_{n_i=0}^{N-1} \ldots n_k \sum_{i \neq k} P(\mathbf{n} + \mathbf{e}_i, t) T(\mathbf{n} + \mathbf{e}_i \to \mathbf{n}) \\
&+ \ldots \sum_{n_k=0}^{N} \sum_{n_i=1}^{N} \ldots n_k \sum_{i \neq k} P(\mathbf{n} - \mathbf{e}_i, t) T(\mathbf{n} - \mathbf{e}_i \to \mathbf{n}) \\
&- \ldots \sum_{n_k=0}^{N} \sum_{n_i=0}^{N} \ldots n_k \sum_{i \neq k} P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_i) \\
&- \ldots \sum_{n_k=0}^{N} \sum_{n_i=0}^{N} \ldots n_k \sum_{i \neq k} P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_i),
\end{aligned} \tag{11}$$

where the first four lines describe birth or death of a microbe of type k and the last four lines describe birth or death of a microbe of type i different from k. Note that by definition at the boundaries $P(\mathbf{n} + \mathbf{e}_i, t)|_{n_i=N} = 0$ and $P(\mathbf{n} - \mathbf{e}_i, t)|_{n_i=0} = 0$, so their summation indices go up to $n_i = N - 1$, or start from $n_i = 1$, respectively.

After appropriate transformations of variables to only deal with $P(\mathbf{n},t)$ and re-indexing, we obtain

$$\begin{aligned}\frac{d\langle n_k\rangle}{dt} =& \ldots \sum_{n_k=0}^{N} \ldots (n_k-1)P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_k)\\ &+ \ldots \sum_{n_k=0}^{N} \ldots (n_k+1)P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_k)\\ &- \ldots \sum_{n_k=0}^{N} \ldots n_k P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_k)\\ &- \ldots \sum_{n_k=0}^{N} \ldots n_k P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_k)\\ &+ \ldots \sum_{n_k=0}^{N}\sum_{n_i=0}^{N} \ldots n_k \sum_{i\neq k} P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_i)\\ &+ \ldots \sum_{n_k=0}^{N}\sum_{n_i=0}^{N} \ldots n_k \sum_{i\neq k} P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_i)\\ &- \ldots \sum_{n_k=0}^{N}\sum_{n_i=0}^{N} \ldots n_k \sum_{i\neq k} P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_i)\\ &- \ldots \sum_{n_k=0}^{N}\sum_{n_i=0}^{N} \ldots n_k \sum_{i\neq k} P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_i),\end{aligned} \tag{12}$$

Note that the last four terms reduce to zero, and that at the boundaries $T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_k)|_{n_k=0}=0$ and $T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_k)|_{n_k=N}=0$, which allows including $n_k=0$ and $n_k=N$ in the summations. Simplifying, we find

$$\frac{d\langle n_k\rangle}{dt} = \ldots \sum_{n_k=0}^{N} \ldots P(\mathbf{n},t)\left(T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_k)-T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_k)\right), \tag{13}$$

and substituting the transition rates $T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_i)$ and $T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_i)$ from Eqs. (10) leads to

$$\frac{d\langle n_k\rangle}{dt} = f_k\left(\langle n_k\rangle - \sum_j \frac{\langle n_k n_j\rangle}{N}\right) + m_k\left(1-\sum_j\frac{\langle n_j\rangle}{N}\right) - \frac{\phi_k}{N}\langle n_k\rangle. \tag{14}$$

For other moments and models similar derivations can be done.

For the second moment, we find

$$\begin{aligned}\frac{d\langle n_k^2\rangle}{dt} =& \ldots \sum_{n_k=0}^{N} \ldots (1+2n_k)P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_k)\\ &+ \ldots \sum_{n_k=0}^{N} \ldots (1-2n_k)P(\mathbf{n},t)T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_k),\end{aligned} \tag{15}$$

which after substituting $T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_i)$ and $T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_i)$ from Eqs. (10) reduces to

$$\begin{aligned}\frac{d\langle n_k^2\rangle}{dt} =& f_k \left(\langle n_k \rangle - \sum_j \frac{\langle n_k n_j \rangle}{N} + 2 \left(\langle n_k^2 \rangle - \sum_j \frac{\langle n_k^2 n_j \rangle}{N} \right) \right) \\ &+ m_k \left(1 - \sum_j \frac{\langle n_j \rangle}{N} + 2 \left(\langle n_k \rangle - \sum_j \frac{\langle n_k n_j \rangle}{N} \right) \right) \\ &+ \frac{\phi_k}{N} \left(\langle n_k \rangle - 2\langle n_k^2 \rangle \right).\end{aligned} \tag{16}$$

