Automated adaptive inference of phenomenological dynamical models

Bryan C. Daniels

*Center for Complexity and Collective Computation, Wisconsin Institute for Discovery, University of Wisconsin, Madison, WI 53715, USA**

Ilya Nemenman

Departments of Physics and Biology, Emory University, Atlanta, GA 30322, USA†

Abstract

Dynamics of complex natural and artificial systems is often driven by large and intricate networks of microscopic interactions, whose sheer size obfuscates understanding. In light of limited experimental data, many parameters of such dynamics are unknown, and thus models built on the detailed, mechanistic viewpoint risk overfitting and making faulty predictions. At the other extreme, simple *ad hoc* models often miss defining features of the underlying systems. Targeting modern biophysics applications, here we develop an approach that instead constructs *phenomenological*, coarse-grained models of network dynamics that automatically adapt their complexity to the amount of available data. Such adaptive models lead to accurate predictions even when microscopic details of the studied systems are unknown due to insufficient data. The approach is computationally tractable, even for a relatively large number of dynamical variables. For example, it correctly infers the phase space structure for simulated planetary motion data, avoids overfitting in a complex biological signaling system, and produces accurate predictions for a yeast glycolysis model with only tens of data points and over half of the interacting species unobserved.

*Electronic address: bdaniels@discovery.wisc.edu

†Electronic address: ilya.nemenman@emory.edu

One can view the physics enterprise as reverse-engineering Nature — using data to infer predictive mathematical models of physical systems, and then finding similarities among such models of distinct systems to identify physical laws. In the era of Big Data, these models are becoming Big Models, which are often as complicated as the data themselves, reflecting the humorous maxim that "the best material model of a cat is another, or preferably the same, cat" [1]. This is especially evident in modern biophysics and systems biology [2], which are the primary focus of this article. Continued success of such approaches that systematize all known details in a combinatorially large mathematical model is uncertain. Indeed, generalizing and generating insight from complex models is difficult. Further, specification of myriads of microscopic mechanistic parameters in such models demands vast data sets and computational resources, and is hard even for very large data sets due to widely varying sensitivities of predictions to the parameters [3]. Finally, the very structures of these models are often unknown because they depend on many yet-unobserved players on the microscopic level. Identification of these structural characteristics is labor intensive and does not scale up easily. Thus it is unlikely that mathematical models based solely on a detailed microscopic representation will be able to account accurately for the observed dynamics of many complex systems. More importantly, even if they could, the resulting models would be too unwieldy to bring about understanding of the modeled systems. Model reduction may alleviate some of these problems, but it still suffers from the difficulty of needing an exact, detailed model as an intermediate step [4–7].

Because of these difficulties, the need to predict responses of complex systems to dynamical perturbations has led to a resurgence of research into automated inference of dynamical systems from time series data, which had been attempted since the early days of the field of nonlinear dynamics [8, 9]. Approaches have been developed using linear dynamic models [10], Bayesian Networks (see *Supplementary Information (SI)*), recurrent neural networks [11], evolved regulatory networks [12], and symbolic regression [13, 14]. The latter two produce models that are more mechanistically accurate and interpretable. However, because of the focus on microscopic accuracy, these approaches require searching through an extremely large space of all possible microscopic dynamics. In general, this leads to very long search times [12, 14], especially if some underlying variables are unobserved, and dynamics are coupled and cannot be inferred one variable at a time.

To move forward, we note that microscopic and macroscopic complexity are not neces-

sarily related [15, 16]. Thus complex living systems may realize rather simple dynamics, at least in typical experimental setups. For example, activation of a combinatorially complex receptor can be specified with only a handful of effective parameters, including the dynamic range, cooperativity, and time delay [17–19], and the purpose of microscopic structural complexity can be in making the simple macroscopic functional output robust in the face of perturbations [18, 20]. Similarly, in engineering [21], effective models are often sufficient for forward (but not reverse) engineering of complex systems, as illustrated by the ubiquity of the purely phenomenological Kalman filter. These considerations suggest that macroscopic prediction does not necessarily require microscopic accuracy even in systems biology [22], and that a complementary approach is needed, one in which we seek phenomenological, coarse-grained models of cellular processes that are simple and inferable, and nonetheless predictive and useful in limited domains [23].

Here we propose an adaptive approach for inference of dynamics from time series data that does not attempt to find the single best microscopically "correct" model, but rather a phenomenological, effective model that is "as simple as possible, but not simpler" than needed to account for the experimental data. Deemphasizing microscopic accuracy means that we do not have to search through all possible microscopic dynamics, and we can focus on a much smaller hierarchy of models. By choosing a hierarchy that is nested and complete, we gain theoretical guarantees of statistical consistency, meaning the approach is able to adaptively fit any smooth dynamics with enough data, yet is able to avoid problems with overfitting that can happen without restrictions on the search space [24]. While similar complexity control methods are well established in statistical inference [25] and in choosing a systems biology model for data from a finite set of models [26–28], we believe that they have not been used yet in the context of inferring complex, nonlinear dynamics from an infinite, complete set of all possible dynamics. Importantly, this adaptive approach requires testing a number of models that scales only polynomially with the number of dynamical variables. Further, it uses computational resources that asymptotically scale linearly with the number of observations. This allows us to construct models with much smaller computational effort and fewer experimental measurements, even when many dynamical variables are unobserved. While our main goal is effective dynamical modeling in systems biology, our approach works for general physical dynamical systems. In fact, we call it *Sir Isaac* due to its success in discovering the law of universal gravity from simulated data.

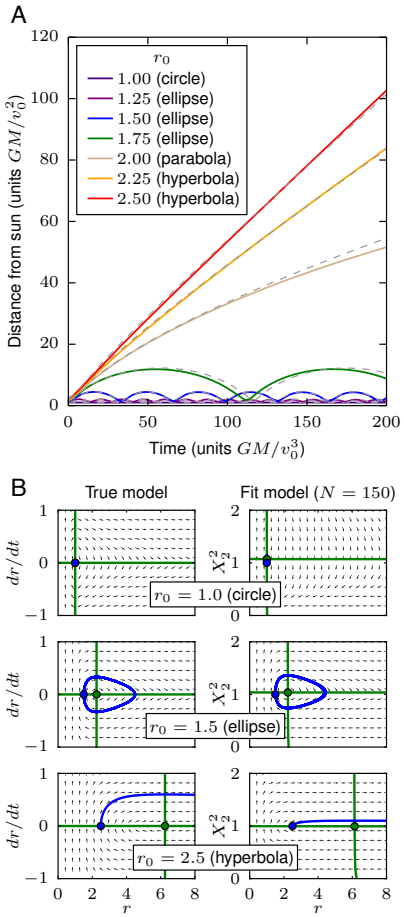


FIG. 1: The law of gravity: an example of dynamical inference. A particle is released with velocity v_0 perpendicular to the line connecting it to the sun, with varying initial distance r_0 from the sun. (a) With only $N = 150$ examples (each consisting of just a single noisy observation of r at a random time t after the release; see *SI*), we infer a single dynamical model in the S-systems class that reproduces the data. With no supervision, adaptive dynamical inference produces bifurcations that lead to qualitatively different behavior: in this case, a single model produces both oscillations (elliptical orbits) and monotonic growth (hyperbolic trajectories). Inferred trajectories are shown with solid colored lines, and the corresponding true trajectories are shown with dashed lines. (b) Like the true model (left), the inferred model (right) contains a single hidden variable X_2 and works using a similar phase space structure. Specifically, the location of nullclines (green lines) and a single fixed point (green circle) as a function of r_0 are recovered well by the fit. Note that the hidden variable is defined up to a power (see *SI*), and we choose to plot X_2^2 here.

I. RESULTS

We seek a phenomenological model of dynamics in the form

$$\frac{d\vec{x}}{dt} = \vec{F}_x(\vec{x}, \vec{y}, \vec{I}), \quad \frac{d\vec{y}}{dt} = \vec{F}_y(\vec{x}, \vec{y}, \vec{I}), \tag{1}$$

where $\vec{x}$ are observed variables, $\vec{y}$ are hidden variables, and $\vec{I}$ are inputs or other parameters to the dynamics. We neglect intrinsic stochasticity in the dynamics (either deterministic chaotic, or random thermal), and focus on systems for which repeated observations with nearly the same initial conditions produce nearly the same time series, save for measurement noise. The goal is then to find a phenomenological model of the force fields $\vec{F}_x, \vec{F}_y$ [8]. The same dynamics may produce different classes of trajectories $\vec{x}(t)$ dependent on initial conditions (e. g., ellipses and hyperbolas in gravitational motion). *Dynamical* inference rather than more familiar statistical modeling of trajectories is needed to represent these multiple functional forms within a single dynamical system.

Since our primary focus is on complex cellular processes, we construct two classes of nested and complete model hierarchies, both well matched to properties of biochemistry that underlies cellular network dynamics. We build the first with S-systems [29] and the second with continuous time sigmoidal networks [30]. The S-systems use production and degradation terms for each dynamical variable formed by products of powers of all involved variables (chemical species concentrations); this is a natural generalization of biochemical mass-action laws. Specifically, an S-system consists of J dynamical variables x_i and K inputs $I_k = x_{J+k}$, with each dynamical variable governed by an ordinary differential equation [29]

$$\frac{dx_i}{dt} = G(\mathbf{x})_i - H(\mathbf{x})_i, \tag{2}$$

where production G and degradation H terms have the form

$$G(\mathbf{x})_i = \alpha_i \prod_{j=1}^{J+K} x_j^{g_{ij}}, \; H(\mathbf{x})_i = \beta_i \prod_{j=1}^{J+K} x_j^{h_{ij}}. \tag{3}$$

Secondly, the sigmoidal class represents interactions using linear combinations of saturating functions of species concentrations, similar to saturation in biochemical reaction rates:

$$\frac{dx_i}{dt} = -x_i/\tau_i + \sum_{j=1}^{J} W_{ij} \; \xi(x_j + \theta_j) + \sum_{k=1}^{K} V_{ik} I_k, \tag{4}$$

where the sigmoidal function $\xi(y) = 1/(1 + e^y)$. Importantly, both classes are complete and are able to represent *any* smooth, nonlinear dynamics with a sufficient number of (hidden) dynamical variables [29, 31, 32]. They can also each efficiently represent the types of sharp nonlinearities typically found in biophysical systems (see *SI*).

To perform adaptive fitting within a model class, a specific ordered hierarchy of models is chosen *a priori* that simultaneously varies both the degree of nonlinearity (the number of factors in Eq. 3 or terms in Eq. 4) and the number of hidden variables (additional x_i; see FIG. S1 and *SI*). Within this restricted model space, Bayesian inference is then used to select a single best model (see *Methods*).

A. The law of gravity

Before applying the approach to complex dynamics where the true model may not be expressible simply within the chosen search hierarchy, we test it on a simpler system with a known exact solution. We choose the iconic law of gravity, inferred by Newton based on empirical observations of trajectories of planets, the Moon, and, apocryphally, a falling apple. Crucially, the inverse-squared-distance law of Newtonian gravity can be represented exactly within the S-systems power-law hierarchy for elliptical and hyperbolic trajectories, which do not go to zero radius in finite time. It requires a hidden variable, the velocity, to completely specify the dynamics of the distance of an object from the sun (see *SI*).

FIG. 1 displays the result of adaptive inference using the S-systems class. Given data about the distance of an object from the sun over time, we discover a model that reproduces the underlying dynamics, including the necessary hidden variable and the bifurcation points. Since the trajectories include hyperbolas and ellipses, this example displays the advantage of inferring a single set of dynamical equations of motion, rather than statistical fits to trajectories themselves, which would be different in the two cases. This adaptive dynamical inference is comparable to other recent methods [13], and it successfully treats a hidden dynamical variable. FIG. S4 additionally shows inference of the law of gravity using the sigmoidal model class. While accurate, the fits are worse than those using S-systems, illustrating the importance of understanding basic features of the studied system when conducting automated model inference.

Empowered by the success of the adaptive inference approach in this case, we chose to

name it *Sir Isaac*. The software implementation can be found under this name on GitHub.

B. Multi-site phosphorylation model

When inferring models for more general systems, we do not expect the true dynamics to be perfectly representable by any specific model class: even the simplest biological phenomena may involve combinatorially many interacting components. Yet for simple macroscopic behavior, we expect to be able to use a simple approximate model that can produce useful predictions. To demonstrate this, we envision a single immune receptor with n modification sites, which can exist in 2^n microscopic states [33], yet has simple macroscopic behavior for many underlying parameter combinations. Here, we test a model receptor that can be phosphorylated at each of $n = 5$ sites arranged in a linear chain. The rates of phosphorylation and dephosphorylation at each site are affected by the phosphorylation states of its nearest neighboring sites. With Michaelis-Menten kinetics and independence of kinetic rates for different states, this produces a complicated model with 32 coupled ODEs specified by 52 parameters, which we assume are unknown to the experimenter.

We imagine an experimental setup in which we can control one of these parameters, e.g., by changing concentrations of various kinases. We are interested in effects of such changes on the time evolution of the total phosphorylation of all 5 sites. Here we arbitrarily treat as input I the maximum rate of cooperative phosphorylation of site 2 due to site 3 being occupied, V. This is inspired, for example, by being able to measure or control concentrations of the SRC-family kinases (input), which mediate immune signaling conditional on the previous steps in the receptor activation sequence being completed [17]. We then "measure" the resulting time course of total phosphorylation starting from the unphosphorylated state. Experimental measurements are corrupted with noise at the scale of 10% of their values (see *SI* for details).

A straightforward approach to modeling this system is to fit the 52 parameters of the known model to the data. A second approach is to rely on intuition to manually develop a functional parameterization that captures the most salient features of the timecourse data. In this case, we can write a simple 5 parameter model (see *SI*) that captures exponential saturation in time with an asymptotic value that depends sigmoidally on the input V. A third approach, advocated here, is to use automated model selection to create a model with

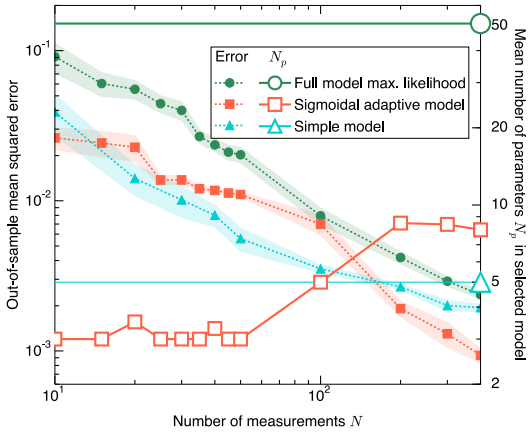


FIG. 2: Multi-site phosphorylation model selection as a function of the number of measurements N. The sizes of errors made by three models (filled symbols; left axis) decrease as the amount of data increases. Adaptive sigmoidal models (orange squares) outperform a maximum likelihood fit to the full 52-parameter model (green circles) in this range of N (although we expect that it will eventually outperform all other models as $N \to \infty$). A simple 5-parameter model (blue triangles) that is custom-made to match salient features of the true behavior is the best performer for a moderate amount of data, but is outperformed by adaptive models when given more data. The mean over 10 sets of input data are shown, with shaded regions indicating the standard deviation of the mean. The full and simple models each use a fixed number of parameters (open symbols; right axis), while the sigmoidal model adapts to use more parameters when given more data.

complexity that matches the amount and precision of the available data.

In FIG. 2, we compare these three approaches as the amount of available data is varied, and FIG. 3(a) shows samples of fits done by different procedures. With limited and noisy data, fitting the parameters of the full known model risks overfitting, and in the regime we test, it is the worst performer on out-of-sample predictions. The simple model performs best when fitting to less than 100 data points, but for larger amounts of data it saturates in performance, as it cannot fit more subtle effects in the data. In contrast, an adaptive model remains simple with limited data and then grows to accommodate more subtle behaviors once enough data is available, eventually outperforming the simple model. Even when given up to 400 data points, the adaptive model remains relatively simple, avoiding using as many degrees of freedom as the full model (see also FIG. S5). Crucially, this performance stays robust when various assumptions of the adaptive inference approach are violated (such as

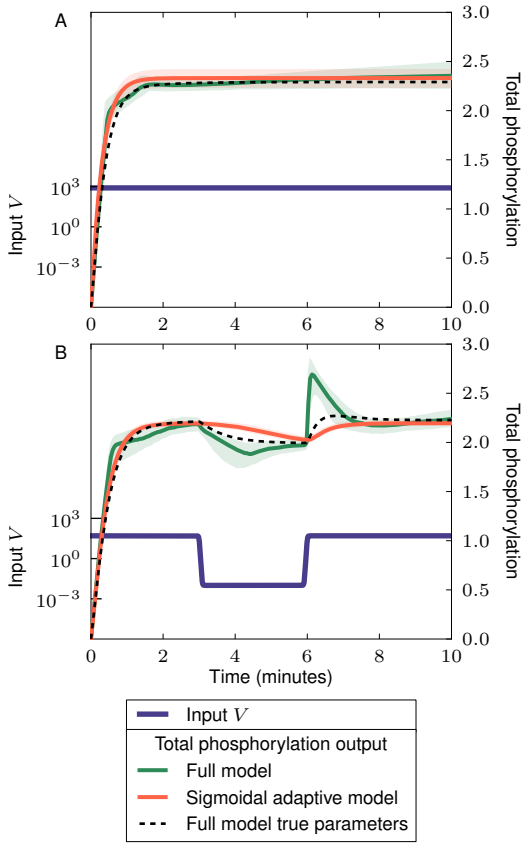


FIG. 3: Response (right axis) to (a) out-of-sample constant and (b) time-varying input (left axis, blue lines) in the models of multi-site phosphorylation. Fit to $N = 300$ constant input data points, the full known model (green) produces erratic behavior typical of overfitting (especially evident in (b)), while the adaptive sigmoidal model (orange) produces more stable out-of-sample predictions with median behavior that is closer to the true dynamics. Plotted is the median behavior over 100 samples from each model's parameter posterior (see *SI*), with shaded regions indicating 90% confidence intervals, which are in some cases smaller than the width of the line.

the model of the measurement noise, cf. FIGS. S2A, S2B). And it barely depends on details of the approach such as the ordering with which parameters are added into the model (cf. FIG. S2C).

The multi-site phosphorylation example also demonstrates that dynamical phenomenological models found by *Sir Isaac* are more than fits to the existing data, but rather they uncover the true nature of the system in a precise sense: they can be used to predict responses to some classes of inputs that are qualitatively different from those used in the inference. For example, as seen in FIG. 3(b), an adaptive sigmoidal model inferred using

temporally constant signals produces a reasonable extrapolated prediction for response to a *time-varying* signal. At the same time, overfitting is evident when using the full, detailed model, even when one averages responses over the posterior distribution of the inferred model parameters.

C. Yeast glycolysis model

A more complicated test of the method is to reproduce nonlinear oscillatory dynamics, such as that describing yeast glycolysis, for which there has been recent interest in automated inference [14]. A recent model for the system [34, 35], informed by detailed knowledge of metabolic pathways, consists of coupled ODEs for 7 species whose concentrations oscillate with a period near 1 minute. The system dynamic is simpler than its structure in the sense that some complexity is used to stabilize oscillations to perturbations. On the other hand, the oscillations are not smooth (see FIG. 4) and hence are hard to fit with simple methods. These aspects make this model an ideal test case for *Sir Isaac*.

