

A data-driven framework to model the organism–environment system

Lisandro Milocco  | Tobias Uller 

Department of Biology, Lund University, Lund, Sweden

Correspondence

Lisandro Milocco, Department of Biology, Lund University, Lund, Sweden.
Email: lisandro.milocco@biol.lu.se

Funding information

John Templeton Foundation,
Grant/Award Number: #62220

Abstract

Organisms modify their development and function in response to the environment. At the same time, the environment is modified by the activities of the organism. Despite the ubiquity of such dynamical interactions in nature, it remains challenging to develop models that accurately represent them, and that can be fitted using data. These features are desirable when modeling phenomena such as phenotypic plasticity, to generate quantitative predictions of how the system will respond to environmental signals of different magnitude or at different times, for example, during ontogeny. Here, we explain a modeling framework that represents the organism and environment as a single coupled dynamical system in terms of *inputs* and *outputs*. Inputs are external signals, and outputs are measurements of the system in time. The framework uses time-series data of inputs and outputs to fit a nonlinear black-box model that allows to predict how the system will respond to novel input signals. The framework has three key properties: it captures the dynamical nature of the organism–environment system, it can be fitted with data, and it can be applied without detailed knowledge of the system. We study phenotypic plasticity using *in silico* experiments and demonstrate that the framework predicts the response to novel environmental signals. The framework allows us to model plasticity as a dynamical property that changes in time during ontogeny, reflecting the well-known fact that organisms are more or less plastic at different developmental stages.

KEYWORDS

environment, evo-devo, modeling, organism, plasticity

1 | INTRODUCTION

Models in biology are important tools that organize experimental observations, reveal salient aspects of the biological system, and provide falsifiable predictions. A challenge when representing organisms in such models

is to accurately capture their capacity to actively modify their development and function in response to their environment, or to “act on their own behalf” (Kauffman, 2003), a key feature of biological agency (Sultan et al., 2022). Moreover, since those organismal responses can modify the environment (Odling Smee

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Evolution & Development* published by Wiley Periodicals LLC.

et al., 2003), the result is a complex interaction between organism and environment that makes the outcome (e.g., new phenotypic state) of environmental signals difficult to predict.

At any time point, the organism (or environment) is both affecting and being affected by the environment (or organism). The organism–environment system is then dynamic, in the sense that it is in constant change, with its behavior at a given time depending on its state at that time. Models must capture this dynamical and contingent aspect to be able to predict the system's temporal behavior, and potentially control it. Such predictive models can have a wide range of important applications in ecological and evolutionary developmental biology (Gilbert et al., 2015; Sultan, 2007), including the study of how interactions between embryos and their environment give rise to tissues and organs, how organisms modify their behavior and physiology in response to stress, and, for longer time-scales, how such environmental responsiveness interacts with selective pressures, resulting in evolutionary change through niche construction.

One way to model the relationship between organismal features and environmental factors is through statistical techniques, such as regression or covariance analyses, that can capture certain patterns from measured empirical data. These representations include reaction norms, which are commonly defined in a cartesian plane, with some measure of the environment in the x -axis, a measure of the phenotypic trait on the y -axis, and a statistical regression associating the two (Schlichting & Pigliucci, 1988). For simplicity, reaction norms are typically (but not always, see Pigliucci et al., 1996) done with phenotypic measures at a fixed developmental stage, commonly adulthood. Such representation of the organism–environment system provides a statistical snapshot of the constructive and contingent process described above. While it can be useful to study certain aspects of plasticity, such as its between-individual variation and whether it is likely to evolve in response to natural selection (Dingemanse et al., 2010; Nussey et al., 2007), a static representation that focusses on a fixed moment in development is not suitable to address aspects such as how plasticity can change during the organism's lifetime, or how the plastic response depends on the timing of the environmental change. These aspects may have important implications for development and evolution, but such implications cannot be studied with representations of the organism–environment system that do not explicitly incorporate developmental time. Indeed, because organism and environment interact dynamically in time, statistical measures of covariance and regression can also

change through ontogeny, and a single value at a fixed developmental time could be a misleading representation of the system (Milocco & Salazar-Ciudad, 2022; Sugihara et al., 2012). Furthermore, while it is commonly observed that reaction norms or trait integration indeed change over ontogeny (Mitteroecker & Bookstein, 2009; Nilsson-Örtman et al., 2015; Pottier et al., 2022), there remains substantial ambiguity and controversy over what such patterns imply about development and evolution (e.g., Armbruster et al., 2014; Hubbe et al., 2023; Pigliucci et al., 1996).

The dynamical properties of the organism–environment system could be captured through models that explicitly represent the causal interactions between the components of the system, known as process-based or mechanistic models. While some examples of this type of modeling exist for well-studied gene regulatory networks (e.g., the gap gene network, Jaeger & Monk, 2014; Verd et al., 2019) and some organs (e.g., teeth, Salazar-Ciudad & Jernvall, 2010), they rarely incorporate the role of environment. This is not surprising given that extensive detail on the causal relations between organism and environment would be needed to build such models. This severely limits the applicability of this type of models to empirical data.

The objective of this paper is to explain a modeling framework for the organism–environment system that can be fitted with empirical time-series data, while still capturing the dynamical nature of the system. This data-driven approach enables the study of the organism–environment system in time, beyond relying solely on statistical associations at a fixed moment. The identified model can be used to make predictions about how focal features of the system (e.g., a phenotypic state) will respond to perturbations or signals of different magnitudes or at different points in time. Importantly, the framework proposed here can represent nonlinear dynamics, which is the general case for biological systems.

The central idea of the framework is to represent the organism–environment coupled dynamical system in terms of *inputs* and *outputs* (Ljung & Glad, 2016). Inputs are the signals received by the system and that affect the internal states, while outputs are functions of these internal states of the system that we can measure in time. Here we propose to relate the inputs and outputs of the system through a black-box model, which does not require knowledge of the underlying biological mechanisms and network of interactions of the system, but can be fitted, or identified, with time-series data of known inputs and outputs (Ljung & Glad, 2016). In this way, the parameters of these models do not have direct biological or physical meaning, but rather describe the properties of the input–output relationships of the system.

