Integral Bayesian symbolic regression for optimal discovery of governing equations from scarce and noisy data

Oriol Cabanas-Tirapu[a], Sergio Cobo-Lopez[a], Savannah E. Sanchez[b], Forest L. Rohwer[c], Marta Sales-Pardo[a,*], and Roger Guimerà[a,e,*]

[a]Department of Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia
[b]Department of Microbiology and Immunology, Virginia Commonwealth University School of Medicine. 1101 East Marshall Street P.O. Box 980678, Richmond, VA 23298-0678, USA
[c]Department of Biology, San Diego State University, San Diego, CA, USA
[e]ICREA, 08007 Barcelona, Catalonia
[*]Corresponding authors: Marta Sales-Pardo (E-mail: marta.sales@urv.cat); Roger Guimerà (E-mail: roger.guimera@urv.cat)

ABSTRACT

Understanding how systems evolve over time often requires discovering the differential equations that govern their behavior. Automatically learning these equations from experimental data is challenging when the data are noisy or limited, and existing approaches struggle, in particular, with the estimation of unobserved derivatives. Here, we introduce an integral Bayesian symbolic regression method that learns governing equations directly from raw time-series data, without requiring manual assumptions or error-prone derivative estimation. By sampling the space of symbolic differential equations and evaluating them via numerical integration, our method robustly identifies governing equations even from noisy or scarce data. We show that this approach accurately recovers ground-truth models in synthetic benchmarks, and that it makes quasi-optimal predictions of system dynamics for all noise regimes. Applying this method to bacterial growth experiments across multiple species and substrates, we discover novel growth equations that outperform classical models in accurately capturing all phases of microbial proliferation, including lag, exponential, and saturation. Unlike standard approaches, our method reveals subtle shifts in growth dynamics, such as double ramp-ups or non-canonical transitions, offering a deeper, data-driven understanding of microbial physiology.

Introduction

Closed-form mathematical models are crucial to understand the behavior of natural and human-made systems across scientific and engineering disciplines. They provide explicit analytical descriptions of the mechanisms that govern phenomena and enable prediction of future outcomes, control, and optimization. Closed-form models are particularly valuable because they are interpretable, computationally efficient, and facilitate the derivation of fundamental principles. Symbolic regression (also known as equation discovery) plays a key role in elucidating such governing equations by leveraging data-driven machine learning approaches to identify mathematical expressions that best describe observed relationships[1,2]. Unlike traditional regression methods, symbolic regression explores the space of possible mathematical forms, combining machine learning with symbolic mathematics, to uncover parsimonious and physically meaningful equations. This makes it a powerful tool for uncovering hidden patterns and deriving interpretable models from empirical data, bridging the gap between statistical and mechanistic modeling[3–5].

Within the symbolic regression framework, identifying the governing equations of dynamical systems in the form of differential equations[6,7] presents especial challenges, particularly when it comes to dealing with the derivatives. These derivatives are usually not directly observable and must be estimated numerically from the observed data. However, numerical estimation of derivatives

introduces biases and leads to spurious correlations, which significantly impact the performance of most symbolic regression techniques when applied to learning differential equations[8].

Perhaps the best known and most widely used method for learning governing equations of dynamical systems from data is sparse identification of nonlinear dynamical systems (SINDy)[6,9]. SINDy uses sparsity-promoting techniques to identify parsimonious differential equations in the form of linear combinations of predefined sets of library functions. However, SINDy is very sensitive to noise because of its reliance on numerically estimated derivatives. To alleviate this shortcoming, several SINDy extensions have been proposed, including those based on smoothing of derivatives[10] and estimating confidence intervals for model parameters[11], those based on Bayesian estimates of the parameters[12,13], those based on bagging and ensembles of models to increase robustness[14,15], and, most significantly for the purpose of our work, those based on weak[16,17] and integral formulations[18] of the equation discovery problem. In the latter, integrated versions of the differential governing equation are used to completely avoid derivative estimation. Furthermore, to mitigate the issues raised by the need of expert input to predefine library functions, combinations with machine learning methods have also been proposed to address the problem[19–21].

Here, we introduce an integral Bayesian approach to identify governing equations from scarce and noisy data. Like integral versions of SINDy and other approaches that use numerical integration of the differential equations[22–24], it does not require numerical estimation of derivatives. However, unlike SINDy-based approaches, our approach does not require that the differential equation is a linear combination of previously known library functions; rather, it can have any arbitrary form. Additionally, also unlike SINDy, it does not require any hyperparameter tuning. Finally, because it is based on rigorous probabilistic arguments, our approach provides optimal treatment of the uncertainty associated with data scarcity and observational noise.

To validate our approach, we test it in well-known model dynamical systems, in a regime (scarce data and high noise) in which even state-of-the-art approaches typically fail to identify the correct governing equation. We find that the integral Bayesian approach does discover the true generating models in this regime, up to high levels of noise and even with only tens of data points. We also discuss the transition by which, due to observational noise, governing equations become unlearnable by any method[25]; and show that the integral Bayesian approach makes optimal predictions of system trajectories both above and below this transition. Finally, we apply our approach to learn the governing equations of bacterial growth, and show that the models discovered perform significantly better than logistic and Gompertz models typically used to model these phenomena.

Results

Probabilistic formulation of the problem of discovering governing equations from noisy data

Let us consider a system whose evolution $\mathbf{x}(t)$ is governed by the differential equation

$$\dot{\mathbf{x}} = \mathbf{f}^*(\mathbf{x}, \boldsymbol{\theta}), \quad \mathbf{x} \in \mathbb{R}^n, \qquad \mathbf{f}^* : \mathbb{R}^n \to \mathbb{R}^n, \tag{1}$$

where $\boldsymbol{\theta}$ are some fixed but unknown parameters. If the time derivatives $\dot{\mathbf{x}}$ are observed directly, along with the coordinates $\mathbf{x}$ themselves, the problem of discovering the governing (vector) function $\mathbf{f}^*(\mathbf{x}, \boldsymbol{\theta})$ can be addressed by standard symbolic regression. In this scenario, one typically assumes that the observations of the derivatives have some measurement error

$$\dot{x}_{ik} = f_i^*(\mathbf{x}_k, \boldsymbol{\theta}) + \epsilon_{ik}, \tag{2}$$

where $\dot{x}_{ik}$ is the k-th observation of the time derivative of the i-th component of $\mathbf{x}$, f_i^* is the i-th component of $\mathbf{f}$, $\mathbf{x}_k$ is the k-th observation of $\mathbf{x}$, and ϵ_{ik} is a Gaussian random variable with zero mean and unknown variance σ. Under these conditions, the complete and optimal[25] probabilistic solution of the problem is encapsulated in the posterior distribution $p(f_i|\dot{D}_i)$[25,26], which gives the probability that an arbitrary function $f_i(\mathbf{x},\boldsymbol{\theta})$ is the true governing function $f_i^*(\mathbf{x},\boldsymbol{\theta})$, given the observed data $\dot{D}_i = \{(\dot{x}_{ik},\mathbf{x}_k)\}$. (Note that the dot on $\dot{D}_i$ just indicates that time derivatives are observed directly.) Without loss of generality, this posterior can be written as[25,26]