If we were given abundant time series data from all 7 species and were confident that there were no other important hidden species, we may be in a position to infer a "true" model detailing interactions among them. If we are instead in the common situation of having limited data on a limited number of species, we may more modestly attempt to make predictions about the inputs and outputs that we have measured. This is conceptually harder since an unknown number of hidden variables may need to be introduced to account for the dynamics of the observed species. We demonstrate our approach by constructing adaptive models using data for only 3 of the 7 coupled chemical species, as their initial conditions are varied.

Depicted in FIG. 4 is the model selection procedure for this case. After selecting an adaptive model fit to noisy data from N single timepoints, each starting from initial conditions sampled from specified ranges, we test the inferred model's ability to predict the time course resulting from out-of-sample initial conditions, including those lying far away from the limit cycle. With data from only $N = 40$ measurements, the selected model is able to predict behavior with mean correlation of over 0.6 for initial conditions chosen from ranges *twice as large* as those used as training data (shown in FIG. 4) and 0.9 for out-of-sample ranges equal to in-sample ranges (shown in FIG. S9). At this point, the model saturates at about

65 nominal and 35 effective parameters (FIG. S11). This is larger than in the true model and does not necessarily reflect its topology. However, since discovering the functional form of the true model (including hidden nodes) would require a search through a much larger space of models, complexity here should not be measured just by the number of parameters. This is illustrated, in part, by the admirable predictive performance of the phenomenological model for a relatively small N.

We can compare this to the performance of a hand-constructed "simple" 9 parameter harmonic oscillator model (an analog of the simple model in the multi-site phosphorylation case). The simple model, for which the numbers of nominal and effective parameters are equal (FIG. S11), does not have the exploratory power to resolve the sharp peaks and obtain good predictions (see SI and FIG. S9). In another comparison, the true model that generated the data has 16 parameters, which is more than the result of Sir Isaac. However, the functional form of the dynamics for this exact model should also be counted as inferred parameters, making such comparisons harder. In fact, because of this, previous work that inferred the exact equations of the original 7-dimensional model (including also an unexpected conservation law) [14] had to use roughly 500 times as many measurements of all 7 variables and 200 times as many model evaluations. While *Sir Isaac* is somewhat aided by an appropriate choice of sigmoidal basis functions, and has not been designed to look for conservation laws, this example illustrates how focusing on a simpler problem, namely finding an approximate, phenomenological model of the process, can decrease data requirements by orders of magnitude. This example also demonstrates that adaptive modeling can hint at the complexity of the hidden dynamics beyond those measured: the best performing sigmoidal model requires three hidden variables, for a total of six chemical species, which is exactly what one would expect for a seven-dimensional system with a (hidden) conservation law [14]. Crucially, the computational complexity of *Sir Isaac* still scales linearly with the number of observations, even when a large fraction of variables remains hidden (see *SI* and FIG. S10). We anticipate that using advanced approaches to identify and conduct the most informative experiments and efficiently search the model hierarchy using genetic algorithms, as in [14], may improve performance further.

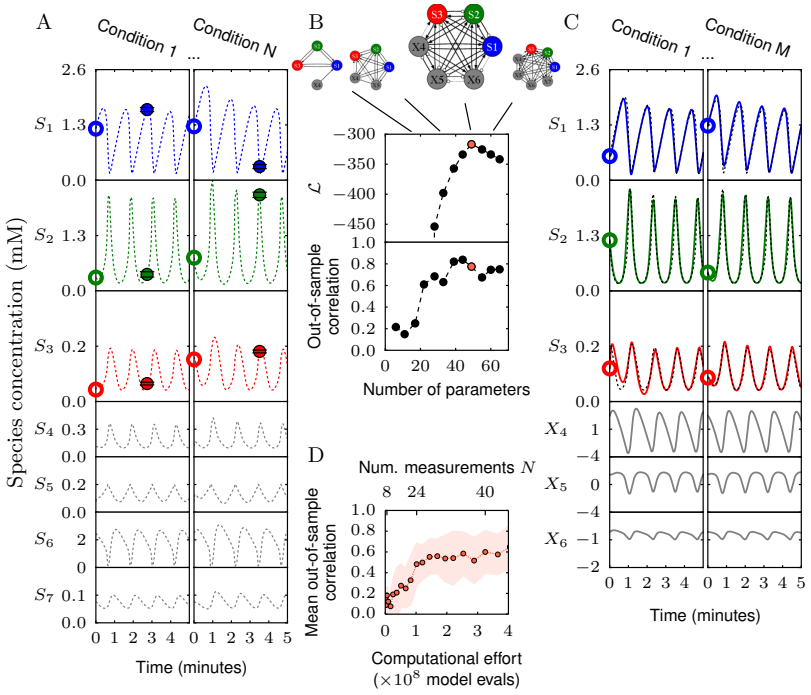


FIG. 4: An example of the model selection process using measurements of timecourses of three metabolites in yeast glycolysis as their initial concentrations are varied. (a) For each set of initial conditions (open circles), a noisy measurement of the three observable concentrations (filled circles) is made at a single random time. Hidden variables (in gray) are not measured. In this example, we fit to $N = 40$ in-sample conditions. (b) Models from an ordered class, with the illustrated connectivity, are fit and tested sequentially until $\mathcal{L}$, an approximation of the relative log-likelihood, decreases sufficiently from a maximum. (c) The selected model (large connectivity diagram) is used to make predictions about out-of-sample conditions. Here, we compare the output of the selected model (solid lines) to that of the model that created the synthetic data (dashed lines). (d) Performance versus computational and experimental effort. The mean out-of-sample correlation for 3 measured biochemical species from the range of initial conditions twice that used in training rises to over 0.6 using less than 5×10^8 model evaluations and 40 in-sample measurements. In Ref. [14], inferring an exact match to the original 7-dimensional model used roughly 500 times as many measurements of all 7 species (with none hidden). The approach also uses 200 times as many model evaluations (see *SI*). Nonetheless, the accuracy of both approaches is comparable, and *Sir Isaac* additionally retains information about the phase of the oscillations. This illustrates that the problem of adaptively finding an approximation to the dynamics is, in fact, much simpler than the problem of inferring the detailed equations describing the dynamics.

II. DISCUSSION

The three examples demonstrate the power of the adaptive, phenomenological dynamical modeling approach. *Sir Isaac* models are inferred without an exponentially complex search over model space, which would be impossible for systems with many variables. These models are as simple or complex as warranted by data and are guaranteed not to overfit even for small data sets. Thus they require orders of magnitude less data and computational resources to achieve the same predictive accuracy as methods that infer a pre-defined, large number of mechanistic parameters in the true model description.

These advantages require that the inferred models are phenomenological, and are designed for efficiently predicting system dynamics at a given scale, determined by the available data. While FIG. 1 shows that *Sir Isaac* will infer the true model if it is within the searched model hierarchy and enough data is available, more generally the inferred dynamics may be quite distinct from the true microscopic mechanisms, as shown by a different number of chemical species in the true and the inferred dynamics in FIG. 4. What is then the utility of the approach if it says little about underlying mechanisms?

First, there is the obvious advantage of being able to predict responses of systems to yet-unseen experimental conditions, including those qualitatively different from the ones used for inference. This is trivially useful in the context of engineering and control, where predictive, usable models are often necessarily far removed from microscopic precision [21]. Second, some general mechanisms, such as the necessity of feedback loops or hidden variables, are easily uncovered even in phenomenological models. However, more importantly, we draw the following analogy. When in the 17th century Robert Hooke studied the force-extension relations for springs, a linear model for a specific spring did not tell much about the force generation. However, the observation that *all* springs exhibit such linear relations for small extensions allowed him to combine the models into a law — Hooke's law, the first of many phenomenological physical laws that followed. It instantly became clear that experimentally measuring just one parameter, the Hookean stiffness, provided an exceptionally precise description of the spring's behavior. And yet the mechanistic understanding of how this Hooke's constant is related to atomic interactions within materials is only now starting to emerge. Similarly, by studying related phenomena across complex living systems (e.g., chemotactic behavior in *E. coli* [36] and *C. elegans* [37], or behavioral bet hedging, which

can be done by a single cell [38] or a behaving rodent [39]), we hope to build enough *models* of specific systems, so that general physical *laws* describing how nature implements them become apparent.

If successful, our search for phenomenological, emergent dynamics should allay some of the most important skepticism regarding the utility of automated dynamical systems inference in science [40], namely that such methods typically start with known variables of interest and known underlying physical laws, and hence cannot do transformative science and find new laws of nature. Indeed, we demonstrated that, for truly successful predictions, the model class used for automated phenomenological inference must match basic properties of the studied dynamics (contrast, for example, FIG. 1 to FIG. S4, and see FIG. S6). Thus fundamental properties of the underlying mechanisms, such as the power-law structure of the law of gravity, or the saturation of biochemical kinetic rates, can be inferred from data even if unknown *a priori*. Finally, we can contrast our approach with a standard procedure for producing coarse-grained descriptions of physical systems: starting from mechanistically accurate dynamics, and then mapping them onto one of a small set of universality classes [22, 41]. This procedure is possible due to symmetries of physical interactions that are not typically present in living systems. Without such symmetries, the power of universality is diminished, and microscopic models may result in similarly different macroscopic ones. Then specifying the microscopic model in order to coarse-grain it later becomes an example of solving a harder problem to solve a simpler one [42]. Thus for living systems, the direct inference of phenomenological dynamics, such as done by *Sir Isaac*, may be the optimal way to proceed.

III. MATERIALS AND METHODS

A. Classes of phenomenological models used by Sir Isaac

To create a model in the form of (1), we would like to gradually increase the complexity of F until we find the best tradeoff between good fit and sufficient robustness, essentially extending traditional Bayesian model selection techniques to the realm of an infinite set of possible dynamical models. Ideally, this process should progress much like a Taylor series approximation to a function, adding terms one at a time in a hierarchy from simple to more

complex, until a desired performance is obtained. To guarantee that this is possible, the hierarchy of models must be nested (or ordered) and complete in the sense that any possible dynamics can be represented within the hierarchy [24] (see *SI*). Any model hierarchy that fits these criteria may be used, yet specification of the hierarchy is nontrivial in that it requires choosing an ordering of models that gradually adds both nonlinearities and unobserved variables. Further, different model hierarchies may naturally perform differently on the same data, depending on whether the studied dynamics can be represented succinctly within a hierarchy. Our results suggest that the choice of model class, specifying the functional forms used to model the dynamics, is more important to performance than the subsequent choice of the order of adding parameters within that class (see FIG. S2C).

Our first model class is the S-system power-law class, defined in Eqs. 2 and 3. In a process called "recasting," any set of differential equations written in terms of elementary functions can be rewritten in the power-law form by defining new dynamical variables in the correct way [29]. Since any sufficiently smooth function can be represented in terms of a series of elementary functions (e. g., Taylor series), a power-law network of sufficient size can describe any such deterministic dynamical system. Note that, since exponents are not constrained to be positive or integer-valued, dynamics in this class are generally ill-defined when variables are not positive. We find that the S-systems model class works well for planetary motion, which has an exact representation in the class (see *SI*). For our biological test examples, the S-systems class is outperformed by the sigmoidal class (see below). This may be indicating that behavior common in the S-systems class is not common in typical biological systems (e. g., real production and degradation terms cannot grow without bounds). It may also stem from the positivity constraint: since the condition that variables remain positive is not easily determined from parameter values, we are forced in our model selection process to simply discard any tested parameters that lead to zero or negative values.

The second model hierarchy that we construct is the sigmoidal network class. In this class, we use the fact that the interactions among biological components often take the form of a sigmoidal function to define the system of ODEs in Eq. 4. This class of models has also been shown to approximate any smooth dynamics arbitrarily well with a sufficient number of dynamical variables [30–32, 43]. Note that natural variations of this class to be explored in future work include rescaling of the arguments of the sigmoids ξ or switching the order of operations to apply the sigmoidal function to a linear combination of state variables in

order to more closely match traditional neural network models [44].

It is possible that both the S-system and sigmoidal classes can be unified into power-law dynamical systems with algebraic power-law constraints among the dynamical variables [29], but this will not be explored in this report. Other than these two model classes and their modifications described above, the authors are not aware of other biologically relevant dynamical representations that are currently known to be complete. Yet others certainly exist and could be developed into alternate model hierarchies in future work.

B. Description of model selection procedure

For each model in the hierarchy, its parameters are fit to the data using a two step process akin to simulated annealing (see *SI*), with best-fit parameters from the next simplest model in the hierarchy used as a starting point to speed convergence. The resulting fit model is evaluated by calculating an estimate of the Bayesian log-likelihood $\mathcal{L}$. This estimate makes use of a generalized version of the Bayesian Information Criterion [45] (BIC), which is described in detail in *SI*. We believe that this is the first time BIC has been adopted for use with automated nonlinear dynamical systems inference over an infinite set of models. As models increase in complexity, $\mathcal{L}$ first grows as the quality of fit increases, but eventually begins to decrease, signifying overfitting. Since, statistical fluctuations aside, there is just one peak in $\mathcal{L}$ [24], one can be certain that the global maximum has been observed once it has decreased sufficiently. The search through the hierarchy is then stopped, and the model with maximum $\mathcal{L}$ is "selected" (see FIG. 4(b)).

Acknowledgments

We thank William Bialek and Michael Savageau for discussions, Andrew Mugler and Fereydoon Family for critical comments, and the hospitality of the Center for Nonlinear Studies at Los Alamos National Laboratory. This research was supported in part by the James S. McDonnell foundation Grant No. 220020321 (I. N.), a grant from the John Templeton Foundation for the study of complexity (B. D.), the Los Alamos National Laboratory Directed Research and Development Program (I. N. and B. D.), and NSF Grant No. 0904863 (B.

D.).

[1] Rosenblueth, A. & Wiener, N. The role of models in science. *Phil Science* **12**, 316–321 (1945).

[2] Hlavacek, W. How to deal with large models? *Mol Syst Biol* **5**, 240 (2009).

[3] Gutenkunst, R. *et al.* Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol* **3**, 1871–1878 (2007).

[4] Feret, J., Danos, V., Krivine, J., Harmer, R. & Fontana, W. Internal coarse-graining of molecular systems. *Proc Natl Acad Sci (USA)* **106**, 6453–6458 (2009).

[5] Borisov, N., Chistopolsky, A., Faeder, J. & Kholodenko, B. Domain-oriented reduction of rule-based network models. *IET Syst Biol* **2**, 342–351 (2008).

[6] Danos, V., Feret, J., Fontana, W., Harmer, R. & Krivine, J. Abstracting the differential semantics of rule-based models: exact and automated model reduction. In *Logic in Computer Science*, 362–381 (IEEE Computer Society, Edinburgh, United Kingdom, 2010).

[7] Dokoumetzidis, A. & Aarons, L. Proper lumping in systems biology models. *IET Syst Biol* **3**, 40–51 (2009).

[8] Crutchfield, J. & McNamara, B. Equations of motion from a data series. *Complex Systems* **1**, 417 (1987).

[9] Packard, N., Crutchfield, J., Farmer, J. & Shaw, R. Geometry from a Time Series. *Physical Review Letters* **45** (1980).

[10] Friston, K., Harrison, L. & Penny, W. Dynamic causal modelling. *NeuroImage* **19**, 1273–1302 (2003).

[11] Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–57 (2009).

[12] François, P., Hakim, V. & Siggia, E. D. Deriving structure from evolution: metazoan segmentation. *Molecular systems biology* **3**, 154 (2007).

[13] Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81 (2009).

[14] Schmidt, M. *et al.* Automated refinement and inference of analytical models for metabolic networks. *Phys Biol* **8**, 055011 (2011).

[15] Anderson, P. W. More is different. *Science (New York, NY)* **177**, 393–396 (1972).

[16] May, R. Simple mathematical models with very complicated dynamics. *Nature* **261**, 459 (1976).

[17] Goldstein, B., Faeder, J. & Hlavacek, W. Mathematical and computational models of immune-receptor signalling. *Nat Rev Immunol* **4**, 445–456 (2004).

[18] Bel, G., Munsky, B. & Nemenman, I. The simplicity of completion time distributions for common complex biochemical processes. *Phys Biol* **7**, 016003 (2010).

[19] Cheong, R., Rhee, A., Wang, Nemenman, I. & Levchenko, A. Information transduction capacity of noisy biochemical signaling networks. *Science* **334**, 354–358 (2011).

[20] Lander, A. Pattern, growth, and control. *Cell* **144**, 955–969 (2011).

[21] LeDuc, P. R., Messner, W. C. & Wikswo, J. P. How do control-based approaches enter into biology? *Ann Rev Biomed Eng* **13**, 369 (2011).

[22] Machta, B. B., Chachra, R., Transtrum, M. K. & Sethna, J. P. Parameter space compression underlies emergent theories and predictive models. *Science* **342**, 604–7 (2013).

[23] Wiggins, C. & Nemenman, I. Process pathway inference via time series analysis. *Experim Mech* **43**, 361 (2003).

[24] Nemenman, I. Fluctuation-dissipation theorem and models of learning. *Neural Comput* **17**, 2006 (2005).

[25] MacKay, D. *Information theory, inference, and learning algorithms* (Cambridge UP, 2003).

[26] Vyshemirsky, V. & Girolami, M. Bayesian ranking of biochemical system models. *Bioinform* **24**, 833–839 (2008).

[27] Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interf* **6**, 187–202 (2009). 0901.1925.

[28] Lillacci, G. & Khammash, M. Parameter estimation and model selection in computational biology. *PLoS Comput Biol* **6** (2010).

[29] Savageau, M. A. & Voit, E. O. Recasting Nonlinear Differential Equations as S-Systems: A Canonical Nonlinear Form. *Math Biosci* **87**, 83–115 (1987).

[30] Beer, R. D. Parameter space structure of continuous-time recurrent neural networks. *Neural Comput* **18**, 3009–51 (2006).

[31] Funahashi, K.-I. & Nakamura, Y. Approximation of Dynamical Systems by Continuous Time Recurrent Neural Networks. *Neural networks* **6**, 801–806 (1993).

[32] Chow, T. W. & Li, X.-D. Modeling of continuous time dynamical systems with input by recurrent neural networks. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* **47**, 575–578 (2000).

[33] Hlavacek, W. S. *et al.* Rules for modeling signal-transduction systems. *Sci. STKE* **2006**, re6 (2006).

[34] Wolf, J. & Heinrich, R. Effect of cellular interaction on glycolytic oscillations in yeast: a theoretical investigation. *Biochem J* **334**, 321–334 (2000).

[35] Ruoff, P., Christensen, M., Wolf, J. & Heinrich, R. Temperature dependency and temperature compensation in a model of yeast glycolytic oscillations. *Biophys Chem* **106**, 179 (2003).

[36] Berg, H. E. coli *in Motion* (Springer, 2004).

[37] Ryu, W. & Samuel, A. Thermotaxis in Caenorhabditis elegans analyzed by measuring responses to defined thermal stimuli. *J Neurosci* **22**, 5727–5733 (2002).

[38] Kussell, E. & Leibler, S. Phenotypic diversity, population growth, and information in fluctuating environments. *Science* **309**, 2075–2078 (2005).

[39] Gallistel, C., Mark, T., King, A. & Latham, P. The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *J Exp Psychol: Anim Behav Process* **27**, 354–372 (2001).

