PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Martin BT, Munch SB, Hein AM. 2018 Reverse-engineering ecological theory from data. *Proc. R. Soc. B* **285**: 20180422.
http://dx.doi.org/10.1098/rspb.2018.0422

Received: 27 February 2018
Accepted: 16 April 2018

Subject Category:
Ecology

Subject Areas:
theoretical biology, ecology, computational biology

Keywords:
theoretical ecology, population dynamics, ecological prediction, forecasting, time-series analysis, multi-model inference

Authors for correspondence:
Benjamin T. Martin
e-mail: benjamin.martin@noaa.gov
Andrew M. Hein
e-mail: andrew.hein@noaa.gov

†These authors contributed equally to this work.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4078910.

THE ROYAL SOCIETY PUBLISHING

Reverse-engineering ecological theory from data

Benjamin T. Martin[1,2,†], Stephan B. Munch[1,2] and Andrew M. Hein[1,2,†]

[1]Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 110 McAllister Way, Santa Cruz, CA 95060, USA
[2]Institute of Marine Sciences, University of California 1156 High St, Santa Cruz, CA 95064, USA

BTM, 0000-0003-3927-0449

Ecologists have long sought to understand the dynamics of populations and communities by deriving mathematical theory from first principles. Theoretical models often take the form of dynamical equations that comprise the ecological processes (e.g. competition, predation) believed to govern system dynamics. The inverse of this approach—inferring which processes and ecological interactions drive observed dynamics—remains an open problem in ecology. Here, we propose a way to attack this problem using a machine learning method known as *symbolic regression*, which seeks to discover relationships in time-series data and to express those relationships using dynamical equations. We found that this method could rapidly discover models that explained most of the variance in three classic demographic time series. More importantly, it reverse-engineered the models previously proposed by theoretical ecologists to describe these time series, capturing the core ecological processes these models describe and their functional forms. Our findings suggest a potentially powerful new way to merge theory development and data analysis.

1. Introduction

Since the early 1800s, theoretical ecologists have sought to understand ecological dynamics by deriving mathematical models from assumptions about how living things grow, reproduce and interact [1–5]. The practice of building such models has changed little since the early days of theoretical ecology; one begins with hypotheses about which biological processes influence the system at hand, and refines this intuition by expressing it in mathematical form, for example, as a set of differential equations or scaling relationships. An important feature of such models is their modularity: an equation describing the dynamics of a given population typically contains distinct components, each of which represents a different biological process. By formulating models in this way and comparing them to data, ecologists have discovered how core processes such as species competition [6], ontogenetic growth [4] and predation [2,3,5] govern the dynamics of a wide range of ecological systems. Moreover, mathematical analysis of these models has led to the discovery of emergent dynamics such as chaos, the paradox of enrichment, apparent competition and biomass overcompensation [7–10].

One of the challenges with developing theoretical models in this way is that doing so requires knowledge of the kinds of interactions that are most important in governing dynamics of a system of interest and the functional forms that describe those interactions. Developing a model in this way might be described as *forward-engineering*. By contrast, in many ecological systems, the structure and nature of interactions are not known *a priori* and the primary challenge is *reverse-engineering* these relationships from data. Recently, researchers in the field of artificial intelligence have proposed methods to infer dynamical relationships among variables in a system directly from time-series data [11–14]. For example, Schmidt & Lipson [11] showed that a method known as *symbolic regression* could discover the fundamental equations of Newtonian mechanics directly from measurements

of moving objects. Similar methods have been used to reverse-engineer the dynamics of other physical systems [12–14].

In contrast to traditional regression, where the analyst specifies a functional form and optimizes its free parameters, symbolic regression searches the vast space of possible functional forms to arrive at a parsimonious description of the data. Though this method has rarely been applied to biological data [15,16], its success in physical systems suggests the intriguing possibility that it might be useful for inferring ecological relationships from demographic time series. Because symbolic regression can describe relationships using dynamical equations, it also has the potential to bridge an important gap between statistical analysis and theoretical ecology; it shares the flexibility of semi-parametric methods that are familiar to ecologists (e.g. General Additive Models (GAMs) [17], Gaussian processes [18] and neural networks [19]), while retaining the intuition one can gain by posing biological hypotheses as dynamical equations—the language of theoretical ecology. The value of such equations is that they can potentially be deconstructed to determine which core ecological processes they describe, and they can be analysed to gain insights about emergent properties that may not be obvious from the data alone.

In this paper, we apply symbolic regression to several classic demographic datasets whose dynamics are thought to be driven by canonical ecological processes including self-regulation, predation and cannibalism. We test whether symbolic regression can infer these processes and their functional forms directly from time-series data. We find the following: (1) symbolic regression discovers dynamical models that explain most of the variance in all three population dynamic datasets, (2) the predictive ability of models generated through this procedure begins to saturate with increasing model complexity, defined as the number of free parameters in the model, at a surprisingly small number of free parameters in all datasets, (3) for each of the three datasets, the model occupying the saturation point, where increases in model complexity cease to substantially improve predictive capacity, contained the same ecological processes as models previously proposed by theoretical ecologists. In fact, the models occupying the saturation point were precisely the logistic growth equation for *Paramecium* growing in isolation [1,20], the Lotka–Volterra predator–prey equations for *Paramecium* and *Didinium* in co-culture [2,3,20], and a chaotic stage-structured population model for *Tribolium* flour beetles [21].

2. Material and methods

(a) How symbolic regression works

Symbolic regression is used to find mathematical equations that best describe the relationships among variables in a dataset [22,23]. To begin, a 'population' of equations is generated from a set of mathematical building blocks (i.e. mathematical operators [e.g. $+, -, \times, \div$, log, exp], variables and constants). These building blocks are combined using a tree-like network, where each node in the tree represents an operator, a state variable or a free parameter (figure 1). Operator nodes accept a specific number of input arguments [23], for example, multiplication requires two arguments ($A \times B$), whereas exponentiation requires only one argument (exp[A]). Variables and free parameters (also referred to as 'constants' in the symbolic regression literature) are terminal nodes in the tree (figure 1). In this way, building blocks can be combined to build mathematical expressions that vary widely in form and complexity.

