



Research



Cite this article: Guimerà R, Sales-Pardo M.

2026 Bayesian symbolic regression: automated equation discovery from a physicist's perspective. *Phil. Trans. R. Soc. A* **384**: 20250089.

<https://doi.org/10.1098/rsta.2025.0089>

Received: 22 July 2025

Accepted: 2 September 2025

One contribution of 15 to a discussion meeting issue 'Symbolic regression in the physical sciences'.

Subject Areas:

statistical physics, applied mathematics, mathematical modelling, computational physics

Keywords:

symbolic regression, Bayesian inference, computational science, equation discovery

Author for correspondence:

Roger Guimerà

e-mail: roger.guimera@urv.cat

Bayesian symbolic regression: automated equation discovery from a physicist's perspective

Roger Guimerà^{1,2,3} and Marta Sales-Pardo^{1,2}

¹Department of Chemical Engineering, and ²Center for Computational Science and Applied Mathematics (ComSCIAM), Universitat Rovira i Virgili, Tarragona, Catalonia, Spain

³ICREA, Barcelona, Catalonia, Spain

RG, 0000-0002-3597-4310; MS-P, 0000-0002-8140-6525

Symbolic regression automates the process of learning closed-form mathematical models from data. Standard approaches to symbolic regression, as well as newer deep learning approaches, rely on heuristic model selection criteria, heuristic regularization and heuristic exploration of model space. Here, we discuss the probabilistic approach to symbolic regression, an alternative to such heuristic approaches with direct connections to information theory and statistical physics. We show how the probabilistic approach establishes model plausibility from basic considerations and explicit approximations, and how it provides guarantees of performance that heuristic approaches lack. We also discuss how the probabilistic approach compels us to consider model ensembles, as opposed to single models.

This article is part of the discussion meeting issue 'Symbolic regression in the physical sciences'.

1. Introduction

It took 4 years for Kepler to establish that Mars' orbit is elliptical, in 1609; and it was not until 1687 that Newton unified his empirical observations into a mathematical model. Can we design computer programmes and theoretical frameworks to automate this process and make it faster? Can machine learning revolutionize science as it is revolutionizing other areas of our lives? [1,2]. At least since the 1970s, some researchers have thought so and have tried to develop algorithms that automatically learn

© 2026 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, provided the original author and source are credited.

closed-form models from data [3,4]. Under diverse names such as computational scientific discovery, equation discovery or, more recently, symbolic regression, this field has grown, matured and is becoming established within machine learning.

Traditional symbolic regression uses genetic algorithms to evolve populations of expressions that are ‘well adapted’ to the data at hand [5–7], that is, closed-form expressions that (i) describe the data well and (ii) are reasonably simple. To operationalize these two criteria, two loss functions are heuristically defined, one for error and one for complexity. They are eventually combined, again heuristically, leading to a single unified model-selection criterion. More modern approaches have recently been proposed based on sparse regression [8], recurrent neural networks [9], variational autoencoders [10] or a combination of neural networks with physics-inspired techniques [11], among others (for systematic reviews, see [12,13]). However, despite their differences, these modern methods share with traditional symbolic regression (at least to some degree) the need to define loss functions and model selection criteria heuristically. The way they explore model space and come up with specific models is also heuristic. In addition, both in traditional and modern methods, performance is typically evaluated based on benchmark problems or datasets, which is problematic because: (i) no benchmark can possibly include all situations that are potentially of interest and, therefore, any benchmark will be biased in certain ways; (ii) benchmarks lead to human-in-the-loop overfitting [14] because, once a benchmark is well established in the community, algorithms are designed and refined based on the performance on *those* benchmarks.

Here, we argue that symbolic regression should move away from heuristics and methods should be built to conform to some basic desiderata. In particular, we demand from symbolic regression approaches the following properties:

- (i) They must establish the plausibility of models based on rigorous arguments and, when necessary, explicit—and hence scrutinizable—assumptions and/or approximations.
- (ii) They must integrate goodness of fit and model complexity into a single measure of plausibility, so that no ad hoc parameters or thresholds need to be fixed to balance them.
- (iii) They must be consistent; that is, they must select the true model with probability approaching one as the sample size grows to infinity.
- (iv) They must account for the uncertainty inherent in the model discovery process.

We show that a probabilistic (or Bayesian) approach to symbolic regression [15–17] satisfies all of these conditions. This approach draws from probability theory, information theory and statistical physics and, we believe, provides a solid foundation for future developments in the area.

2. Bayesian symbolic regression

In symbolic regression, we aim to identify the closed-form mathematical model $m^*(\mathbf{x}, \theta^*)$ that is responsible for the generation of some observed dependent variables $\{y^k\}$ through the process

$$y^k = m^*(\mathbf{x}^k, \theta^*) + \epsilon^k. \quad (2.1)$$

Here, \mathbf{x}^k is the k th observation of the features or independent variables, θ^* are some parameters of the model m^* and ϵ^k is an observational noise, which is typically assumed to be Gaussian-distributed with zero mean and unknown variance σ^2 . The space of candidate models \mathcal{M} comprises, in principle, all possible closed-form models $m_i(\mathbf{x}, \theta_i)$, although in some situations, one may want to restrict the space to certain subsets of models.

Given that there is uncertainty in both the data generation process and in the model selection itself, the most complete description of the symbolic regression problem is probabilistic [18]. Indeed, given some observed data $D = \{(y^k, \mathbf{x}^k), k = 1, \dots, N\}$, the complete solution to the problem is given by the conditional probability $p(m_i|D)$ of m_i being the true generating model given the data D . Indeed, given this posterior distribution $p(m_i|D)$ over models $m_i \in \mathcal{M}$, we can answer any

model-selection question (for example, what is the most plausible model?) or make any prediction (for example, what is the probability that y is larger than a certain value at some point \mathbf{x}). The practical question is, then, whether $p(m_i|D)$ can be computed or, at least, approximated.