[40] Anderson, P. W. & Abrahams, E. Machines fall short of revolutionary science. *Science* **324**, 1515–1516 (2009).

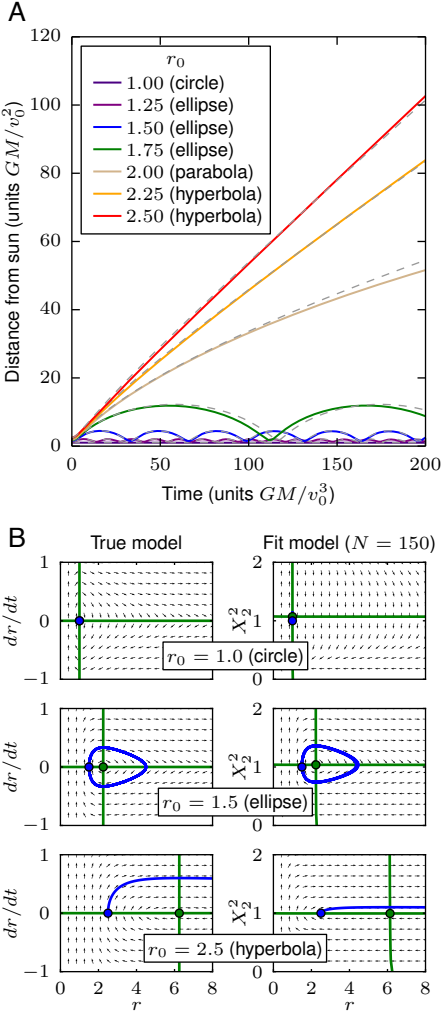
[41] Wilson, K. Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture. *Phys Rev B* **4**, 3174 (1971).

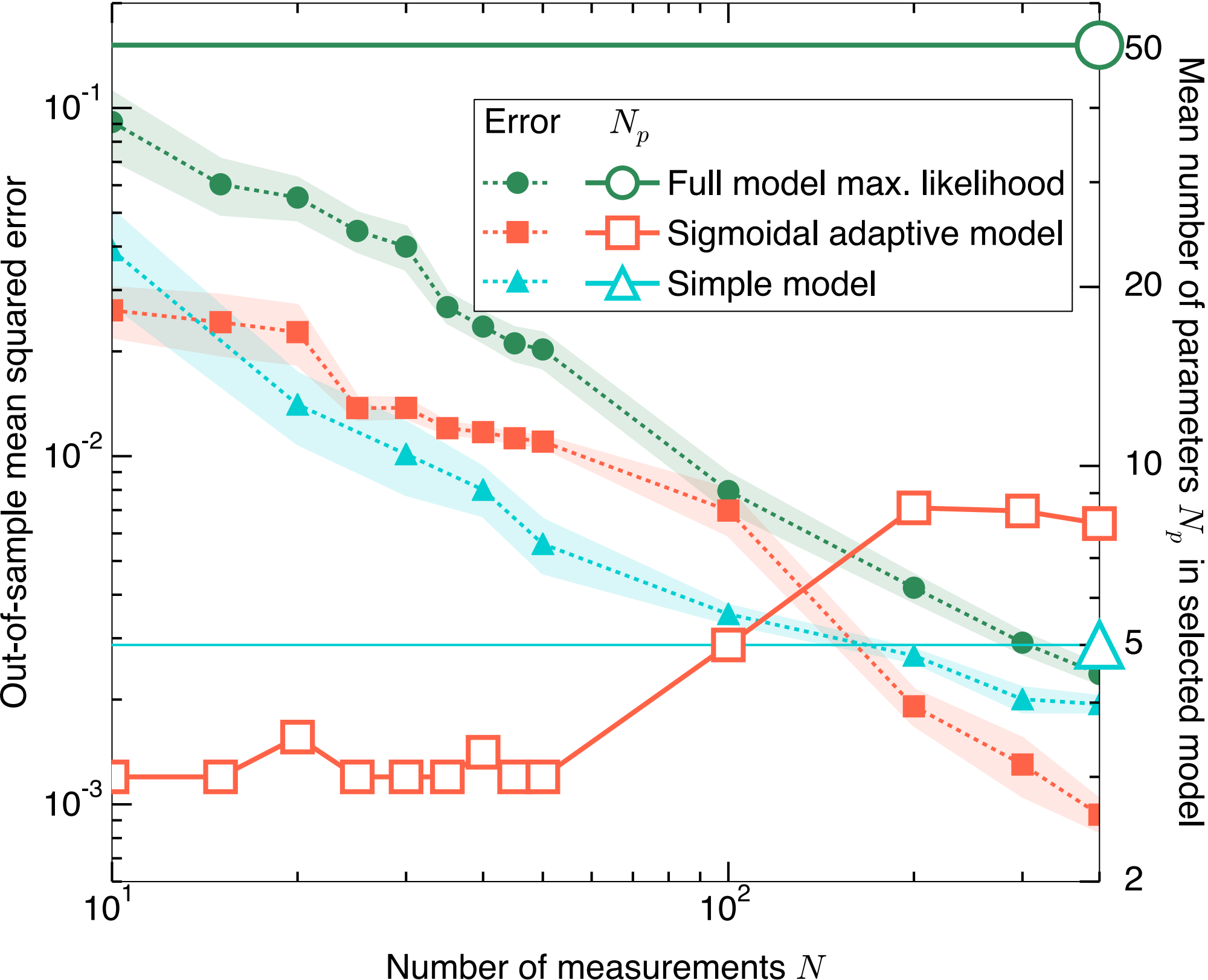
[42] Vapnik, V. *The nature of statistical learning theory* (Springer, New York, NY, 2000), 2nd edn.

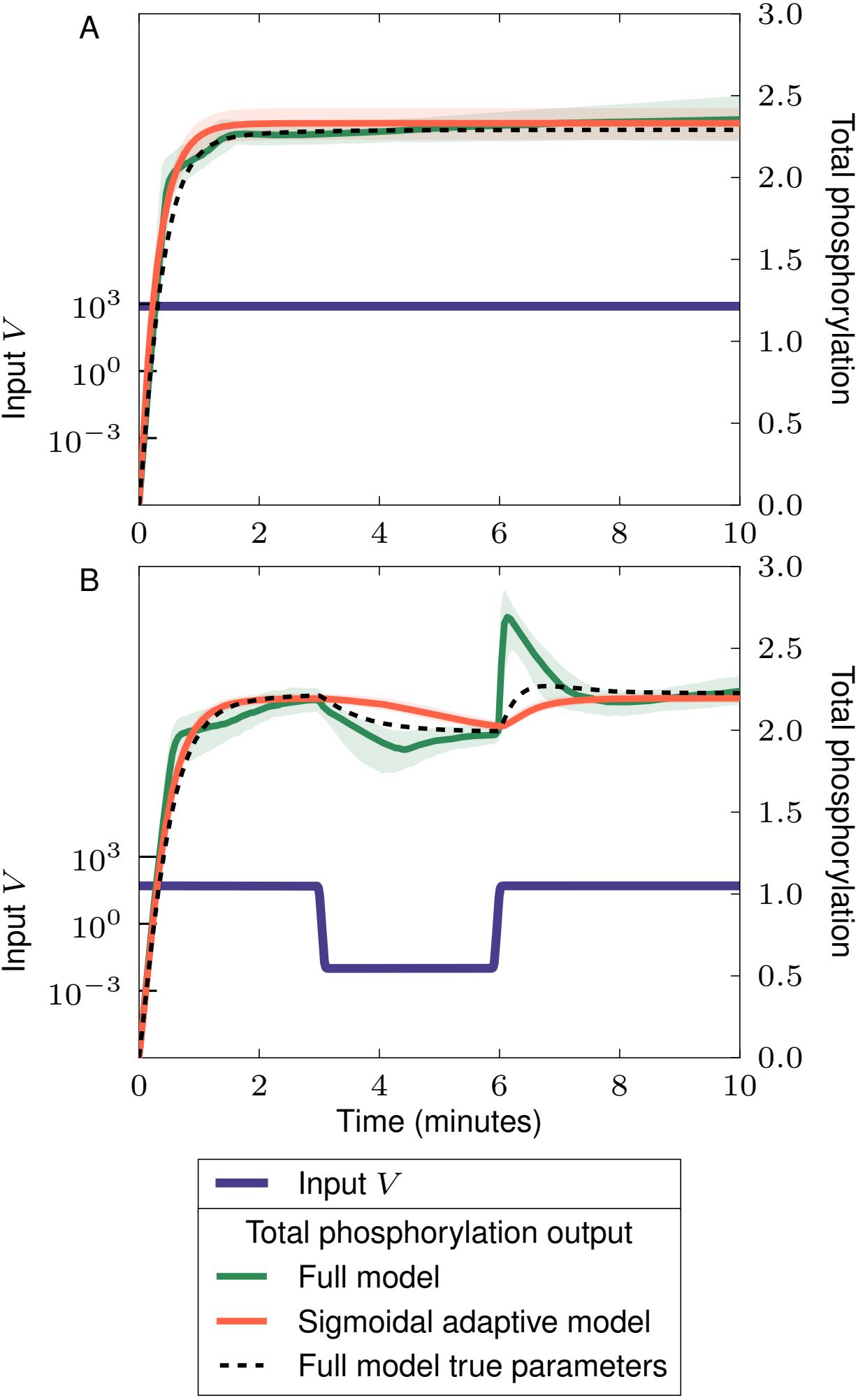
[43] Beer, R. D. & Daniels, B. Saturation Probabilities of Continuous-Time Sigmoidal Networks. *arXiv preprint arXiv:1010.1714* 856–873 (2010).

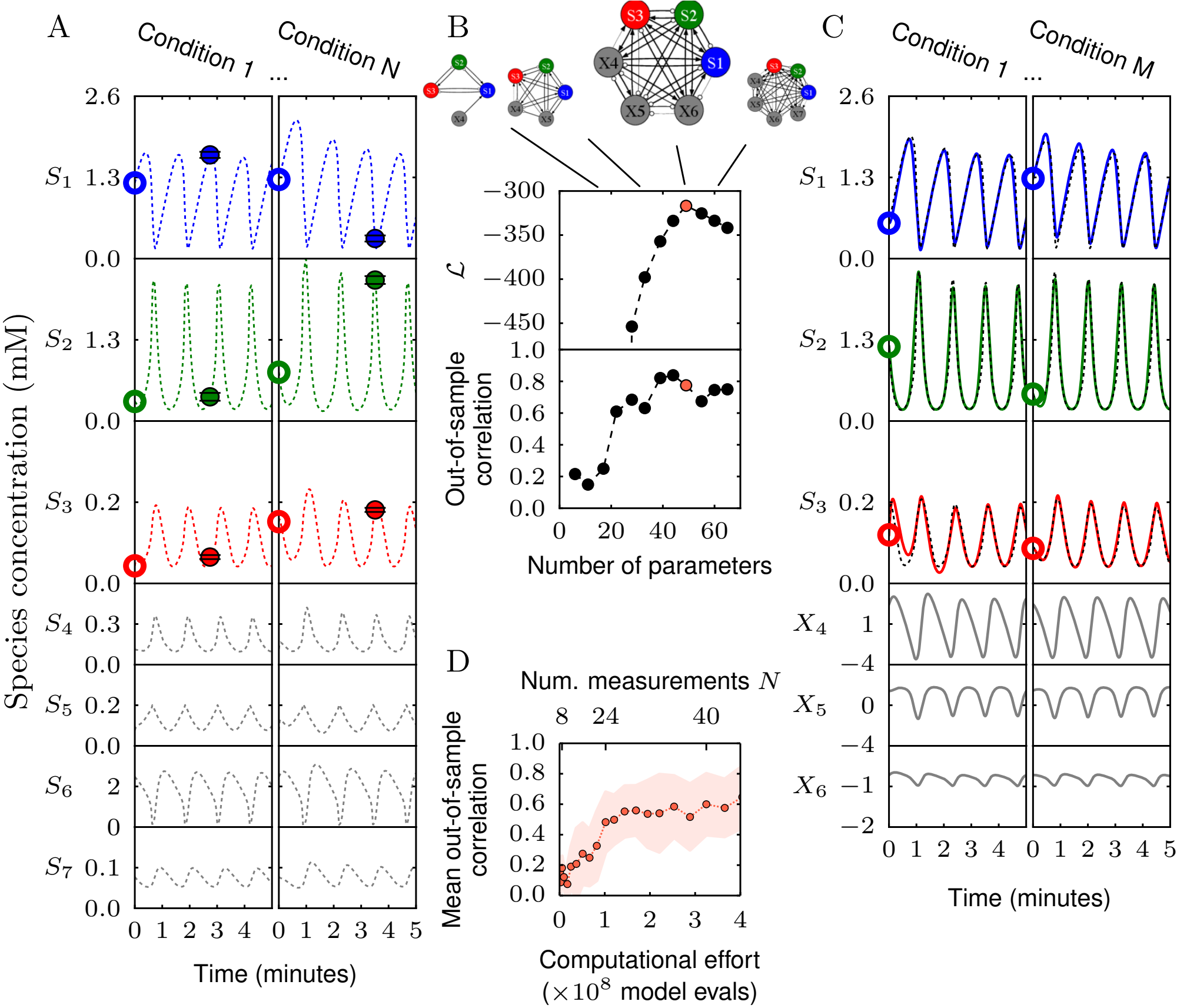
[44] Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. *Nature* **323**, 533 (1986).

[45] Schwarz, G. Estimating the dimension of a model. *Annals Stat* **6**, 461 (1978).


A
r_0
1.00 (circle)
1.25 (ellipse)
1.50 (ellipse)
1.75 (ellipse)
2.00 (parabola)
2.25 (hyperbola)
2.50 (hyperbola)
Distance from sun (units GM/v_0^2)
Time (units GM/v_0^3)
B
True model
Fit model ($N = 150$)
dr/dt
X_2^2
$r_0 = 1.0$ (circle)
$r_0 = 1.5$ (ellipse)
$r_0 = 2.5$ (hyperbola)
r


Error
N_p
Full model max. likelihood
Sigmoidal adaptive model
Simple model
Out-of-sample mean squared error
Mean number of parameters N_p in selected model
Number of measurements N
10^{-1}
10^{-2}
10^{-3}
10^{1}
10^{2}
50
20
10
5
2


A
B
Input V
10^3
10^0
10^{-3}
Total phosphorylation
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
Time (minutes)
Input V
Total phosphorylation output
Full model
Sigmoidal adaptive model
Full model true parameters


A
Condition 1
...
Condition N
Species concentration (mM)
S_1
S_2
S_3
S_4
S_5
S_6
S_7
Time (minutes)
B
S1
S2
S3
X4
X5
X6
X7
$\mathcal{L}$
Out-of-sample correlation
Number of parameters
C
Condition 1
...
Condition M
X_4
X_5
X_6
Time (minutes)
D
Num. measurements N
Mean out-of-sample correlation
Computational effort
($\times 10^8$ model evals)

Automated adaptive inference of phenomenological dynamical models: Supporting Information

Bryan C. Daniels[1] & Ilya Nemenman[2]

[1]Center for Complexity and Collective Computation,

Wisconsin Institute for Discovery, University of Wisconsin, Madison, WI 53716, USA

[2]Departments of Physics and Biology, Emory University, Atlanta, GA 30322, USA

(Dated: June 4, 2015)

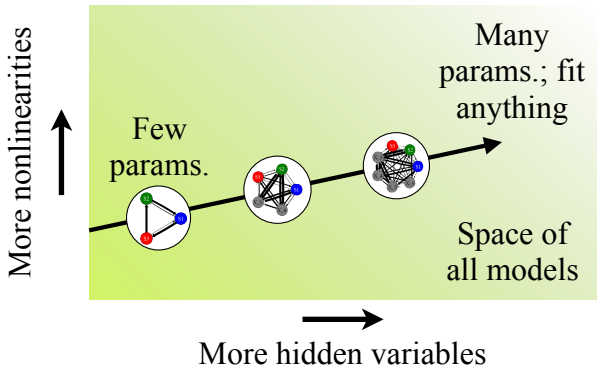


FIG. S1. Hierarchical model selection follows a single predefined path through model space.

I. HIERARCHICAL BAYESIAN MODEL SELECTION

For consistent inference, we need a hierarchy of models that satisfies criteria laid out in Ref. [1]. First, we desire a model hierarchy that will produce a single maximum in $\mathcal{L}$, up to statistical fluctuations, as we add complexity. For this, the hierarchy should be nested (but not necessarily regular or self-similar), meaning that once a part of the model is added, it is never taken away. Second, the hierarchy should be complete, meaning it is able to fit any data arbitrarily well with a sufficiently complex model. Intuitively, instead of searching a large multidimensional space of models, hierarchical model selection follows a single predefined path through model space (FIG. S1). While the predefined path may be suboptimal for a particular instance (that is, the true model may not fall on it), even then the completeness guarantees that we will still eventually learn any dynamical system F given enough data, and nestedness assures that this will be done without overfitting along the way [26].

A. Ordering of hierarchies

An advantage of the S-systems and sigmoidal representations is the existence of a natural scheme for creating a one-dimensional model hierarchy: simply adding dynamical variables x_i. The most general network is fully connected, such that every variable x_i has an interaction term in every other dx_j/dt. Our hierarchy starts with a fully-connected network consisting of the necessary number of input and output variables, and adds "hidden" dynamical variables to add complexity.

With each additional x_i, we add parameters in a predetermined order.

In the S-systems class, without connections, variable x_i's behavior is specified by 5 parameters: $x_i^{\text{init}}, \alpha_i, \beta_i, g_{ii}$, and h_{ii}. Each connection to and from x_j is specified by 4 parameters: g_{ij}, g_{ji}, h_{ij}, and h_{ji}. When adding a new dynamic variable, we first fix its parameters (to zero for the exponential parameters and one for the multiplicative parameters), and then allow them to vary one at a time in the following order: $g_{ii}, g_{ji}, h_{ji}, g_{ij}, h_{ij}, \beta_i, h_{ii}, \alpha_i$ (adding connections to every other x_j one at a time). An example is shown in Table I.

The sigmoidal class is similar: without connections, variable x_i's behavior is specified by 4 parameters: $x_i^{init}, W_{ii}, \tau_i$, and θ_i. Each connection to and from x_j is specified by 2 parameters: W_{ij} and W_{ji}. When adding a new dynamic variable, we first fix its parameters (to zero for W and θ and one for τ), and then allow them to vary one at a time in the following order: $W_{ij}, W_{ji}, W_{ii}, \tau_i, \theta_i$ (adding connections to every other x_j one at a time). An example is shown in Table II.

For every adaptive fit model and the full multi-site phosphorylation model,[27] we use the same prior for every parameter α_k, which we choose as a normal distribution $\mathcal{N}(0, 10^2)$ with mean 0 and standard deviation $\varsigma = 10$.[28]

B. Representation of sharp nonlinearities

Both the sigmoidal and S-systems classes can represent arbitrary dynamics. However, it is important that they can *efficiently* represent sharp nonlinearities that are often present in biological systems, such as those typically represented by large Hill coefficients. While this is straightforward for the S-systems class [2], it is less obvious for sigmoidal models.