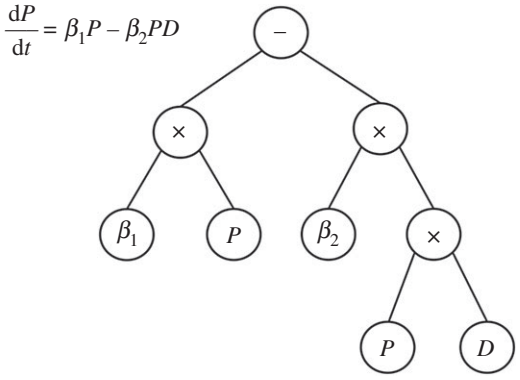

Figure 1. Generating dynamical equations from primitive operators using a tree-based encoding. Here, the Lotka–Volterra equation for a prey population is represented as a tree expression, which can be encoded as a function by evaluating the expression tree sequentially from the terminal nodes up. The equation can also be represented as a string, and mutated or recombined with other strings to generate enormous sets of mathematical equations from simple building blocks of operators, constants and independent variables.

After a large initial population of expressions is randomly generated, the ability of each to accurately describe data is assessed and the next generation of models is created using an evolutionary algorithm: models are selected based on a pre-defined measure of goodness-of-fit; then a reproduction stage occurs, in which selected models generate new models either via crossover, where two selected models produce two new models by exchanging random subtrees, or mutation, where a randomly selected subtree of a model is replaced with a new subtree (see 'Implementing Symbolic Regression' for further details).

Because goodness-of-fit tends to increase with model complexity, modern model selection criteria typically use a penalty for complexity when identifying the best model (e.g. Akaike information criterion (AIC), Bayesian information criterion (BIC), etc.). Although it is possible to identify the optimal penalization scheme for a limited class of models (e.g. linear regressions with normal errors) it is not possible to identify the optimal trade-off for an arbitrary class of models. Rather, symbolic regression seeks to find what is known as the Pareto front—the set of models that have the best goodness-of-fit for each level of model complexity. A major advantage of symbolic regression is that it allows one to thoroughly sample models that occur near this front, thereby defining (rather than assuming) the trade-off between model performance and complexity.

(b) Datasets

We applied symbolic regression to three classic ecological datasets. The first dataset consisted of four time series, each from a *Paramecium* population grown at one of four different nutrient (Cerophyl) concentrations (0.1, 0.375, 0.5 and 1.0 g l^{-1}, [20]). In each experiment, *Paramecium* populations were initialized at small population sizes and monitored for 7 days. Population density was recorded every 12 h. We numerically estimated time-derivatives of *Paramecium* population density using a cubic spline with a roughness penalty on the integral of the squared second derivative determined by generalized cross validation (smooth.spline, [24]). *Paramecium* population density and nutrient concentration were used as predictor variables. These in combination with free parameters made up the 'terminal nodes' in the symbolic regression.

In the second dataset, *Paramecium* and a predator, *Didinium*, were grown in co-culture for a 35 day period at a nutrient concentration of 0.5 g l^{-1} [20]. The densities of *Paramecium* and *Didinium* were recorded every 12 h. Again, we numerically estimated derivatives for both populations using cubic splines. We