$$p(f_i|\dot{D}_i) = \frac{1}{Z}\exp\left[-\mathscr{L}\left(f_i,\dot{D}_i\right)\right], \tag{3}$$

where $\mathscr{L}(f_i,\dot{D}_i)$ is the description length of model f_i, measured in nats, and results from integrating the likelihood of $f_i(\mathbf{x},\boldsymbol{\theta})$ over its parameters $\boldsymbol{\theta}$ (Methods). Under mild assumptions[26,27], it can be approximated as

$$\mathscr{L}(f_i,\dot{D}_i) = \frac{B(f_i,\dot{D}_i)}{2} - \log p(f_i), \tag{4}$$

where $B(f_i,\dot{D}_i)$ is the Bayesian information criterion[27] of model f_i, and $p(f_i)$ is a suitable prior distribution over models[26]. The normalization constant in Eq. (3) (or partition function) is $Z = \sum_{f_i}\exp[-\mathscr{L}(f_i,\dot{D}_i)]$, where the sum is over all possible functions $f_i(\mathbf{x},\boldsymbol{\theta})$. We call Bayesian machine scientist (BMS) any algorithm that selects models by maximizing the posterior in Eq. (3) (or, equivalently, minimizing the description length in Eq. (4))[26]. When the derivatives $\dot{\mathbf{x}}$ are observed directly, the BMS is Bayesian-optimal for learning the corresponding governing equations from data[25].

However, in most situations of interest, only the coordinates $\mathbf{x}$ are observed and measured, but not the time derivatives $\dot{\mathbf{x}}$. The simplest approach in this situation is to estimate the derivatives $\dot{\mathbf{x}}$ numerically from $\mathbf{x}$ using finite differences, and then proceed as in a regular symbolic regression problem, that is, as if the derivatives themselves had been observed. However, this approach leads to problems arising from the instability and spurious correlations that result from numerical estimates of the derivatives. Smoothing the derivatives reduces the instabilities, but introduces further spurious correlations, which, as we show below, are also problematic.

To solve this limitation, we present an integral formulation of the probabilistic BMS framework outlined so far. We start by integrating Eq. (1) between the initial time $t = 0$ and some arbitrary time t

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}^*(\mathbf{x},\boldsymbol{\theta})\,\mathrm{d}t' \equiv \mathbf{F}^*(t,\boldsymbol{\theta},\mathbf{x}_0). \tag{5}$$

For any given component x_i, we can then assume that at time t_k the data are generated as

$$x_{ik} = F_i^*(t_k,\boldsymbol{\theta},\mathbf{x}_0) + \epsilon_{ik}, \tag{6}$$

and the posterior over integral governing functions F_i (and, thus, over the corresponding differential governing functions f_i) is

$$p(f_i|D) = p(F_i|D) = \frac{1}{Z}\exp[-\mathscr{L}(F_i,D)], \tag{7}$$

where the data are now a set of observations of the times and coordinates $D = \{(t_k,\mathbf{x}_k)\}$, but not the time derivatives. The description length is as in Eq. (4), with the prior calculated on the differential

function f_i but the Bayesian information criterion calculated by comparing F_i to the integrated data x_{ik}. Therefore, in this formulation there is no need to estimate the derivatives at any point. The integral governing function F_i corresponding to a given governing function f_i may or may not have a closed-form expression, but it can always be estimated numerically[22–24] for given parameter values $\boldsymbol{\theta}$ and initial conditions $\mathbf{x}_0$ (which can be regarded as additional parameters of the integral governing function), so that these parameters can be estimated and the description length $\mathscr{L}(F_i, D)$ can be computed (Methods). In this formulation, which we call integral Bayesian machine scientist (I-BMS), the most plausible governing function f_i is the maximum a posterior or, equivalently, the one that minimizes the description length $\mathscr{L}(F_i, D)$ of the integral governing function.

The integral Bayesian machine scientist recovers ground-truth governing equations up to high levels of noise

To evaluate the ability of the integral probabilistic approach to identify ground-truth governing equations, we start by generating synthetic data with varying levels of noise and number of observations, and for two different systems: the one-dimensional logistic equation and the two-dimensional Lotka-Volterra system (Fig. 1; Methods). In each scenario, we select the model $\mathbf{f}^e$ whose integral governing function $\mathbf{F}^e$ minimizes the integral description length $\mathscr{L}(\mathbf{F}, D)$ as defined in Eq. (7), and compare this model to the governing equation that truly generated the data. Each comparison is performed over 40 datasets to determine the frequency with which the approach converges to the exact ground-truth expression.

We benchmark these results against two sets of algorithms. First, we compare with the (non-integral) BMS given by Eq. (3) under two conditions: using finite-difference estimates of the derivatives $\dot{\mathbf{x}}$ (FD-BMS), and smoothing those estimates by means of the derivative of a polynomial fit to the time series[8] (SD-BMS). Second, we compare to ensemble-SINDy[6,14] in its weak/integral formulation[9,16–18]. SINDy uses sparse regression techniques to identify the relevant terms in a prescribed linear combination of certain library functions. In the context of discovering governing equations, these library functions are typically powers or products of powers of the components x_i (for example, x_1, x_1^3 or $x_1 x_2^2$). Ensembling (ESINDy) adds robustness to SINDy by considering ensembles of models (obtained by resampling the data) as opposed to one single model[14].

Weak and integral formulations of SINDy and ESINDy are similar in spirit to the I-BMS, in the sense that they also integrate the SINDy models of the governing equations to avoid numerical estimation of the derivatives; but, like SINDy and ESINDy, they are restricted to consider models that are linear combinations of certain predefined functions. Therefore, for the comparisons in this section, we similarly restrict our approach to consider the same linear combinations allowed by weak ESINDy (W-ESINDy). However, instead of using sparse regression to choose among models (that is, to choose which terms belong to the model), we exhaustively evaluate the description lengths $\mathscr{L}(F_i, D)$ of each allowed model[28], and select the model $\mathbf{f}^e$ whose corresponding $\mathbf{F}^e$ has components such that $F_i^e = \arg\min_{F_i} \mathscr{L}(F_i, D)$, that is, that minimize the integral description length.

For the range of noise and data sparsity levels considered, we find that W-ESINDy and ESINDy approximate the ground-truth governing equation relatively well (Fig. 2A-C)), but are rarely able to identify it exactly when the function library is expanded beyond quadratic terms (Fig. 2D-G; Supporting Text). Bayesian symbolic regression methods identify the ground-truth governing equations to different degrees in this situation. The regular BMS (Eq. (3)) with finite-differences derivatives (FD-BMS) yields the poorest performance, being the most sensitive to noise (Fig. 2D,F) and to data scarcity (Fig. 2E,G). Smoothing of the derivatives (SD-BMS) leads to a much more robust identification, but sometimes generates serious problems for low levels of noise (Fig. 2F) or abundant data (Fig. 2E).