The sigmoidal model class relies on $\xi(y)$, which has the largest derivative $\xi'(0) = -1$. Thus it may seem that sharp nonlinearities could be hard to produce. In fact, the introduction of hidden variables that perform multiple transformations can produce arbitrarily sharp production rate laws. As an example, we show here that the nonlinearity captured by the Hill equation,

$$f = S^n/(S^n + K), \tag{S1}$$

(where S is the substrate concentration, K is the dissociation constant, and f is the fraction of

Model No. i	Num. parameters N_p	Form of power-law ODEs
0	3	$x_1(0) = x_1^{\text{init}}$ $\frac{dx_1}{dt} = x_I^{g_{10}} x_1^{g_{11}} - \beta_1$
1	4	$x_1(0) = x_1^{\text{init}}$ $\frac{dx_1}{dt} = x_I^{g_{10}} x_1^{g_{11}} - \beta_1 x_I^{h_{10}}$
2	5	$x_1(0) = x_1^{\text{init}}$ $\frac{dx_1}{dt} = x_I^{g_{10}} x_1^{g_{11}} - \beta_1 x_I^{h_{10}} x_1^{h_{11}}$
3	6	$x_1(0) = x_1^{\text{init}}$ $\frac{dx_1}{dt} = \alpha_1 x_I^{g_{10}} x_1^{g_{11}} - \beta_1 x_I^{h_{10}} x_1^{h_{11}}$
4	8	$x_1(0) = x_1^{\text{init}}$ $x_2(0) = x_2^{\text{init}}$ $\frac{dx_1}{dt} = \alpha_1 x_I^{g_{10}} x_1^{g_{11}} x_2^{g_{12}} - \beta_1 x_I^{h_{10}} x_1^{h_{11}}$ $\frac{dx_2}{dt} = x_2^{g_{22}} - 1$
5	9	$x_1(0) = x_1^{\text{init}}$ $x_2(0) = x_2^{\text{init}}$ $\frac{dx_1}{dt} = \alpha_1 x_I^{g_{10}} x_1^{g_{11}} x_2^{g_{12}} - \beta_1 x_I^{h_{10}} x_1^{h_{11}} x_2^{h_{12}}$ $\frac{dx_2}{dt} = x_2^{g_{22}} - 1$
6	10	$x_1(0) = x_1^{\text{init}}$ $x_2(0) = x_2^{\text{init}}$ $\frac{dx_1}{dt} = \alpha_1 x_I^{g_{10}} x_1^{g_{11}} x_2^{g_{12}} - \beta_1 x_I^{h_{10}} x_1^{h_{11}} x_2^{h_{12}}$ $\frac{dx_2}{dt} = x_1^{g_{21}} x_2^{g_{22}} - 1$

TABLE I. The first seven models of an example hierarchy in the S-systems class with one input x_I and fixed initial conditions x_1^{init} and x_2^{init}.

Model No. i	Num. parameters N_p	Form of sigmoidal ODEs
0	3	$x_1(0) = x_1^{\text{init}}$ $\frac{dx_1}{dt} = -x_1/\tau_1 + W_{11}\xi(x_1) + W_{10}x_I$
1	4	$x_1(0) = x_1^{\text{init}}$ $\frac{dx_1}{dt} = -x_1/\tau_1 + W_{11}\xi(x_1 + \theta_1) + W_{10}x_I$
2	6	$x_1(0) = x_1^{\text{init}}$ $x_2(0) = x_2^{\text{init}}$ $\frac{dx_1}{dt} = -x_1/\tau_1 + W_{11}\xi(x_1 + \theta_1) + W_{12}\xi(x_2) + W_{10}x_I$ $\frac{dx_2}{dt} = -x_2$
3	7	$x_1(0) = x_1^{\text{init}}$ $x_2(0) = x_2^{\text{init}}$ $\frac{dx_1}{dt} = -x_1/\tau_1 + W_{11}\xi(x_1 + \theta_1) + W_{12}\xi(x_2) + W_{10}x_I$ $\frac{dx_2}{dt} = -x_2 + W_{20}x_I$
4	8	$x_1(0) = x_1^{\text{init}}$ $x_2(0) = x_2^{\text{init}}$ $\frac{dx_1}{dt} = -x_1/\tau_1 + W_{11}\xi(x_1 + \theta_1) + W_{12}\xi(x_2) + W_{10}x_I$ $\frac{dx_2}{dt} = -x_2 + W_{21}\xi(x_1 + \theta_1) + W_{20}x_I$
5	9	$x_1(0) = x_1^{\text{init}}$ $x_2(0) = x_2^{\text{init}}$ $\frac{dx_1}{dt} = -x_1/\tau_1 + W_{11}\xi(x_1 + \theta_1) + W_{12}\xi(x_2) + W_{10}x_I$ $\frac{dx_2}{dt} = -x_2 + W_{22}\xi(x_2) + W_{21}\xi(x_1 + \theta_1) + W_{20}x_I$

TABLE II. The first six models of an example model hierarchy in the sigmoidal class with one input x_I and fixed x_1^{init} and x_2^{init}.

bound receptors) can be represented exactly in the sigmoidal class using two dynamical variables. Treating $I = \log S$ as the input to the system, the sigmoidal system

$$\begin{aligned} \frac{dx_1}{dt} &= -\frac{x_1}{\tau_1} - I, \\ \frac{dx_2}{dt} &= -x_2 + \xi(x_1 + \theta_1), \end{aligned} \tag{S2}$$

where we set $\tau_1 = n$ and $\theta_1 = \log K$, has a steady state solution that reproduces (S1):

$$\lim_{t\to\infty} x_2(t) = \xi(-n \log S + \log K) = f. \tag{S3}$$

C. Robustness of adaptive inference

In FIG. S2, we test the robustness of the performance of adaptive models in the multi-site phosphorylation example (see below) when various assumptions of the modeling framework are violated.

First, the derivation in Section V assumes that the distribution of noise on measured data is Gaussian with known variance. In FIG. S2A, we compare fitting to the same data but using an incorrect standard deviation for noise on the data when calculating the Bayesian log-likelihood. When the data is thought to be noisier than it actually is (purple and red points), performance remains unchanged until large N, when, as expected, simpler than optimal models are chosen, and comparatively more data is required to select complex models that produce better performance. When the data is thought to be less noisy than it actually is (yellow points), more complex models are selected, which in this case yields performance that can be better or worse, depending on N. In FIG. S2B, we compare fitting to data with log-normally distributed noise, keeping the mean and variance fixed. The closely overlapping performance suggests that, in the absence of knowledge about the true noise distribution, a good estimate of σ may be enough to attain consistent inference.

Finally, a somewhat arbitrary choice must be made to define an ordering for adding parameters in the model hierarchy; we chose to use the "node order" that is described in Table II. In FIG. S2C, we instead add parameters for each dynamical variable in random order. This includes orderings that first add parameters controlling only hidden nodes, which may be decoupled from the visible variables and hence cannot improve the fit. To compensate for this and avoid erroneously stopping

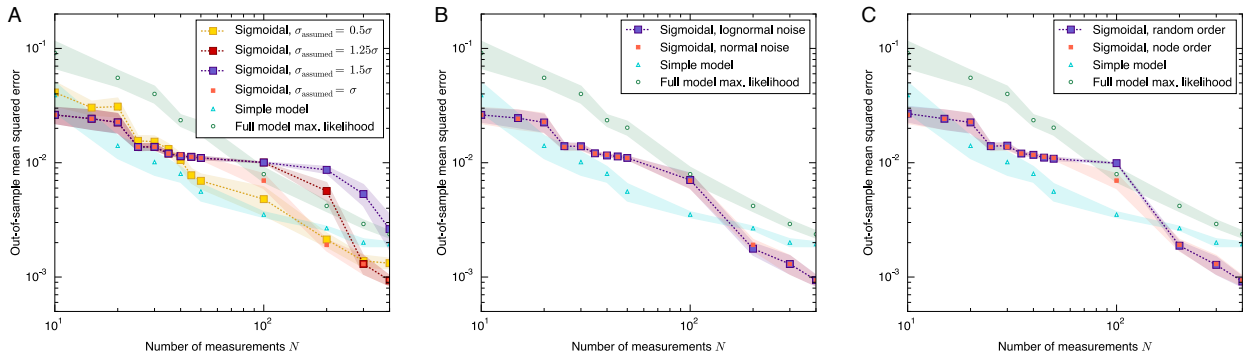


FIG. S2. Testing the robustness of adaptive inference in the multi-site phosphorylation example. In each case, the original performance curves from the main text's Fig. 2 (smaller symbols) are compared to an altered version of the model selection process (larger symbols). (A) Comparing fitting to the same data but using an incorrect standard deviation σ_{assumed} when calculating the Bayesian log-likelihood. (B) Comparing fitting to data with log-normally distributed noise; the two lines overlap and are hard to distinguish on the plot. (C) Comparing to adding parameters in random order, averaged over 10 realizations. See text for details.

fitting due to adding these unproductive parameters, we increase the number of models checked by increasing $i_{\text{overshoot}}$ from 3 to 4 (see Section VI). One could additionally avoid unproductive orderings by checking that each additional parameter has some causal influence on visible variables. But even including these orderings, mean performance is largely unaffected.

II. THE LAW OF GRAVITY MODEL

For a mass m in motion under the influence of the gravitational field of a mass $M \gg m$, the distance r between the two evolves as [3]

$$\frac{d^2r}{dt^2} = \frac{h^2}{r^3} - \frac{GM}{r^2}, \tag{S4}$$

where $h = (\vec{v}_0 \cdot \hat{\theta})r_0$ is the specific angular momentum, $\vec{v}_0$ is the initial velocity, r_0 is the initial distance, $\hat{\theta}$ is the unit vector perpendicular to the line connecting the two masses, and G is the gravitational constant. Setting the initial velocity parallel to $\hat{\theta}$ and measuring distance in units of $\frac{GM}{v_0^2}$ and time in units of $\frac{GM}{v_0^3}$, the dynamics become [29]

$$\frac{d^2r}{dt^2} = \frac{1}{r^2}\left(\frac{r_0^2}{r} - 1\right). \tag{S5}$$

When written as two first-order differential equations, we see that this system can be represented exactly in the S-systems class if the particle does not fall onto the Sun:

$$\begin{aligned}\frac{dr}{dt} &= \chi - 1 \\ \frac{d\chi}{dt} &= r_0^2 r^{-3} - r^{-2},\end{aligned} \tag{S6}$$

where we use the variable $\chi = \frac{dr}{dt} + 1$, so that the resulting system's variables are never negative, a requirement of the S-systems class.

To illustrate constructing an adaptive model for planetary motion, we consider as input the initial distance from the sun r_0. We sample r_0 uniformly between 1 and 3 (in units of GM/v_0^2), which covers the possible types of dynamics: at $r_0 = 1$, the orbit is circular; when $1 < r_0 < 2$ the orbit is elliptical; when $r_0 = 2$ the orbit is parabolic; and when $r_0 > 2$ the orbit is hyperbolic. In this and later examples, to best determine the minimum number of measurements needed for a given level of performance, we sample the system at a single time point for each initial condition (FIG. S3), rather than sampling a whole trajectory per condition. This ensures that samples are independent, which would not be the case for subsequent data points of the same trajectory, and hence allows us to estimate the data requirements of the algorithm more reliably. Further, this is similar to the sampling procedure already used in the literature [4]. In the planetary motion case, we assume only the distance r is measured, meaning the total number of of datapoints $N_D = N$, where N is the number of initial conditions sampled. We choose the time of the observation as a random time uniformly chosen between 0 and 100, with time measured in units of GM/v_0^3. To each measurement we add Gaussian noise with standard deviation equal to 5% of the maximum value of r between $t = 0$ and $t = 100\ GM/v_0^3$.

Typical training data for the model can be seen in FIG. S3. Fits to $N = 150$ data points are shown in FIG. 1. Here our adaptive fitting algorithm selects a model of the correct dimension, with one hidden variable. The selected model ODEs in this case are

$$\begin{aligned}\frac{dr}{dt} &= e^{-3.405} r_0^{3.428} r^{0.049} X_2^{7.372} - e^{-2.980} r_0^{2.936} r^{0.046} X_2 - 4.925 \\ \frac{dX_2}{dt} &= r_0^{-0.651} r^{-3.435} X_2^{-0.014} - e^{-0.006} r_0^{-4.288} r^{-1.595}.\end{aligned} \tag{S7}$$

Note that certain transformations of the hidden variable and parameters can leave the output behavior unchanged while remaining in the S-systems class. First, the initial condition of hidden parameters can be rescaled to 1 without loss of generality, so we remove this degree of freedom and set $X_2(0) = 1$. Second, we have the freedom to let the hidden variable $X_2 \to X_2^\gamma$ for any $\gamma \neq 0$ with appropriate shifts in parameters. To more easily compare the fit model with the perfect model, in the rightmost column of FIG. 1 we plot X_2^2 on the vertical axes instead of X_2 when comparing it to the dynamics of the true hidden variable χ.

Finally, we may compare performance when we fit the gravitation data using sigmoidal models, a model class that we know is not representative of the underlying mechanics. The results are shown in FIG. S4; the selected sigmoidal network, which contains three hidden variables, still provides a good fit to the data, as expected, but it does not generalize as well when r_0 is near the edge of the range contained in the data and timepoints are outside of the range of data to which they were fit. This is expected since forces can diverge in the true law of gravity, and they are necessarily limited in the sigmoidal model.

III. MULTI-SITE PHOSPHORYLATION MODEL

To explore a complicated biological system with relatively simple output behavior, we imagine a situation in which an immune receptor can be phosphorylated at each of five sites arranged in a linear chain. The rates of phosphorylation and dephosphorylation at each site are affected by the phosphorylation states of its nearest neighboring sites. A site can be unphosphorylated (U) or phosphorylated (P), and its state can change via one of two processes. The first process does not depend on states of neighboring sites:

$$U_i \rightleftharpoons P_i, \tag{S8}$$

with on-rate $k_i^{\text{on}}([U_i])$ and off-rate $k_i^{\text{off}}([P_i])$ that depend on the concentration of the corresponding substrate. The second, cooperative process happens only when a neighboring site j is phosphorylated:

$$U_i P_j \rightleftharpoons P_i P_j \tag{S9}$$

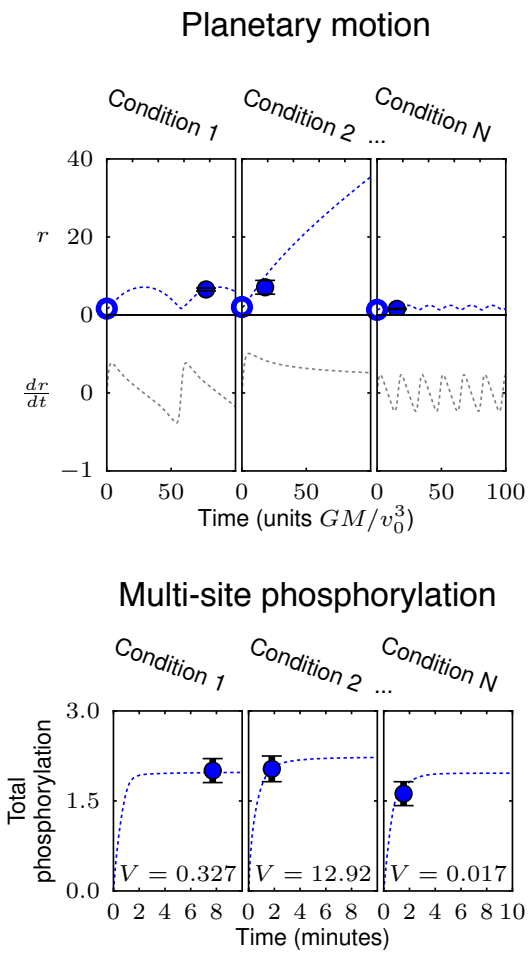


FIG. S3. Typical in-sample data points for the planetary motion and multi-site phosphorylation model examples. For the planetary motion, r_0 is treated as input, and for each in-sample r_0, r is measured, with added noise, at a single randomly chosen time between 0 and 100. For multi-site phosphorylation, the single parameter V is treated as input, and the total phosphorylation is measured, with added noise, at a single randomly chosen time between 0 and 10 minutes. Dotted lines show the original model behavior, filled circles with error bars show the in-sample data, and unfilled circles show the varying initial conditions in the planetary motion case. The original planetary motion model includes a single hidden variable X_2 corresponding to the time derivative of r. (For the yeast glycolysis example, a similar depiction of typical in-sample data is shown in the left panel of FIG. 4.)

with on- and off-rates $k_{ij}^{\text{on}}([U_iP_j])$ and $k_{ij}^{\text{off}}([P_iP_j])$. All rates k are modeled as Michaelis-Menten reactions: $k([S]) = \frac{V[S]}{K_m+[S]}$. With each reaction specified by two parameters (V and K_m) and 26 possible reactions, the phosphorylation model has a total of 52 parameters. To more easily generate the differential equations that govern the multi-site phosphorylation model, we use the

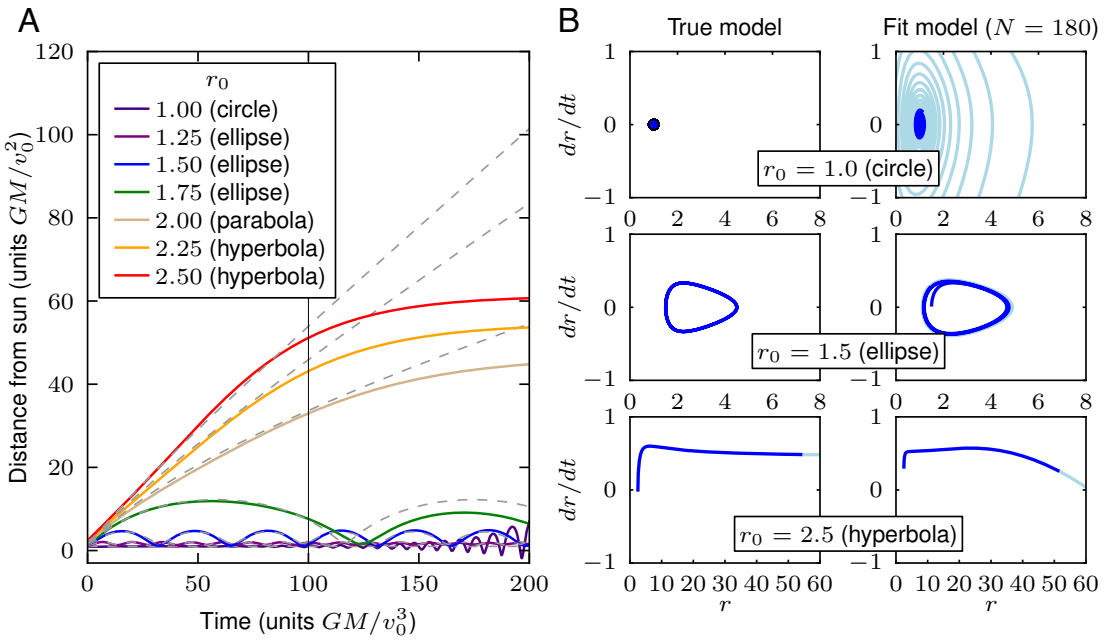


FIG. S4. Fit of sigmoidal model to planetary data. We know that the sigmoidal network model class is not likely to perform as well for the planetary data case because gravitational interactions do not saturate. Here we show the performance of a model fit to $N = 180$ data points, which contains three hidden variables. The model still fits well in the time region where data is given (between 0 and 100 GM/v_0^3, corresponding to the left half of A and the dark blue part of the trajectories in B), but has a larger divergence from the expected behavior at the extremes of the range of given r_0s in the extrapolated time region (corresponding to the right half of A and the light blue part of the trajectories in B).