Black-box modeling of the type we suggest here has a long history in the field of systems identification (Ljung & Glad, 2016; Söderström & Stoica, 1989). Linear modeling dates back to the 1950s, while nonlinear identification has become increasingly popular in recent years to solve problems that cannot be solved with linear approaches (Pillonetto et al., 2022; Schoukens & Ljung, 2019). A limitation of nonlinear methods is that they typically require a large amount of data in the operation region to calibrate the model. Despite this limitation, nonlinear systems identification has been successfully applied in multiple problems, including the modeling of airplanes and telecommunications (Mu et al., 2005; Schoukens & Ljung, 2019). In ecology, some of these ideas are the basis of the framework known as Empirical Dynamical Modeling (EDM; Deyle et al., 2016; Munch et al., 2022; Sugihara et al., 2012). EDM has been used to model species abundance and interactions in ecological networks, mostly in the context of autonomous (i.e., without external inputs) and chaotic systems in an attractor (Munch et al., 2022, but see Brias & Munch, 2021). In this paper, we propose to use the powerful methods of input–output modeling to represent the organism–environment coupled system. We suggest that this has great potential to study phenomena that change over time or that depend on the system’s history, such as plasticity and behavior.

Using *in silico* experiments, we illustrate how this framework can be used to build models that can predict the phenotypic response to new environmental signals of different magnitudes and timing. Further, we use this to define a measurement of plasticity that changes during developmental time. This is useful if, for example, the aim is to detect developmental stages that are particularly sensitive to environmental perturbations, or to potentially control the organism–environment system and bring it to a desired phenotypic value. Finally, we discuss why these objectives would be difficult to obtain using a static description of plasticity, such as a reaction norm at a fixed developmental stage.

2 | RESULTS

2.1 | The general model for the organism–environment system

Beer (1995) suggested that an agent—for example, a responsive organism—and environment can be modeled as a single coupled dynamical system whose mutual interaction is jointly responsible for the agent’s behavior. In keeping with this literature, we refer to the responsive organism (or focal part of an organism) as an

‘agent’ throughout the description of the modeling framework.

We represent the agent and the environment as a coupled dynamical system using the following equations (simplified from Beer, 1995),

$$\frac{d\mathbf{x}_A}{dt} = f_A(\mathbf{x}_A, \mathbf{x}_E, \mathbf{u}), \quad (1)$$

$$\frac{d\mathbf{x}_E}{dt} = f_E(\mathbf{x}_E, \mathbf{x}_A, \mathbf{u}), \quad (2)$$

where \mathbf{x}_A is the state vector of the agent, \mathbf{x}_E is the state vector of the environment, \mathbf{u} is a vector of inputs, the functions f_A and f_E are the dynamical laws, and d/dt represents the time derivative. The states are the minimal set of variables needed to completely describe the system’s behavior. Note that \mathbf{x}_A , \mathbf{x}_E , and \mathbf{u} are variables of time. The equations show that the change in the vector describing the states of the agent depends on the value of the states of the agent, the values of the states of the environment, and some other inputs. Similarly, the change in the states of the environment depends on the states of the environment, the states of the agent, and on other inputs.

Equations (1) and (2) are known as the state-space representation of the system (Beer, 1995; Ljung & Glad, 2016). This representation shows explicitly how the states change in time. By defining appropriate states for the agent and the environment, and with knowledge of the dynamical laws and initial conditions, the state-space equations are a complete representation of a given system. Note that the environmental states \mathbf{x}_E that go into Equations (1) and (2) refer to the local environment of the agent, which can be affected by the activities of the agent. This local environment is also affected by larger-scale environmental factors, which here would enter as external inputs \mathbf{u} . For example, the availability of glucose around a biofilm (\mathbf{x}_E , local environment) is affected by the metabolism of the biofilm (\mathbf{x}_A) as well as the bulk concentration of glucose (\mathbf{u}).

The agent–environment coupled dynamical system as described above can be represented with the simplified block diagram given in Figure 1a, where A is the agent, E is the environment interacting with the agent, \mathbf{u} is a set of inputs and \mathbf{y} is a set of measurements that we can obtain from the system, that we call outputs. The dotted lines represent the coupled agent–environment system. Figure 1b shows a more detailed representation of the coupled system, which is a graphic representation of Equations (1) and (2). Note that the outputs \mathbf{y} are a combination of the states of the agent and the

environment given by a function g . Also note that \mathbf{u} contains both measurable and unmeasurable inputs. The latter are considered process noise.

2.2 | Building an input–output model

The central idea of the framework proposed in this work is to represent the organism–environment system using an input–output model. This type of model relates the inputs and outputs in time of the dynamical system and can be used to make predictions and study the system's dynamics (Ljung & Glad, 2016). Particularly, we focus here on the general case in which nothing about the real network of interactions between organism and environment is known. That means that we only have access to measurements of the inputs and outputs, something that would be true in many real-life situations. For example, we may have data on the amount of solar energy received in a patch of soil, the soil temperature near the egg, and the sex of hatchling turtles emerging from their nests, but limited knowledge of the mechanisms of temperature-dependent sex determination, or how embryos shift their position within the egg in response to natural temperature fluctuations (Ye et al., 2019). In such a general case, we can build “black-box” models (Ljung & Glad, 2016). The parameters of these models lack direct biological meaning but represent the properties of the input–output

relationships of the system. Notably, when some insight about the interactions is known, a “gray-box” model can be built with similar ideas to the ones presented here. These models can therefore be used to directly measure and interpret biological and physical variables (Ljung & Glad, 2016).

Here, we are interested in describing the dynamics of the coupled organism–environment system using an input–output model. For simplicity, we focus on models with one input and one output, but more inputs and outputs can be accommodated as suggested below. Input–output models are typically described in discrete time, since data are collected from the system in sampled form (Ljung & Glad, 2016). This means that, experimentally, one has access to N data points of inputs $u(i)$ and outputs $y(i)$ for $i = 1, 2, 3, \dots, N$. We focus on models that predict the value of the output at time i as a function of both the inputs u and outputs y at times $j < i$,

$$\hat{y}(i) = F(u(i-1), u(i-2), \dots, u(i-n_u), y(i-1), y(i-2), \dots, y(i-n_y)) \quad (3)$$

where F is a function that combines the past n_u inputs and n_y outputs to predict the output at time i , symbolized $\hat{y}(i)$. The difference between $y(i)$ and $\hat{y}(i)$ is the prediction error. Equation (3) is schematically represented in Figure 1c.