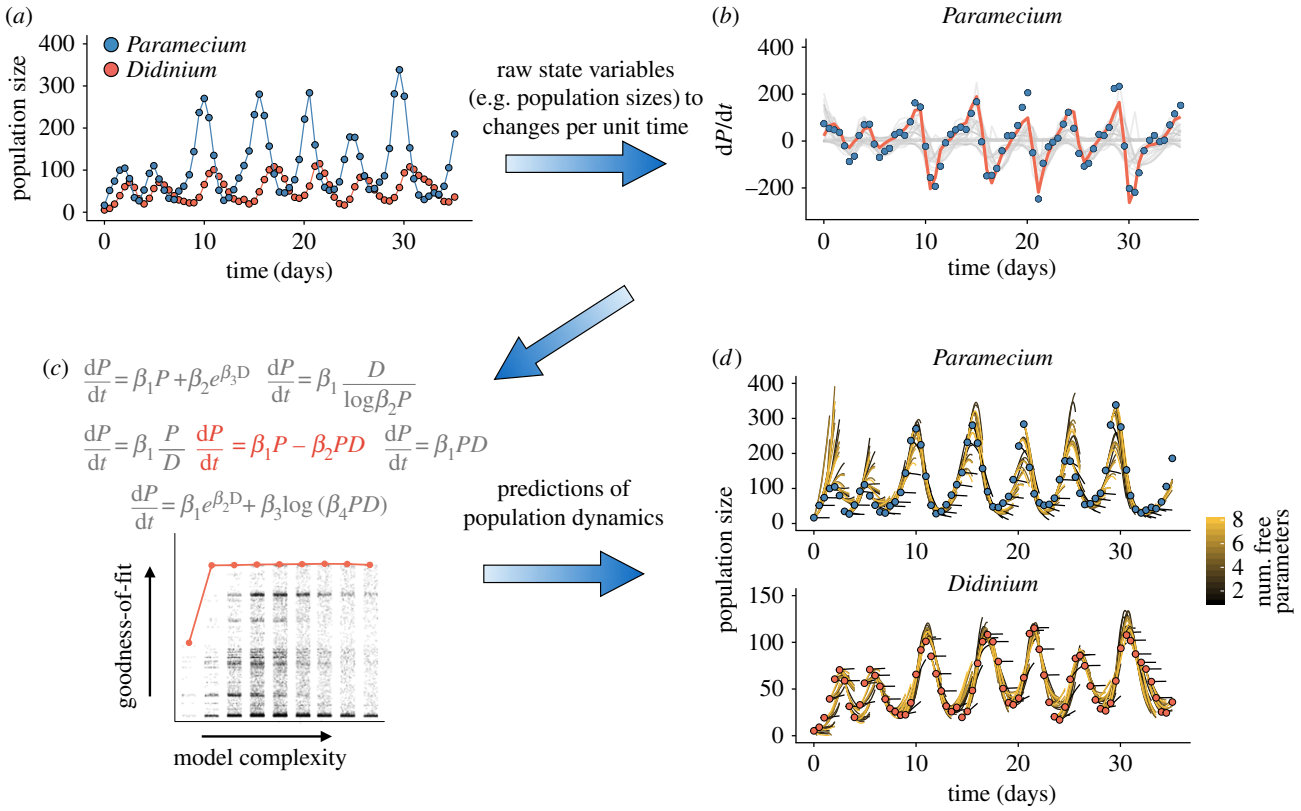

Figure 2. Discovering dynamical relationships from data—a raw time series (e.g. of population counts) is converted to a time series of derivatives or differences (e.g. population growth rates). Then a genetic programme is used to generate dynamical equations that are fitted to the derivative time series (*b*). This routine results in a large set of dynamical expressions that vary in complexity and goodness-of-fit (*c*), allowing one to identify the models that fall along the Pareto front. Finally, a subset of models (e.g. the Pareto front models) can be used to forecast population dynamics (*d*). Panel *d* shows three-step-ahead predictions from the ordinary differential equation (ODE) models along the Pareto front for the *Paramecium* and *Didinium* co-culture datasets. In general, predictions by models in the Pareto front with two or more parameters were similar, consistent with the relationship between goodness-of-fit and model complexity (*c*).

used *Paramecium* and *Didinium* densities as potential predictor variables in the symbolic regression, which along with free parameters can occur as terminal nodes.

The final dataset documented stage-structured population dynamics of the flour beetle, *Tribolium* [21,25]. The data consisted of time series of abundances of three life stages (larvae, pupae and adults) of populations maintained in milk bottles over a period of 80 weeks. In these experiments, Costantino *et al.* [21] experimentally set the adult-dependent pupae-to-adult recruitment rate, c_{pa}, at one of seven levels (0.00, 0.05, 0.10, 0.25, 0.35, 0.50, 1.00), with three replicate time series per treatment, where $\exp(-c_{pa} \times$ adult density) is the probability of a pupa recruiting to an adult in the presence of adults. Both μ_a and c_{pa} were experimentally controlled by removing or adding adults at the time of census, which occurred every two weeks. The two-week measurement interval coincided with the approximate duration of the larvae and pupae stages. Because of the relatively discrete life stages, Costantino *et al.* [21] modelled the dynamics of *Tribolium* using a system of difference equations. We used symbolic regression to predict the number of *Tribolium* larvae in the ith+1 time step as a function of the abundances of the three life stages in the ith time step and the c_{pa} variable. We focus on the larval-stage population dynamics because predicting larval dynamics is not trivial, whereas pupal abundance at the ith+1 time step is nearly a constant proportion of larvae abundance in the ith time step and adult dynamics, as described above, were manipulated as part of the experiment [21].

(c) Reverse-engineering dynamical relationships from data

Using the *Paramecium*–*Didinium* co-culture as an example, we begin with raw population time series (figure 2*a*) and convert population densities to changes in density over time (figure 2*b*). We then apply symbolic regression to generate an initial population of models, fit these models to data and generate subsequent model generations based on model performance. This routine resulted in a large set of dynamical expressions that vary in complexity and goodness-of-fit (figure 2*b*, lines). With these models in hand, we can examine the relationship between goodness-of-fit and model complexity (figure 2*c*) to identify the models that fall along the Pareto front. This provides several valuable pieces of information, including the limits to goodness-of-fit for this dataset, and the minimum level of complexity required to approach this limit (figure 2*c*, orange saturating Pareto front). Finally, we can use the set of Pareto front models (or any other subset of models) to forecast population dynamics (figure 2*d*).

(d) Implementing symbolic regression

We developed a genetic programme to implement symbolic regression and applied this method to the three datasets. Our implementation was based on the GPtips [26] package in MATLAB but differed in several important ways: we enforced dimensional consistency, we compressed unidentifiable parameters when computing model complexity, and we performed nonlinear optimization for each candidate expression using a robust optimization algorithm called restricted back-propagation (described below; [27]).

(i) Dimensional consistency

A key feature of any model of a physical system is that the units of the quantities in the model must be dimensionally consistent in the sense that if two objects are added to one another, they must have the same dimensions; if they are multiplied, the dimension of the

product must change accordingly. To enforce dimensional consistency, we multiplied all state variables by a free parameter. All terms that are inherently non-dimensional (e.g. exponential and logarithmic terms) were also multiplied by a free parameter. The implicit dimensions of free parameters ensure that all models produced by the genetic programme were dimensionally consistent.

(ii) Constant compression

Randomly placing free parameters within strings using tree generation inevitably creates unidentifiable models. For example, applying the addition operator to parameters β_1 and β_2 results in an expression $(\beta_1 + \beta_2)$ that can be satisfied by an infinite number of values of each of these parameters. We developed a string-based routine to reduce the occurrence of unidentifiable expressions by compressing them into a single parameter. Because of the diverse range of models that can be created by the tree routine, unidentifiable parameter combinations still occasionally occur. To ensure that we correctly determined the number of identifiable free parameters for each model, we evaluated the rank of the derivative matrix of each model,

$$\mathbf{A} = \left\{ \frac{\partial \mu_i}{\partial \theta_j} \right\}, \tag{2.1}$$

where μ is a vector of the expected value of the model at each observation, and θ is the vector of parameters in the model. The true number of parameters in a model is equal to the rank of $\mathbf{A}$ [28].

(iii) Optimization

Because free parameters could occur both as linear coefficients and as nonlinear terms, we used a nonlinear optimization scheme to estimate parameter values. Standard Nelder–Mead simplex algorithms performed slowly, particularly for models with many parameters. Instead, we used an alternative optimization scheme—restricted back-propagation (R_{prop}, [27])—which performed more rapidly and more robustly.