BioNetGen package [5, 6].

When fitting this phosphorylation model, we use as input the parameter V_{23}^{on}, which is chosen from a uniform distribution in log-space between 10^{-3} and 10^{3} min^{-1}. The remaining 51 V and K_m parameters we sample randomly from our priors on these parameters. As output, we measure the total phosphorylation of the 5 sites P_{tot} at a single random time uniformly chosen between 0 and 10 minutes. To each measurement we add Gaussian noise with standard deviation equal to 10% of the P_{tot} value at $t = 10$ min.

Typical training data for the model is shown in FIG. S3. The out-of-sample mean squared error, as plotted in FIG. 2, is measured over 100 new input values selected from the same distribution as the in-sample values, each of which is compared to the true model at 100 timepoints evenly spaced from 0 to 10 minutes.

As a simple guess to the functional form of the total phosphorylation timecourse as a function of our control parameter $V = V_{23}^{\text{on}}$ (the "simple model" in FIG. 2), we use an exponential saturation starting at 0 and ending at a value P_∞ that depends sigmoidally on V:

$$P_{\text{tot}} = P_\infty(V)\left[1 - \exp\left(\frac{t}{t_0}\right)\right], \tag{S10}$$

where

$$P_\infty(V) = a + \frac{b}{2}\left[1 + \tanh\left(\frac{\log(V) - d}{c}\right)\right] \tag{S11}$$

and a, b, c, d, and t_0 are parameters fit to the data. FIG. 2shows that this simple *ad hoc* model can fit the data quite well.

For the example shown in FIG. 3, the selected sigmoidal model consists of the ODEs

$$\begin{aligned}
\frac{dP_{\text{tot}}}{dt} &= \frac{-P_{\text{tot}}}{e^{-1.219}} + \frac{0.409}{1+\exp(P_{\text{tot}} - 4.469)} + \frac{7.087}{1+\exp(X_2)} + 0.0005V \\
\frac{dX_2}{dt} &= -X_2 - \frac{2.303}{1+\exp(P_{\text{tot}} - 4.469)} - 0.071V \\
X_2(0) &= 0.101,
\end{aligned} \tag{S12}$$

with $P_{\text{tot}}(0) = 0$.

The selected sigmoidal models contain fewer parameters than the microscopic exact model, even when taking into account that the full model is effectively lower dimensional, with many directions in parameter space unconstrained by typical data; see FIG. S5.

In this multi-site phosphorylation example, the sigmoidal model class is a better performer than the S-systems class. A typical example of performance is depicted in FIG. S6. Though the S-systems class makes predictions that are still qualitatively correct, and its predictions steadily improve as N increases, the sigmoidal class comes closer to the true underlying model with an equal amount of data.

The confidence intervals on the dynamics in FIG. 3correspond to samples from the posterior over parameters given $N = 300$ data points. In the notation of section V, this posterior $P(\alpha \mid \text{data}) \propto \exp\left[-\tilde{\chi}^2(\alpha)/2\right]$. To generate samples from this distribution, we use Metropolis Monte Carlo as implemented in SloppyCell [7, 8]. As a starting point, we use the best-fit parameters from the model

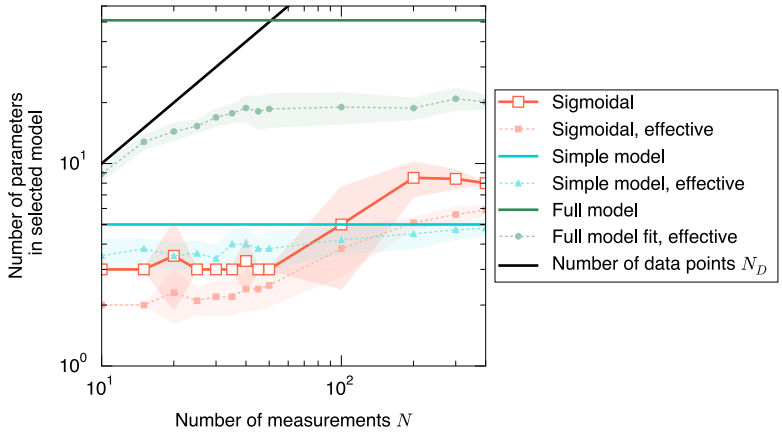


FIG. S5. Selected adaptive sigmoidal models in the phosphorylation example have both fewer total parameters and fewer effective parameters than the full microscopic model. Solid colored lines indicate the total number of parameters in each model, as in Figure 2 in the main text. Solid symbols connected by dotted lines indicate the effective number of parameters, which we define as the number of directions in parameter space that are constrained by the data such that the corresponding Hessian eigenvalue $\lambda > 1$ (compared to parameter priors with eigenvalue 10^{-2}). Shown are the mean and standard deviation of values over 10 data realizations. For comparison, the solid black line indicates the number of data points $N_D = N$ used to infer the model.

selection procedure, and we sample candidate steps in parameter space from a multidimensional Gaussian corresponding to the Hessian at the best-fit parameters.[30] From 10^4 Monte Carlo steps, the first half are removed to avoid bias from the initial condition, and every 50 of the remaining steps are used as 100 approximately independent samples from the parameter posterior.

IV. YEAST GLYCOLYSIS MODEL

As an example of inference of more complicated dynamics, we use a model of oscillations in yeast glycolysis, originally studied in terms of temperature compensation [9] and since used as a test system for automated inference [4]. The model's behavior is defined by ODEs describing the dynamics of the concentrations of seven molecular species (the biological meaning of the species is

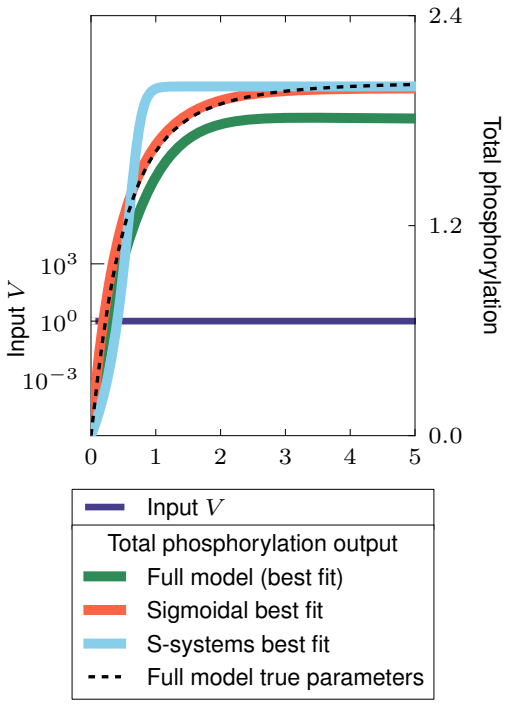


FIG. S6. A typical example of out-of-sample performance in the multi-site phosphorylation example. Here, each model is fit using $N = 50$ datapoints. With this small amount of data, the differences between model classes are more apparent, with the sigmoidal model class clearly better predicting the dynamics than the S-systems model class and the full phosphorylation model.

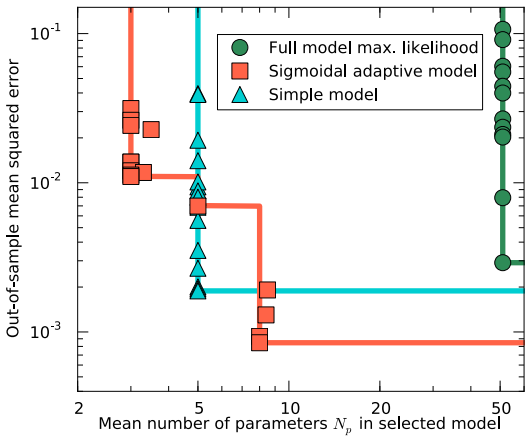


FIG. S7. The performance of models fit to data from the multi-site phosphorylation model as a function of the number of parameters in each model. This is a replotting of the data in Figure 2 in the main text. If we think of a model as more efficient if it can produce the same level of predictive power with fewer parameters, then the best models lie at the Pareto front, drawn in solid lines for each model type.

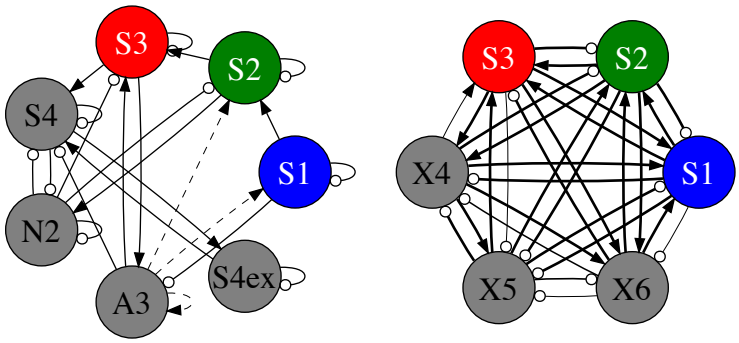


FIG. S8. (Left) Network depicting the yeast glycolysis model defined by Eqns. (S13). Solid arrows represent excitation, solid lines with circles represent inhibition, and dashed arrows represent other types of interaction terms. (Right) Selected sigmoidal network fit to $N = 40$ noisy measurements from the yeast glycolysis model, as shown in FIG. 4. Again, arrows represent excitation and circles inhibition, with the thickness of arrows indicating interaction strength. For clarity, self-inhibitory terms for each variable are not shown.

not important here):

$$\begin{aligned}
\frac{dS_1}{dt} &= J_0 - \frac{k_1 S_1 S_6}{1 + (S_6/K_1)^q} \\
\frac{dS_2}{dt} &= 2\frac{k_1 S_1 S_6}{1 + (S_6/K_1)^q} - k_2 S_2 (N - S_5) - k_6 S_2 S_5 \\
\frac{dS_3}{dt} &= k_2 S_2 (N - S_5) - k_3 S_3 (A - S_6) \\
\frac{dS_4}{dt} &= k_3 S_3 (A - S_6) - k_4 S_4 S_5 - \kappa (S_4 - S_7) \\
\frac{dS_5}{dt} &= k_2 S_2 (N - S_5) - k_4 S_4 S_5 - k_6 S_2 S_5 \\
\frac{dS_6}{dt} &= -2\frac{k_1 S_1 S_6}{1 + (S_6/K_1)^q} + 2k_3 S_3 (A - S_6) - k_5 S_6 \\
\frac{dS_7}{dt} &= \psi \kappa (S_4 - S_7) - k S_7.
\end{aligned} \tag{S13}$$

Parameter values, listed in Table III, are set to match with those used in Ref. [4] and Table 1 of Ref. [9], where our $S_5 = N_2$, our $S_6 = A_3$, and our $S_7 = S_4^{ex}$.

For the yeast glycolysis model, we use as input the initial conditions for the visible species S_1, S_2, and S_3. These are each chosen uniformly from ranges listed in the "In-sample IC" column of Table IV. Each of the three visible species are then measured at a random time uniformly chosen from 0 to 5 minutes, meaning the total number of datapoints $N_D = 3N$ for this system, where N is the number of initial conditions sampled. Gaussian noise is added to each measurement with

J_0	2.5 mM min^{-1}
k_1	100. mM^{-1} min^{-1}
k_2	6. mM^{-1} min^{-1}
k_3	16. mM^{-1} min^{-1}
k_4	100. mM^{-1} min^{-1}
k_5	1.28 min^{-1}
k_6	12. mM^{-1} min^{-1}
k	1.8 min^{-1}
κ	13. min^{-1}
q	4
K_1	0.52 mM
ψ	0.1
N	1. mM
A	4. mM

TABLE III. Parameters for the yeast glycolysis model defined in Eqns. (S13).

standard deviations given in Table IV. To evaluate the model's performance, we test it using 100 new input values selected uniformly from the ranges listed in the "Out-of-sample IC" column of Table IV, each of which is compared to the true model at 100 timepoints evenly spaced from 0 to 5 min. The correlation between the adaptive fit model and the actual model over these 100 timepoints is calculated separately for each visible species, set of initial conditions, and in-sample data, and the average is plotted as the "mean out-of-sample correlation" in FIG. 4. The topology of the selected sigmoidal model in an example with $N = 40$ is illustrated in FIG. S8. The model

Variable	In-sample IC (mM)	Out-of-sample IC (mM)	In-sample σ (mM)
S_1	[0.15, 1.60]	[0.15, 3.05]	0.04872
S_2	[0.19, 2.16]	[0.19, 4.13]	0.06263
S_3	[0.04, 0.20]	[0.04, 0.36]	0.00503
S_4	0.115	0.115	N/A
S_5	0.077	0.077	N/A
S_6	2.475	2.475	N/A
S_7	0.077	0.077	N/A

TABLE IV. Initial conditions (IC) and standard deviations of experimental noise (σ) used in the yeast glycolysis model. Initial conditions for visible species S_1, S_2, and S_3 are chosen uniformly from the given ranges, chosen to match Ref. [4]. Out-of-sample ranges are each twice as large as in-sample ranges. Initial conditions for the remaining hidden species are fixed at reference initial conditions from Refs. [4] and [9]. In-sample noise is set at 10% of the standard deviation of each variable's concentration in the limit cycle, as quoted in Ref. [4].

ODEs in this case are

$$
\begin{aligned}
\frac{dS_1}{dt} &= \frac{-S_1}{e^{2.284}} + \frac{2.520}{1+\exp(S_1-0.4246)} + \frac{14.04}{1+\exp(S_2-0.4943)} - \frac{19.56}{1+\exp(S_3+0.6711)} \\
&\quad - \frac{10.68}{1+\exp(X_4+2.240)} + \frac{6.759}{1+\exp(X_5-0.7566)} - \frac{3.051}{1+\exp(X_6)} \\
\frac{dS_2}{dt} &= \frac{-S_2}{e^{-1.288}} - \frac{3.015}{1+\exp(S_1-0.4246)} + \frac{2.244}{1+\exp(S_2-0.4943)} + \frac{14.55}{1+\exp(S_3+0.6711)} \\
&\quad + \frac{25.77}{1+\exp(X_4+2.240)} - \frac{6.699}{1+\exp(X_5-0.7566)} - \frac{4.380}{1+\exp(X_6)} \\
\frac{dS_3}{dt} &= \frac{-S_3}{e^{1.514}} - \frac{2.463}{1+\exp(S_1-0.4246)} - \frac{10.99}{1+\exp(S_2-0.4943)} + \frac{0.6530}{1+\exp(S_3+0.6711)} \\
&\quad - \frac{0.07038}{1+\exp(X_4+2.240)} - \frac{6.806}{1+\exp(X_5-0.7566)} + \frac{12.61}{1+\exp(X_6)} \\
\frac{dX_4}{dt} &= \frac{-X_4}{e^{1.771}} + \frac{25.77}{1+\exp(S_1-0.4246)} - \frac{50.05}{1+\exp(S_2-0.4943)} - \frac{6.648}{1+\exp(S_3+0.6711)} \\
&\quad - \frac{59.44}{1+\exp(X_4+2.240)} + \frac{52.34}{1+\exp(X_5-0.7566)} + \frac{1.148}{1+\exp(X_6)} \\
\frac{dX_5}{dt} &= \frac{-X_5}{e^{-2.513}} + \frac{16.39}{1+\exp(S_1-0.4246)} + \frac{33.15}{1+\exp(S_2-0.4943)} + \frac{0.6452}{1+\exp(S_3+0.6711)} \\
&\quad - \frac{33.65}{1+\exp(X_4+2.240)} - \frac{8.976}{1+\exp(X_5-0.7566)} + \frac{0.01966}{1+\exp(X_6)} \\
\frac{dX_6}{dt} &= -X_6 + \frac{0.3391}{1+\exp(S_1-0.4246)} - \frac{2.514}{1+\exp(S_2-0.4943)} - \frac{4.479}{1+\exp(S_3+0.6711)} \\
&\quad - \frac{3.396}{1+\exp(X_4+2.240)} + \frac{1.219}{1+\exp(X_5-0.7566)} + \frac{2.313}{1+\exp(X_6)}
\end{aligned}
\tag{S14}
$$

$$X_4(0) = 3.437$$

$$X_5(0) = 1.453$$

$$X_6(0) = -0.7183.$$

Note that our model fitting approach assumes that the model timecourse is fully determined (aside from measurement error) by the concentrations of measured species. To be consistent with this assumption we do not vary the initial conditions of the three hidden variables. In future work it may be possible to relax this assumption, allowing the current state of intrinsic variations in hidden variables to be learned as well.

A. Simple sinusoidal model

As with the multi-state phosphorylation example, we can use a simple *ad hoc* model of yeast glycolysis for comparison to our adaptive models. The long-term behavior of the yeast network

consists of stable oscillations with a roughly fixed period; a minimally complicated model of the measured concentrations S_1, S_2, and S_3 then consists of three sinusoidal oscillators with equal frequency ω and phase relationship fixed by two parameters, ϕ_2 and ϕ_3:

$$\begin{aligned} S_1(t) &= y_1 + A_1 \sin(\omega t + \phi) \\ S_2(t) &= y_2 + A_2 \sin(\omega t + \phi + \phi_2) \\ S_3(t) &= y_3 + A_3 \sin(\omega t + \phi + \phi_3). \end{aligned} \tag{S15}$$

The phase ϕ depends on the initial conditions $S_1(0), S_2(0), S_3(0)$. Specifically, when the initial condition is a valid point on the one-dimensional elliptical curve specified by Eqs. (S15), ϕ can be determined by any two initial values; for instance,

$$\phi = \arctan \frac{x_1 \sin(\phi_2)}{x_2 - x_1 \cos(\phi_2)}, \tag{S16}$$

where $x_i = (S_i(0) - y_i)/A_i$. Because the model is not exact, however, we cannot assume that initial conditions will lie on this curve. Instead, we will assume that transient dynamics infinitely quickly bring the state of the system into the plane defined by the curve. This plane has normal vector $\vec{n} = (\sin(\phi_2 - \phi_3), \sin\phi_3, -\sin\phi_2)$, so that any initial conditions $\vec{x}$ can be projected onto a point on the plane $\vec{x}' = \vec{x} - c\vec{n}$, where $c = (\vec{x} \cdot \vec{n})/(\vec{n} \cdot \vec{n}) = (x_1 \sin(\phi_2 - \phi_3) + x_2 \sin\phi_3 - x_3 \sin\phi_2)/(\sin^2(\phi_2 - \phi_3) + \sin^2\phi_2 + \sin^2\phi_3)$. Thus $\vec{x}'$ is a modified initial condition that is inserted into (S16) to obtain ϕ. Unlike the adaptive model, this simple sinusoidal model does not capture the jagged shape of the yeast glycolysis oscillations, but when its 9 parameters are fit to data, its rough approximation is moderately predictive. Its performance is compared to sigmoidal adaptive models in FIG. S9.

B. Comparing to EUREQa

In Ref. [4], the EUREQa engine is used to infer the same yeast glycolysis model that we use here. We can roughly compare performance as a function of computational and experimental effort by measuring the number of required model evaluations and measurements (FIG. 4). Here we compare the two approaches in more detail. However, we emphasize that they have different goals: EUREQa aims at finding the exact microscopic model of the process, while Sir Isaac strives

for accurate prediction with the simplest phenomenological model. The former is a harder task, and thus one expects it to require more data and computation.