For the co-moments, we find

$$\begin{aligned}\frac{d\langle n_k n_l\rangle}{dt} =& \ldots \sum_{n_k=0}^{N} \sum_{n_l=0}^{N} \ldots n_l P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_k) \\ &+ \ldots \sum_{n_k=0}^{N} \sum_{n_l=0}^{N} \ldots n_k P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_l) \\ &- \ldots \sum_{n_k=0}^{N} \sum_{n_l=0}^{N} \ldots n_l P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_k) \\ &- \ldots \sum_{n_k=0}^{N} \sum_{n_l=0}^{N} \ldots n_k P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_l),\end{aligned} \tag{17}$$

which after substituting $T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_i)$ and $T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_i)$ from Eqs. (10) leads to

$$\begin{aligned}\frac{d\langle n_k n_l\rangle}{dt} =& (f_k + f_l) \left(\langle n_k n_l \rangle - \sum_j \frac{\langle n_k n_l n_j \rangle}{N} \right) \\ &+ m_k \left(\langle n_l \rangle - \sum_j \frac{\langle n_l n_j \rangle}{N} \right) + m_l \left(\langle n_k \rangle - \sum_j \frac{\langle n_k n_j \rangle}{N} \right) \\ &- \frac{\phi_k + \phi_l}{N} \langle n_k n_l \rangle.\end{aligned} \tag{18}$$

Because each equation depends on higher moments, e.g., $d\langle n_k n_l\rangle/dt$ depends on $\langle n_k n_l n_j\rangle$, it is not possible to solve this system of equations without additional assumptions. However, one can find approximate expressions, where lower moments replace higher moments. For example, $\langle n_k^2 n_j\rangle$ and $\langle n_k n_l n_j\rangle$ are approximated as functions of the lower moments: $\langle n_k^2\rangle$, $\langle n_k n_l\rangle$, and $\langle n_j\rangle$. Concretely, we can approximate e.g. $\langle n_k n_l n_j\rangle \approx \langle n_k n_l\rangle\langle n_j\rangle$. This technique, called moment closure approximation, leads to a closed system of ODEs and we use it in our approach. Kuehn [18] makes a thorough review on this technique.

A.2. Lotka-Volterra

Now, for a model with intra- and inter-specific interactions, let us define the transition rates,

$$T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_i) = n_i \left(f_i + \sum_j A_{i,j} n_j \right) \tag{19a}$$

$$T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_i) = n_i \sum_j B_{i,j} n_j, \tag{19b}$$

where A and B are positively defined matrices containing the interactions, satisfying $A_{i,j} = 0$ if $B_{i,j} > 0$, and $B_{i,j} = 0$ if $A_{i,j} > 0$. Ecologically, while interactions in A promote growth, those in B lead to death. Finally, f_i is the intrinsic growth rate.

For the first moment, similarly to Eq. (13), we have

$$\frac{d\langle n_k \rangle}{dt} = \ldots \sum_{n_k=0}^{\infty} \ldots P(\mathbf{n}, t)(T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_k) - T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_k)), \tag{20}$$

which after substituting $T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_i)$ and $T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_i)$ from Eqs. (19),

$$\frac{d\langle n_k \rangle}{dt} = f_k \langle n_k \rangle + \sum_j (A_{k,j} - B_{k,j}) \langle n_k n_j \rangle, \tag{21}$$

which takes the form of the conventional, determininistic Lotka-Volterra equations for the abundance with growth rate f_k and interaction matrix $A_{k,j} - B_{k,j}$.

For the second moment, similarly to Eq. (15)

$$\begin{aligned} \frac{d\langle n_k^2 \rangle}{dt} = &\ldots \sum_{n_k=0}^{\infty} \ldots (1 + 2n_k) P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_k) \\ &+ \ldots \sum_{n_k=0}^{\infty} \ldots (1 - 2n_k) P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_k), \end{aligned} \tag{22}$$

which after substituting $T(\mathbf{n} \to \mathbf{n} + \mathbf{e}_i)$ and $T(\mathbf{n} \to \mathbf{n} - \mathbf{e}_i)$ from Eqs. (19) leads to

$$\begin{aligned} \frac{d\langle n_k^2 \rangle}{dt} = &f_k(\langle n_k \rangle + 2\langle n_k^2 \rangle) \\ &+ \sum_j (A_{k,j} + B_{k,j}) \langle n_k n_j \rangle + 2 \sum_j (A_{k,j} - B_{k,j}) \langle n_k^2 n_j \rangle. \end{aligned} \tag{23}$$