We used a standard evolutionary algorithm for our implementation of symbolic regression [22,26]. We briefly outline the algorithm below and provide the code in the electronic supplementary material, but for a detailed description of the standard evolutionary algorithm used in genetic programming, see Koza [22] and Poli *et al*. [23]. Each generation, the top 5% of models (see 'Evaluating model fit') were directly copied to the next generation. The remaining 95% of models were selected for the next generation using tournament selection [22,23], whereby two models were randomly selected from the population, their goodness-of-fit was compared, and the model with the lowest residual sum of squares was selected to contribute to the next generation. The next generation of models was constructed either through crossover ($p = 0.75$) or via mutation ($p = 0.2$), or direct reproduction (also referred to as 'cloning'; $p = 0.05$). For both *Paramecium* and *Didinium* datasets, preliminary runs indicated a population size of approximately 2500 models and 20 generations was sufficient to generate models across the range of model complexities that interested us (i.e. 1–8 free parameters). More generations (40) were used for the *Tribolium* dataset because it took longer for symbolic regression runs to discover complex models (greater than 6 free parameters) that fit the data well. This is likely because a larger fraction of the initial models generated for that dataset produced models with undefined model predictions (e.g. division or log of independent variables with a value of 0) or likelihoods (predicted densities less than 0 produced undefined likelihoods due to the error model used for the *Tribolium* dataset [see 'Evaluating model fit']). For each dataset, we performed three symbolic regression runs, each starting with a different initial population of models to better explore the space of possible models. We constructed Pareto fronts by combining all models into a single model set. For all analyses, we used the operators $+$, $-$, $\times$, $\div$, ln, and exp as our initial operator set. This operator set is sufficient to generate virtually all commonly used population models except those with periodic forcing.

(e) Evaluating model fit

For each dataset, we computed derivatives or differences as described above, and then selected a random 30% of data points to serve as out-of-sample data. We performed symbolic regression on the remaining 70% of the data. Parameter estimation and model fitness were calculated using these in-sample data only. For the *Paramecium* in isolation and *Paramecium* and *Didinium* in co-culture datasets, we assumed additive, normally distributed errors and used the residual sum-of-squares (RSS) as a measure of model fit. For the *Tribolium* dataset, we used the likelihood derived by Dennis *et al*. [25] in their original analysis of this dataset. That is, we used a normal approximation of a Poisson distribution by square-root transforming the model predictions and observations to normalize and stabilize the variance, and then used the residual RSS on the transformed data as a measure of model fit (see [25] for a more rigorous justification of this error model). For all fitted models, we computed model log likelihood and r^2 to evaluate in-sample performance.

We performed the entire symbolic regression procedure twice, in each case taking a random subset of 70% of the data as in-sample, and the remaining 30% as out-of-sample. Repeating this procedure twice and taking the average out-of-sample performance for models on the Pareto front ensured that results were robust to the particular random subsets of data chosen as in-sample and out-of-sample.

(f) Variable importance and model averaging

Previous applications of symbolic regression tend to focus on a single 'optimal' model structure. While we do make the comparison between ecological theory and specific models on the Pareto front (see below), there is much to be gained by combining the tools of symbolic regression and current approaches to multimodel inference (e.g. [29]). For instance, ecologists are often interested in general inferences about the relationships among state variables; for example, is X useful for predicting Y? If so, what kind of function relates the two? These questions can be answered using the model set generated by symbolic regression. To do this, we define a conditional effect size:

$$\gamma|Y, X = \frac{1}{N}\sum_{i=1}^{N} |f(X_i, y_i) - f(X_i, \bar{y})|, \tag{2.2}$$

where Y is the vector of observed values of the variable of interest, X is a matrix containing observed values of all other state variables in the model, y_i is the ith observation of the state variable of interest, $\bar{y}$ is the mean value of Y over all observations, X_i is a vector of other state variables in the model associated with the ith observation, and f is the model. We refer to this as a *conditional* effect size to emphasize that its value is conditional on the observed values of Y and X. Note that the mean serves only as a reference point; different statistics or quantiles could be used as reference points instead. As a frame of reference, note that if the fitted model were a multiple linear regression with slope parameters $b_1, \ldots b_p$ then $\gamma_k = b_k M_{1k}$ (where γ_k is the conditional effect size for variable k and M_{1k} is the first absolute moment for variable k), which is the expected change in the output for a one standard deviation change in the input. To illustrate how conditional effect size can be used to make inferences about the importance of different candidate predictor variables, we applied it to the *Tribolium* dataset discussed above.

3. Results

In all three datasets, symbolic regression discovered dynamical rules that explained most of the variance in the data (figure 3).