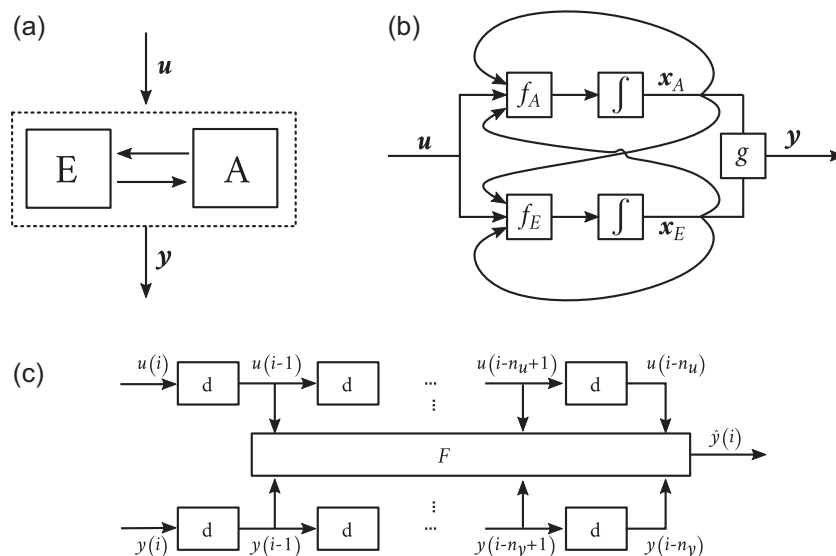


FIGURE 1 Different representations of the framework. (a) is a simplified schematic representing agent A and environment E interacting with each other, forming the coupled system represented by the dotted lines. There are inputs entering the system and outputs that can be measured. (b) represents Equations (1) and (2) in block diagrams. The symbol \int represents the integration with respect to time, since f_A and f_E return the time derivatives of x_A and x_E , respectively. The outputs of the system y are a combination of the state variables of the coupled system. That combination is represented as a function g . (c) represents the general approach proposed in this paper, given in Equation (3). Operator d is the delay operator, which returns a delayed copy of a signal. The parameters n_u and n_y determine the number of past inputs and outputs, respectively, that are used for prediction of the output at time i which is symbolized $\hat{y}(i)$.

An important note is that not all nonlinear systems can be represented in the form given in Equation (3). Leontaritis and Billings (1985) prove that nonlinear input–output representations exist provided that the system is finitely realizable, and a linearized model would exist if the system were operated close to an equilibrium point (Billings, 2013, Ch. 2). Based on these results it has been demonstrated that many real systems can be modeled using this representation (Billings, 2013; Mu et al., 2005; Schoukens & Ljung, 2019), and in the remaining of this paper, we assume that these conditions for existence hold.

The process to build an input–output model is known as system identification (Pillonetto et al., 2022; Söderström & Stoica, 1989) and is composed of four main steps, which we describe in some detail below. There is a very rich literature on input–output modeling spanning several decades, so we refer the interested reader to these more detailed and formal treatments (Ljung, 1987; Schoukens & Ljung, 2019; Söderström & Stoica, 1989). Our focus will be on an experimental setting where the inputs are subject to manipulation, the system's output can be measured, and the resulting measured data is used to fit the model.

The steps to build the model are:

Step 0: determine the inputs and outputs of the system. This requires prior knowledge of the system, and commonly a hypothesis (e.g., that temperature has some effect on sex determination). In an experimental set-up, the input will also be constrained by which variables can be controlled, as well as how they can be controlled. Multiple inputs can be measured and used for prediction by including their delayed values in the prediction Equation (3). Multiple outputs can be predicted by building a separate predictive model following Equation (3) for each output.

Step 1: collecting the data. Obtaining the input–output measurements is a central step to identify a model under this framework, since it is fully data-driven. Experiments should then be thoughtfully designed to obtain data in the domain of interest, introducing input signals that bring out the features of the system that will be relevant for future predictions. Such signals are known as “persistently-exciting signals” (Ljung, 1987). The sampling time for the measurements of inputs and outputs is important and should be short enough to capture the fastest temporal changes in the measurements, but not any shorter to reduce computational costs (Ljung & Glad, 2016). There is no theoretical justification to the minimal amount of data samples needed to identify a model (Munch et al., 2022; Schoukens & Ljung, 2019), since this amount will depend on the dimensionality of the system, the types of nonlinearities present and the quality of the data.

Step 2: choosing the model structure. This refers to what we assume about the function F in Equation (3). If we assume that the function F is linear, the model given by Equation (3) is known as ARX (AutoRegressive with eXogenous input). In these linear models, the output at time i is a linear combination of the inputs and outputs at previous times. However, we do not expect in general that the organism–environment behavior will be correctly captured by a linear model, so we do not assume that F in Equation (3) is linear. Note that if the true dynamics of the system happen to be linear, a nonlinear model will still provide good predictions if it has been correctly fitted and validated (see below).

Step 3: fitting the model structure with the data. Fitting the model means using the measured time-series data to determine the input–output relationships. Multiple algorithms exist for this, including artificial neural networks, Gaussian processes, and k -nearest-neighbor algorithms (Billings, 2013; Friedman et al., 1977; Schoukens & Ljung, 2019). For example, a neural network can learn the relationship between input data and output data through supervised learning (Mu et al., 2005; Schoukens & Ljung, 2019). Particularly for ecological data, k -nearest-neighbor algorithms have been shown to provide good predictions within the EDM framework (Munch et al., 2022; Sugihara et al., 2012). These are nonparametric algorithms that predict the value of new data points based on the values of their closest neighbors in the data set used for fitting. The number of regressors (i.e., n_u and n_y) required for prediction will depend on the details of the system including the number of internal states. Therefore, here we suggest determining these numbers, together with the hyper-parameters of the algorithm used to fit the data, through cross-validation as explained below.

Step 4: model validation. Ultimately, we are interested in a model that predicts the behavior of the system, so it is natural to use prediction error as a measure of the usefulness of the model. Cross-validation is a common and practical tool for model validation that assesses how well a model reproduces the behavior of new datasets not used to fit the model. Given new input–output data, one can compare the predicted output using the model and input data against the actual observed output data, to check if the models meet the desired use. The validation typically consists of inspecting predictions plots as well as measuring prediction error but can also include other tests such as a whiteness test of residuals, cross-correlation test between input and output residuals or higher order moments tests (Schoukens & Ljung, 2019). If the model does not meet the user's requirements, then aspects of the model need to be revised, including number of regressors, the fitting algorithm, and the data

collection. In this way, identification is an iterative problem.

With a fitted and validated model, it is possible to predict how the dynamical system will respond to new input signals. These inputs can be different in terms of magnitude or timing than the data used to fit the model. Specifically, the model allows us to predict the output $\hat{y}(i)$ given information of y and u up to time $i - 1$. This is known as the one-step-ahead predictor. The same model, however, can be used recursively to predict the behavior of the system for more steps, by using the predictions $\hat{y}(i - 1), \hat{y}(i - 2), \dots$ instead of $y(i - 1), y(i - 2), \dots$. This makes it possible to predict the complete trajectory of the output \hat{y} (i.e., several timesteps ahead) under novel experimental conditions (i.e., sequence of inputs) that are not part of the set used for identification, given some initial conditions.