The answer to this question is that, under relatively mild approximations, $p(m_i|D)$ can indeed be computed. First, consider the joint distribution $p(m_i, \theta_i|D)$ of the model m_i and its parameters θ_i , given D . This distribution can be written in terms of the likelihood $p(D|m_i, \theta_i)$ by application of Bayes' theorem:

$$p(m_i, \theta_i|D) = \frac{p(D|m_i, \theta_i)p(m_i, \theta_i)}{p(D)} = \frac{p(D|m_i, \theta_i)p(\theta_i|m_i)p(m_i)}{p(D)}. \quad (2.2)$$

Then, one can calculate the posterior distribution $p(m_i|D)$ by marginalizing over parameter values

$$\begin{aligned} p(m_i|D) &= \int_{\theta_i} d\theta_i p(m_i, \theta_i|D) \\ &= \frac{p(m_i)}{p(D)} \int_{\theta_i} d\theta_i p(D|m_i, \theta_i) p(\theta_i|m_i) \\ &= \frac{\exp[-\mathcal{L}(m_i, D)]}{Z}, \end{aligned} \quad (2.3)$$

where the last step is simply notation and can be regarded as the definition of $\mathcal{L}(m_i, D)$, and $Z = p(D)$ is introduced to make it clear that, in the context of model selection, $p(D)$ is just a model-independent normalizing constant.

In general, the integral in equation (2.3) cannot be evaluated analytically because the model $m_i(\mathbf{x}, \theta_i)$ and the prior $p(\theta_i|m_i)$ may have arbitrarily complex dependencies on the parameters θ_i . However, the integral can be estimated by means of the Laplace approximation [19] by assuming that (i) the likelihood is sufficiently peaked around the parameter values $\hat{\theta}_i$ that maximize the likelihood and (ii) the prior $p(\theta_i|m_i)$ is sufficiently smooth around $\hat{\theta}_i$. Under these conditions, we have that

$$\mathcal{L}(m_i, D) = \frac{B(m_i, D)}{2} - \log p(m_i), \quad (2.4)$$

where $B(m_i, D)$ is given by [19]

$$B(m_i, D) = -2 \log p(D|m_i, \hat{\theta}_i) + (k_i + 1) \log N + \log |\mathcal{I}(\hat{\theta}_i)|. \quad (2.5)$$

Here, k_i is the number of parameters in model m_i (that is, the dimension of θ_i),¹ $\log p(D|m_i, \hat{\theta}_i)$ is the log-likelihood calculated at the maximum likelihood estimator of the parameters (including σ) and $\mathcal{I}(\hat{\theta}_i)$ is the Fisher information matrix, calculated at the maximum likelihood estimators of the parameters. The Fisher information matrix represents the curvature of the likelihood around its maximum at $\hat{\theta}_i$, so that, other things being equal, models with small curvature (that is, models for which changes in parameter values produce small changes in model predictions for observed data points) are preferred over models with a large curvature.

Since the last term in equation (2.5) is the only one whose absolute value does not grow with N , it is asymptotically small compared to the other terms and, therefore, it is often ignored. Within this additional approximation, we have that

$$B(m_i, D) = -2 \log p(D|m_i, \hat{\theta}_i) + (k_i + 1) \log N, \quad (2.6)$$

and B is then the so-called Bayesian information criterion (BIC) [20]. The BIC is often used in model selection in general [19] and in Bayesian symbolic regression in particular [15–17,21], although the approximation in equation (2.5) has also been used, more recently, within this context [22].

¹Note that the +1 in the term $(k_i + 1)$ arises from the other parameter of the probabilistic model, that is, the observational noise σ . If this value is known, then this term should be $k_i \log N$. See also footnote 2.

(a) Likelihood function

The exact form of the likelihood function $p(D|m_i, \theta_i)$ depends on the precise modelling assumptions. Under the standard assumptions in regression and symbolic regression, which we have outlined in equation (2.1) (namely, that the observed dependent variables $\{y^k\}$ are subject to additive, unbiased, Gaussian noise with unknown variance σ^2), the likelihood is given by

$$\begin{aligned} p(D|m_i, \theta_i) &= \prod_{k=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y^k - m_i(\mathbf{x}^k, \theta_i))^2}{2\sigma^2}\right] \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{\sum_{k=1}^N (y^k - m_i(\mathbf{x}^k, \theta_i))^2}{2\sigma^2}\right]. \end{aligned} \quad (2.7)$$

The corresponding term in the BIC is then

$$-\log p(D|m_i, \hat{\theta}_i) = \frac{N}{2} \left[\log 2\pi + \log \frac{\sum_k (y^k - m_i(\mathbf{x}^k, \hat{\theta}_i))^2}{N} + 1 \right], \quad (2.8)$$

where we have already replaced the maximum likelihood estimator $\hat{\sigma}^2 = \frac{\sum_k (y^k - m_i(\mathbf{x}^k, \hat{\theta}_i))^2}{N}$ for the variance.

Under different modelling assumptions, the likelihood may take a different form. For example, consider a situation in which each y^k measurement has an uncertainty σ_k associated with it. In that case, a reasonable likelihood would be²

$$p(D|m_i, \theta_i) = \left(\prod_{k=1}^N \frac{1}{\sqrt{2\pi\sigma_k^2}} \right) \exp\left[-\sum_{k=1}^N \frac{(y^k - m_i(\mathbf{x}^k, \theta_i))^2}{2\sigma_k^2}\right].$$

Similarly, in some situations, it may be appropriate to use likelihoods corresponding to non-Gaussian or non-additive noise, or other suitable likelihood functions.