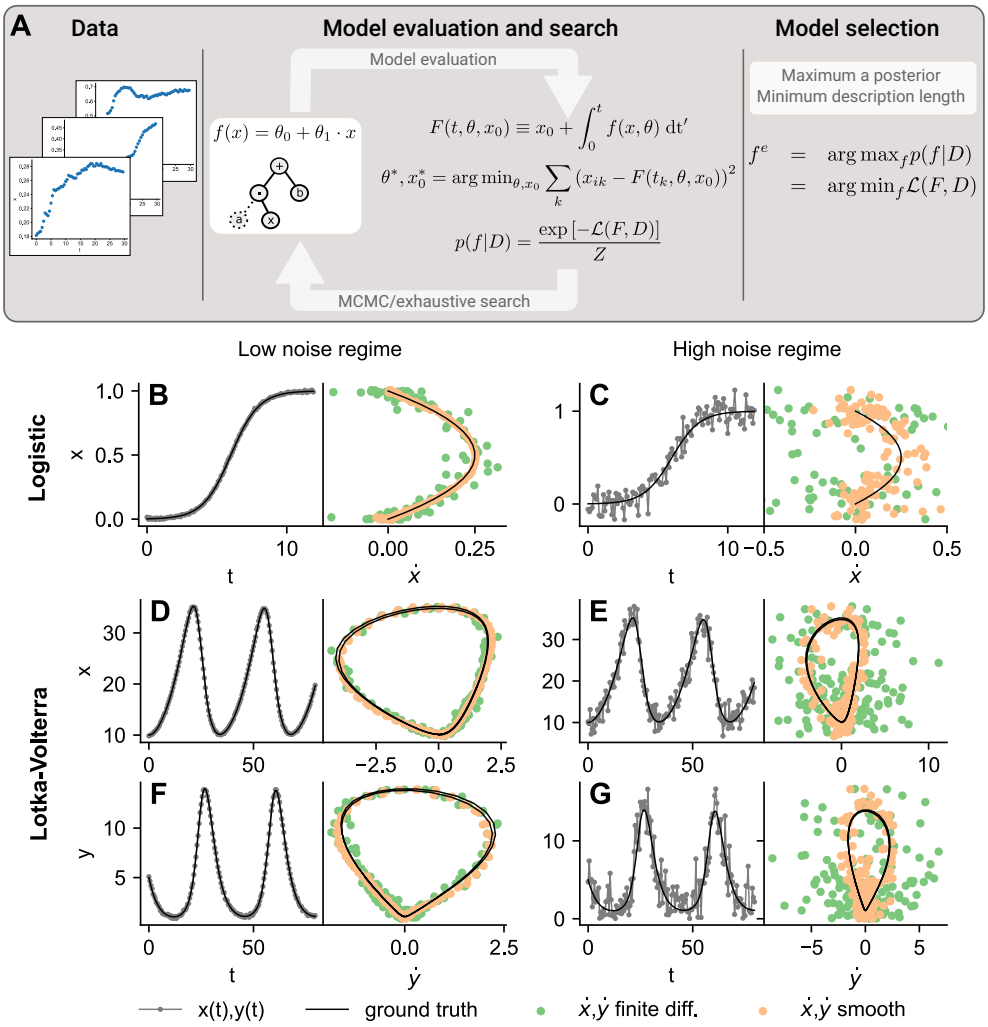

Figure 1. Integral Bayesian symbolic regression and benchmark data. (A) Schematic representation of the approach. We start with scarce and noisy measured data D from a system driven by the equation $\dot{x} = f^*(x, \boldsymbol{\theta})$. Given the observed data and any model f, we can evaluate the posterior probability $p(f|D)$ of the model without needing to estimate numerically the derivatives $\dot{x}$. This involves optimizing model parameters $\boldsymbol{\theta}$ and initial conditions x_0 on the integrated form, and calculating the description length $\mathscr{L}$ (see text). We select the model with the highest posterior (minimum description length), either by searching exhaustively within a predefined set of models or by sampling models through MCMC. **(B-G)** Left panels show the noisy synthetic data used in our validations, which we represent with gray lines; the black line corresponds to the noiseless ground truth behavior. Right panels show the phase space (measured variable against its derivative), with green dots representing the finite difference estimations of the derivative, yellow dots representing the smoothed estimate, and black lines representing the ground truth. **(B-C)** Logistic model. **(D-G)** Lotka-Volterra model. **(B, D, F)** Low noise regime. **(C, E, G)** High-noise regime.

In these situations, the probabilistic approach is overconfident about the data quality and tends to overfit. Altogether, the integral Bayesian symbolic regression (I-BMS) leads to the best results, that is, maximum resilience to noise and minimum requirements on the number of data points.