2.3 | Application to phenotypic plasticity

We use the general framework described in the previous section to study how individuals modify their phenotype in response to their environment (i.e., phenotypic plasticity). Particularly, we show how the framework allows us to study plasticity as a dynamical property, that is, a property that changes in time during the development and lifetime of the organism.

In our example, we have an organism inside a bioreactor. The variables are summarized in Table 1. The environmental state variable is the temperature inside the perfectly mixed bioreactor, which is affected by the temperature of fluid entering the bioreactor. This effect is modeled as a first-order differential equation, which represents the fact that the temperature inside the reactor will change gradually as fluid of a different temperature enters (see equation in Figure 2a). The organism is simulated as a gene regulatory network with two genes (Gardner et al., 2000; see Figure 2a). Note that

this gene network is not trying to represent the biology of any given “real” system, but rather it is used as a generic representation of an agent with a dynamical behavior. The states of the agent are the expression levels of the two genes. Gene expression levels are also the outputs of the system (i.e., the measurements we sample from the system). The gene network has a parameter that is affected by the temperature inside the bioreactor (Figure 2a). The expression of gene 2 has an effect on temperature (e.g., by generating heat), creating a two-way interaction between organism and environment as represented in Figure 1a.

To study the system, we perform 10 *in silico* experiments at constant input values. In such experiments, we measure the output of the system in time (i.e., the gene expression for genes 1 and 2). We also have the value of the input (i.e., the constant temperature of the incoming fluid). Figure 2b shows the results of these experiments, each plotted as a trajectory with arrows of different colors in the gene 1–gene 2 plane (i.e., the state-space of the agent states). The arrows in this figure represent the direction of change in the level of expression for genes 1 and 2 at different time points of the experiments. The results in Figure 2b shows that, depending on the input temperature u , the system shows two qualitatively distinct behaviors. For some input values, the system goes to a state of high expression of gene 1 and low expression of gene 2. For other values of u , the system reaches a steady state with low expression of gene 1 and high expression of gene 2.

We now turn to the goal of building a black-box input–output model to predict the behavior of the system. Here, it is important to clarify the distinction between the organism–environment system, here simulated using the equations in Figure 2a, and the predictive input–output model. The latter is built, as explained in the previous section, only using input and output data points and without a priori knowledge of the details of the system (i.e., without knowledge of the equations in Figure 2a).

General framework	Plasticity example	Symbols
States of the agent (\mathbf{x}_A)	Gene expression of genes 1 and 2 (G_1, G_2)	$\mathbf{x}_A = (G_1, G_2)$
States of the environment (\mathbf{x}_E)	Temperature in the reactor (T)	$\mathbf{x}_E = T$
Input (\mathbf{u})	Temperature of incoming fluid (u)	$\mathbf{u} = u$
Output (\mathbf{y})	Gene expression of genes 1 and 2 (G_1, G_2)	$\mathbf{y} = (G_1, G_2)$

TABLE 1 Definition of the variables for the plasticity example.

Note: The table presents a summary of the variables as they appear in the description of the general framework, along with specific definitions and symbols relevant to the plasticity example.

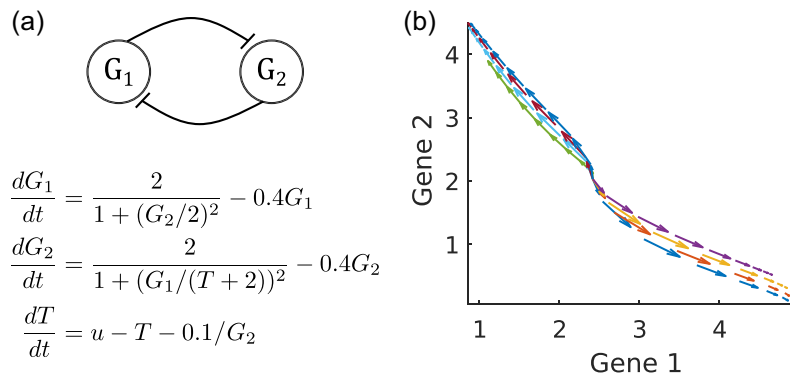


FIGURE 2 An organism–environment system to study plasticity. (a) shows the gene network used in this study, where G_1 and G_2 are genes 1 and 2, respectively (Gardner et al., 2000). The equations represent the dynamics of the system, including the change in the states of the agent (i.e., the expression levels G_1 and G_2) and the change in the state of the environment (i.e., the temperature T). The equations for the change in G_1 and G_2 represent the gene network. The change in the temperature inside the bioreactor (T) depends on the temperature of the fluid entering the bioreactor (u) and the expression of gene 2 (G_2). (b) shows the dynamics of the outputs for different constant inputs. In this example, the outputs are also the states of the agent, namely the gene expression levels of genes 1 and 2.

As explained in the previous section, the first step to build the model is to obtain the data. For this, we perform 100 in silico simulations where the input is a random walk with a fixed step size of 0.02 added or subtracted every 10 sample intervals. Sample interval was 0.5 units of time, which was short enough to capture the fastest dynamics of the system (see Supporting Information: Figure 1). In practice, this means introducing fluid at a constant temperature for 10 sample intervals, and then changing the temperature of the fluid by an amount of 0.02. This type of signal brings out all the relevant features of the system (i.e., persistently-exciting; see Ljung, 1987).

For steps 2 and 3, we build a black-box model as shown in Equation (3) and use a k -nearest-neighbor algorithm to fit the model (Friedman et al., 1977). The algorithm works by first building a reference table of the training data, where the input–output sequences $u(i-1), u(i-2), \dots, u(i-n_u), y(i-1), y(i-2), \dots, y(i-n_y)$ are assigned the target value $y(i)$, following Equation (3). To make predictions of new input–output data, the algorithm searches the k most similar time sequences of inputs and outputs in the training data and provides the weighted average of the corresponding target values as the predicted value for that time point. The measure of “similarity” that we use to determine which are the neighbors is Euclidean distance, but other metrics can be used. The number of neighbors used for prediction (i.e., k) is a hyperparameter that needs to be defined by the user and cross-validated. Note that the algorithm is analogous to “kernels” in machine learning and control theory (Schoukens & Ljung, 2019).