For the co-moments, similarly to Eq. (17), we derive

$$\begin{aligned}\frac{d\langle n_k n_l\rangle}{dt} =& \ldots \sum_{n_k=0}^{\infty}\sum_{n_l=0}^{\infty} \ldots n_l P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_k) \\ &+ \ldots \sum_{n_k=0}^{\infty}\sum_{n_l=0}^{\infty} \ldots n_k P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_l) \\ &- \ldots \sum_{n_k=0}^{\infty}\sum_{n_l=0}^{\infty} \ldots n_l P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_k) \\ &- \ldots \sum_{n_k=0}^{\infty}\sum_{n_l=0}^{\infty} \ldots n_k P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_l),\end{aligned} \tag{24}$$

which after substituting $T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_i)$ and $T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_i)$ from Eqs. (19) reduces to

$$\frac{d\langle n_k n_l\rangle}{dt} = (f_k+f_l)\langle n_k n_l\rangle + \sum_j (A_{k,j} - B_{k,j} + A_{l,j} - B_{l,j})\langle n_k n_l n_j\rangle. \tag{25}$$

As previously, a moment closure approximation is required to solve the system of equations.

A.3. From absolute to relative abundance

The former equations account for the change of absolute abundances. To focus on relative abundance data, we define the relative abundance as

$$x_k = \frac{n_k}{\sum_j n_j}, \tag{26}$$

and

$$n_\Sigma = \sum_j n_j, \tag{27}$$

to serve as a scaling factor. Thus,

$$n_k = x_k n_\Sigma. \tag{28}$$

Let us find the transformation to relative abundances for the first moment. Using the definition of the covariance $\langle x_k, n_\Sigma\rangle = \langle x_k n_\Sigma\rangle - \langle x_k\rangle\langle n_\Sigma\rangle$, such that $\langle x_k n_\Sigma\rangle = \langle x_k\rangle\langle n_\Sigma\rangle + \langle x_k, n_\Sigma\rangle$, we have

$$\begin{aligned}\frac{d\langle n_k\rangle}{dt} &= \frac{d\langle x_k n_\Sigma\rangle}{dt} \\ &= \langle n_\Sigma\rangle\frac{d\langle x_k\rangle}{dt} + \langle x_k\rangle\frac{d\langle n_\Sigma\rangle}{dt} + \frac{d\langle x_k, n_\Sigma\rangle}{dt}.\end{aligned} \tag{29}$$

Rearranging, the transformation is given by

$$\frac{d\langle x_k\rangle}{dt} = \frac{1}{\langle n_\Sigma\rangle}\left(\frac{d\langle n_k\rangle}{dt} - \langle x_k\rangle\frac{d\langle n_\Sigma\rangle}{dt} - \frac{d\langle x_k, n_\Sigma\rangle}{dt}\right). \tag{30}$$

For second order moments, we use that $\langle x_k x_l n_\Sigma^2\rangle = \langle x_k x_l\rangle\langle n_\Sigma^2\rangle + \langle x_k x_l, n_\Sigma^2\rangle$ and approximate $\langle n_\Sigma^2\rangle \approx \langle n_\Sigma\rangle^2$. Then, using the chain rule

$$\begin{aligned}\frac{d\langle n_k n_l\rangle}{dt} &= \frac{d\langle x_k x_l n_\Sigma^2\rangle}{dt} \\ &= \langle n_\Sigma\rangle^2 \frac{d\langle x_k x_l\rangle}{dt} + 2\langle x_k x_l\rangle\langle n_\Sigma\rangle \frac{d\langle n_\Sigma\rangle}{dt} + \frac{d\langle x_k x_l, n_\Sigma^2\rangle}{dt}.\end{aligned} \tag{31}$$