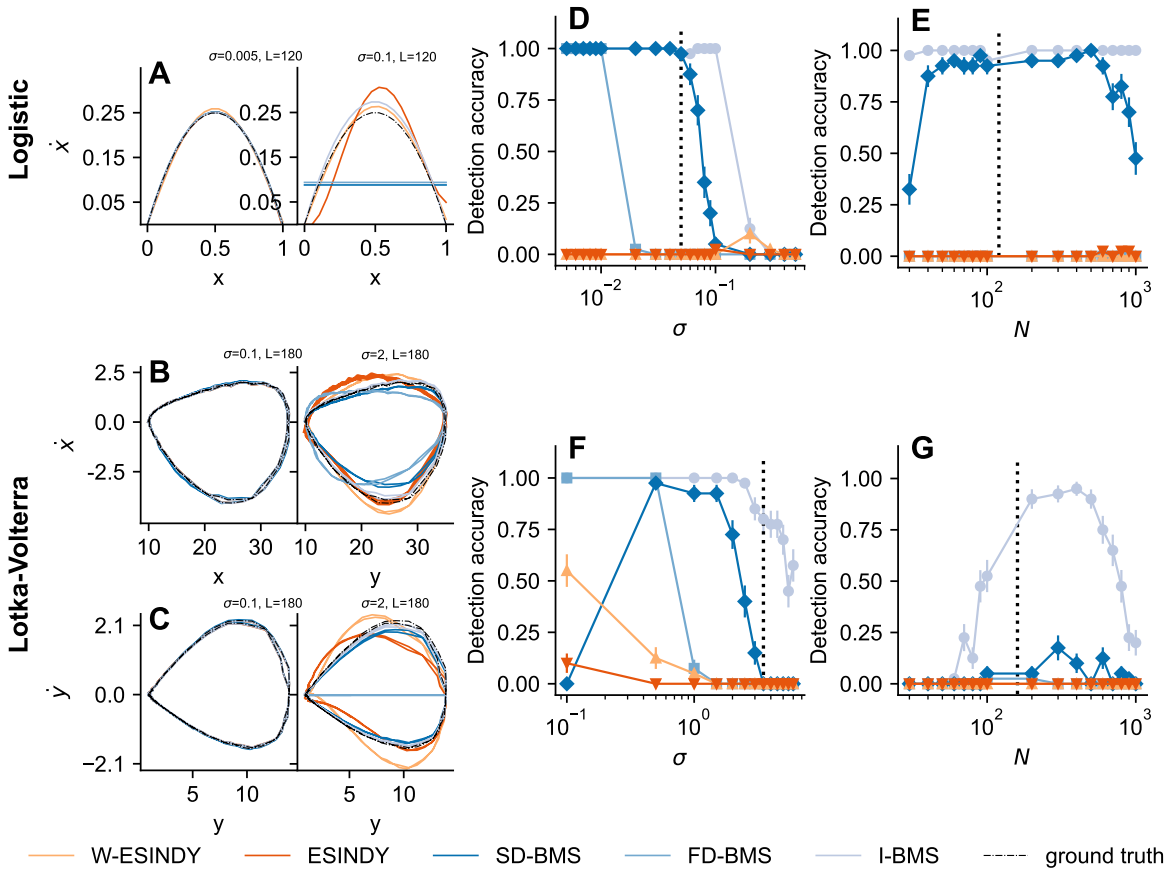

Figure 2. Validation on synthetic data. We generate synthetic data using the logistic and Lotka-Volterra models, with different levels of noise and different number of data points. We then explore exhaustively the space of all polynomial expressions (see "Exhaustive search of linear terms" in Methods) for $f_i(\mathbf{x}, \boldsymbol{\theta})$, and use the BMS to select the model with the shortest description length. We do this for the integral BMS (I-BMS), as well as for the standard BMS with finite difference-estimated derivatives (FD-BMS) and with smoothed derivatives (SD-BMS); and we benchmark these algorithms against ensemble SINDy (ESINDy) and weak ensemble SINDY (W-ESINDy), with the same library functions as the BMS. **(A-C)** Phase space trajectories predicted by the governing equations obtained using each of the approaches for one particular realization of the noise. Left panels correspond to the low-noise regime, while right panels correspond to the high-noise regime. **(D-G)** To quantify the ability of each approach to identify the true governing equation, we show the detection accuracy, that is, the fraction of times that the true governing equation is exactly recovered (each data point corresponds to an average over 40 datasets D): **(D)** as a function of noise level, for the logistic model and fixed number of observed points ($N = 120$); **(E)** as a function of the number of points, for the logistic model, fixed noise ($\sigma = 0.05$) and fixed total time range; **(F)** as a function of noise level, for the Lotka-Volterra model and fixed number of observed points ($N = 180$); **(G)** as a function of the number of points, for the Lotka-Volterra model, fixed noise ($\sigma = 3.5$) and fixed time spacing between consecutive observations.

Fundamental limits of governing equation discovery and optimality of the integral approach

We have shown how the BMS and especially the I-BMS clearly outperform state-of-the-art approaches in the identification of governing equations from noisy data. To make the comparison meaningful,

this benchmarking was limited to the space of library functions considered by SINDy and its variants (in our case, polynomials). We now go a step further and investigate: (i) to what extent is it possible to identify governing equations when we lift the restriction of considering only linear combinations of certain library functions; (ii) how close are the identified governing equations to the ground truth, in terms of their dynamics.

To answer these questions, we run a BMS that uses Markov chain Monte Carlo[25,26,29] (MCMC) to sample governing equations from the posterior distributions given by Eq. (3), in the non-integral formulation, and by Eq. (7), in the integral formulation. Note that the sampling process is not restricted to any specific form of the governing function $\mathbf{f}(\mathbf{x},\boldsymbol{\theta})$. For a given dataset D, the best estimate of the governing function is the $\mathbf{f}^e(\mathbf{x},\boldsymbol{\theta})$ with the shortest description length sampled by the BMS (integral or non-integral, depending on the case). We then define the ground truth as learnable for a given dataset D if the description length of $\mathbf{f}^e(\mathbf{x},\boldsymbol{\theta})$ is larger or equal than that of the ground truth governing function $\mathbf{f}^*(\mathbf{x},\boldsymbol{\theta})$. Conversely, a model is unlearnable if $\mathbf{f}^e(\mathbf{x},\boldsymbol{\theta})$ has a shorter description length than the ground truth governing function for that dataset. In that case $\mathbf{f}^e(\mathbf{x},\boldsymbol{\theta})$ is more plausible than the true model because $p(\mathbf{f}^e|D) > p(\mathbf{f}^*|D)$ and the ground-truth function cannot possibly be identified as the true model from the data alone[25].

As we show in Fig. 3A-B, the learnability of non-integral formulations is sensitive to noise, and the ground-truth governing function becomes unlearnable at low levels of noise. As observed before, estimating derivatives with smoothing leads to more resilience to noise than working directly with finite differences, but that comes at the cost of considerably reducing learnability at lower levels of noise. This, again, is due to the fact that smoothing makes data look more reliable (less fluctuating) than it really is, which leads the probabilistic approach to be overconfident and, thus, to overfit. As in the previous section, the integral formulation outperforms the non-integral approaches, and makes the ground truth learnable up to very high levels of noise.

To assess the predictive performance of all methods (regardless of whether the ground-truth function is learnable or not), we calculate the root mean squared error (RMSE) of the predictions $\mathbf{x}^e(t)$ of the most plausible model $\mathbf{f}^e(\mathbf{x},\boldsymbol{\theta})$ with respect to the observed values of $\mathbf{x}(t)$, where the predictions are given by integrating the governing function

$$\mathbf{x}^e(t) = \mathbf{x}_0^e + \int_0^t \mathbf{f}^e(\mathbf{x},\boldsymbol{\theta})\,\mathrm{d}t'.$$

Because $\mathbf{x}(t)$ has a measurement error σ, optimal predictions are such that $\mathrm{RMSE}/\sigma = 1$. In Fig. 3C-D, we show that the models identified by the I-BMS are close to this optimal behavior for all levels of noise. Above the learnability threshold, this does not mean that the true model has been identified, but rather that any error related to the identification of incorrect models is small compared to the measurement error. In any case, our results suggest that, at least for the two systems considered, the I-BMS finds governing equations that are optimally predictive. By contrast, non-integral formulations lead to suboptimal predictions, especially in the transition region between the learnable and the unlearnable phases.