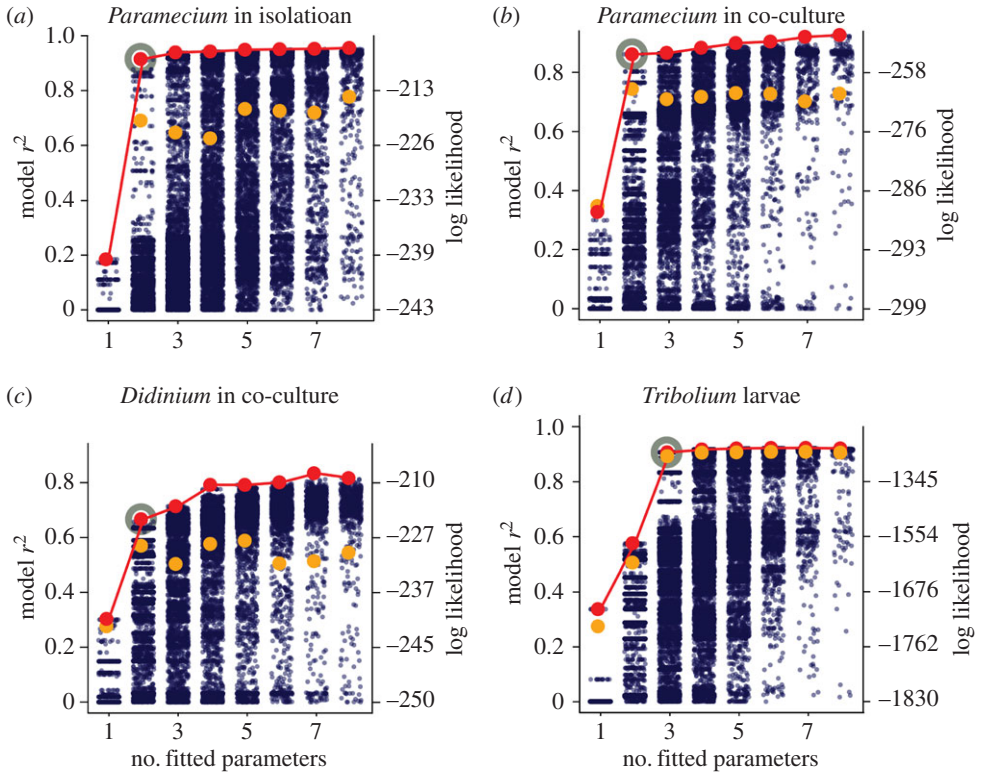

Figure 3. Model r^2 (left axis) and log likelihood (right axis) for fitted models. Each blue point is the fit of a single model. Points are jittered in the horizontal direction for visualization. Red points denote the best model at each level of complexity. For each of the three datasets, the model occupying the saturation point where increases in model complexity cease to substantially improve predictive capacity (grey circle) was the model previously proposed by theoretical ecologists: the logistic growth equation for *Paramecium* growing in isolation (*a*), the Lotka–Volterra predator–prey equations for *Paramecium* (*b*) and *Didinium* (*c*) in co-culture, and a chaotic stage-structured population model for *Tribolium* flour beetles (*d*). Orange points are mean r^2 of Pareto front models computed from out-of-sample data.

In the *Paramecium* monoculture dataset, the best model explained 95% of the variance. In the *Paramecium* and *Didinium* datasets, the best models explained 92 and 84% of the variance, respectively. For the *Tribolium* stage-structured population dynamics dataset, the best model explained 92% of the variance in larval abundances in the next time step. Model goodness-of-fit increased rapidly in all three datasets as we increased the number of parameters up to two or three parameters. Beyond this point, increasing model complexity resulted in relatively small improvements in model performance. For example, in the case of the *Tribolium* larvae, moving from the best model with two free parameters to the best model with three free parameters increased r^2 from 0.58 to 0.91 and model log-likelihood of 463.7 units. While the best model, with eight free parameters, only improved r^2 to 0.92 and increased the log likelihood by only 35.5 units compared to the best three-parameter model.

The fact that model performance saturated quickly with increasing model complexity in all of the datasets to which we applied our method demonstrates that there exist simple models that describe population dynamics nearly as well as much more complex ones. This implies that a relatively small number of core processes determine population dynamics in these systems. The next natural question, then, is whether the simple models that capture dynamics represent biologically plausible hypotheses about the mechanisms that underlie these dynamics. For *Paramecium* growing at different nutrient concentrations without a predator, model fit saturates at roughly two parameters (figure 3*a*). The two-parameter model on the Pareto front, $\mathrm{d}P/\mathrm{d}t = \beta_1 NP - \beta_2 P^2$, states that *Paramecium* dynamics are governed by two terms; the first term on the right-hand side of the equation can be interpreted as a *per capita* growth rate that increases linearly with nutrient level, N. The second term can be interpreted as a linear decrease in *per capita* growth rate as *Paramecium* density, P, increases. This model is an alternative parametrization of the resource-dependent logistic equation where the maximum *per capita* growth rate, r is $\beta_1 N$, and the carrying capacity K is $\beta_1 N/\beta_2$. Similarly, for *Paramecium* and *Didinium* in co-culture, model fit began to saturate at two parameters, and the two-parameter models occupying the Pareto front were the Lotka–Volterra equations for predator–prey populations: $\mathrm{d}P/\mathrm{d}t = \beta_1 P - \beta_2 \mathrm{PD}$, and $\mathrm{d}D/\mathrm{d}t = \beta_3 PD - \beta_4 D$, which include terms that represent reproduction, predation and natural mortality. In the case of both datasets, these relatively simple models have similar performance to the more complex Pareto front models when predicting out-of-sample data (figure 3, orange points).

Since the logistic and Lotka–Volterra models describe reproduction, self-regulation, predation and mortality as linear combinations of polynomial terms, one could argue that these models could have been reverse-engineered through more conventional means such as polynomial regression. The dynamics of *Tribolium* populations are more complex as this species exhibits multiple life stages and chaotic dynamics. Nevertheless, model fit began to saturate at three parameters (figure 3*d*). The best three-parameter model, $L_{t+1} = \beta_1 A_t \exp(-\beta_2 A_t - \beta_3 L_t)$, is strongly nonlinear and cannot be expressed as a sum of polynomials. Costantino *et al.* [21] derived precisely this model structure from a set of experiments on individual life stages to determine which processes dominated population dynamics and to identify the functional forms that describe these processes. In their interpretation of