Rearranging, the transformations are given by

$$\frac{d\langle x_k x_l\rangle}{dt} = \frac{1}{\langle n_\Sigma\rangle^2}\left(\frac{d\langle n_k n_l\rangle}{dt} - 2\langle x_k x_l\rangle\langle n_\Sigma\rangle\frac{d\langle n_\Sigma\rangle}{dt} - \frac{d\langle x_k x_l, n_\Sigma^2\rangle}{dt}\right), \tag{32}$$

and

$$\frac{d\langle x_k^2\rangle}{dt} = \frac{1}{\langle n_\Sigma\rangle^2}\left(\frac{d\langle n_k^2\rangle}{dt} - 2\langle x_k^2\rangle\langle n_\Sigma\rangle\frac{d\langle n_\Sigma\rangle}{dt} - \frac{d\langle x_k^2, n_\Sigma^2\rangle}{dt}\right), \tag{33}$$

where the differential equation for $\langle n_\Sigma\rangle$ is given by

$$\frac{d\langle n_\Sigma\rangle}{dt} = \frac{d\langle \sum_j n_j\rangle}{dt} = \sum_j \frac{d\langle n_j\rangle}{dt}. \tag{34}$$

A close look at the dynamics of the covariances shows their contribution is negligible in large populations. To see this, let us write

$$\begin{aligned}\frac{d\langle x_k, n_\Sigma\rangle}{dt} =& \ldots \sum_{n_k=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle\frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle)\, P(\mathbf{n}+\mathbf{e}_k, t) T(\mathbf{n}+\mathbf{e}_k \to \mathbf{n}) \\ &+ \ldots \sum_{n_k=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle\frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle)\, P(\mathbf{n}-\mathbf{e}_k, t) T(\mathbf{n}-\mathbf{e}_k \to \mathbf{n}) \\ &- \ldots \sum_{n_k=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle\frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle)\, P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n}+\mathbf{e}_k) \\ &- \ldots \sum_{n_k=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle\frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle)\, P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n}-\mathbf{e}_k) \\ &+ \ldots \sum_{n_k=0}^{\infty}\sum_{n_i=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle\frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle) \sum_{i\neq k} P(\mathbf{n}+\mathbf{e}_i, t) T(\mathbf{n}+\mathbf{e}_i \to \mathbf{n}) \\ &+ \ldots \sum_{n_k=0}^{\infty}\sum_{n_i=1}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle\frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle) \sum_{i\neq k} P(\mathbf{n}-\mathbf{e}_i, t) T(\mathbf{n}-\mathbf{e}_i \to \mathbf{n}) \\ &- \ldots \sum_{n_k=0}^{\infty}\sum_{n_i=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle\frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle) \sum_{i\neq k} P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n}+\mathbf{e}_i) \\ &- \ldots \sum_{n_k=0}^{\infty}\sum_{n_i=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle\frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle) \sum_{i\neq k} P(\mathbf{n}, t) T(\mathbf{n} \to \mathbf{n}-\mathbf{e}_i),\end{aligned} \tag{35}$$

after the appropriate transformations of variable to only deal with $P(\mathbf{n},t)$ and re-indexing, we find

$$\begin{aligned}
\frac{d\langle x_k, n_\Sigma\rangle}{dt} =& \ldots \sum_{n_k=0}^{\infty} \ldots \left(\frac{n_k-1}{n_\Sigma-1} - \left\langle \frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - 1 - \langle n_\Sigma\rangle)\, P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_k) \\
&+ \ldots \sum_{n_k=0}^{\infty} \ldots \left(\frac{n_k+1}{n_\Sigma+1} - \left\langle \frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma + 1 - \langle n_\Sigma\rangle)\, P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_k) \\
&- \ldots \sum_{n_k=0}^{N} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle \frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle)\, P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_k) \\
&- \ldots \sum_{n_k=0}^{N} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle \frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle)\, P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_k) \\
&+ \ldots \sum_{n_k=0}^{\infty}\sum_{n_i=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma-1} - \left\langle \frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - 1 - \langle n_\Sigma\rangle) \sum_{i\neq k} P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_i) \\
&+ \ldots \sum_{n_k=0}^{\infty}\sum_{n_i=1}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma+1} - \left\langle \frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma + 1 - \langle n_\Sigma\rangle) \sum_{i\neq k} P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_i) \\
&- \ldots \sum_{n_k=0}^{\infty}\sum_{n_i=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle \frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle) \sum_{i\neq k} P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}+\mathbf{e}_i) \\
&- \ldots \sum_{n_k=0}^{\infty}\sum_{n_i=0}^{\infty} \ldots \left(\frac{n_k}{n_\Sigma} - \left\langle \frac{n_k}{n_\Sigma}\right\rangle\right)(n_\Sigma - \langle n_\Sigma\rangle) \sum_{i\neq k} P(\mathbf{n},t) T(\mathbf{n}\to\mathbf{n}-\mathbf{e}_i).
\end{aligned} \tag{36}$$