Reference [4] attempts to match time derivatives of species concentrations as a function of species concentrations, instead of species concentrations as a function of time as we do. This means that each model evaluation[31] is more computationally costly for us, since it requires an integration of the ODEs over time. It also means, however, that we are able to match well the phases of oscillations, which remain unconstrained in Ref. [4]. The fitting of time courses instead of derivatives also makes our method focus on the fitting of dynamics near the attractor, rather than attempting to constrain dynamics through the entire phase space.

To consistently infer exact equations for the full 7-dimensional model, Ref. [4] used $20,000$ datapoints and roughly 10^{11} model evaluations. We contrast this with our method that produces reasonable inferred models using 40 datapoints and less than 5×10^8 model evaluations (FIG. 4).

Finally, in the main text we test the performance of our yeast glycolysis models for out-of-sample ranges of initial conditions that are twice as large as the in-sample ranges from which data is taken, as in Ref. [4], in order to more directly test their ability to extrapolate to regimes that were not tested in training. In FIG. S9, we compare this to performance when out-of-sample initial conditions are chosen from the same ranges as in-sample data (note that, nonetheless, none of the test examples has appeared in the training set). Here we see that the mean correlation can reach 0.9 using $N = 40$ measurements.

V. DERIVATION OF BAYESIAN LOG-LIKELIHOOD ESTIMATE $\mathcal{L}$

Multiple previous approaches have used approximate sampling methods to perform Bayesian model selection on a small number of alternate models in the context of systems biology; e. g., [10–12]. For our approach that relies on a search over an infinite set of models, even such approximate sampling is slow. Yet with sufficiently large N, an expansion resembling that used to derive the Bayesian Information Criterion produces good performance without sampling. The derivation here largely follows Refs. [13, 14], but can be traced to the 1970s [15].

For a given model M that depends on parameters α, our model selection algorithm requires

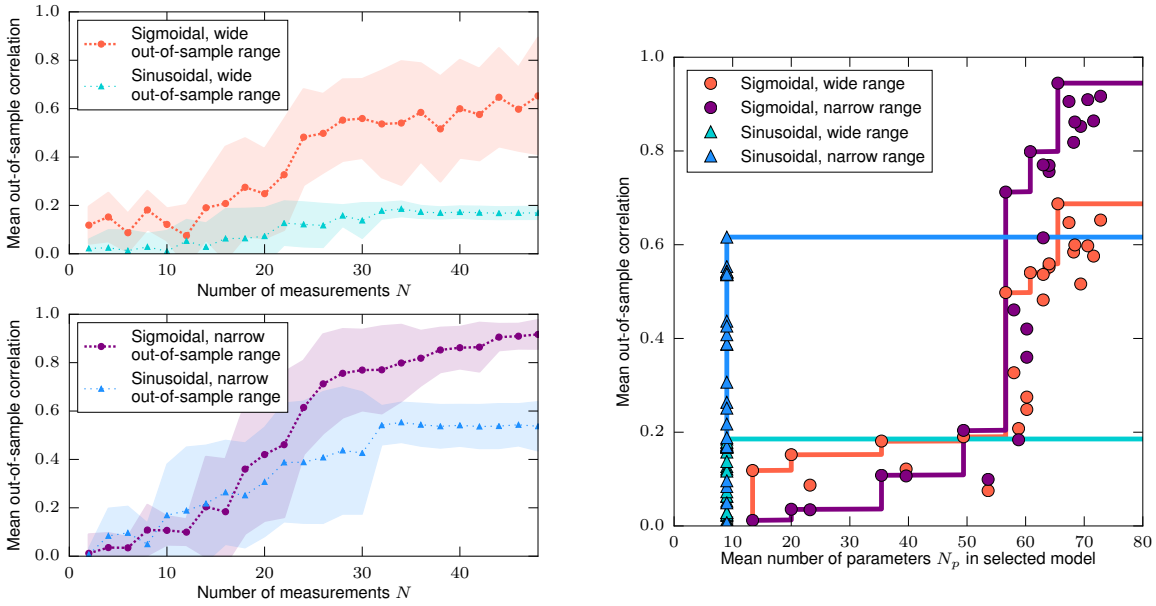


FIG. S9. Performance of inferred models of yeast glycolysis as a function of the number of measurements N (left) and the mean number of parameters N_p in the selected model (right). The given sigmoidal model hierarchy requires about 30 measurements (corresponding to 90 datapoints) and 60 parameters to produce reasonable predictions. Here we compare mean correlations produced for out-of-sample initial conditions chosen from ranges twice as large as in-sample ranges ("wide ranges," plotted in red, listed in the "out-of-sample" column of Table IV) to when out-of-sample conditions are chosen from the same ranges as in-sample ranges ("narrow ranges," plotted in purple, listed in the "in-sample" column of Table IV). For comparison, the simple sinusoidal model defined in (S15) is shown in shades of blue. The mean and standard deviation over 5 realizations of in-sample data are shown by filled symbols and shaded regions. Also plotted are the Pareto fronts for each model (solid lines on right plot) indicating the maximal correlation for a given mean N_p.

an estimate of the probability that M is the model that produced a given set of data $\{y_i\}$ with corresponding error estimates $\{\sigma_i\}$ (measured at a set of timepoints $\{t_i\}$), and $i = 1, \ldots, N$, so that there are N measurements. Since the parameters α are unknown aside from a prior distribution $P(\alpha)$, we must integrate over all possible values:

$$P(M \mid \text{data}) = P(M \mid \{y_i, \sigma_i, t_i\}) \tag{S17}$$

$$= Z_\alpha^{-1} \int d^{N_p}\alpha \; P(M \mid \{y_i, \sigma_i, t_i\}; \alpha) \; P(\alpha), \tag{S18}$$

where the normalization constant $Z_\alpha = \int d^{N_p}\alpha \; P(\alpha)$ and N_p is the number of parameters. In

terms of the output given the model, Bayes rule states

$$P\left(M \mid \{y_i, \sigma_i, t_i\}; \alpha\right) = \frac{P\left(M\right)}{P\left(\{y_i\}\right)} P\left(\{y_i\} \mid M(\alpha); \{\sigma_i, t_i\}\right). \tag{S19}$$

Assuming that the model output has normally distributed measurement errors,

$$\begin{aligned} P\left(\{y_i\} \mid M(\alpha); \{\sigma_i, t_i\}\right) &= \prod_{i=1}^{N} P\left(y_i \mid M(\alpha); \sigma_i; t_i\right) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - M(t_i, \alpha)}{\sigma_i}\right)^2\right] \\ &= Z_\sigma^{-1} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - M(t_i, \alpha)}{\sigma_i}\right)^2\right] \\ &= Z_\sigma^{-1} \exp\left[-\frac{1}{2}\chi^2(M(\alpha), \{y_i, \sigma_i, t_i\})\right], \end{aligned} \tag{S20}$$

where χ^2 is the usual goodness-of-fit measure consisting of the sum of squared residuals, and Z_σ is the normalization constant $\prod_{i=1}^{N}\sqrt{2\pi\sigma_i^2}$. Thus we have:[32]

$$P(M \mid \text{data}) = C Z_\alpha^{-1} \int d^{N_p}\alpha \quad \exp\left[-\frac{1}{2}\tilde{\chi}^2(\alpha)\right], \tag{S21}$$

where $C \equiv 2P(M)/Z_\sigma P\left(\{y_i\}\right)$ and $\tilde{\chi}^2(\alpha) = \chi^2(\alpha) - 2\log P(\alpha)$. Since we will be comparing models fitting the same data, and we assume all models have the same prior probability $P(M)$, C will be assumed constant in all further comparisons (but see Ref. [16] for the discussion of this assumption).

If there are enough data to sufficiently constrain the parameters (as is the case for ideal data in the limit $N \to \infty$), then the integral will be dominated by the parameters near the single set of best-fit parameters α_{best}. To lowest order in $1/N$, we can approximate the integral using a saddle-point approximation [14]:

$$P(M \mid \text{data}) \approx C Z_\alpha^{-1} \exp\left[-\frac{1}{2}\tilde{\chi}^2(\alpha_{\text{best}})\right] \int d^{N_p}\alpha \ \exp\left[-(\alpha - \alpha_{\text{best}})\mathcal{H}(\alpha - \alpha_{\text{best}})\right], \tag{S22}$$

where $\mathcal{H}$ is the Hessian:[33]

$$\mathcal{H}_{k\ell} = \frac{1}{2}\frac{\partial^2 \tilde{\chi}^2(\alpha)}{\partial\alpha_k d\alpha_\ell}\bigg|_{\alpha_{\text{best}}}. \tag{S23}$$

If we assume normally distributed priors on parameters with variances ς_k^2, the log posterior probability becomes

$$\log P(M \mid \text{data}) \approx \text{const} - \frac{1}{2}\tilde{\chi}^2(\alpha_{\text{best}}) - \frac{1}{2}\sum_{\mu=1}^{N_p}\log\lambda_\mu - \frac{1}{2}\sum_{k=1}^{N_p}\log\varsigma_k^2, \tag{S24}$$

where λ_μ are the eigenvalues of $\mathcal{H}$, and the last term comes from Z_α. We thus use as our measure of model quality

$$\mathcal{L} \equiv -\frac{1}{2}\tilde{\chi}^2(\alpha_{\rm best}) - \frac{1}{2}\sum_\mu \log\lambda_\mu - \frac{1}{2}\sum_k \log\varsigma_k^2. \tag{S25}$$

Eq. (S25) is a generalization of the Bayesian Information Criterion (BIC) [15] when parameter sensitivities and priors are explicitly included.[34] The first term is the familiar χ^2 "goodness of fit," and the last two terms constitute the fluctuation "penalty" for overfitting or complexity. Note that here the goodness of fit and the complexity penalty are both functions of the entire dynamics, rather than individual samples, which is not a common application of Bayesian model selection techniques.

VI. FITTING ALGORITHM

We are given N data points $\mathbf{x}_i$ at known times t_i and known exogenous parameters I_i, and with known or estimated variances σ_i^2. We are approximating the functions $\vec{F}_X$ and $\vec{F}_Y$ in Eq. (1), where $\mathbf{y}$ are hidden dynamic model variables, and $\mathbf{x} = \mathbf{x}(t, I)$ and $\mathbf{y} = \mathbf{y}(t, I)$ in general depend on time t and inputs I. As described in Section V, we fit to the data $\mathbf{x}_i$ using a combination of squared residuals from the data and priors $P(\alpha)$ on parameters α, which we assume to be Gaussian and centered at zero:

$$\tilde{\chi}^2 = \sum_{i=1}^{N}\left(\frac{\mathbf{x}_i - \mathbf{x}(t_i, I_i)}{\sigma_i}\right)^2 + 2\sum_{k=1}^{N_p}\left(\frac{\alpha_k}{\varsigma_k}\right)^2, \tag{S26}$$

where F's are integrated to produce the model values $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{x}(t, I) = \mathbf{x}_0(I) + \int_0^t \vec{F}_X(\mathbf{x}(s, I), \mathbf{y}(s, I))\, ds \tag{S27}$$

$$\mathbf{y}(t, I) = \mathbf{y}_0(I) + \int_0^t \vec{F}_Y(\mathbf{x}(s, I), \mathbf{y}(s, I))\, ds. \tag{S28}$$

To fit parameters, we use a two step process akin to simulated annealing that uses samples from a "high temperature" Monte Carlo ensemble as the starting points for local optimization performed using a Levenberg-Marquardt routine. The phenomenological models are implemented using SloppyCell [7, 8] in order to make use of its parameter estimation and sampling routines.

Following is a high-level description of the fitting algorithm, with choices of parameters for the examples in the main text listed in Table V.

1. Choose a model class, consisting of a sequence of nested models indexed by i, where the number of parameters N_p monotonically increases with i. Choose a step size Δp.

2. Given data at N_{total} timepoints, fit to data from the first N timepoints, where N is increased to N_{total} in steps of ΔN.

3. At each N, test models of increasing number of parameters N_p (stepping by Δp) until $\mathcal{L}$ consistently decreases (stopping when the last $i_{\text{overshoot}}$ models tested have smaller $\mathcal{L}$ than the maximum). For each model, to calculate $\mathcal{L}$:

   (a) Generate an ensemble of starting points in parameter space using Metropolis-Hastings Monte Carlo to sample from $P(\alpha) \propto \exp(-\tilde{\chi}^2(\alpha)/2TN_D)$ with $\tilde{\chi}^2$ from (S26). The temperature T is set large to encourage exploration of large regions of parameter space, but if set too large can result in a small acceptance ratio. Infinities and other integration errors are treated as $\tilde{\chi}^2 = \infty$.

      i. Use as a starting point the best-fit parameters from a smaller N_p if a smaller model has been previously fit, or else default parameters.

      ii. As a proposal distribution for candidate steps in parameter space, use an isotropic Gaussian with standard deviation $\sqrt{TN_D}/\lambda_{\max}$, where N_D is the total number of data residuals and $\lambda_{\max}$ is the largest singular value of the Hessian [Eq. (S23)] at the starting parameters.

      iii. If this model has previously been fit to less data, use those parameters as an additional member of the ensemble.

   (b) Starting from each member of the ensemble, perform a local parameter fit, using Levenberg-Marquardt to minimize $\tilde{\chi}^2$ from (S26). Stop when convergence is detected (when the L1 norm of the gradient per parameter is less than `avegtol`) or when the

number of minimization steps reaches `maxiter`. The best-fit parameters α^* are taken from the member of the ensemble with the smallest resulting fitted $\tilde{\chi}^2$.

(c) At α^*, calculate $\mathcal{L}$ from (S25).

4. For each N, the model with largest log-likelihood $\mathcal{L}$ is selected as the best-fit model.

Δp (gravitation and phosphorylation examples)	2
Δp (yeast example)	5
$i_{\text{overshoot}}$	3
Ensemble temperature T (full phosphorylation model)[a]	10
Ensemble temperature T (all other models)	10^3
Total number of Monte Carlo steps (full phosphorylation model)[a]	10^2
Total number of Monte Carlo steps (all other models)	10^4
Number of ensemble members used	10
`avegtol`	10^{-2}
`maxiter`	10^2

TABLE V. Adaptive inference algorithm parameters. [1]In the full phosphorylation model, we fit parameters in log-space since they are known to be positive. This makes the model more sensitive to large changes in parameters, meaning that we are forced to be more conservative with taking large steps in parameter space to achieve reasonable acceptance ratios.

VII. SCALING OF COMPUTATIONAL EFFORT

In FIG. S10, we plot the number of model evaluations used in each search for the best-fit phenomenological model. We define a model evaluation as a single integration of a system of ODEs. (Note that the amount of necessary CPU time per integration is dependent on the size and stiffness of the system.) This includes both integration of model ODEs and the derivatives of model ODEs, used in gradient calculations [35]. Note that in FIG. 4, to indicate the total number

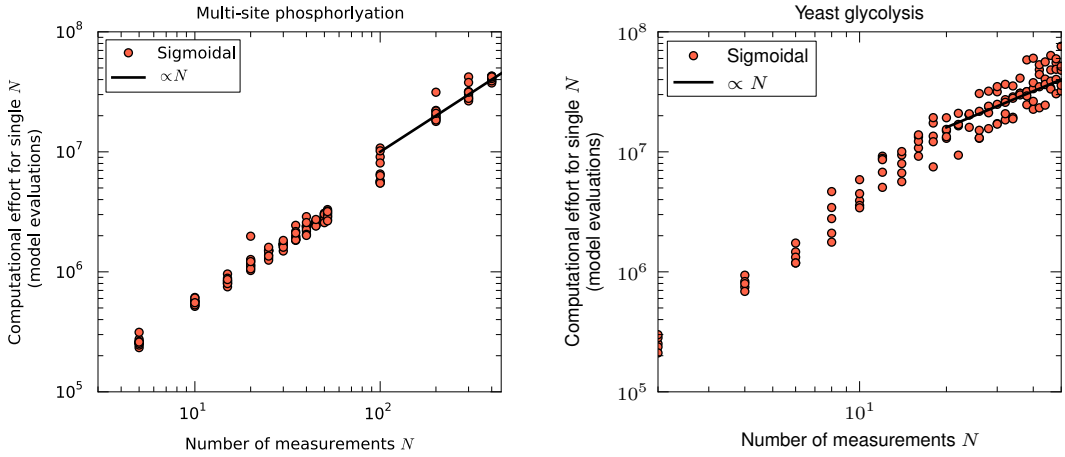


FIG. S10. The number of model evaluations (integrations) used at each N, for the multi-site phosphorylation and yeast glycolysis examples. Once the size of model has saturated, we expect the number of evaluations to scale linearly with N (black lines). If the selected model size is growing with N, as in the yeast glycolysis example below $N = 20$ (see FIG. S11), we expect faster than linear growth.

of evaluations used as N is gradually increased to its final value, we plot the cumulative sum of the number of model evaluations depicted in FIG. S10. We see that the number of model evaluations scales superlinearly with N if the selected model size is growing with N, as is the case in the yeast glycolysis model below about $N = 20$ (FIG. S10 and FIG. S11). When the model size saturates, the number of model evaluations scales roughly linearly with N.

VIII. COMPARISON TO BAYESIAN NETWORK APPROACHES

A related set of methods for inferring causal structure from time series data comes from the field of Bayesian Networks (BN), and specifically Dynamic Bayesian Networks (dBN). Implementations typically make the following assumptions:

1. Variables are updated at a discrete set of times, rather than continuously.

2. Latent variables are not allowed, or their number is known *a priori*.

3. The state space of dynamical variables is itself discrete (and often of low cardinality, such as binary or ternary).

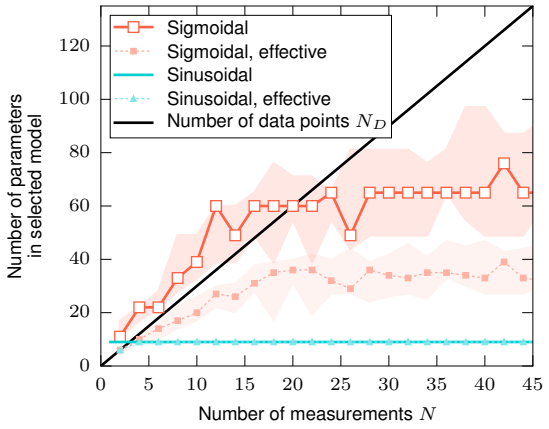


FIG. S11. Fitting sigmoidal models to the yeast glycolysis oscillation data, the number of total parameters in the selected model, plotted with red open squares, saturates to roughly 65. Plotted with red solid squares is the effective number of parameters, which we define as the number of directions in parameter space that are constrained by the data such that the corresponding Hessian eigenvalue $\lambda > 1$ (compared to parameter priors with eigenvalue 10^{-2}). Corresponding values for the simple sinusoidal model are plotted in blue. Since the blue curve does not grow for $N \geq 5$, we conclude that the simple model does not have the statistical power to fit the data and is too simple for this case. For comparison, the solid black line indicates the number of data points $N_D = 3N$ used to infer the model. We expect the optimal effective number of parameters to stay below N_D. Shown are the median and full range of values over 5 data realizations.