3. Interpretations of the Bayesian approach

(a) Probabilistic interpretation

The probabilistic interpretation of the symbolic regression approach outlined above should be clear—each expression m_i has a probability $p(m_i|D)$ of being the true generating model, and the most plausible model \hat{m} is the maximum a posteriori

$$\hat{m} = \arg \max_{m_i} p(m_i|D). \quad (3.1)$$

In the probabilistic interpretation, the posterior probability $p(m_i|D)$ is obtained by updating our prior expectations $p(m_i)$ about models with the marginal likelihood

$$\int_{\theta_i} d\theta_i p(D|m_i, \theta_i) p(\theta_i|m_i),$$

which can be approximated, leading to equations (2.4)–(2.6).

Two important considerations, in this respect. First, the prior $p(m_i)$ does play a role in estimating the posterior $p(m_i|D)$ —ignoring the prior amounts to assuming that all models m_i are, in

²Note that, in this case, the variance σ^2 of the observational noise would *not* be a parameter of the probabilistic model. Therefore, in this case, the $(k+1) \log N$ term in the BIC would become $k_i \log N$.

principle, equally plausible; and since there are exponentially many more complex models than simple models, it amounts to assuming that, in principle, complex models are more plausible than simple models. That being said, since the prior is fixed (intensive) and the marginal likelihood grows linearly with the number of observations N (is extensive), in the limit $N \rightarrow \infty$ the prior washes out; that is, asymptotically, our prior expectations do not matter (as long as we do not assign $p(m_i) = 0$ to any model).

Second, our prior expectations get updated by the integrated marginal likelihood, not the maximum likelihood or any other point estimate of the likelihood. This is important because a model m_i may fit the data well for a specific choice $\hat{\theta}_i$ of parameters, but poorly for other choices; and, since we are not certain about the exact values of the parameters, all values, good and bad, should be taken into consideration when evaluating the plausibility of the model. This is what happens, for example, to models with many parameters—one may find a good combination $\hat{\theta}_i$, but the volume of models with poor fit grows with the dimension of the parameter space. This is the origin of the term that scales with the number of parameters $k_i + 1$ in B , which is sometimes interpreted as an heuristic regularization term but is, as we have seen, an unavoidable consequence of the application of probability theory.

(b) Information-theoretic interpretation

From equation (2.3), one can see that

$$\mathcal{L}(m_i, D) = -\log p(m_i, D) = -\log p(D|m_i) - \log p(m_i), \quad (3.2)$$

where $p(D|m_i) = \int_{\Theta_i} d\theta_i p(D|m_i, \theta_i) p(\theta_i|m_i)$ is the marginal likelihood.

In information-theoretic terms, $\mathcal{L}(m_i, D)$ is the description length, that is, the number of nats (or bits, if we used base-2 logarithms instead of natural logarithms) necessary to convey m_i and the data D to a receiver by means of an optimal code [23]. Then, from equation (2.3), it is clear that the most plausible model $\hat{m} = \arg \max_{m_i} p(m_i|D) = \arg \min_{m_i} \mathcal{L}(m_i, D)$ is the one with the minimum description length. This means that \hat{m} is the model that allows the sender to convey that data most efficiently, that is, the model that best compresses the data.

The description length has two parts. The term $-\log p(m_i)$ corresponds to the number of nats necessary to convey model m_i among all possible models. The more plausible the model is *a priori*, the smaller this term. The second term $-\log p(D|m_i)$ corresponds to the number of nats necessary to convey the data once model m_i is specified. If m_i provides a better description of the data, then we need fewer nats in addition to the model itself to describe the data. Importantly, both terms are in the same ‘units’ and are therefore comparable—description length nats provide a unified measure of model complexity and goodness of fit.

(c) Statistical physics interpretation

Finally, the Bayesian approach can also be interpreted in the context of the canonical ensemble in statistical physics. Indeed, consider a physical system whose configurations are c_i . The probability of observing configuration c_i in such a system is given by the Boltzmann distribution

$$p(c_i) = \frac{\exp[-\beta \mathcal{H}(c_i)]}{Z}, \quad (3.3)$$

with $Z = \sum_j \exp[-\beta \mathcal{H}(c_j)]$ being the partition function, $\mathcal{H}(c_i)$ being the energy of configuration c_i and β being the inverse of the temperature.

Then, by equation (2.3), each model m_i in symbolic regression can be interpreted as a configuration whose energy is $\mathcal{L}(m_i, D)$, for a system at temperature $\beta = 1$. In the context of information field theories, \mathcal{L} is called the information Hamiltonian [24]. In this picture, the most plausible model $\hat{m} = \arg \max_{m_i} p(m_i|D) = \arg \min_{m_i} \mathcal{L}(m_i, D)$ is the one with the lowest energy, that is, the ground state of the system.

4. Arguments for a Bayesian approach

Consider the problem of choosing among states $s_i \in \mathcal{S}$, for example the sides of a biased dice. If an oracle establishes $p(s_i|X)$ using probability theory and all imaginable information X about the dice and the dice-rolling process, then it is impossible (on average) to beat the oracle at predicting dice outcomes. For example, betting on the outcome of a single dice roll, the oracle would always predict $\arg \max_{s_i} p(s_i|X)$ and no player could do better than that on average, no matter what algorithm they used.