Note that if $n_\Sigma \gg 1$, $n_\Sigma \pm 1 \approx n_\Sigma$, then, if either $n_k \gg 1$, such that $n_k \pm 1 \approx n_k$ or $\frac{n_\Sigma - \langle n_\Sigma\rangle}{n_\Sigma} \approx 0$, the terms from the previous equation simplify, leading to

$$\frac{d\langle x_k, n_\Sigma\rangle}{dt} \approx 0. \tag{37}$$

Similar arguments lead to conclude that,

$$\frac{d\langle x_k x_l, n_\Sigma^2\rangle}{dt} \approx 0, \tag{38}$$

and

$$\frac{d\langle x_k^2, n_\Sigma^2\rangle}{dt} \approx 0. \tag{39}$$

These approximations of the covariances are sensible in microbiomes, where $n_\Sigma, n_k \gg 1$ is often the case. Moreover, in the infinite population limit, covariances must be zero.

Putting all together, the change of the first moment of relative abundance in large populations is

given by

$$\begin{aligned}\frac{d\langle x_k\rangle}{dt} &= \frac{1}{\langle n_\Sigma\rangle}\left(\frac{d\langle n_k\rangle}{dt} - \langle x_k\rangle \sum_j \frac{d\langle n_j\rangle}{dt}\right) \\ &= \frac{1}{\langle n_\Sigma\rangle}\left(\frac{d\langle n_k\rangle}{dt} - \langle x_k\rangle \frac{d\langle n_\Sigma\rangle}{dt}\right),\end{aligned} \tag{40}$$

while for the second moments of relative abundance

$$\begin{aligned}\frac{d\langle x_k x_l\rangle}{dt} &= \frac{1}{\langle n_\Sigma\rangle^2}\left(\frac{d\langle n_k n_l\rangle}{dt} - 2\langle x_k x_l\rangle \sum_j \frac{d\langle n_j\rangle}{dt}\right) \\ &= \frac{1}{\langle n_\Sigma\rangle^2}\left(\frac{d\langle n_k n_l\rangle}{dt} - 2\langle x_k x_l\rangle\langle n_\Sigma\rangle \frac{d\langle n_\Sigma\rangle}{dt}\right),\end{aligned} \tag{41}$$

and

$$\begin{aligned}\frac{d\langle x_k^2\rangle}{dt} &= \frac{1}{\langle n_\Sigma\rangle^2}\left(\frac{d\langle n_k^2\rangle}{dt} - 2\langle x_k^2\rangle \sum_j \frac{d\langle n_j\rangle}{dt}\right) \\ &= \frac{1}{\langle n_\Sigma\rangle^2}\left(\frac{d\langle n_k^2\rangle}{dt} - 2\langle x_k^2\rangle\langle n_\Sigma\rangle \frac{d\langle n_\Sigma\rangle}{dt}\right).\end{aligned} \tag{42}$$

Finally, to solve these equations in terms of relative abundance, the change of variable $n_i = x_i n_\Sigma$ is needed all along. As Joseph et al. [14], we see that the second term of each equation serves as "correction factor" due to the fact that relative abundances must add up to one at all times.

B. True parameters in simulations

	growth f_i	death ϕ_i	immigration m_i	shared carrying capacity N	initial population size
$i = 1$	0.5	100	2000	10^5	2800
$i = 2$	0.8	130	4000	10^5	1400
$i = 3$	1.4	160	3000	10^5	700

Table 1: **Parameters in simulated logistic growth with immigration and death (Fig. 2 A).** The growth and death rates as well as the immigration parameters were only chosen for illustration, thus, time units are arbitrary.

		interaction $I_{i,j}$			
	growth f_i	$j = 1$	$j = 2$	$j = 3$	initial population size
$i = 1$	0.9	$-4.8 \cdot 10^{-5}$	$2.7 \cdot 10^{-5}$	$-5 \cdot 10^{-5}$	700
$i = 2$	1.7	$-2.9 \cdot 10^{-5}$	$-6.3 \cdot 10^{-5}$	$3.3 \cdot 10^{-5}$	7000
$i = 3$	1.3	$2 \cdot 10^{-5}$	$-4.1 \cdot 10^{-5}$	$-5.4 \cdot 10^{-5}$	12000

Table 2: **Parameters in simulated Lotka-Volterra (Fig. 2 B).** The interaction parameters as well as the growth rates were only chosen for illustration, thus, time units are arbitrary.