Discovering governing equations for bacterial growth

Having validated the predictive ability of the I-BMS on synthetic data, we apply it to a dataset comprising empirical growth curves of bacteria in various media, so as to obtain an equation governing bacterial growth. We use six data sets containing data from laboratory experiments on different species[30,31]: *C. sedlakii*, *E. coli*, *E. aerogenes*, *K. pneumaniae*, *P. vulgaris* and *S. aureus*. The bacteria

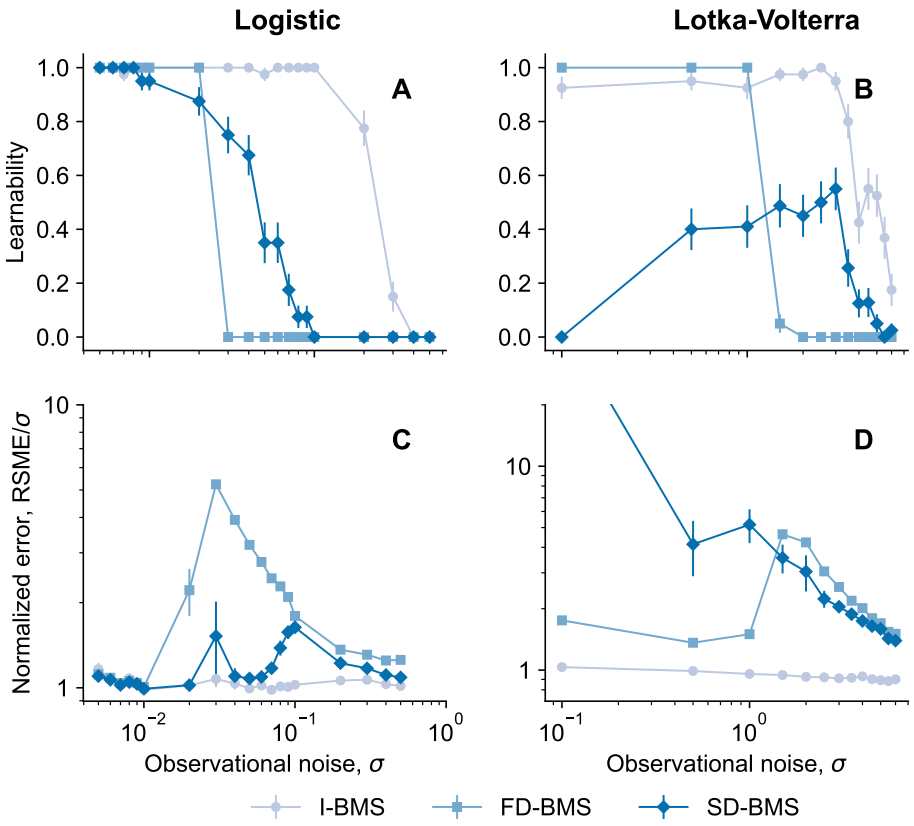

Figure 3. Learnability and model predictive accuracy across noise levels. We generate synthetic data for the logistic and Lotka-Volterra models, as in Fig. 2. We then use MCMC to sample models from the posterior $p(f_i|D)$ (for each dataset, we run two independent MCMC processes with 3,000 steps each, except for the I-BMS on Lotka-Volterra data, for which we use 4,000 steps), and consider the most plausible model (equivalently, the model with the minimum description length). All points are averages over 40 datasets D. **(A-B)** Learnability as a function of noise level for the logistic and Lotka-Volterra datasets, respectively. **(C-D)** Root mean squared error (RMSE) between the ground truth data $x(t)$ and the predictions of the minimum description length model $x^e(t)$, normalized by the noise level σ for the logistic and Lotka-Volterra datasets, respectively.

were grown for 30 to 31 hours, and optical density OD_{600} measurements of their populations were taken every 10 to 15 minutes. Each species was cultivated on 96 different substrates; here we only consider substrates that yielded substantial growth (exponential growth followed by a saturation phase). In total, we kept 29, 40, 24, 35, 28, and 13 substrates, respectively for each of the species. The training set comprises half of the experiments of five of the six bacterial species, randomly selected, whereas the test set includes the other half of the experiments on the five bacterial species and all the experiments on the remaining species.

We use the integral BMS to identify the most plausible model using MCMC, exploring model space with different inductive biases. First, we explore model space without imposing any constraints on the form of the mathematical model. Second, given the understanding that bacteria replicate through division, we consider a physics-informed I-BMS that focuses on models that incorporate a linear growth term, $\dot{B} = rB + g(B)$, where B is the bacterial population size, r is the growth rate and $g(B)$ is the arbitrary function that we aim to identify. Finally, we consider a scenario where

growth rate is itself a function of bacterial population $h(B)$, so we ask the I-BMS to explore models including a term $\dot{B} = \alpha B\, h(B)$, where α is an arbitrary scaling parameter (see Methods). We run three independent sampling processes, each with 3,000 MCMC steps, and select the governing function with the minimum description length for each run.

We compare the results of the I-BMS to two classic and widely used models for bacterial growth: the logistic growth mode[32] and the Gompertz model[33]. The logistic growth model is defined as

$$\frac{dB}{dt} = rB\left(1-\frac{B}{K}\right),$$

where, B is the population size (in practice, the optical density gives a measure of B shifted by a quantity, which can be interpreted as the optical density measured when there are no bacteria in the plate). The first term $r\,B$ gives rise to the exponential growth of bacterial population with a growth rate r, the maximum growth rate when nutrients are unlimited. The factor $1-B/K$ introduces the competition among bacteria for nutrients, so that the larger the population the larger the competition, and K is the carrying capacity, that is, the theoretical maximum concentration of bacteria achieved in the steady state.

The Gompertz model

$$\frac{dB}{dt} = rB\log\left(\frac{K}{B}\right) = rB\left(\log K - \log B\right),$$

can be interpreted in a similar fashion, the same exponential growth as the logistic model and a logarithmic term that reduces growth when the population approaches the carrying capacity K.

Recent studies that attempt to model the growth of cell populations in microbial communities use a generalized logistic model[34] that interpolates between the logistic and Gompertz curves, making it possible for sublinear growths observed in microbial communities[35]. Nonetheless, we decided to compare against the classical models typically used for population growth of single cell types.

We find that the most plausible model identified by the I-BMS has the form

$$\frac{dB}{dt} = B\,r\left(1+c_0\left(c_1 B e^{c_2 B}\right)^{B^3}\right), \tag{8}$$

where r, c_0 c_1 and c_2 are model constants, that are different for each bacterial species and growth medium in the training set (that is, we find a single model for all datasets, but allow for model parameters to be dataset-specific[29]). Formally, like in the logistic and Gompertz models, this model has two terms with opposite effects, as we find that c_0 and r have opposite signs. The first term is proportional to growth, and is typically associated to exponential growth for $r > 0$. The second term is a competition term typically reflecting competition for nutrients and other effects that limit growth. However, the dependence of this term on population size is more complicated than on benchmark models. Still, similarly to the other models, there is limiting population size for which $\frac{dB}{dt} = 0$, although this carrying capacity cannot be expressed as a simple combination of parameters. Importantly, this higher complexity of the competition term increases model expressiveness, as this model outperforms the logistic and Gompertz models at describing bacterial growth (Fig. 4). Indeed, the I-BMS model successfully captures the dynamics of bacterial growth in all phases (lag phase, growth phase, and stationary phase), even allowing for non-standard behaviors in the lag and growth phases (double ramps, faster and slower growths, etc.) (Fig. 4A-D). Besides fitting the empirical data accurately, the model does not overfit spurious fluctuations. By contrast, the benchmark models often struggle to