Analogously, in the context of model selection and symbolic regression, consider a situation in which a model $m_i \in \mathcal{M}$ and its parameters θ_i are drawn from the known prior distribution $p(m_i, \theta_i) = p(\theta_i|m_i)p(m_i)$ and data D are generated according to the known process described by equation (2.1). Then, the product rule of probability theory [18] dictates that $p(m_i, \theta_i|D) \propto p(D|m_i, \theta_i)p(m_i, \theta_i)$, with the likelihood $p(D|m_i, \theta_i)$ given by equation (2.7). The sum rule [18], on the other hand, dictates that $p(m_i|D) = \int d\theta_i p(m_i, \theta_i|D)$. Therefore, in this scenario (and aside from the Laplace approximation, which becomes exact as the number N of data points grows), there is no better estimation of $p(m_i|D)$ than in the probabilistic approach outlined above. In this sense, the proposed approach is Bayes optimal—it achieves the best possible performance, and no other algorithm can outperform it on average. Given the different axiomatizations of probability theory, this statement translates into different arguments for the use of the Bayesian approach as opposed to heuristic approaches.

Cox-type argument: Cox's theorem establishes that any system of reasoning under uncertainty that satisfies certain basic consistency and common sense requirements must necessarily follow the laws of probability theory [18,25]. Therefore, it justifies the use of probability as the unique consistent framework for quantifying degrees of belief. Therefore, any way to assign plausibilities to models not conforming to equation (2.3) must violate some of the very basic commonsensical conditions assumed by Cox.

Consistency argument: Related to the previous argument, the Bayesian approach outlined above is consistent, that is, in the large N limit, it will select the true model with probability approaching one. In fact, this is true even if the prior $p(m)$ is unknown, because the marginal likelihood is extensive in N , whereas the prior is intensive. Therefore, any alternative that does not coincide with the Bayesian approach in this limit is virtually guaranteed to select the wrong model.

Minimum description length argument: As discussed above, the Bayesian approach selects the model with the shortest description length, that is, the model that maximally compresses the data. Any alternative way of selecting models will lead to models that compress the data less, that is, models that are objectively less parsimonious than those selected by the Bayesian approach.

Dutch book argument: In de Finetti's axiomatization, a probability is one's degree of belief in an event's occurrence, quantified as the price they would be willing to spend on a fair bet that pays one unit on the occurrence of the event [26]. In this context, a Dutch book is a set of bets constructed to exploit non-probabilistic beliefs, guaranteeing a sure loss to those not using probability theory, no matter how the events unfold. Betting on symbolic regression models using any assignment of plausibility other than the Bayesian approach results in Dutch books, that is, in certain loss.

5. Traditional heuristic approaches under the light of the Bayesian approach

We hope that the reader finds the arguments in favour of the Bayesian approach in the previous sections convincing. However, one may still wonder how important these considerations are in practice. Here, we address this question by comparing the Bayesian approach to traditional heuristic approaches, both on theoretical grounds and in two simple scenarios that illustrate the failure of heuristic approaches to fulfil the basic desiderata outlined in the introduction.

We start by outlining how traditional symbolic regression works. First, an arbitrary loss function is defined, typically the squared error, which is equivalent to maximizing the likelihood $p(D|m_i, \hat{\theta}_i)$ in equation (2.8). Second, some algorithm, typically a genetic algorithm, is used to

find models that minimize the loss. However, given a dataset D , the likelihood can always be made arbitrarily close to 1 by considering arbitrarily complex models. To escape this ‘structural overfitting’, [15] traditional symbolic regression proceeds by defining an heuristic measure of complexity, typically related to the number of operations and/or parameters in the model. Based on this complexity, one defines a Pareto front comprising the models that have minimum loss at each value of complexity. Finally, among all models in the Pareto front, one is typically selected by identifying (again, heuristically) the ‘elbow’ of the front, that is, the point at which, somehow, the loss increases maximally for a fixed reduction in complexity.

All in all, the traditional approach involves three heuristic choices: loss, complexity and model selection criterion within the Pareto front (elbow). By contrast, the Bayesian approach does not require an heuristic definition of loss—the description length $\mathcal{L}(m_i, D)$ is the quantity to minimize (or, equivalently, the posterior $p(m_i|D)$ is the quantity to maximize) as prescribed by probability theory. Similarly, there is no need to select models within the Pareto front because, as we have argued, $\mathcal{L}(m_i, D)$ already combines goodness of fit and complexity within a single metric—the term quantifying goodness of fit is the marginal likelihood $-\log p(D|m_i)$,³ whereas the term quantifying complexity is the prior $-\log p(m_i)$.

Here, we follow previous work using a maximum entropy prior [15] (that is, a prior that maximizes the entropy $S = -\sum_{m_i} p(m_i) \log p(m_i)$) subject to (i) having the mean number $\langle n_o \rangle = \sum_{m_i} n_o(m_i) p(m_i)$ of operations of each type $o \in \{+, \times, \exp, \log, \sin, \cos \dots\}$ as in an empirical corpus of mathematical formulas;⁴ (ii) having the mean squared number of each operation $\langle n_o^2 \rangle = \sum_{m_i} n_o^2(m_i) p(m_i)$ as in the empirical corpus. The maximum entropy prior satisfying these conditions is

$$p(m_i) = \exp \left[-\sum_o (\alpha_o n_o(m_i) + \beta_o n_o^2(m_i)) \right]. \quad (5.1)$$

This is the most agnostic prior with the empirically observed frequencies of each operator and the empirically observed fluctuations of these frequencies. Alternatively, one could select other reasonable priors, including more informative priors encoding available background knowledge in a given context [27,28]. The important consideration here is that, unlike traditional approaches, the choice of prior is explicit and transparent, and therefore scrutinizable.