The next step to build the model is the validation process. For this, we try different values of n_u , n_y , and k

and find parameter combinations that yield the best predictions for new test data. The prediction errors are given in Supporting Information: Table 1. We found the best performance in the validation data for $n_u = 3$, $n_y = 3$, and $k = 3$. This means that we have the recursive expression,

$$\hat{y}_j(i) = F_j(u(i-1), u(i-2), u(i-3), \hat{y}_j(i-1), \hat{y}_j(i-2), \hat{y}_j(i-3)), j = 1, 2, \quad (4)$$

where j indexes genes 1 and 2. The equation above expresses that the output at time i (i.e., the expression level of gene j at time i) is a function F_j of the inputs and outputs at times $i-1$, $i-2$, and $i-3$. Note that both $\hat{y}_j(i-1)$, $\hat{y}_j(i-2)$, and $\hat{y}_j(i-3)$ are themselves a function of previous inputs and outputs. This means that we can predict the complete trajectory several steps ahead of the output y using the recursive expression (4). Finally, $k = 3$ means that for each prediction, we average the three data points in the training data that are closest to the new data we want to predict.

We now use the fitted model to predict the trajectory of the outputs for novel input signals. The first row of Figure 3 shows the predicted trajectories for constant input signals (i.e., where the input temperature is held constant during the experiment after a short period of zero input). As shown in the figure, despite constant signals not being a part of the data set used to fit the model, it is able to provide good predictions for the trajectories of the outputs.

The previous simulations use a constant input. However, one of the main strengths of the proposed modeling framework is that it allows us to study

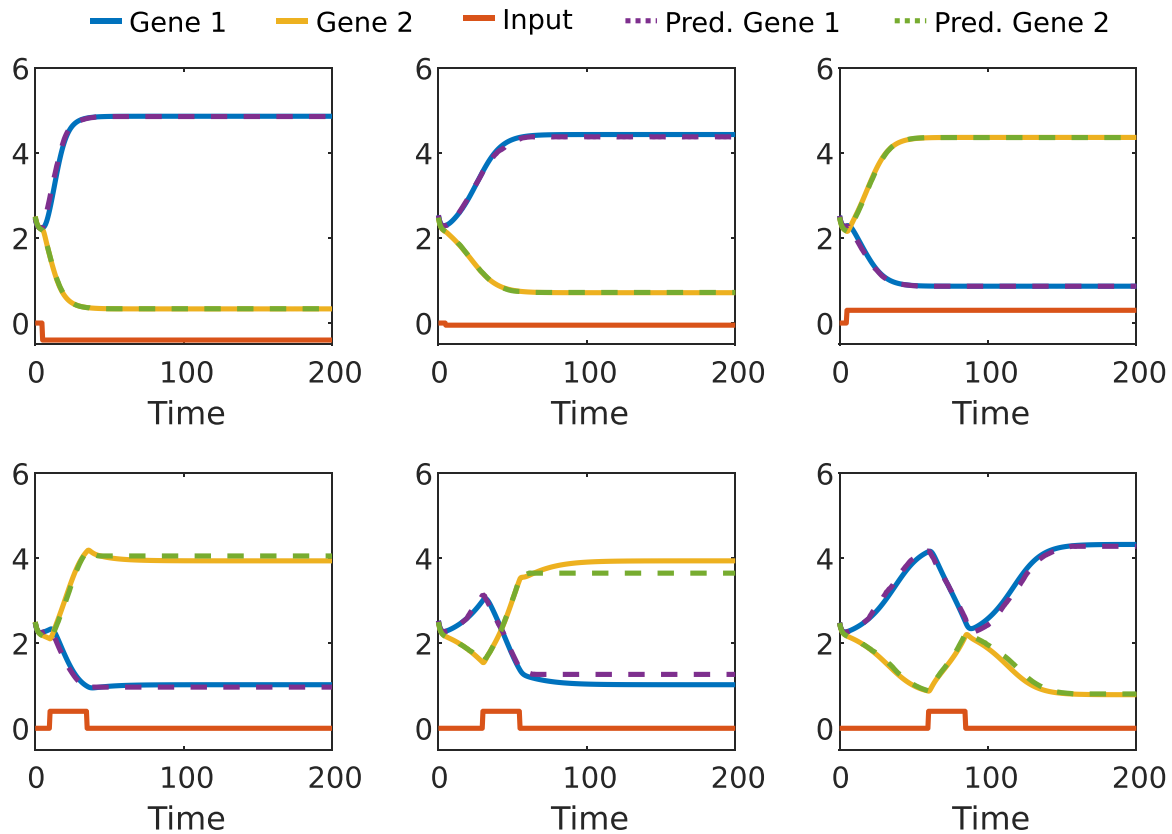


FIGURE 3 Input–output modeling for phenotypic plasticity for constant and pulse inputs. Observed predicted outputs for different input signals. The first row shows the outputs (i.e., expression of genes 1 and 2) in time for constant input signals (i.e., constant temperature of the incoming fluid). The second row shows the outputs for a pulse signal of the input. Note that both constant and pulse signals are not signals used in the training set (see Supporting Information: Figure 1 and main text).

plasticity as a dynamical property. In this way, we can study how a pulse of temperature affects the agent, depending on the timing of that pulse. To study this, we perform experiments in which the input temperature of the fluid is raised for some time, and then lowered to its basal value. In other words, the input signal is now a function of time, in particular, a rectangle function.

When exploring this system, we see that depending on the timing of the pulse, the effect on the agent will be different. As shown in the second row of Figure 3, if the pulse occurs early enough, the agent will go to a state with high expression of gene 2 and low expression of gene 1. However, if the pulse is done later, the agent reaches the alternative state with low expression of gene 2 and high expression of gene 1. This highlights the important dynamical properties of the system: it is not only the magnitude, but also the timing of the perturbation that will influence the behavior of the system. This is a well-known behavior of developmental systems, and the period during which the system is responsive to perturbation is commonly referred to as the critical or sensitive window (Burggren, 2020; Crews, 1996; Fawcett & Frankenhuis, 2015). As shown in Figure 3, the fitted

input–output model can predict the trajectory of the outputs in time for these types of signals.

The modeling framework allows us to measure plasticity as a dynamical property. The idea behind this is that we can measure, at each time point, the number of places where the system can be driven by modifying the input. This gives us an idea of how “plastically responsive” the system is at different time points. Essentially, this involves predicting the system’s behavior for different constant inputs in a window of future time that we refer to as the prediction horizon (see Figure 4a). The variance of the predicted outputs at the end of the prediction horizon measures the spread of final output values that the system can be driven to by modifying the input. If the output is some property of the agent (i.e., a phenotype), this is a measure of phenotypic plasticity. This idea is shown schematically in Figure 4b.