C. Inference settings

parameter	units	prior
growth f_i	time^{-1}	$\mathcal{U}(0, 2)$
death ϕ_i	cells/time	$\mathcal{U}(0, 200)$
immigration m_i	cells/time	$\mathcal{N}(3000, 1000)$
shared carrying capacity N	cells	$\mathcal{N}(10^5, 10^4)$
initial scaling factor $\bar{n}_\Sigma(0)$	cells	$\mathcal{U}(4000, 6000)$

Table 3: **Priors for simulated logistic growth with immigration and death (Fig. 2 A).** A combination of uninformative (uniform) and informative (normal) priors were used for illustration. These priors span a wide range of values to test the ability of the inference workflow to find the true parameters in simulations (Table 1). $\mathcal{U}(a, b)$ indicates a uniform distribution in the range from a to b. $\mathcal{N}(a, b)$ indicates a normal distribution with mean a and standard deviation b.

parameter	units	prior
growth f_i	time^{-1}	$\mathcal{N}(1.5, 0.5)$
intra-specific $I_{i,i}$	(cells $\cdot$ time)$^{-1}$	$\mathcal{U}(-10^{-4}, 0)$
inter-specific $I_{i,j}$	(cells $\cdot$ time)$^{-1}$	$\mathcal{N}(0, 5 \cdot 10^{-5})$
initial scaling factor $\bar{n}_\Sigma(0)$	cells	$\mathcal{U}(1.5 \cdot 10^4, 2.5 \cdot 10^4)$

Table 4: **Priors for simulated Lotka-Volterra (Fig. 2 B).** A combination of uninformative (uniform) and informative (normal) priors were used for illustration. These priors span a wide range of values to test the ability of the inference workflow to find the true parameters in simulations (Table 2). $\mathcal{U}(a, b)$ indicates a uniform distribution in the range from a to b. $\mathcal{N}(a, b)$ indicates a normal distribution with mean a and standard deviation b.

parameter	units	prior	motivation
growth f_i	day^{-1}	$\mathcal{U}(0.25, 72)$	cell division from 4 days to 20 minutes [19]
death ϕ_i	cells/day	$\mathcal{U}(0, 2 \cdot 10^6)$	from 0 to ≈ 23 cells/second*
immigration m_i	cells/day	$\mathcal{U}(0, 2 \cdot 10^6)$	from 0 to ≈ 23 cells/second*
shared carrying capacity N	cells	$\mathcal{U}(1.4 \cdot 10^7, 1.6 \cdot 10^7)$	max. number of cells in data [12]

Table 5: **Priors for logistic growth with immigration and death in empirical mouse data (Fig. 3).** Available evidence and back-of-the-envelope calculations (marked by *) were used to propose wide priors. $\mathcal{U}(a, b)$ indicates a uniform distribution in the range from a to b. $\mathcal{N}(a, b)$ indicates a normal distribution with mean a and standard deviation b.

	simulations (Fig. 2)		mouse data (Fig. 3)
	logistic	Lotka-Volterra	logistic
numeric integration (numpy)			
method	LSODA		LSODA
ABC-SMC (pyABC)			
calibration samples (to get first ε)	2000		400
$*$ max. number of generations	15		10
accepted samples per generation	AdaptivePopulationSize (CV $= 0.4$)		APS (CV $= 0.25$)
samples in first generation	750		2000
min. samples per generation	500		500
max. samples per generation	1000		1000
mismatch threshold (ε) update	QuantileEpsilon (best 10%)		QE (10%)
strategy to sample parameters	MultivariateNormalTransition		MNT
$*$ mismatch threshold (ε) min.	0.046	0.071	*NA*
$*$ min. ε change btw. generations	*NA*		10^{12}

Table 6: **Settings for Approximate Bayesian Computation - Sequential Monte Carlo (ABC-SMC) code.** These settings were chosen to decrease the computing time, but still robustly minimize the distance between data and model. We used tools from the Python package pyABC [16], mainly *ABCSMC*. The maximum number of generations, mismatch threshold (ε) minimum, and minimum ε change between generations are all stopping criteria (marked by *). LSODA is a numerical solver capable of selectively adapting to the stiffness of a system of differential equations. *NA*: Not Applicable.