To compare traditional and Bayesian symbolic regression in practice, we consider two simple scenarios. In the first one, we generate data through the process $y^k = \theta_0^* + \theta_1^* x + \epsilon^k$, so that $m^*(x^k, \theta^*) = \theta_0^* + \theta_1^* x$, with $\theta_0^* = -2.3$ and $\theta_1^* = 4.1$. We then use traditional symbolic regression (using PySR [7] with default parameters) and Bayesian symbolic regression (using the Bayesian machine scientist available at <https://bitbucket.org/rguimera/machine-scientist>)⁵ to learn m^* . We repeat the process for varying numbers of data points $N \in \{10, 100, 1000\}$ and different levels of observational noise $\sigma \in \{0.05, 0.5, 5, 50\}$ (figure 1).

This is a very simple model, where one may expect symbolic regression to work. Indeed, the Bayesian approach generally identifies the correct model—although in the high-noise regime, it underfits the data, in those cases, the error between the identified model and the true model (reducible error) is very small compared to the noise level σ (irreducible error), so underfitting is actually reasonable. In practice, with such observational noise, making predictions with the true model or the underfit model would lead to almost identical errors, because error is dominated by the irreducible error σ .

The traditional approach also learns the correct model in the low-noise regime. However, when noise is high, it overfits the data very significantly, even when the number of points is large; and, in

³The term $(k_i + 1) \log N$ in equations (2.6) and (2.5) comes from approximating the marginal likelihood. Therefore, although often interpreted as a penalty to parametric complexity, it seems more appropriate to consider this term as part of the goodness of fit.

⁴Here, $n_o(m_i)$ is the number of times that operator o appears in m_i .

⁵In situations in which model space can be somehow restricted, one may use exhaustive symbolic regression [22]. However, here we do not restrict models in any way, so exhaustive search is unfeasible.

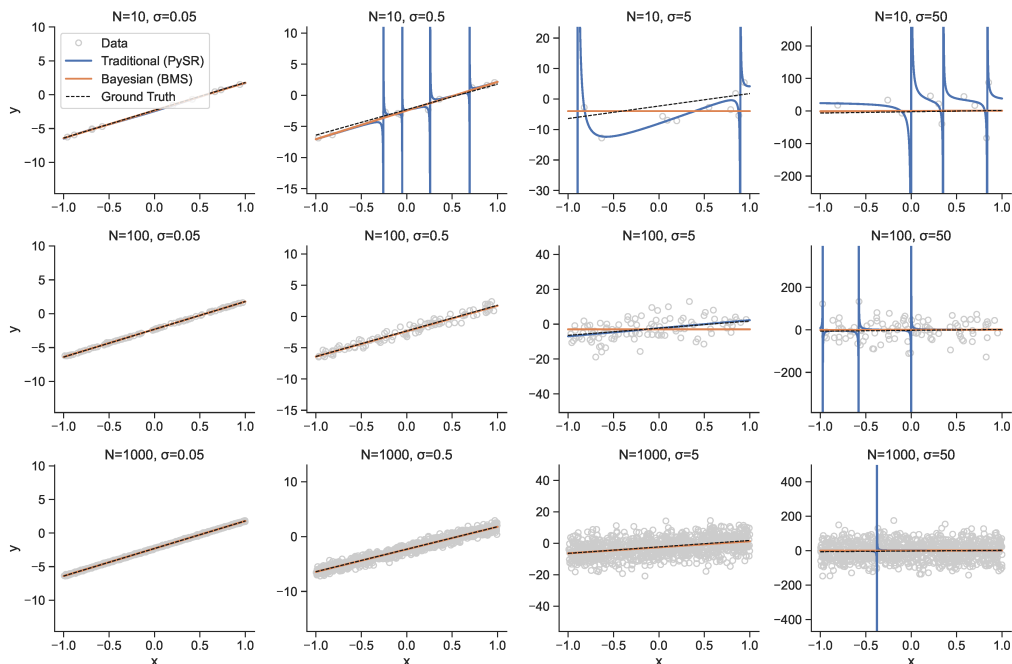


Figure 1. For varying numbers of data points $N \in \{10, 100, 1000\}$ and different levels of observational noise $\sigma \in \{0.05, 0.5, 5, 50\}$, we generate data (grey symbols) through the process $y^k = \theta_0^* + \theta_1^* x + \epsilon^k$, with $\theta_0^* = -2.3$ and $\theta_1^* = 4.1$. We then use traditional heuristic symbolic regression (using PySR [7]; blue lines) and Bayesian symbolic regression (using the Bayesian machine scientist (BMS) [15]; orange lines) to learn m^* . We run PySR with default parameters: $\sigma \in \{+, -, \times, /\}$, maximum equation complexity 30, 100 iterations, 31 populations with 27 individuals in each population and 380 cycles in each iteration. Model selection was also the default in PySR, that is, we used a quadratic loss and defined the score as minus the derivative of the log-loss with respect to complexity; and we selected the model with the highest score among expressions with a loss lower than 1.5 times the loss of the most accurate model. We ran the BMS for 1000 Markov chain Monte Carlo steps, with 20 temperatures for the parallel tempering (see [15] for details).

this case, the reducible error is not necessarily small compared to the observational noise. The tendency of the traditional approach to overfit can be understood in light of the Bayesian approach. Indeed, as we have argued, probability theory dictates that we select models by minimizing the description length

$$\mathcal{L}(m_i, D) = -\log p(m_i) - \log p(D|m_i, \hat{\theta}_i) + \frac{(k_i + 1)}{2} \log N + \dots, \quad (5.2)$$

where the dots indicate additional terms in the approximation of the exact marginal likelihood. In practice, traditional symbolic regression aims to minimize squared errors and, thus, to maximize the likelihood, so that the loss is

$$\mathcal{L}_{trad}(m_i, D) = -\log p(D|m_i, \hat{\theta}_i). \quad (5.3)$$

By comparing the last two expressions, we note that the prior effectively being used by the traditional approach is

$$\log p_{trad} = \frac{(k_i + 1)}{2} \log N + \dots, \quad (5.4)$$

that is, the traditional approach is favouring *a priori* models with *more*, rather than fewer, parameters. In fact, traditional approaches favour everything that the successive approximate terms of the marginal likelihood *penalize*.