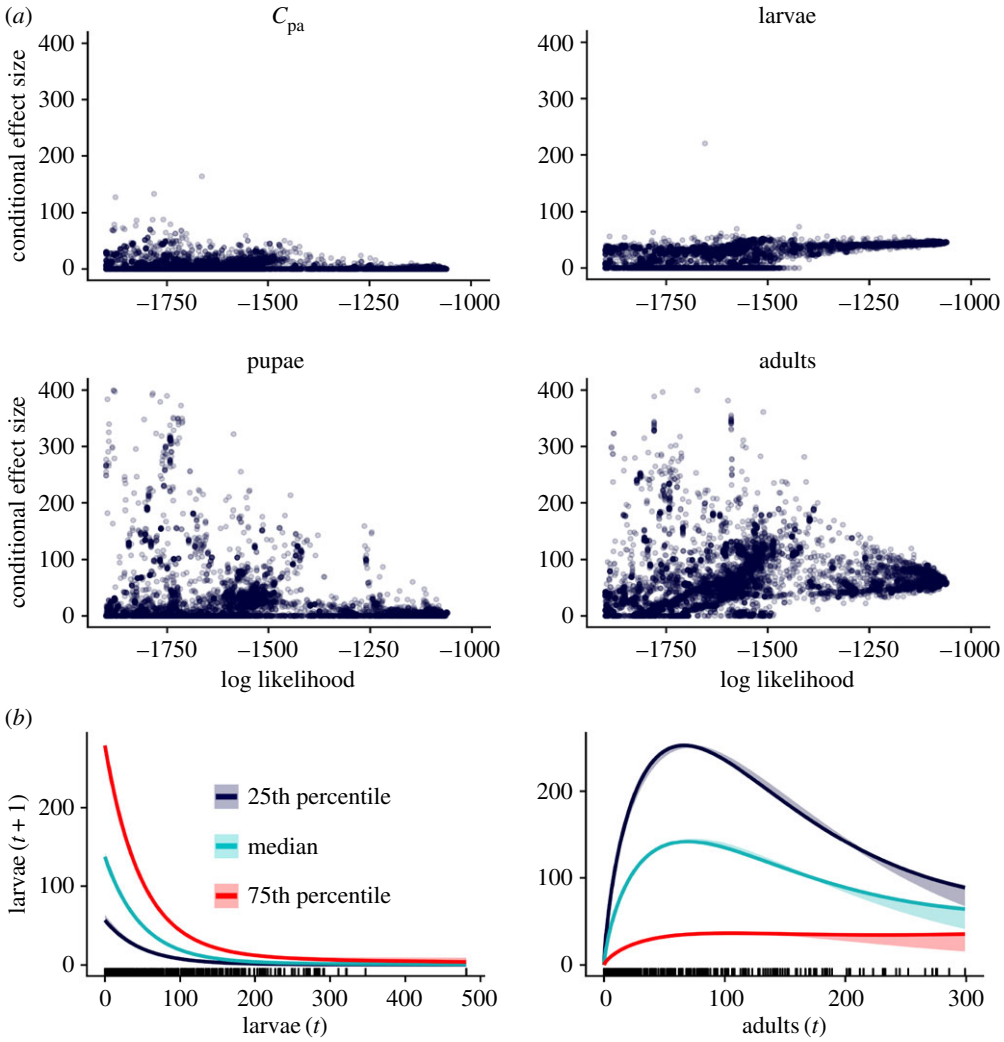

Figure 4. Conditional effect size (see text) and the shape of functional relationships for the *Tribolium* dataset. (*a*) Conditional effect size of each of the four state variables as a function of model log likelihood for over 40 000 fitted models. For two of the four variables (the pupa-to-adult recruitment rate, c_{pa} and pupa abundance), effect size is highly variable in poorly performing models, and converges to low values in the best models. For the other two variables (larvae abundance and adult abundance), conditional effect sizes converge to positive values in models with the highest log likelihoods, revealing that models with different structures agree on the importance of these variables and the magnitude of their effects on larval recruitment dynamics. (*b*) Solid curves show AIC-weighted prediction of larval abundance at time $t + 1$ from the full set of fitted models. In the left panel the three curves show the model predictions when adult (left panel) or larvae (right panel) abundance at time t is held to its 25th (blue), 50th (teal) or 75th (red) percentile value. Pupae abundance and c_{pa} are held to their median values in both panels. Envelopes show upper and lower bounds on predictions of models with AIC weights greater than 0.01 (19 models). Agreement in model shape indicates that the best performing models predicted similar structural relationships between larval dynamics and state variables except at very high values of adult density, where there was limited data.

the model, the number of larvae at time, $t + 1$, depends on the product of the number of eggs produced by adults at time, t, and the survival probability of those eggs to time, $t + 1$. The survival probability of eggs is governed by the exponentiation of instantaneous mortality rates that are linearly dependent on the abundance of adults and larvae, consistent with the hypothesis that both adults and larvae are cannibalistic. As in the other datasets, this relatively simple model performed as well on out-of-sample data as more complex models (figure 3*d*, orange points).

In addition to identifying the mathematical expressions that best capture dynamics, we used the entire set of models generated by the symbolic regression to evaluate the overall importance of each candidate predictor variable in governing the observed population dynamics. In the *Tribolium* dataset (figure 4*a*), models that fit the data well have small conditional effect sizes (equation (2.2)) for c_{pa} and pupae, and large effect sizes for larvae and adults, indicating that larval dynamics are not strongly influenced by c_{pa} or pupae regardless of the functional form assumed to link these variables to larval population growth. The conditional effect sizes of adult and larval abundance converge to stable values among models that fit the data well (figure 4*a*). This indicates two important features of the model set: (1) good models agree that these two variables are the ones that drive larval population dynamics, and (2) they also agree on the magnitude of the effect of these variables for the dataset at hand. In addition to agreeing on which variables were most important in governing *Tribolium* dynamics, models also agree on the shape of the relationship that relates state variables to one another (figure 4*b*).

4. Discussion

By extensively searching the space of plausible models, symbolic regression quantifies how the amount of explainable

variance scales with model complexity along the Pareto front, and allows us to identify the ecological processes that appear to drive dynamics of these populations. For all three demographic datasets, the model that occurred at the point on the Pareto front where model fit began to saturate with increasing model complexity was exactly the model previously proposed by theoretical ecologists to describe these dynamics. It is important to point out that this is not by any means guaranteed. The theoretical models were derived under an additional constraint that cannot be enforced in the symbolic regression: these models only contain mathematical expressions that represent biologically plausible hypotheses. The set of models that meet this constraint is a small fraction of the universe of models that are algebraically and dimensionally consistent. Given this, it is encouraging that models with biologically meaningful structures occurred on the Pareto front.

(a) Is symbolic regression a useful method for reverse-engineering dynamical relationships in ecological data?

By applying symbolic regression to well-studied single-species and predator–prey systems, we were able to test whether this method could discover the core biological processes such as self-regulation, predation and cannibalism that are central to classical theories of population dynamics. The fact that the method was successful, suggests that it may be useful in other biological systems where the processes that drive dynamics are not well understood. To put it another way, had the logistic population model, Lotka–Volterra equations, and stage-structured population growth model not yet been developed, the routine we applied would have reverse-engineered them without any foreknowledge of the underlying biological processes these models represent. Several obvious areas where this approach could be useful are (1) discovering interactions in food webs from population data [30], (2) inferring the behavioural rules animals use to respond to sensory stimuli [31], and (3) relating temperature and other environmental variables to vital rates in wild populations [32].