We test our dynamical measure of plasticity for the expression of gene 2 in the example gene regulatory network. We use a prediction horizon of 50 sample time steps and test a series of constant input signals ranging from -0.5 to 0.5 with step size of 0.02 . Figure 4c shows the variance of the distribution of outputs at the end of

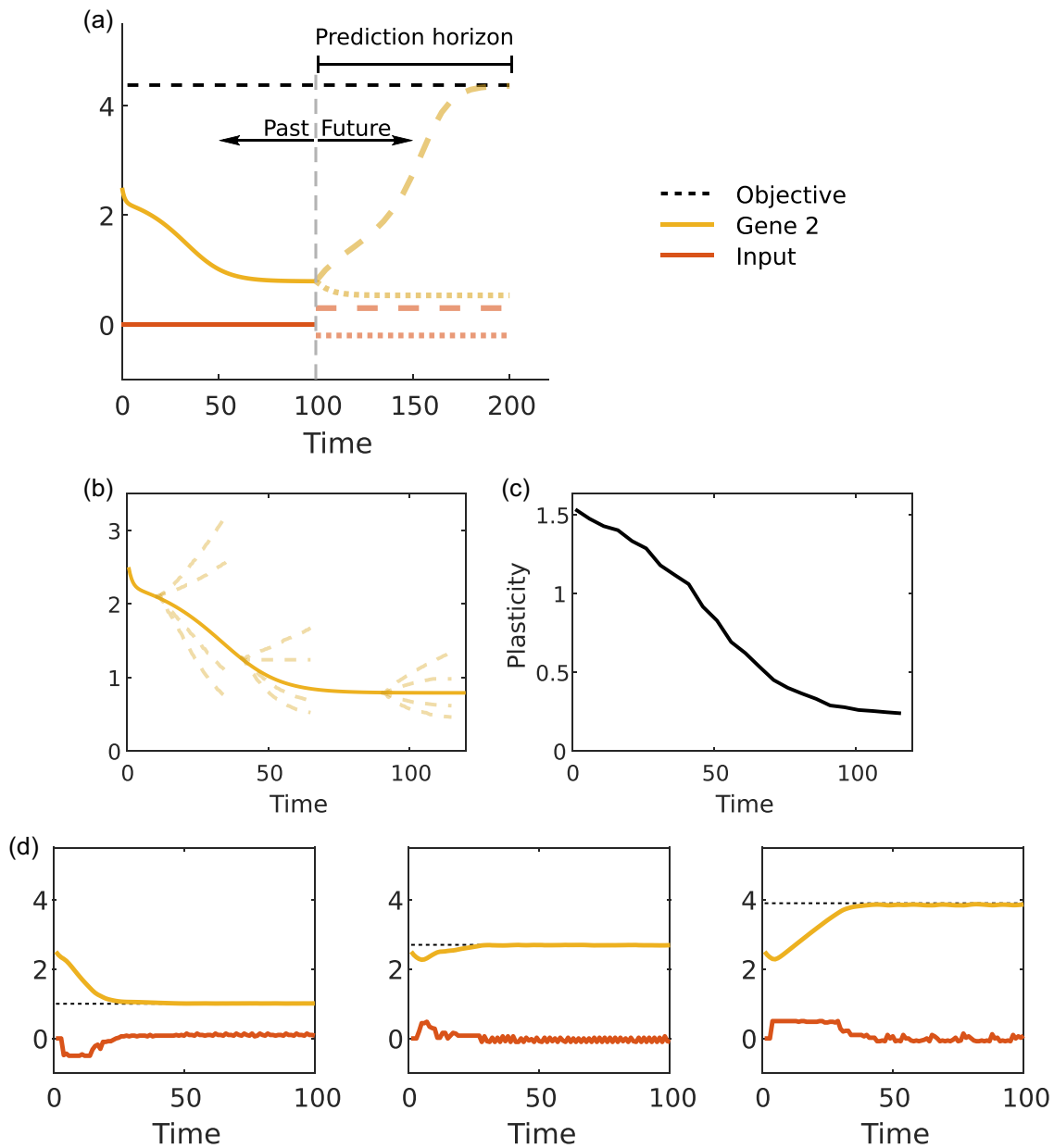


FIGURE 4 Using input–output modeling to measure plasticity dynamically in time and control the system. (a) illustrates a general approach to measuring plasticity over time and controlling a system. Using the fitted predictive input–output model, the output variable can be forecasted for different possible inputs at each time point, over a “prediction horizon.” In the diagram, a positive input value (represented by dashed lines) drives the predicted output (gene 2, also in dashed lines) toward a higher value, while a negative input value drives the predicted output toward values closer to 0 (both represented by corresponding dotted lines). Thus, if the objective is to drive the output to a specific value (represented by dashed lines), a positive input value is chosen. This is the basis of model predictive control, which we use to control the expression of gene 2 toward different objective values, as shown in Panel (d). (b) represents the idea behind the dynamic measure of plasticity. At each time, different input values can be used to predict the various values that the output can be driven to in the prediction horizon, shown as dashed lines. The variance of the predicted values at the end of the prediction horizon provides a measure of plasticity in time. (c) illustrates this measure of plasticity in time for gene 2, indicating that it is higher in the early stages of development but decreases over time.

the prediction horizon for all constant inputs tried. It is clear from this measure that the system is highly plastic at the start of the experiment, but its ability to respond to environmental inputs decreases toward the end. This reflects the results shown before for the pulse

experiments, where the pulse could drive the system to a new state only when introduced early in development.

The idea of predicting the future behavior of the system at each time point can also be used to control the system. This is the basis of model predictive control

(Camacho & Alba, 2013). Essentially, from the various tested input sequences at a given time point, we can identify the one that brings the system's output closest to an objective state that we define a priori (see dashed line in Figure 4a,d). We then apply the most effective control input to the system and allow it to evolve to the next sample point. The process is then repeated to calculate the next most effective input for the system. We use this idea and show that the expression of gene 2 can be controlled by modifying the input in Figure 4d. The figure shows how model predictive control allows to drive the output to three different objective values defined a priori, by deciding the most appropriate input value at each time point. Note that to keep gene expression at the objective value, the system must be permanently excited by a non-zero input.

3 | DISCUSSION

The framework we introduce combines desirable properties of both process-based and statistical models. From the first, it takes the dynamical description of the organism–environment coupled system. From the second, it takes the fact that these models can be fitted with measured empirical data without detailed knowledge of the internal causal interactions. Notably, the framework proposed here is not in contradiction with the alternatives mentioned above. That is, a process-based model can be readily used to predict the output signal for a given input signal. Furthermore, it is straightforward to calculate the covariation structure of a set of outputs at a given time point.