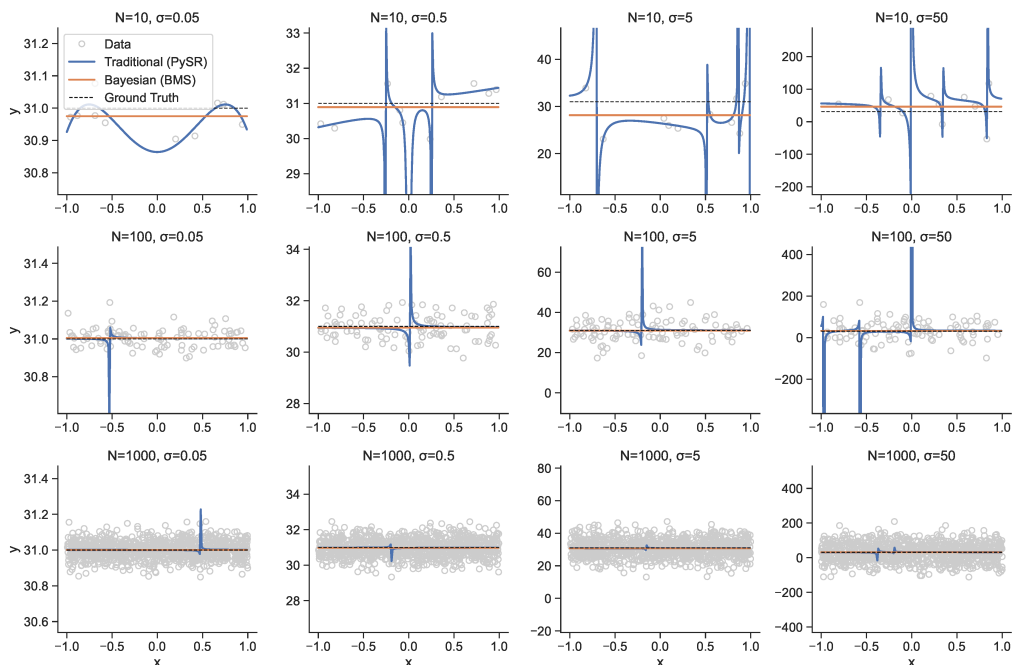


Figure 2. For varying numbers of data points $N \in \{10, 100, 1000\}$ and different levels of observational noise $\sigma \in \{0.05, 0.5, 5, 50\}$, we generate data (grey symbols) through the process $y^k = \theta_0^* + \epsilon^k$, with $\theta_0^* = 31$. We then use traditional heuristic symbolic regression (using PySR [7]; blue lines) and Bayesian symbolic regression (using the Bayesian machine scientist (BMS) [15]; orange lines) to learn m^* . The parameters of PySR and BMS are the same as in figure 1.

In the second experiment, we generate data through an even simpler process $y^k = \theta_0^* + \epsilon^k$, so that $m^*(x^k, \theta^*) = \theta_0^* = \text{const.}$, with $\theta_0^* = 31$ (figure 2). Bayesian symbolic regression always identifies the correct model, although noise leads to estimates of the parameter $\hat{\theta}_0$ that deviate from the exact real value. By contrast, traditional symbolic regression fails to identify the correct model in every single instance and systematically overfits the data, even when noise is low and the number of points is high.

Besides the *a priori* preference for more complex models discussed in the previous experiment, the reason for the overfitting in this case is the heuristic used to select the best model within the Pareto front—since the constant model is the simplest possible model, it sits at the edge of the front and can never be considered an elbow. However, a linear model with a very small slope is also in the front and could, in principle, be selected. Rather, much more complex models are chosen in all cases, which means that, in this popular implementation of traditional symbolic regression, some complex model is always selected, even when no relationship whatsoever exists between dependent and independent variables. This example is thus sufficient to prove that the heuristics chosen do not lead to consistent model selection.

6. From single models to posterior distributions over models

So far, in line with traditional symbolic regression, we have focused on discussing how to get a single best model for a given dataset. In the Bayesian approach, we have identified this best model with the maximum *a posteriori*/minimum description length $\hat{m} = \arg \max_m p(m_i|D) = \arg \min_{m_i} \mathcal{L}(m_i, D)$. However, the Bayesian approach does not only give the most plausible model; it gives the whole posterior $p(m_i|D)$, which contains much more information than any single model $m_i \in \mathcal{M}$.

(a) Model averaging

Consider a situation in which one wants to predict the value of y for a certain value of \mathbf{x} , given the observed data D . A common approach to do this is to use the most plausible model $\hat{m}(\mathbf{x}, \hat{\theta})$ and predict $y = \hat{m}(\mathbf{x}, \hat{\theta})$. However, it is important to note that this is just an approximation because, in general, $p(\hat{m}|D) \ll 1$, that is, we have no certainty whatsoever that $\hat{m}(\mathbf{x}, \hat{\theta})$ is the true generating model. The statistical physics interpretation helps understand how incorrect this point estimate typically is—trying to predict y with model $\hat{m}(\mathbf{x}, \hat{\theta})$ alone is like trying to predict the properties of a physical system at some finite temperature using only the ground-state configuration.