Many generalizations of (d)BNs have been presented that lift each of these assumptions. Below we include a brief literature review of current implementations of (d)BNs that address each issue. However, our method is distinct from (d)BNs in that it is designed to perform inference simultaneously for *continuous variables*, in *continuous time*, with potentially a *very large number of unknown hidden nodes*, and we are not aware of an approach that is able to lift all three assumptions in order to analyze the type of data handled by *Sir Isaac*.

Continuous time: It is known that exact inference is intractable in continuous time versions of dBNs because calculating a node's distribution at a given time step does not easily factor into conditionally independent subsets, as it does in cases with discrete time. Instead, each node's distribution will in general depend on the entire history of all other variables [17]. Approximate methods have been developed to deal with such Continuous Time Bayesian Networks (CTBNs) [17, 18]. Conversely, converting a set of ODEs, such as those explored by *Sir Isaac*, into the dBN

framework generally leads to an exponentially large model [19] that cannot be readily inferred from data.

Continuous states: Most implementations deal with discrete state variables, to avoid the need to infer multidimensional distributions over continuous variables, which can require very large data sets. It is also relatively common to use continuous variables by specifying the state of nodes as finite-parameter continuous distributions, such as Gaussians. However, these differ from *Sir Isaac* in that they are typically parameterized with means that are simply linear combinations of parent nodes (e. g., [20, 21]). One approach uses biochemically inspired functions relating means of continuous-valued nodes [22], but does not use continuous time.

Unspecified network size: Though some approaches attempt to discover that hidden nodes are necessary for a better description of a system (e. g., [22–24]), this is not a typical feature of Bayesian network implementations. Approaches that are complete, in the sense that they allow, in principle, infinitely many latent variables, are relatively rare (e. g., [25]), and do not address continuous space-time requirements.

---

[1] Nemenman, I. Fluctuation-dissipation theorem and models of learning. *Neural Comput* **17**, 2006 (2005).

[2] Savageau, M. A. & Voit, E. O. Recasting Nonlinear Differential Equations as S-Systems: A Canonical Nonlinear Form. *Math Biosci* **87**, 83–115 (1987).

[3] Landau, L. & Lifshitz, E. *Mechanics* (Butterworth-Heinemann, 1976), 3rd edn.

[4] Schmidt, M. *et al.* Automated refinement and inference of analytical models for metabolic networks. *Phys Biol* **8**, 055011 (2011).

[5] Hlavacek, W. S. *et al.* Rules for modeling signal-transduction systems. *Sci. STKE* **2006**, re6 (2006).

[6] Bionetgen. http://bionetgen.org.

[7] Myers, C. R., Gutenkunst, R. N. & Sethna, J. P. Python unleashed on systems biology. *Computing in Science and Engineering* **9**, 34 (2007).

[8] Gutenkunst, R. N. *et al.* Sloppycell. http://sloppycell.sourceforge.net.

[9] Ruoff, P., Christensen, M., Wolf, J. & Heinrich, R. Temperature dependency and temperature compensation in a model of yeast glycolytic oscillations. *Biophys Chem* **106**, 179 (2003).

[10] Vyshemirsky, V. & Girolami, M. Bayesian ranking of biochemical system models. *Bioinform* **24**, 833–839 (2008).

[11] Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interf* **6**, 187–202 (2009). 0901.1925.

[12] Lillacci, G. & Khammash, M. Parameter estimation and model selection in computational biology. *PLoS Comput Biol* **6** (2010).

[13] Balasubramanian, V. Statistical inference, occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comput* **9**, 349 (1997).

[14] Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput* **13**, 2409 (2001).

[15] Schwarz, G. Estimating the dimension of a model. *Annals Stat* **6**, 461 (1978).

[16] Wolpert, D. & Macready, W. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**, 67–82 (1997).

[17] Nodelman, U., Shelton, C. & Koller, D. Continuous time Bayesian networks. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)* 378–387 (2002). URL `http://dl.acm.org/citation.cfm?id=2073921`.

[18] Nodelman, U. *Continuous time Bayesian networks*. Ph.D. thesis, Stanford University (2007).

[19] Chatterjee, S. & Russell, S. Why are DBNs sparse? In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 81–88 (Sardinia, Italy, 2010). URL `http://machinelearning.wustl.edu/mlpapers/paper\_files/AISTATS2010\_ChatterjeeR10.pdf`.

[20] Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian Networks to Analyze Expression Data. *J Comput Biol* **7**, 601–620 (2000).

[21] Markowetz, F. & Spang, R. Inferring cellular networks–a review. *BMC Bioinf* **8 Suppl 6**, S5 (2007).

[22] Nachman, I., Regev, a. & Friedman, N. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics* **20 Suppl 1**, i248–56 (2004).

[23] Elidan, G., Lotner, N., Friedman, N. & Koller, D. Discovering hidden variables: A structure-based approach. In *Advances in Neural Information Processing Systems (NIPS 2000)* (2000).

[24] Bagrow, J. *et al.* Shadow networks: Discovering hidden nodes with models of information flow. *arXiv preprint arXiv:1312.6122* 1–12 (2013).

[25] Doshi-Velez, F., Wingate, D., Roy, N., Tenenbaum, J. & Roy, N. Infinite dynamic bayesian networks. In *International Conference on Machine Learning (ICML)* (2011).

[26] In general, we are not guaranteed good predictive power until $N \to \infty$, but we can hope that the assumptions implicit in our priors (consisting of the specific form of the chosen model hierarchy and the priors on its parameters) will lead to good predictive power even for small N.

[27] For the simple model fit to the phosphorylation data, parameters are always well-constrained and priors are unimportant, and we therefore do not use explicit priors.

[28] Some parameters (α and β in the S-systems model class, τ in the sigmoidal model class, and k and K parameters in the full phosphorylation model) are restricted to be positive, which we accomplish

by optimizing over the log of each parameter. The priors are still applied in non-log space, effectively creating a prior that is zero for negative parameter values and $2N(0, 10)$ for positive parameter values.

[29] Note that r_0 sets the (conserved) angular momentum: $h = \frac{GM}{v_0} r_0$ with r_0 in rescaled units.

[30] Unconstrained parameter directions in the proposal distribution, corresponding to singular values smaller than $\lambda_{\text{cut}} = \lambda_{\max}/10$, where $\lambda_{\max}$ is the largest singular value, are cut off to λ_{cut} to produce reasonable acceptance ratios (near 0.5).

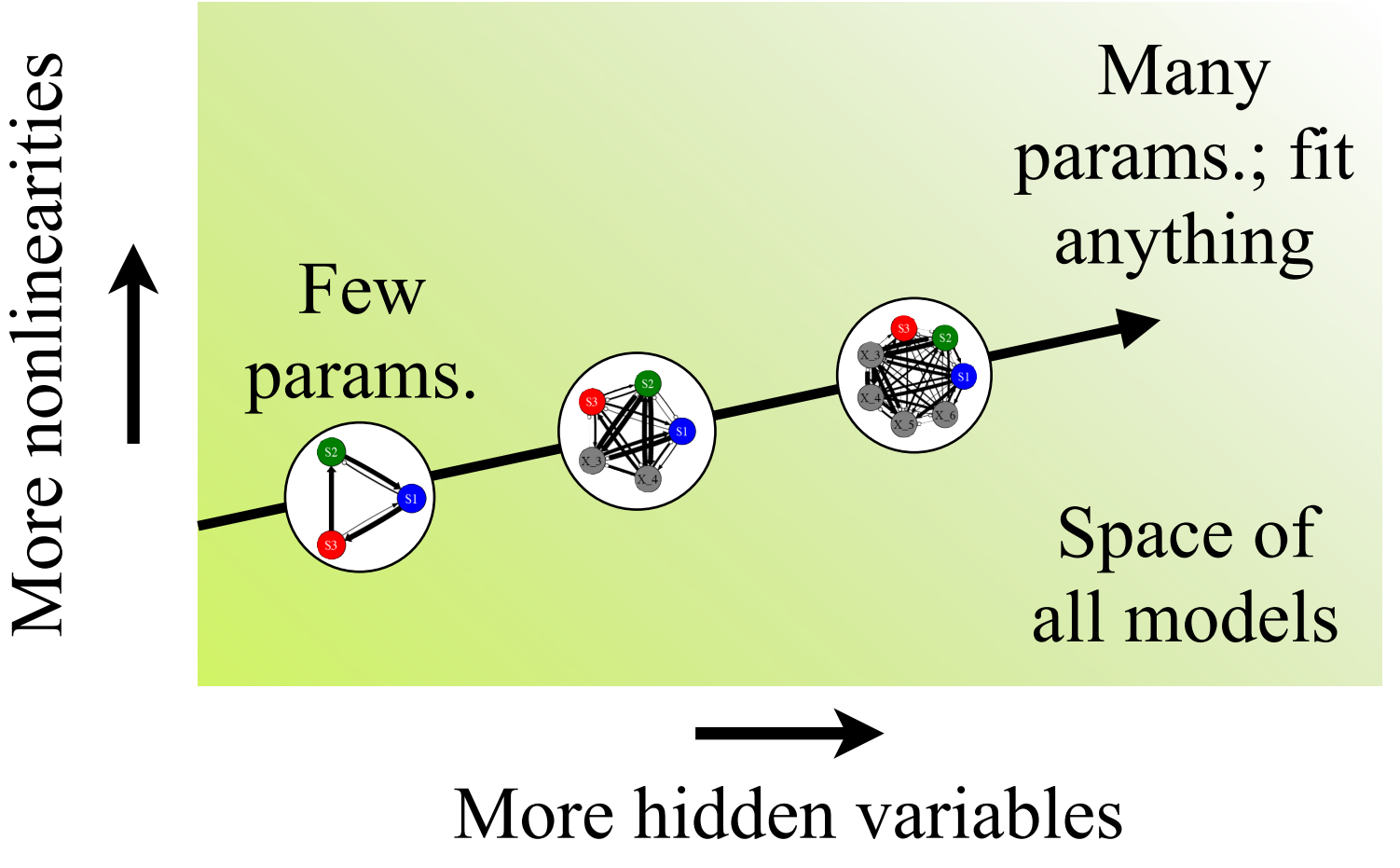
[31] In our setup, we define a model evaluation as a single integration of the model ODEs (see Section VII).

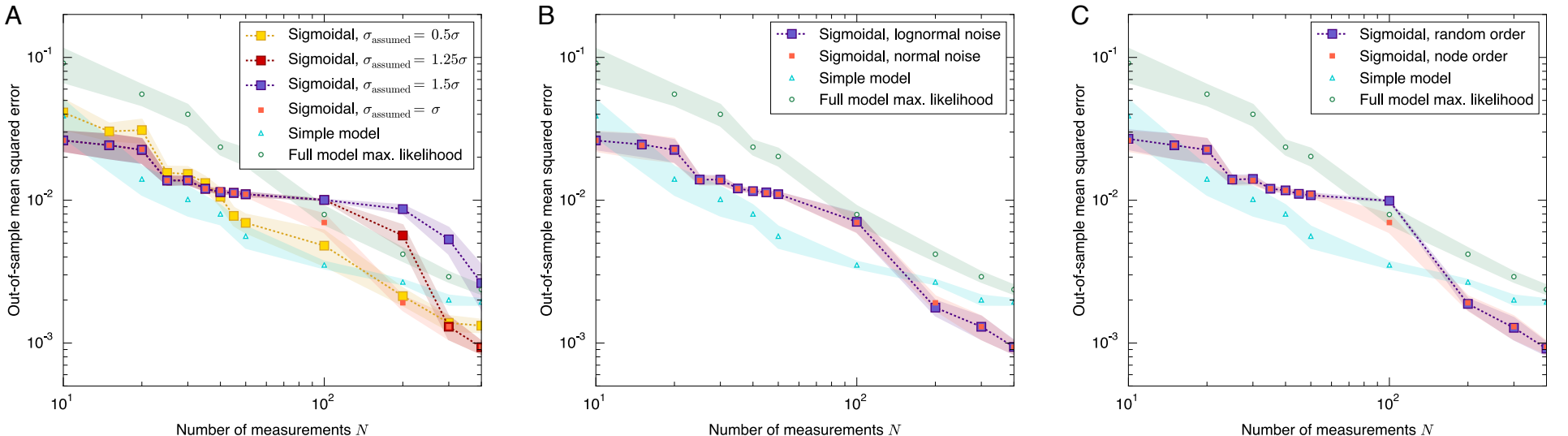
[32] We simplify notation by letting $\chi^2(\alpha) = \chi^2(M(\alpha), \{y_i, \sigma_i, t_i\})$.

[33] Near the best-fit parameters where residuals are small, and when priors are Gaussian, $\mathcal{H}$ is approximated by the Fisher Information Matrix, which depends only on first derivatives of model behavior: $\mathcal{H} \approx J^T J + \Sigma^{-2}$, where the Jacobian $J_{i\ell} = \frac{1}{\sigma_i} \frac{\partial M_i}{\partial \alpha_\ell}$ and the diagonal matrix $\Sigma^{-2}_{k\ell} = \delta_{k\ell} \varsigma_k^{-2}$ expresses the effects of parameter priors.

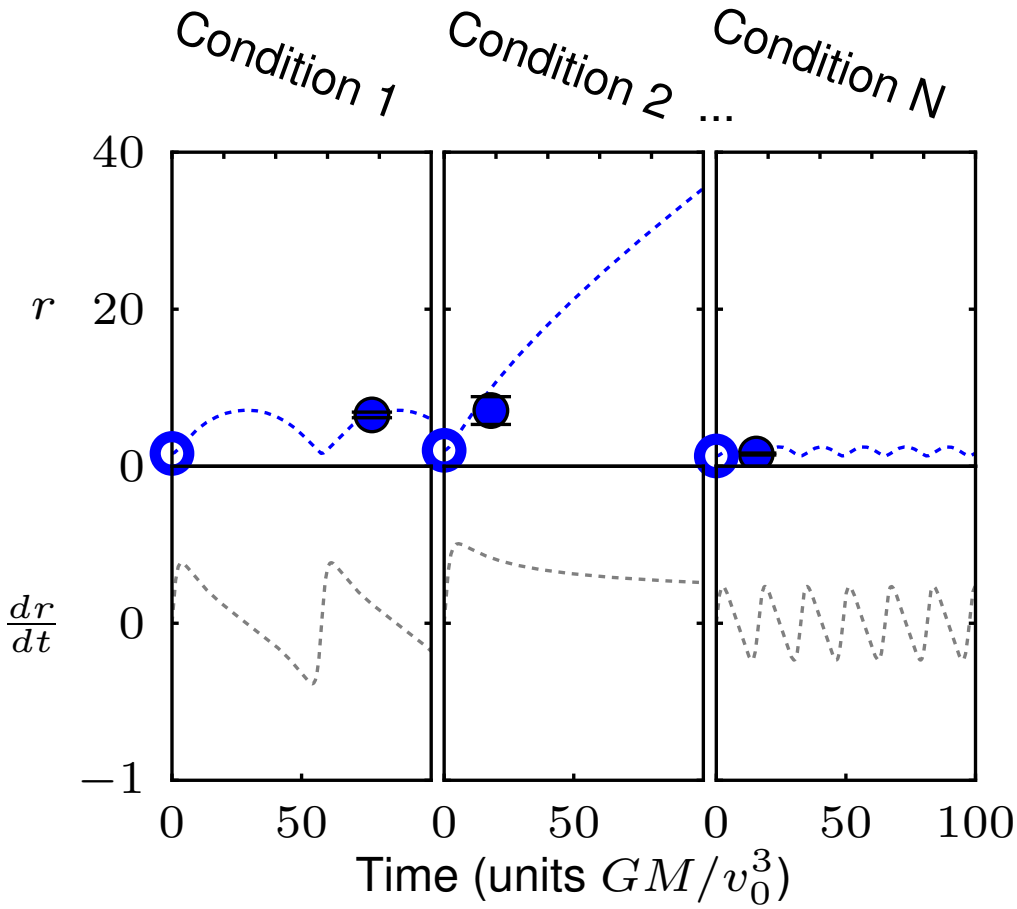
[34] For well-constrained parameters, we expect, to lowest order in $1/N$, our result to be equal to the BIC result of $-\frac{1}{2}\tilde{\chi}^2(\alpha_{\text{best}}) + \frac{1}{2} N_p \log N$.

[35] The number of integrations per gradient calculation is proportional to the number of parameters. This means that the computational effort used to fit large models is dominated by gradient calculations.

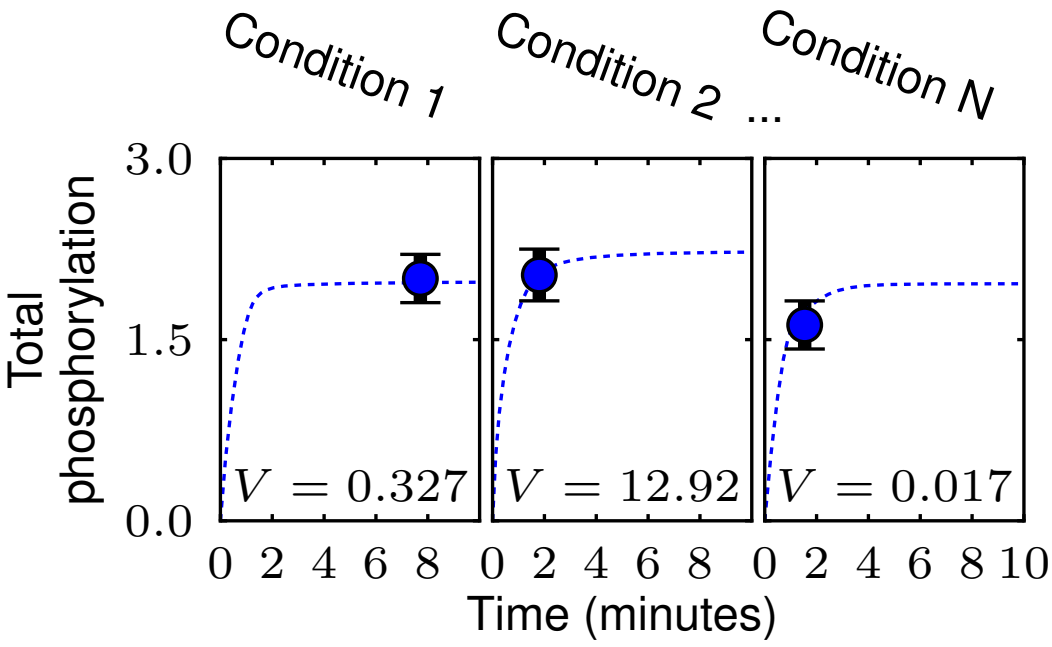

More nonlinearities
Few params.
Many params.; fit anything
Space of all models
More hidden variables
S1
S2
S3
X_3
X_4
X_5
X_6