The framework proposed here can be interpreted as machine learning for input–output relationships of an organism–environment system. This means that the objective is to learn the relationship between the output at time i , $y(i)$, with delayed copies of the input and output signals, that is $y(i-1)$, $y(i-2)$, ..., and $u(i-1)$, $u(i-2)$, ... as given in Equation (3) (Schoukens & Ljung, 2019). In this way, we do not suggest learning a static relationship between a measurement of the environment (e.g., a given temperature) and a measurement of the organism (e.g., the adult phenotype). Instead, we suggest learning the dynamical relationship between the inputs and outputs of the system in time. This key distinction is what allows us to model dynamical properties of the system.

Preserving the dynamical nature of the organism–environment system is crucial as it accurately represents the fact that the system has memory: past state values affect current state values. This feature is characteristic of any “generative” or “constructive”

process and is captured by the general Equation (3) of the predictive framework. The co-construction of the organism and its environment at each moment in time is fundamental to phenomena such as development, plasticity, learned behavior, and niche construction. Here we illustrate the use of this type of modeling for a simple gene regulatory network, and its efficiency to model more complex organism–environment systems is left as an open question.

The framework proposed here makes it possible to study plasticity as a dynamical property, a property that changes in developmental time. Specifically, it allowed us to define a dynamic measure of plasticity as the variance of the possible values that the output can be driven to by modifying inputs (Figure 4b,c). This is useful to detect stages in development that are especially sensitive to environmental effects. For the gene regulatory network we study, we find that the system is particularly plastic in earlier stages of development. The presence of sensitive periods is well-known in biology. The causes and consequences of sensitive periods are of substantial interest to both developmental and evolutionary biology, which is well illustrated by the literature on temperature-dependent sex determination and other polyphenisms (Nijhout, 2003), cognitive development in humans and other animals (Frankenhuis & Walasek, 2020), and the developmental origins of health and disease (Gluckman & Hanson, 2004). The latter considers adult disease, and the effectiveness of preventive intervention, to depend heavily on the timing of environmental insult early in life (Kuijper et al., 2019).

It is because plasticity is a dynamic property that such phenomena are difficult to capture satisfactorily with a representation of plasticity that does not incorporate developmental time. This is often the case with reaction norms, when used to study the relationship between an environmental variable and the phenotype at a fixed stage of an organism's lifetime (e.g., adult). A notable exception explicitly incorporating ontogenic time is developmental reaction norms (Pigliucci et al., 1996), which have also been studied as types of function-valued traits (Gomulkiewicz et al., 2018). Bayesian models of development are a different type of representation that is also able to capture certain dynamical aspects of plasticity during development (Stamps & Frankenhuis, 2016). Like the framework we propose, these approaches are dynamic but suitable to address a different set of questions. In particular, the framework explained here provides a way to predict how the system will respond to novel inputs and provides an opportunity to control the output to a desired state (e.g., healthy). In this way, while the dynamic approach explained here requires more data to fit models when compared to static descriptions, its potential applications suggest that the

collection of such data will be worthwhile. Note that, while preserving the dynamic aspect, the framework proposed here makes its own set of simplifications about the complexity of the biological system (e.g., arguably treating the organism as a machine). However, the framework provides a pragmatic tool that can be useful for certain applications.

While black-box modeling is uninformative about the underlying biological mechanisms, causal interaction among internal components of the black-box can be found under certain conditions. Sugihara et al., 2012 used EDM—a type of black-box modeling—to infer causal relationships based on the predictability of a system's dynamics. Briefly, the method tests for causation by measuring the extent to which the historical record of one state can reliably estimate another state, in which case the first state causally influences the second. While these ideas were mostly tested in the context of species interactions in chaotic systems in an attractor, they can be extended to the systems studied here if internal states can be measured, to study causal interactions among organism and environment states. Making data-driven methods more interpretable is currently an active field of research, and recent efforts also exist to reconstruct the state-space equations directly from measured time-series data (e.g., sparse identification of nonlinear dynamics algorithm, SINDy; Brunton & Kutz, 2022; Brunton et al., 2016). While these methods are currently limited in the dimensionality of the systems they can be applied to, they represent a promising venue for discovering the underlying equations of the coupled organism–environment system.

The ability of organisms to modify their development in changing environments is fundamental to many key questions in evolutionary developmental biology and ecology (Gilbert et al., 2015). However, such questions cannot be adequately studied with models that do not capture the dynamical nature of the interaction between the organism and environment. Therefore, despite the challenge this implies, it should be a central objective to develop modeling frameworks that correctly represent the organism–environment system and can be fitted with data. The framework presented here is a step in this direction.

ACKNOWLEDGMENTS

The authors thank Ruben H. Milocco for insightful discussions on modeling and system identification, Willem Frankenhuis, Nicole Walasek, one anonymous reviewer, Sonia Sultan, and the Reading Group from the Agency, Directionality, and Function project for comments on the manuscript. The authors thank the organizers and participants of the KLI workshop on Agency in Living Systems for discussion and the John Templeton Foundation (#62220) for financial support.

The opinions expressed in this paper are those of the authors and not those of the John Templeton Foundation.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data for simulations available in <https://github.com/millisan/InputOutput>.

ORCID

Lisandro Milocco  <http://orcid.org/0000-0003-3953-0407>

Tobias Uller  <http://orcid.org/0000-0003-1293-5842>

REFERENCES

- Armbruster, W. S., Pélabon, C., Bolstad, G. H., & Hansen, T. F. (2014). Integrated phenotypes: Understanding trait covariation in plants and animals. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 369(1649), 20130245.
- Beer, R. D. (1995). A dynamical systems perspective on agent–environment interaction. *Artificial Intelligence*, 72(1–2), 173–215.
- Billings, S. A. (2013). *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons.
- Brias, A., & Munch, S. B. (2021). Ecosystem based multi-species management using Empirical Dynamic Programming. *Ecological Modelling*, 441, 109423.
- Brunton, S. L., & Kutz, J. N. (2022). *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press.
- Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113(15), 3932–3937.
- Burggren, W. W. (2020). Phenotypic switching resulting from developmental plasticity: Fixed or reversible? *Frontiers in Physiology*, 10, 1634.
- Camacho, E. F., & Alba, C. B. (2013). *Model predictive control*. Springer Science & Business Media.
- Crews, D. (1996). Temperature-dependent sex determination: The interplay of steroid hormones and temperature. *Zoological Science*, 13(1), 1–13.
- Deyle, E. R., Maher, M. C., Hernandez, R. D., Basu, S., & Sugihara, G. (2016). Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 113(46), 13081–13086.
- Dingemans, N. J., Kazem, A. J., Réale, D., & Wright, J. (2010). Behavioural reaction norms: Animal personality meets individual plasticity. *Trends in Ecology & Evolution*, 25(2), 81–89.
- Fawcett, T. W., & Frankenhuis, W. E. (2015). Adaptive explanations for sensitive windows in development. *Frontiers in Zoology*, 12(1), S3.