Rather, the most complete description of y at \mathbf{x} is given by the posterior obtained through model averaging (or ensemble averaging) [29]:

$$p(y|D, \mathbf{x}) = \sum_{m_i} \int_{\theta_i} d\theta_i \delta(y - m_i(\mathbf{x}, \theta_i)) p(m_i, \theta_i|D), \quad (6.1)$$

which is hard to calculate but can be approximated computationally by

$$p(y|D, \mathbf{x}) \approx \sum_{m_i} \delta(y - m_i(\mathbf{x}, \hat{\theta}_i)) p(m_i|D) \approx \frac{1}{K} \sum'_{m_i} \delta(y - m_i(\mathbf{x}, \hat{\theta}_i)). \quad (6.2)$$

Here as before, $\hat{\theta}_i$ is the maximum likelihood estimator of the parameters of model m_i , and the primed sum \sum'_{m_i} indicates that the sum is over K models sampled from $p(m_i|D)$ using, for example, Markov chain Monte Carlo [15].

(b) Fundamental limits and Rashomon sets

Thinking about model averaging and model ensembles in the terms we have just discussed opens the door to deep and important questions about model space and the description length landscape. For example, is the true generating model always the ground state? And under what conditions is there a single model \hat{m} that is overwhelmingly more plausible than any other model? Or, conversely, when do we have multiple models with description lengths similar to the ground state?

Regarding the first question, analysis of the description length landscape leads to the conclusion that the ground truth generating model m^* does not always coincide with the ground state \hat{m} [17]. Let us see why. As we have argued above, the probabilistic approach is consistent; that is, it identifies the true generating model with probability tending to one in the limit $N \rightarrow \infty$; in this limit, we do have $\hat{m} = m^*$. However, for finite N , we can increase the observational noise σ and, intuitively, it seems reasonable that, at some point, m^* will become undetectable. This is indeed the case; and, in fact, this learnability transition (from a phase in which the true model can be learnt to a phase in which it cannot) is properly described by considering only two minima in the description length landscape, namely, the ground truth model m^* and the trivial model $m^0 = \arg \max_{m_i} p(m_i)$ that maximizes the prior over models [17]. Indeed, the mean description length of these two models as a function of the noise level σ can be estimated analytically, and the point where they cross provides a tight upper bound to the learnability transition [17].

Since, as we have argued, the probabilistic approach is Bayes optimal, proving the existence of such learnable-to-unlearnable transition amounts to proving that, above a certain noise level, no algorithm can possibly identify the ground truth model from observational data alone. This is a result that only an approach that meets the desiderata that we have previously outlined could have possibly obtained.

With regard to the other questions, it turns out that often there exist many models with description lengths similar to the ground state \hat{m} for a given dataset. In other contexts, such collections of similarly plausible and explanatory models have been called Rashomon sets [30].⁶ In the learnable phase, all models with description length similar to \hat{m} are similar to the ground truth, so the

⁶After the movie *Rashomon*, in which Akira Kurosawa concatenates a series of contradictory accounts of the same crime.

Rashomon set does not add much to the single best model \hat{m} . However, close to the learnability transition, a Rashomon set of diverse models emerges, which provides non-congruent descriptions of the same data [17]. In empirical datasets, this situation seems to be the norm rather than the exception [15,16,21].

7. Conclusion

Luís A. N. Amaral has recently argued that research on ‘artificial intelligence needs a scientific method-driven reset’ based on reliable use of ‘prior knowledge, falsifiable hypotheses, and rigorous experimentation’ [31]. This, we believe, is true in general but especially for applications of artificial intelligence in science and for symbolic regression in particular.

Here, we have compared probabilistic to traditional symbolic regression. Of course, in recent years, there has been an explosion of new symbolic regression methods based on large language models, variational autoencoders and a variety of other deep learning approaches. However, the main limitations we have identified and discussed here for traditional symbolic regression remain in these newer approaches, namely: (i) the need to define goodness of fit (or loss) and complexity measures heuristically; (ii) the need to choose models heuristically based on fit and complexity; and (iii) the need to explore model space heuristically. The probabilistic approach provides concrete and easy-to-implement alternatives to (i) and (ii), so we see no reason why all other approaches should not adopt them. With regard to (iii), heuristic search is acceptable for practical applications, but one must always keep in mind that, for certain advanced applications (such as model averaging for prediction, or analysis of model space for theoretical results like those related to learnability), sampling over the posterior distribution provides the most detailed description of the problem.

Symbolic regression can revolutionize the scientific process by automating the learning of closed-form mathematical models from data. However, for symbolic regression to advance on solid grounds, and to help other fields also advance on solid grounds by identifying models that are defensible, it must aim for the maximum levels of mathematical and conceptual rigour. In this article, we have argued that the interface between probability theory, information theory and statistical physics provides the ideal framework for this.

Data accessibility. Code for the Bayesian machine scientist is available from <https://bitbucket.org/rquimera/machine-scientist/>. Code for PySR is available from <https://github.com/MilesCranmer/PySR>.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. R.G.: conceptualization, formal analysis, funding acquisition, investigation, methodology, software, validation, visualization, writing—original draft; M.S.-P.: conceptualization, formal analysis, funding acquisition, investigation, methodology, validation, visualization, writing—original draft.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This research was funded by project PID2022-142600NB-I00 from MCIN/AEI/10.13039/501100011033 FEDER, UE and by the Government of Catalonia (2021SGR-633).

References

1. Wang H *et al.* 2023 Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60. (doi:10.1038/s41586-023-06221-2)
2. Cornelio C, Dash S, Austel V, Josephson TR, Goncalves J, Clarkson KL, Megiddo N, El Khadir B, Horesh L. 2023 Combining data and theory for derivable scientific discovery with AI-Descartes. *Nat. Commun.* **14**, 1777. (doi:10.1038/s41467-023-37236-y)
3. Langley PW. 1987 *Scientific discovery: computational explorations of the creative processes*. Cambridge: The MIT Press.