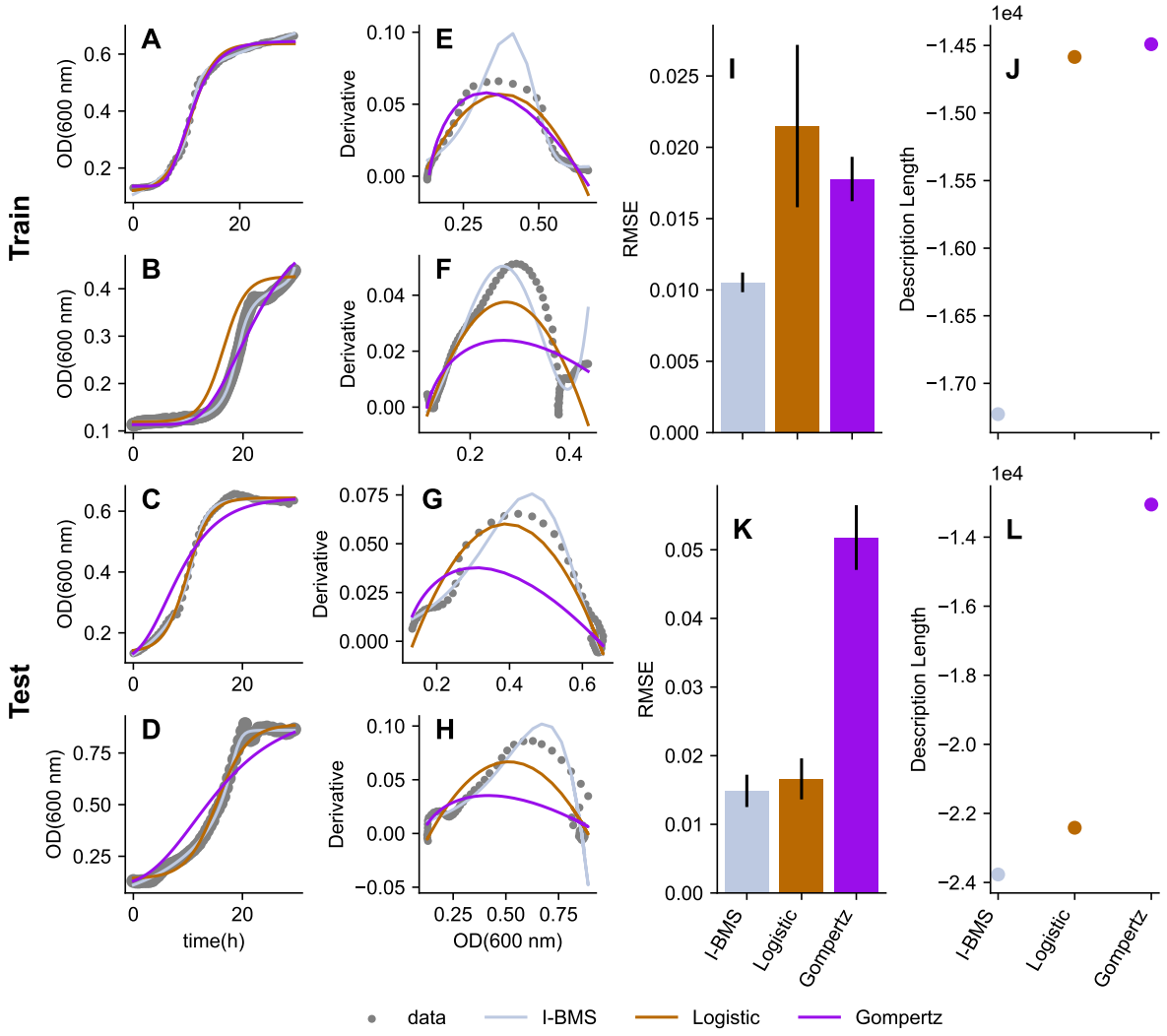

Figure 4. I-BMS model and reference growth models for bacterial growth. We show results for two bacteria-substrate pairs from the training set, and two from the test set. **(A-D)** Empirical groth curves and numerically integrated curves $x^e(t)$ for each model. **(E-H)** Derivative values plotted against different measured optical densities for each model. **(I,K)** Root mean squared error (RMSE) of the integrated curve relative to the observed data, computed for all datasets in the training and test sets, respectively. **(J-L)** Description length of the models for all training and test datasets.

correctly capture the initial phase and the exponential growth phase (Supplementary Figs. S1 and S3). Furthermore, the I-BMS model provides a reasonable estimation of the empirical derivative curve (Fig. 4E-H), despite the fact that the empirical derivative is not used in any way during training and model selection. The benchmark models often fail, again, to approximate the derivative curve (Supplementary Figs. S2 and S4).

In Figs. 4I-L, we show the quality metrics for both the training and test datasets. The I-BMS model demonstrates superior performance in both root mean squared error (although the difference is not significant when compared to the logistic model on test data) and description length, indicating a more parsimonious representation of the observed data.

Discussion

Learning the differential equations governing the behavior of dynamical systems is fundamental, not only to predict their evolution, but also to elucidate the relevant underlying mechanisms. In recent years, this problem has been increasingly addressed through symbolic regression, which aims to automatically infer mathematical expressions that describe a system's dynamics. However, existing methods have struggled with overcoming two main limitations: fitting numerically-estimated derivatives—which introduces bias and amplifies noise; and restricting search space of to a narrow predefined set of model structures—which simplifies the search but risks missing the relevant models. Recent deep learning-based approaches alleviate these limitations but have common caveats in that they need to define heuristic cost functions and heuristic search algorithms, and it is rarely tested under which circumstances they attain the desired (ground-truth) results.

Here, we have introduced an integral formulation of existing Bayesian symbolic regression approaches. The integral formulation addresses the first limitation mentioned above by numerically integrating candidate models rather than numerically estimating derivatives. By working directly with observed or measured system primitives, we avoid the amplification of noise typically introduced by numerical differentiation, which allows for more accurate model evaluation and increases the chances of recovering the true underlying dynamics. Consequently, the models discovered are more often correct and better aligned with the system's real behavior. Indeed, we find that, for synthetic data, the integral Bayesian approach recovers the true governing equations in regions where, due to noise or lack of data, all benchmark approaches fail.

The integral formulation also addresses the second limitation in that it is able to explore the whole space of symbolic expressions, without being restricted to linear combinations of predefined library functions. Therefore, our framework flexibly accommodates complex closed-form equations and is able to propose intricate models that are inaccessible to methods such as SINDy or to other classical methods for system discovery, which rely on linear combinations of library functions. We have confirmed this by studying empirical data of bacterial growth. For this system we discover a relatively simple but nonlinear differential equation, which is a more parsimonious description of the empirical data than existing models.

Last but not least, our integral Bayesian approach is built on solid probabilistic arguments about data and model uncertainty and is thus Bayes optimal. This means that, if models where drawn from a known prior distribution $p(f)$, then our approach would lead to optimal selection of models and optimal predictions of behavior. In practice, we do not know the prior distribution. However, our experiments for synthetic data for which we have ground truth dynamics show that our approach makes optimal predictions about system dynamics, that is, predictions that have as little error as one may reasonably expect. Therefore, we argue that the proposed approach marks a significant step forward in data-driven discovery of governing equations and demonstrates the potential of integral Bayesian methods in advancing the automation of scientific modeling.