- Frankenhuis, W. E., & Walasek, N. (2020). Modeling the evolution of sensitive periods. *Developmental Cognitive Neuroscience*, *41*, 100715.
- Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, *3*(3), 209–226.
- Gardner, T. S., Cantor, C. R., & Collins, J. J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, *403*(6767), 339–342.
- Gilbert, S. F., Bosch, T. C. G., & Ledón-Rettig, C. (2015). Eco-Evo-Devo: Developmental symbiosis and developmental plasticity as evolutionary agents. *Nature Reviews Genetics*, *16*(10), 611–622.
- Gluckman, P. D., & Hanson, M. A. (2004). Living with the past: Evolution, development, and patterns of disease. *Science*, *305*(5691), 1733–1736.
- Gomulkiewicz, R., Kingsolver, J. G., Carter, P. A., & Heckman, N. (2018). Variation and evolution of function-valued traits. *Annual Review of Ecology, Evolution, and Systematics*, *49*, 139–164.
- Hubbe, A., Machado, F. A., Melo, D., Garcia, G., Sebastião, H., Porto, A., Cheverud, J., & Marroig, G. (2023). Morphological integration during postnatal ontogeny: Implications for evolutionary biology. *Evolution*, *77*(3), 763–775.
- Jaeger, J., & Monk, N. (2014). Bioattractors: Dynamical systems theory and the evolution of regulatory processes. *The Journal of Physiology*, *592*(11), 2267–2281.
- Kauffman, S. (2003). Molecular autonomous agents. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *361*(1807), 1089–1099.
- Kuijper, B., Hanson, M. A., Vitikainen, E. I. K., Marshall, H. H., Ozanne, S. E., & Cant, M. A. (2019). Developing differences: Early-life effects and evolutionary medicine. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *374*(1770), 20190039.
- Leontaritis, I. J., & Billings, S. A. (1985). Input–output parametric models for non-linear systems part I: Deterministic non-linear systems. *International Journal of Control*, *41*(2), 303–328.
- Ljung, L. (1987). *System identification: Theory for the user*. Prentice Hall.
- Ljung, L., & Glad, T. (2016). *Modeling and identification of dynamic systems* (1st ed.). Studentlitteratur.
- Milocco, L., & Salazar-Ciudad, I. (2022). A method to predict the response to directional selection using a Kalman filter. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(28), e2117916119.
- Mitteroecker, P., & Bookstein, F. (2009). The ontogenetic trajectory of the phenotypic covariance matrix, with examples from craniofacial shape in rats and humans. *Evolution*, *63*(3), 727–737.
- Mu, J., Rees, D., & Liu, G. P. (2005). Advanced controller design for aircraft gas turbine engines. *Control Engineering Practice*, *13*(8), 1001–1015.
- Munch, S. B., Rogers, T. L., & Sugihara, G. (2022). Recent developments in empirical dynamic modelling. *Methods in Ecology and Evolution*, *14*(3), 732–745.
- Nijhout, H. F. (2003). Development and evolution of adaptive polyphenisms. *Evolution & Development*, *5*(1), 9–18.
- Nilsson-Örtman, V., Rogell, B., Stoks, R., & Johansson, F. (2015). Ontogenetic changes in genetic variances of age-dependent plasticity along a latitudinal gradient. *Heredity*, *115*(4), 366–378.
- Nussey, D. H., Wilson, A. J., & Brommer, J. E. (2007). The evolutionary ecology of individual phenotypic plasticity in wild populations. *Journal of Evolutionary Biology*, *20*(3), 831–844.
- Odling Smee, F. J., Laland, K. N., & Feldman, M. (2003). *Niche construction*. Princeton University Press.
- Pigliucci, M., Schlichting, C. D., Jones, C. S., & Schwenk, K. (1996). Developmental reaction norms: The interactions among allometry, ontogeny and plasticity. *Plant Species Biology*, *11*(1), 69–85.
- Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G., & Ljung, L. (2022). *Regularized system identification: Learning dynamic models from data*.
- Pottier, P., Burke, S., Zhang, R. Y., Noble, D. W. A., Schwanz, L. E., Drobnjak, S. M., & Nakagawa, S. (2022). Developmental plasticity in thermal tolerance: Ontogenetic variation, persistence, and future directions. *Ecology Letters*, *25*(10), 2245–2268.
- Salazar-Ciudad, I., & Jernvall, J. (2010). A computational model of teeth and the developmental origins of morphological variation. *Nature*, *464*(7288), 583–586.
- Schlichting, C. D., & Pigliucci, M. (1998). *Phenotypic evolution: A reaction norm perspective*. Sinauer Associates Incorporated.
- Schoukens, J., & Ljung, L. (2019). Nonlinear system identification: A user-oriented road map. *IEEE Control Systems*, *39*(6), 28–99.
- Söderström, T., & Stoica, P. (1989). *System identification*. Prentice Hall.
- Stamps, J. A., & Frankenhuis, W. E. (2016). Bayesian models of development. *Trends in Ecology & Evolution*, *31*(4), 260–268.
- Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*(6106), 496–500.
- Sultan, S. E. (2007). Development in context: The timely emergence of eco-devo. *Trends in Ecology & Evolution*, *22*(11), 575–582.
- Sultan, S. E., Moczek, A. P., & Walsh, D. (2022). Bridging the explanatory gaps: What can we learn from a biological agency perspective? *BioEssays*, *44*(1), 2100185.
- Verd, B., Monk, N. A., & Jaeger, J. (2019). Modularity, criticality, and evolvability of a developmental gene regulatory network. *eLife*, *8*, e42832.
- Ye, Y. Z., Ma, L., Sun, B. J., Li, T., Wang, Y., Shine, R., & Du, W. G. (2019). The embryos of turtles can influence their own sexual destinies. *Current Biology*, *29*(16), 2597–2603.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Milocco, L., Uller, T. (2023). A data-driven framework to model the organism–environment system. *Evolution & Development*, 1–12.

<https://doi.org/10.1111/ede.12449>