4. Džeroski S, Todorovski L (eds). 2007 Computational discovery of scientific knowledge. In *Lecture notes in artificial intelligence* Berlin, Germany: Springer.
5. Koza JR. 1992 *Genetic programming: on the programming of computers by means of natural selection*. Cambridge, MA: MIT Press.
6. Schmidt M, Lipson H. 2009 Distilling free-form natural laws from experimental data. *Science* **324**, 81–85. (doi:10.1126/science.1165893)
7. Cranmer M. 2023 Machine learning for science with PySR and symbolicregression.jl arXiv:2305.01582. (doi:10.48550/arXiv.2305.01582)
8. Brunton SL, Proctor JL, Kutz JN. 2016 Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA* **113**, 3932–3937. (doi:10.1073/pnas.1517384113)
9. Petersen BK, Larma ML, Mundhenk TN, Santiago CP, Kim SK, Kim JT. 2021 Deep symbolic regression: recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*.
10. Mežnar S, Džeroski S, Todorovski L. 2023 Efficient generator of mathematical expressions for symbolic regression. *Mach. Learn.* **112**, 4563–4596. (doi:10.1007/s10994-023-06400-2)
11. Udrescu SM, Tegmark M. 2020 AI Feynman: a physics-inspired method for symbolic regression. *Sci. Adv.* **6**, eaay2631. (doi:10.1126/sciadv.aay2631)
12. La Cava W, Burlacu B, Virgolin M, Kommenda M, Orzechowski P, de França FO, Jin Y, Moore JH. 2021 Contemporary symbolic regression methods and their relative performance. *Adv. Neural Inf. Process. Syst.* **2021**, 1–16. (doi:10.48550/arXiv.2107.14351)
13. Makke N, Chawla S. 2024 Interpretable scientific discovery with symbolic regression: a review. *Artif. Intell. Rev.* **57**, 2. (doi:10.1007/s10462-023-10622-0)
14. Hofman JM, Sharma A, Watts DJ. 2017 Prediction and explanation in social systems. *Science* **355**, 486–488. (doi:10.1126/science.aal3856)
15. Guimerà R, Reichardt I, Aguilar-Mogas A, Massucci FA, Miranda M, Pallarès J, Sales-Pardo M. 2020 A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Sci. Adv.* **6**, eaav6971. (doi:10.1126/sciadv.aav6971)
16. Reichardt I, Pallarès J, Sales-Pardo M, Guimerà R. 2020 Bayesian machine scientist to compare data collapses for the Nikuradse dataset. *Phys. Rev. Lett.* **124**, 084503. (doi:10.1103/PhysRevLett.124.084503)
17. Fajardo-Fontiveros O, Reichardt I, De Los Ríos HR, Duch J, Sales-Pardo M, Guimerà R. 2023 Fundamental limits to learning closed-form mathematical models from data. *Nat. Commun.* **14**, 1043. (doi:10.1038/s41467-023-36657-z)
18. Jaynes ET. 2003 *Probability theory: the logic of science*. Cambridge, UK: Cambridge University Press.
19. Ando T. 2010 *Bayesian model selection and statistical modeling*. Boca Raton, FL: CRC Press.
20. Schwarz G. 1978 Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464. (doi:10.1214/aos/1176344136)
21. Cabanas-Tirapu O, Danús L, Moro E, Sales-Pardo M, Guimerà R. 2025 Human mobility is well described by closed-form gravity-like models learned automatically from data. *Nat. Commun.* **16**, 1336. (doi:10.1038/s41467-025-56495-5)
22. Bartlett DJ, Desmond H, Ferreira PG. 2024 Exhaustive symbolic regression. *IEEE Trans. Evol. Computat.* **28**, 950–964. (doi:10.1109/TEVC.2023.3280250)
23. Grünwald PD. 2007 *The minimum description length principle*. Cambridge, Massachusetts: The MIT Press.
24. Enßlin TA. 2019 Information theory for fields. *Ann. Phys.* **531**, 1800127. (doi:10.1002/andp.201800127)
25. Cox RT. 1946 Probability, frequency and reasonable expectation. *Am. J. Phys.* **14**, 1–13. (doi:10.1119/1.1990764)
26. Vineberg S. 2022 Dutch book arguments. In *The stanford encyclopedia of philosophy* (eds EN Zalta, U Nodelman), 2022 edition. Stanford, CA: Metaphysics Research Lab, Stanford University Fall.
27. Bartlett D, Desmond H, Ferreira P. 2023 Priors for symbolic regression. In *Proc. of the Companion Conference on Genetic and Evolutionary Computation GECCO '23 Companion*, pp. 2402–2411. New York, NY, USA: Association for Computing Machinery. (doi:10.1145/3583133.3596327)

28. Fox C, Tran ND, Nacion FN, Sharlin S, Josephson TR. 2024 Incorporating background knowledge in symbolic regression using a computer algebra system. *Mach. Learn.* **5**, 025057. (doi:[10.1088/2632-2153/ad4a1e](https://doi.org/10.1088/2632-2153/ad4a1e))
29. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. 1999 Bayesian model averaging: a tutorial. *Stat. Sci.* **14**, 382–417. (doi:[10.1214/ss/1009212519](https://doi.org/10.1214/ss/1009212519))
30. Rudin C. 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215. (doi:[10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x))
31. Amaral LAN. 2024 Artificial intelligence needs a scientific method-driven reset. *Nat. Phys.* **20**, 523–524. (doi:[10.1038/s41567-024-02403-5](https://doi.org/10.1038/s41567-024-02403-5))