(b) How can symbolic regression be used alongside more traditional methods of theory development?

In ecology, as in other disciplines, theory is judged on several grounds at the same time. Does the theory rest on well-established or at least plausible descriptions of how the system works? Is the theory self-consistent in the sense that it does not contain elements or assumptions that contradict one another? How good is the theory at describing data? How well does the theory perform relative to competing theories? Our approach does not attempt to answer the first two questions, which, we emphasize, are still fundamental when evaluating any theoretical description of a natural system. However, symbolic regression does offer a powerful and objective way of answering the second two questions. Comparing models that represent alternative descriptions of the data has become standard practice in ecology, due in large part to a well-deserved backlash against significance testing [33]. However, a shortcoming of most multi-model comparisons is the curse of relativism: every set of models contains a *best model*, rendering judgments about whether a model is *good* or *bad* highly dependent on the set of models being considered. By placing theoretical models in a performance space comprised tens-of-thousands of alternative models, we were able to evaluate them against the maximum performance such models can achieve. This is useful when evaluating whether added model complexity is warranted. For example, Chen *et al*. applied symbolic regression to the same *Paramecium* and *Didinium* dataset used here [16] and focused on a model containing four time lags and 16 free parameters to describe coupled dynamics of the two species in co-culture. While such complex models might be required for some purposes, our analyses suggest that far simpler models can explain the majority of variance in this and other datasets. More generally, measuring the trade-off between complexity and goodness-of-fit across a large and diverse model set can be used to answer the question of 'how good is good' in a more objective way.

(c) How can symbolic regression be used in ecological data analysis?

The fraction of ecological systems for which theoreticians have derived putative governing equations is vanishingly small. As a result, ecologists often use statistical models to make inferences about the patterns present in data. Modern statistical methods provide a multitude of new tools for doing this [17–19]; semi-parametric methods including GAMs, Gaussian processes and neural networks can accommodate nonlinear relationships among state variables that are unknown in advance. The flexibility of these methods is appealing, because in most ecological systems, we do not know the functional forms that describe the dynamical relationships among variables, and assuming the incorrect form can strongly bias inference. The disadvantage of these methods, however, is that breaking their output into ecologically meaningful components—for example, components that represent predation, competition, cannibalism, etc.—can be extremely difficult. In contrast to these methods, symbolic regression describes data-driven functional relationships using explicit equations, rather than networks or basis expansions. Our analyses illustrate one of the major advantages of symbolic regression over alternative methods: if simple nonlinear expressions that accurately describe dynamics exist, symbolic regression can discover them.

Even if the goal of analysis is prediction or determining variable importance rather than identifying functional relationships, symbolic regression has some distinct advantages over more traditional methods of multi-model inference. For example, analysing effect sizes of variables across the full model set provides a way of evaluating variable importance without preconditioning on a particular model structure (figure 4*a*). This is a generalization of the idea of measuring effect sizes in linear models using standardized coefficients. Similarly, the estimates from each model in the full set of models can be averaged after weighing each model by its goodness-of-fit (figure 4*b*) in the same way that model averaging is often applied in more traditional multi-model inference [29]. However, the wide range of model structures created by the genetic programme makes it more likely that the set will include models that fit the data well.

(d) Important considerations when using the method

We do not expect symbolic regression to automatically extract concise mathematical relationships for any arbitrary ecological dataset. As with most regression methods, symbolic regression

is unlikely to extract meaningful relationships from a large list of highly correlated predictors. Given the extremely flexible forms that symbolic regression can generate, one must take particular care to ensure that models that perform well on one set of data are robust. Model cross-validation, forecasting, and other methods for evaluating out-of-sample performance should be used (figure 3). As with any method of fitting and comparing models, symbolic regression can lead to spurious conclusions if it is not used carefully.

While observation errors are minimal in the laboratory datasets analysed here, it can be an important issue when analysing field data. In cases where the observation and process errors are approximately Gaussian, embedding symbolic regression in an extended Kalman filter (e.g. [34]), and using symbolic differentiation to evaluate the Jacobian matrices could provide a means of partitioning noise among process and observation errors. There are also many datasets for which non-Gaussian likelihoods would be preferable (e.g. when making inferences for small populations with many zero counts). To allow for non-Gaussian process noise requires only that we change the likelihood that serves at the 'fitness function' used to compare models and fit parameters. Finally, ecological analyses frequently include random effects to account for complex correlation structures. Random effects could be included in symbolic regression using a hierarchical modelling approach [35], albeit at the expense of considerable additional computation.

(e) Conclusions

Gains in computational power have fuelled an explosion of statistical methods that were developed, in part, to move beyond the limitations of traditional linear statistical tools. The allure of such methods is that they are flexible enough to capture the broad range of functional relationships that are possible in ecological systems. However, this flexibility often comes at the cost of interpretability; these methods allow one to mine ecological datasets for relationships but rarely supply the intuition or insight derived from theoretical ecology. At the same time, simple theoretical models continue to form the conceptual backbone of our discipline, despite the apparent complexity of ecological systems. Symbolic regression has the potential to bridge these schools of thought; it takes advantage of computationally intensive methods to discover structural relationships in data, but it describes those relationships using the mathematical language of theoretical ecology. The resulting equations can be studied independently of the data to which they were fit, potentially leading to insight about unobserved regimes, sensitivity to perturbation, and other dynamical properties that may not be evident from the data alone.

Data accessibility. The datasets and code used for our analyses can be downloaded from https://github.com/btmarti25/Symbolic_Regression_PRSB_Martin_Munch_Hein_2018.

Authors' contributions. A.M.H., B.T.M. and S.B.M. conceived the study. A.M.H. and B.T.M. performed the analyses. A.M.H., B.T.M. and S.B.M. wrote the manuscript.

Competing interests. We declare we have no competing interests.

Funding. We received no funding for this study.

Acknowledgements. The authors thank Massimo Vergassola, Roger Nisbet and Antoine Brias for feedback and discussions that improved the manuscript.