Methods

Synthetic data

We generated synthetic datasets for the one-dimensional logistic equation and the two-dimensional Lotka-Volterra model. For the logistic model we use the integrated equation:

$$x(t) = \frac{A}{B + e^{C - Dt}} \tag{9}$$

where $A=1$, $B=1$, $C=0$ and $D=1$. To generate data according to the model, we evaluated Eq. (9) for $t\in[-6,6]$ with a step $\delta t=0.1$, so that the number of observed points is $N=120$. For experiments with different noise levels, we generated 40 data sets for each Gaussian noise level $\mathcal{N}(0,\sigma)$, where σ is the standard deviation of the distribution. In experiments with a constant noise level but varying number of points N, we evaluated the logistic equation for $t\in[-6,6]$ with datasets comprising equally spaced points, so that the larger the number of points the smaller δt. For each value of N, we generated 40 datasets, as before.

The Lotka-Volterra model is defined as a system of coupled differential equations

$$\begin{cases}\dot{x}=ax-bxy\\ \dot{y}=dxy-cy\end{cases} \tag{10}$$

with $a=0.1$, $b=0.02$, $c=0.02$ and $d=0.4$. We integrate the equations from $t_0=0$ to $t_1=80$ with a time step of $\delta t=0.1$ and initial condition $x_0=10$, $y_0=5$. Then, we generated 40 datasets for the different levels of noise by adding Gaussian noise $\mathcal{N}(0,\sigma)$.

Differentiation methods For numerical differentiation we use the central difference method. This involves computing the derivative at each observed time point t_i using the values at the preceding and following time points, so that

$$x'(t_i)=\frac{x(t_{i+1})-x(t_{i-1})}{2h}, \tag{11}$$

where $x'(t)$ is the numerical derivative, $x(t)$ is the value of the observable at time t, and $h=t(i)-t(i+1)$ is the time step. For the first and last data points, we use the forward and backward finite difference methods, respectively.

To obtain smoothed derivatives, we use the Python package *PyNumDiff*[8], which involves fitting a polynomial to the data and differentiating the fitted curve. Specifically, we used a second-order polynomial and a window size of 21 data points to smooth the data before computing the derivative.

MCMC Model Sampling

To sample the posterior distribution over the space of models, we employ a Markov Chain Monte Carlo (MCMC) approach. The objective is to explore the space of possible mathematical models $\mathbf{f}$, and identify the model $\mathbf{f}^*$ with the highest posterior probability, which corresponds to the one that minimizes the description length.

The general procedure follows the standard BMS approach[26]. At each MCMC step, we attempt an expression update followed by a parallel tempering swap process. We begin by setting an initial expression as the starting configuration, typically a one-node tree. This initial node is usually a parameter but may also be a variable. It is important to note that the operations that we can use in the expression tree are: exp(), pow2(), pow3(), -, +, *, / and **. To ensure a better sampling of model space, we use parallel tempering. To that end, we clone the initial model (at temperature $T_0=1$) and assign a temperature $T_k=1.02^k$ for $k\in[1,20]$ to each replica. At each MCMC step, we propose a formula modification for each replica. This involves randomly altering the tree structure and computing the corresponding posterior of the modified expression. The description length associated with the posterior depends on whether we use the standard BMS[26] approach in Eq. (3) or the integrated I-BMS formulation Eq. (7).

Following the expression update, we propose a series of swaps adjacent replicas at (T_k,T_{k+1}). Finally, at the end of each MCMC step, we compare the description length of the model at T_0 and update the minimum description length model if a lower value is found.

Initial Parameter Guess In the I-BMS approach, we first fit the numerical derivative to obtain an initial estimate for the parameters of the differential equation. This estimate is then used as an initial guess for the optimization of the numerically integrated equations.

2-Dimensional I-BMS In the case of the 2D I-BMS, our objective is to identify a system of differential equations. Since the differential equations for each variable are coupled, their estimation must be handled simultaneously, both during parameter fitting and when computing the sum of squared errors. In each MCMC step, we first update the states for one variable across all temperature levels before proceeding to the other variable. When updating a single variable and computing the description length, we first determine the optimal parameters. These parameters are adjusted such that the numerical integration of the system of equations minimizes the discrepancy with the observed data. Next, we compute the sum of squared errors on the trajectory of the updated variable. If the proposed modification to $\mathbf{f}$ is accepted, we update the optimal parameters for both equations in the system.

Note that at a fixed fr replica k with $\mathbf{f}_k = (f_x, f_y)$, we propose changes for f_x and f_y separately, but that the computation of the description length associated to the proposed model implies the calculation of a new set of parameters for both f_x and f_y. For swaps between adjacent replicas in the temperature sequence we also propose separate swaps for f_x and f_y.

Exhaustive search of linear terms

We explored polynomial expressions up to a given degree and up to a given number of terms. For each dataset, we evaluated all possible linear combinations using both numerical differentiation and integration approaches. Sepcifically, our search space was:

- **Logistic Equation:** Linear terms up to order 4 ($S = 1, x, x^2, x^3, x^4$), generating polynomials with combinations up to 4 terms.
- **Lotka-Volterra Model:** Linear terms up to order 3 ($S = 1, x, y, x^2, y^2, xy, x^2y, xy^2$), generating polynomials with combinations up to 4 terms.

Physics-informed models

For modeling bacterial growth, we use the I-BMS approach to explore arbitrary models that incorporate mathematical constraints derived from phenomenological principles. For instance, we could enforce that component i includes an additive linear term so that $f_i = a\,x_i + f_{1i}(\mathbf{x})$. In implementation terms, each BMS instance maintains an expression tree f_{1i} that does not explicitly include these constraints but the full model needs to be used to compute the description length. Note that in order to assess whether the models found using these constraints have a lower description length than models obtained without constraints, we must use the full model f_i (and not f_{1i}) to compute the prior probability $p(f_i)$.

References

1. La Cava, W. *et al.* Contemporary symbolic regression methods and their relative performance. *Adv. Neural Inf. Process. Syst.* **2021**, 1 (2021).
2. Makke, N. & Chawla, S. Interpretable scientific discovery with symbolic regression: a review. *Artif. Intell. Rev.* **57**, 2, 10.1007/s10462-023-10622-0 (2024).
3. Evans, J. & Rzhetsky, A. Machine science. *Science* **329**, 399–400 (2010).

4. Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60, 10.1038/s41586-023-06221-2 (2023).

5. Cornelio, C. *et al.* Combining data and theory for derivable scientific discovery with AI-Descartes. *Nat. Commun.* **14**, 1777 (2023).

6. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci.* **113**, 3932–3937, 10.1073/pnas.1517384113 (2016). https://www.pnas.org/doi/pdf/10.1073/pnas.1517384113.

7. Quade, M., Abel, M., Shafi, K., Niven, R. K. & Noack, B. R. Prediction of dynamical systems by symbolic regression. *Phys. Rev. E* **94**, 012214, 10.1103/PhysRevE.94.012214 (2016).

8. van Breugel, F., Nathan Kutz, J. & Brunton, B. W. Numerical differentiation of noisy data: A unifying multi-objective optimization framework. *IEEE Access* 10.1109/ACCESS.2020.3034077 (2020).

9. de Silva, B. *et al.* PySINDy: A Python package for the sparse identification of nonlinear dynamical systems from data. *J. Open Source Softw.* **5**, 2104, 10.21105/joss.02104 (2020).

10. He, Y., Kang, S.-H., Liao, W., Liu, H. & Liu, Y. Robust identification of differential equations by numerical techniques from a single set of noisy observation. *SIAM J. on Sci. Comput.* **44**, A1145–A1175, 10.1137/20M134513X (2022).

11. Egan, K., Li, W. & Carvalho, R. Automatically discovering ordinary differential equations from data with sparse regression. *Commun Phys* **7** (2024).

12. Niven, R. K., Mohammad-Djafari, A., Cordier, L., Abel, M. & Quade, M. Bayesian identification of dynamical systems. *Proceedings* **33**, 10.3390/proceedings2019033033 (2019).

13. Niven, R. K., Cordier, L., Mohammad-Djafari, A., Abel, M. & Quade, M. Dynamical system identification, model selection, and model uncertainty quantification by bayesian inference. *Chaos: An Interdiscip. J. Nonlinear Sci.* **34**, 083140, 10.1063/5.0200684 (2024).

14. Fasel, U., Kutz, J. N., Brunton, B. W. & Brunton, S. L. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proc. R. Soc. A* **478**, 20210904, 10.1098/rspa.2021.0904 (2022). Publisher: Royal Society.

15. Mangan, N. M., Kutz, J. N., Brunton, S. L. & Proctor, J. L. Model selection for dynamical systems via sparse regression and information criteria. *Proc. Royal Soc. A: Math. Phys. Eng. Sci.* **473**, 20170009, 10.1098/rspa.2017.0009 (2017).

16. Reinbold, P. A. K., Gurevich, D. R. & Grigoriev, R. O. Using noisy or incomplete data to discover models of spatiotemporal dynamics. *Phys. Rev. E* **101**, 010203, 10.1103/PhysRevE.101.010203 (2020).

17. Messenger, D. A. & Bortz, D. M. Weak SINDy: Galerkin-based data-driven model selection. *Multiscale Model. & Simul.* **19**, 1474–1497, 10.1137/20M1343166 (2021).

18. Schaeffer, H. & McCalla, S. G. Sparse model selection via integral terms. *Phys. Rev. E* **96**, 023302, 10.1103/PhysRevE.96.023302 (2017).

19. Cranmer, M. *et al.* Discovering symbolic models from deep learning with inductive biases. *Adv. Neural Inf. Process. Syst.* **33**, 17429–17442 (2020).

20. Liu, B., Luo, W., Li, G., Huang, J. & Yang, B. Do we need an encoder-decoder to model dynamical systems on networks? In *Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI*, 2178–2186 (2023).

21. Hu, J., Cui, J. & Yang, B. Learning interpretable network dynamics via universal neural symbolic regression. *Nat Commun* **16**, 6226, https://doi.org/10.1038/s41467-025-61575-7 (2025).

22. Stolle, R. & Bradley, E. *Communicable Knowledge in Automated System Identification*, 17–43 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007).

23. Mangiarotti, S. & Huc, M. Can the original equations of a dynamical system be retrieved from observational time series? *Chaos* **29**, 023133 (2019).

24. Omejc, N., Gec, B., Brence, J., Todorovski, L. & Džeroski, S. Probabilistic grammars for modeling dynamical systems from coarse, noisy, and partial data. *Mach. Learn.* **113**, 7689–7721 (2024).

25. Fajardo-Fontiveros, O. *et al.* Fundamental limits to learning closed-form mathematical models from data. *Nat. Commun.* **14**, 1043, 10.1038/s41467-023-36657-z (2023).

26. Guimerà, R. *et al.* A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Sci. Adv.* **6**, eaav6971, 10.1126/sciadv.aav6971 (2020). https://www.science.org/doi/pdf/10.1126/sciadv.aav6971.

27. Schwarz, G. Estimating the Dimension of a Model. *The Annals Stat.* **6**, 461 – 464, 10.1214/aos/1176344136 (1978).

28. Bartlett, D. J., Desmond, H. & Ferreira, P. G. Exhaustive symbolic regression. *IEEE Transactions on Evol. Comput.* **28**, 950–964, 10.1109/TEVC.2023.3280250 (2024).

29. Reichardt, I., Pallarès, J., Sales-Pardo, M. & Guimerà, R. Bayesian machine scientist to compare data collapses for the Nikuradse dataset. *Phys. Rev. Lett.* **124**, 084503, 10.1103/PhysRevLett.124.084503 (2020).

30. Cuevas, D. *et al.* Elucidating genomic gaps using phenotypic profiles [version 2; peer review: 1 approved, 1 approved with reservations]. *F1000Research* **3**, 10.12688/f1000research.5140.2 (2016).

31. Sanchez, S. E. *et al.* Phage phenomics: Physiological approaches to characterize novel viral proteins. *JoVE* e52854, doi:10.3791/52854 (2015).

32. Wachenheim, D. E., Patterson, J. A. & Ladisch, M. R. Analysis of the logistic function model: derivation and applications specific to batch cultured microorganisms. *Bioresour. Technol.* **86**, 157–164, https://doi.org/10.1016/S0960-8524(02)00149-9 (2003).

33. Wang, J. & Guo, X. The gompertz model and its applications in microbial growth and bioproduction kinetics: Past, present and future. *Biotechnol. Adv.* **72**, 108335, https://doi.org/10.1016/j.biotechadv.2024.108335 (2024).

34. Richards, F. J. A flexible growth function for empirical use. *J. Exp. Bot.* **10**, 290–301.

35. Camacho-Mateu, J., Lampo, A., Castro, M. & Cuesta, J. A. Microbial populations hardly ever grow logistically and never sublinearly. *Phys. Rev. E* **111**, 044404, 10.1103/PhysRevE.111.044404 (2025).

Acknowledgments

This research was supported by project PID2022-142600NB-I00 (MS-P and RG), and FPI grant PRE2020-095552 (OC-T) from MCIN/AEI/10.13039/501100011033, and by the Government of Catalonia (2021SGR-633) (MS-P and RG).

Author contributions statement

OCT, MSP and RG designed the research. OCT wrote code and performed experiments. OCT, SCL, MSP and RG analyzed all results. SCL, SES and FLR collected and processed data and analyzed results on bacterial growth. All authors wrote the paper.

Data and code availability

The datasets and code to replicate the experiments are available from Github.

Competing interests

The authors declare no conflict of interest.