References

1. Verhulst PF. 1838 Notice sur la loi que la population suit dans son accroissement. *Corr. Math. Phys.* **10**, 113–121.
2. Lotka AJ. 1925 *Principles of physical biology*. Baltimore, MD: Waverly.
3. Volterra V. 1926 Fluctuations in the abundance of a species considered mathematically. *Nature* **118**, 558–560. (doi:10.1038/118558a0)
4. von Bertalanffy L. 1938 A quantitative theory of organic growth (inquiries on growth laws. II). *Hum. Biol.* **10**, 181–213.
5. May RM. 1972 Limit cycles in predator-prey communities. *Science* **177**, 900–902. (doi:10.1126/science.177.4052.900)
6. Tilman D. 1982 *Resource competition and community structure*. Princeton, NJ: Princeton University Press.
7. May RM. 1974 Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. *Science* **186**, 645–647. (doi:10.1126/science.186.4164.645)
8. Rosenzweig ML. 1971 Paradox of enrichment: destabilization of exploitation ecosystems in ecological time. *Science* **171**, 385–387. (doi:10.1126/science.171.3969.385)
9. Holt RD. 1977 Predation, apparent competition, and the structure of prey communities. *Theor. Popul. Biol.* **12**, 197–229. (doi:10.1016/0040-5809(77)90042-9)
10. de Roos AM, Persson L. 2013 *Population and community ecology of ontogenetic development*. Princeton, NJ: Princeton University Press.
11. Schmidt M, Lipson H. 2009 Distilling free-form natural laws from experimental data. *Science* **324**, 81–85. (doi:10.1126/science.1165893)
12. 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)
13. Quade M, Abel M, Shafi K, Niven RK, Noack BR. 2016 Prediction of dynamical systems by symbolic regression. *Phys. Rev. E* **94**, 012214. (doi:10.1103/PhysRevE.94.012214)
14. Rudy SH, Brunton SL, Proctor JL, Kutz JN. 2017 Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614. (doi:10.1126/sciadv.1602614)
15. Cardoso P, Borges PA, Carvalho JC, Rigal F, Gabriel R, Cascalho J, Correia L. 2016 Automated discovery of relationships, models and principles in ecology. *bioRxiv* 027839.
16. Chen Y, Angulo MT, Liu YY. 2016 Revealing complex ecological dynamics via symbolic regression. *bioRxiv* 074617.
17. Guisan A, Edwards Jr TC, Hastie T. 2002 Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Modell* **157**, 89–100. (doi:10.1016/S0304-3800(02)00204-1)
18. Boettiger C, Mangel M, Munch S. 2015 Avoiding tipping points in fisheries management through Gaussian process dynamic programming. *Proc. R. Soc. B* **282**, 20141631. (doi:10.1098/rspb.2014.1631)
19. Crisci C, Ghattas B, Perera G. 2012 A review of supervised machine learning algorithms and their applications to ecological data. *Ecol. Model.* **240**, 113–122. (doi:10.1016/j.ecolmodel.2012.03.001)
20. Veilleux BG. 1979 An analysis of the predatory interaction between *Paramecium* and *Didinium*. *J. Anim. Ecol.* **48**, 787–803. (doi:10.2307/4195)
21. Costantino RF, Desharnais RA, Cushing JM, Dennis B. 1997 Chaotic dynamics in an insect population. *Science* **275**, 389–391. (doi:10.1126/science.275.5298.389)
22. Koza JR. 1992 *Genetic programming: on the programming of computers by means of natural selection*, vol. 1. Cambridge, MA: MIT press.
23. Poli R, Langdon WB, McPhee NF. 2008 Field guide to genetic programming. Published via http://lulu.

com and freely available at http://www.gp-field-guide.org.uk, 2008. (With contributions by J. R. Koza). GPBiB

24. R Core Team. 2015 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See URL https://www.R-project.org/.
25. Dennis B, Desharnais RA, Cushing JM, Henson SM, Costantino RF. 2001 Estimating chaos and complex dynamics in an insect population. *Ecol. Monogr.* **71**, 277–303. (doi:10.1890/0012-9615(2001)071[0277:ECACDI]2.0.CO;2)
26. Searson DP, Leahy DE, Willis MJ. 2010 GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. In *Proceedings of the International multiconference of engineers and computer scientists*, IMECS 2010, 17–19 March, Hong Kong, vol. 1, pp. 77–80. Hong Kong: International Association of Engineers (IAENG).
27. Riedmiller M, Braun H. 1992 RPROP-A fast adaptive learning algorithm. In *Proc. of ISCIS VII), Universitat*.
28. Catchpole EA, Morgan BJ, Viallefont A. 2002 Solving problems in parameter redundancy using computer algebra. *J. Appl. Stat.* **29**, 625–636. (doi:10.1080/02664760120108601)
29. Johnson JB, Omland KS. 2004 Model selection in ecology and evolution. *Trends Ecol. Evol.* **19**, 101–108. (doi:10.1016/j.tree.2003.10.013)
30. Fahimipour A, Hein AM. 2014 The dynamics of assembling food webs. *Ecol. Lett.* **17**, 606–613. (doi:10.1111/ele.12264)
31. Hein AM, Carrara F, Brumley DR, Stocker R, Levin SA. 2016 Natural search algorithms as a bridge between organisms, evolution, and ecology. *Proc. Natl Acad. Sci.* **113**(34), 9413–9420.
32. Dell AI, Pawar S, Savage VM. 2011 Systematic variation in the temperature dependence of physiological and ecological traits. *Proc. Natl Acad. Sci. USA* **108**, 10 591–10 596. (doi:10.1073/pnas.1015178108)
33. Burnham KP, Anderson DR. 2002 *Model selection and multimodel inference: a practical information—theoretic approach*. New York, NY: Springer.
34. Pastres R, Ciavatta S, Solidoro C. 2003 The Extended Kalman Filter (EKF) as a tool for the assimilation of high frequency water quality data. *Ecol. Modell.* **170**, 227–235. (doi:10.1016/S0304-3800(03)00230-8)
35. Royle JA, Dorazio RM. 2008 *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Amsterdam, the Netherlands: Elsevier.