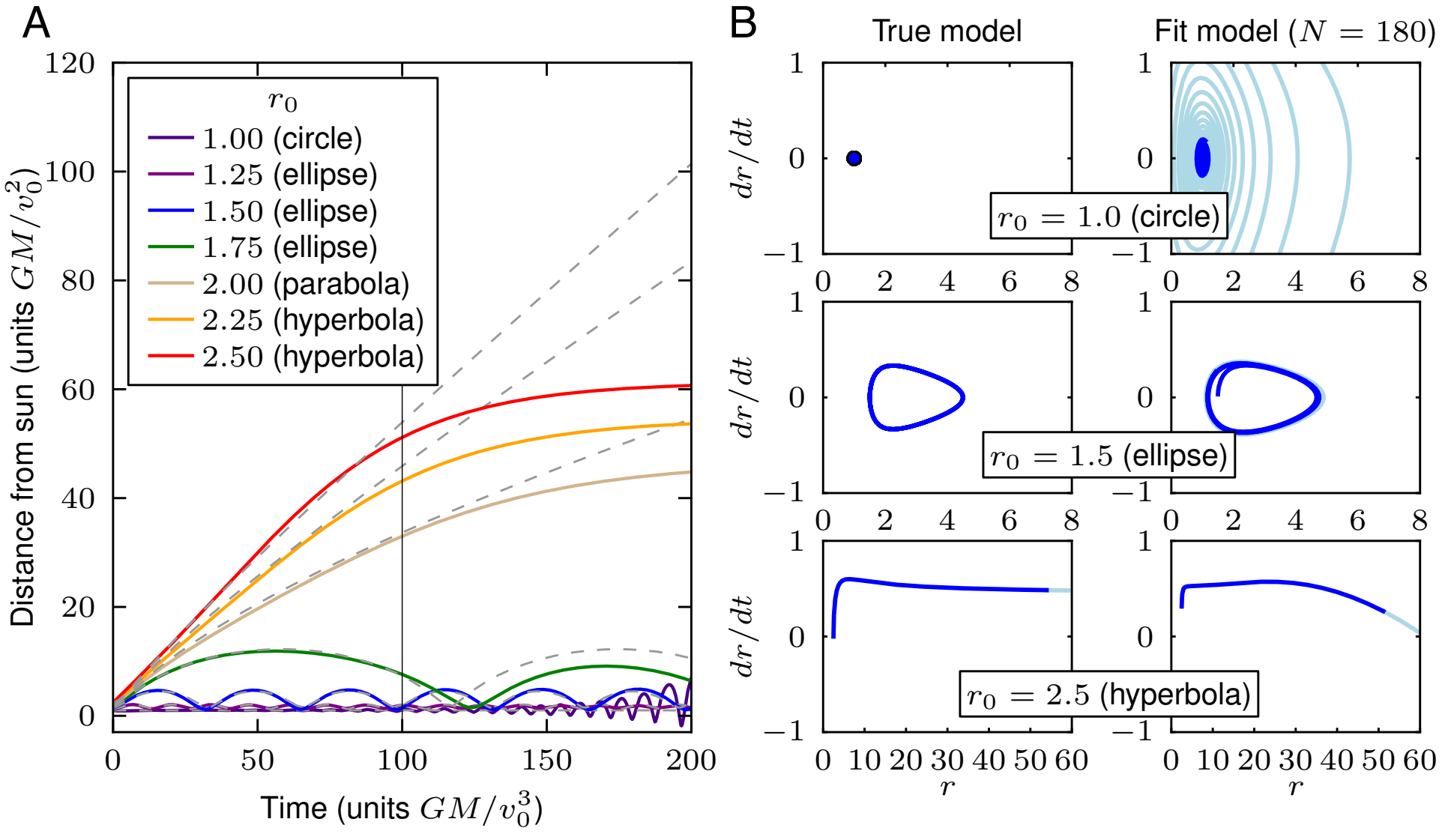

A
B
C
Out-of-sample mean squared error
Number of measurements N
10^{-1}
10^{-2}
10^{-3}
10^{1}
10^{2}
Sigmoidal, $\sigma_{\text{assumed}} = 0.5\sigma$
Sigmoidal, $\sigma_{\text{assumed}} = 1.25\sigma$
Sigmoidal, $\sigma_{\text{assumed}} = 1.5\sigma$
Sigmoidal, $\sigma_{\text{assumed}} = \sigma$
Simple model
Full model max. likelihood
Sigmoidal, lognormal noise
Sigmoidal, normal noise
Simple model
Full model max. likelihood
Sigmoidal, random order
Sigmoidal, node order
Simple model
Full model max. likelihood

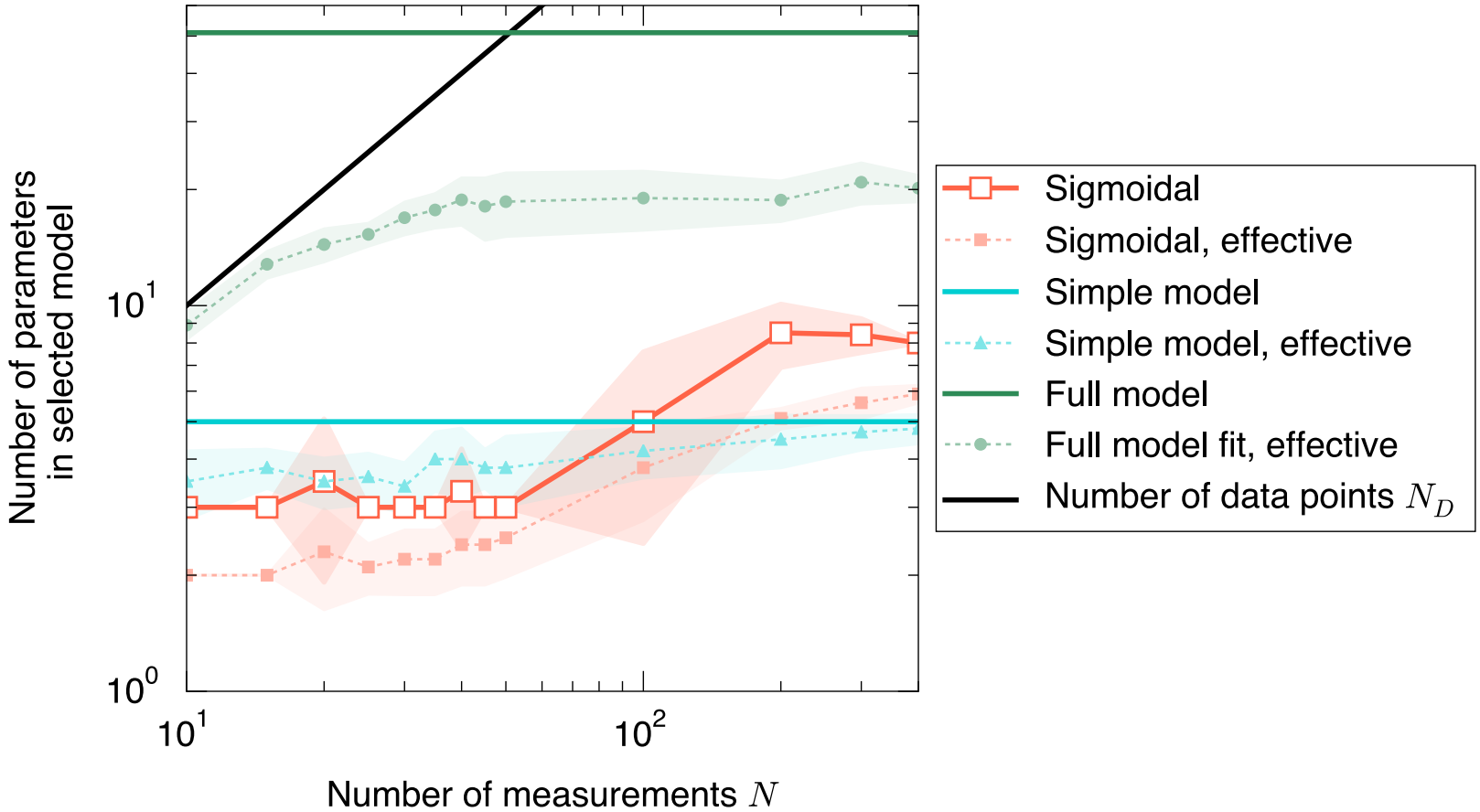
Planetary motion

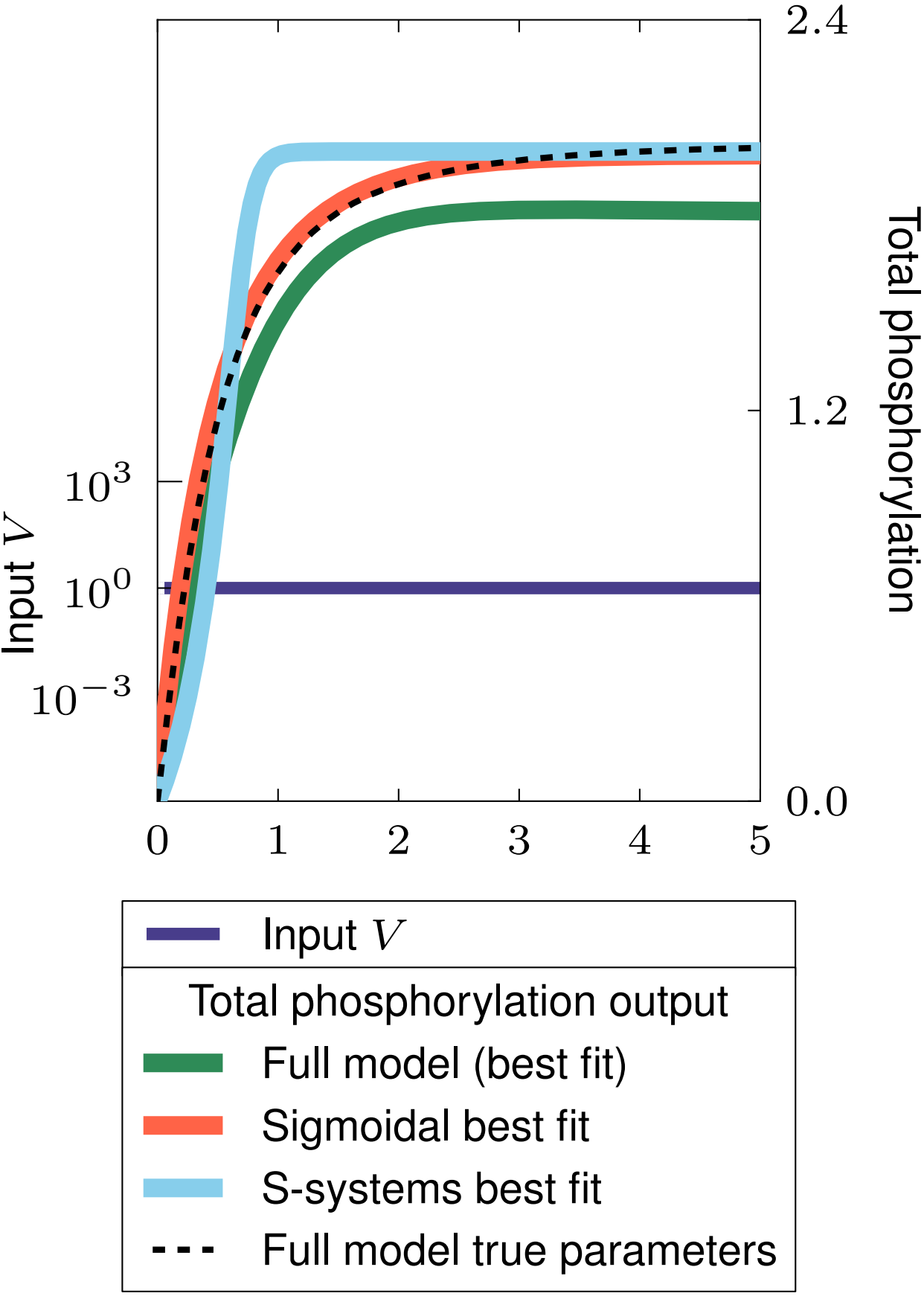


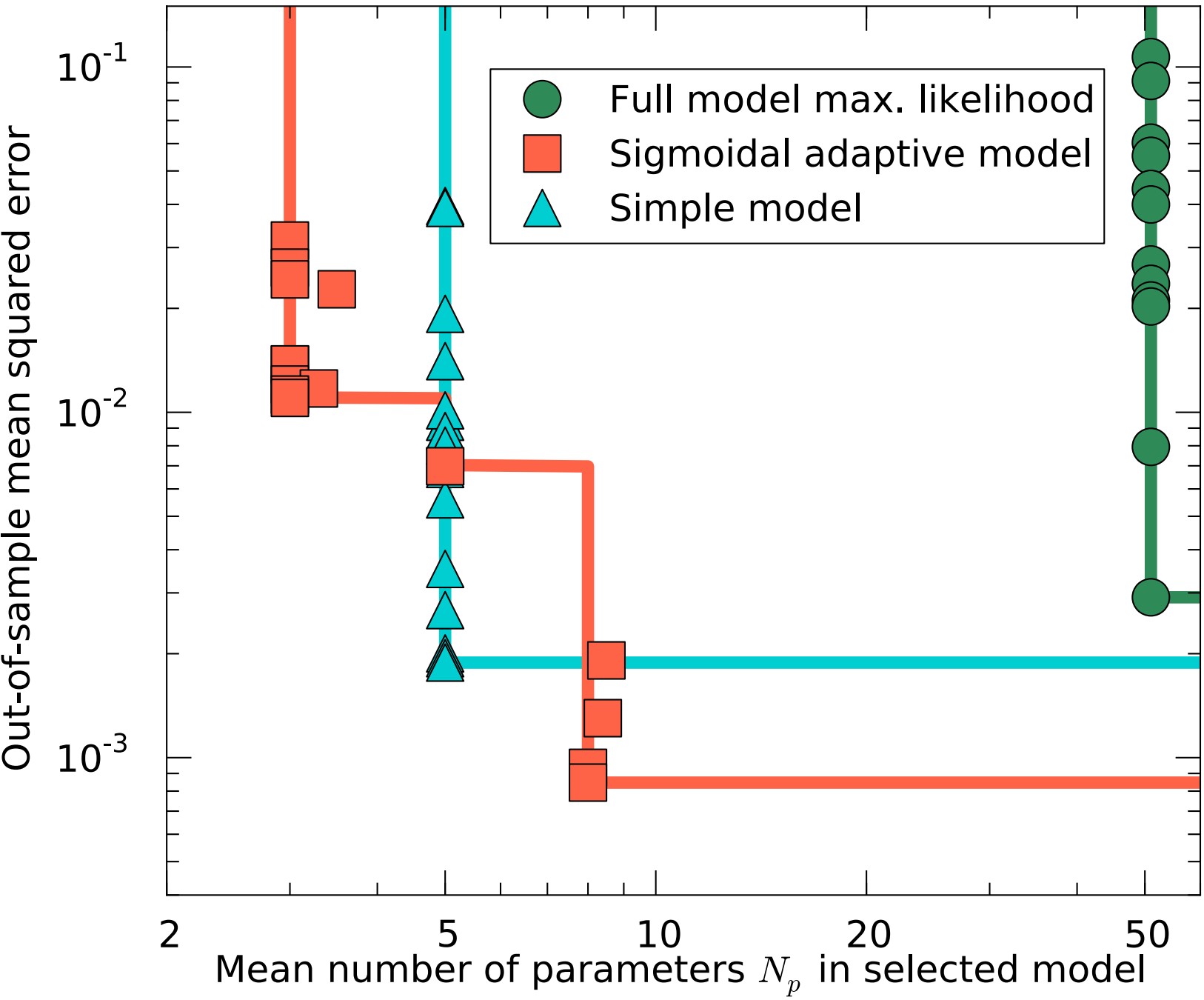
Multi-site phosphorylation

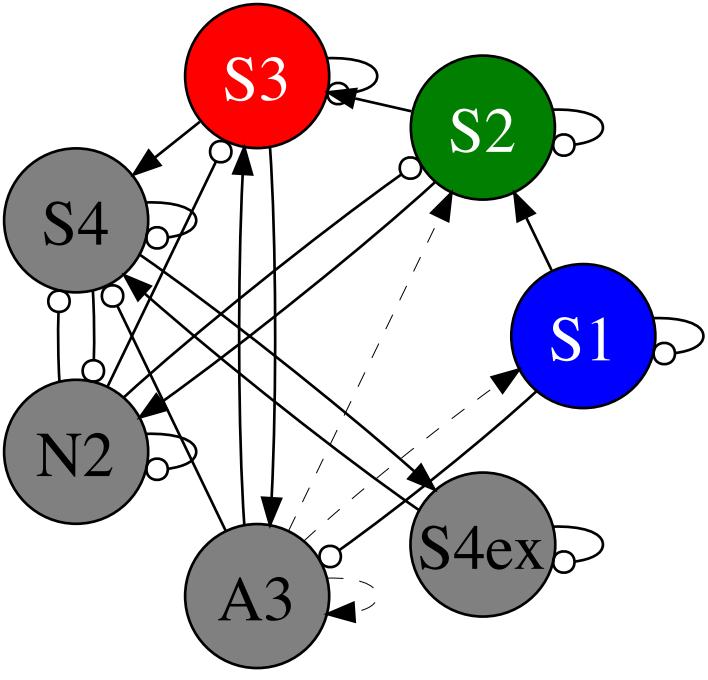


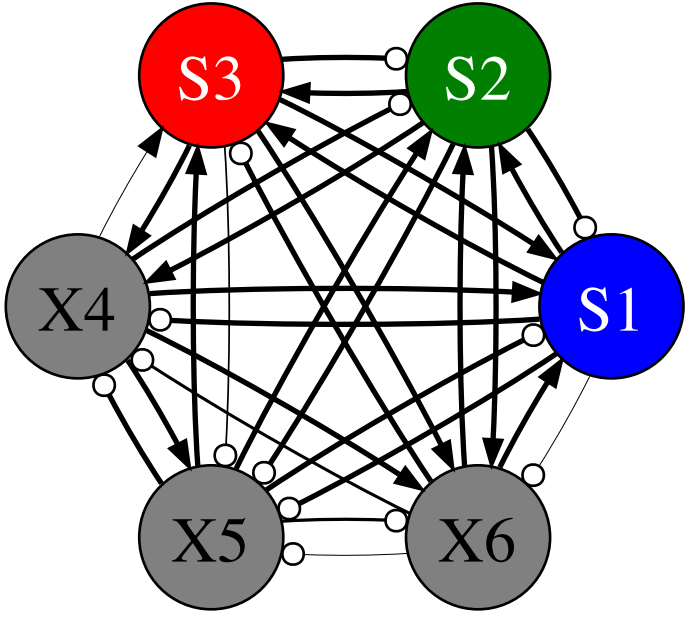

A
B
True model
Fit model ($N = 180$)
r_0
1.00 (circle)
1.25 (ellipse)
1.50 (ellipse)
1.75 (ellipse)
2.00 (parabola)
2.25 (hyperbola)
2.50 (hyperbola)
Distance from sun (units GM/v_0^2)
Time (units GM/v_0^3)
dr/dt
r
$r_0 = 1.0$ (circle)
$r_0 = 1.5$ (ellipse)
$r_0 = 2.5$ (hyperbola)

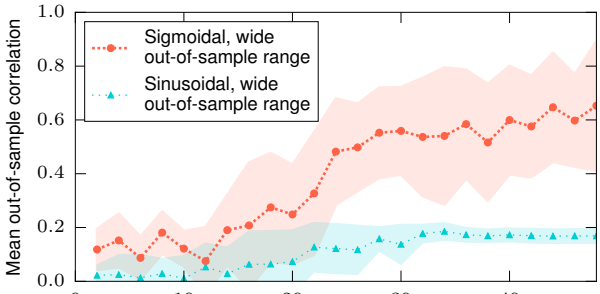

Number of parameters in selected model
10^1
10^0
10^1
10^2
Number of measurements N
Sigmoidal
Sigmoidal, effective
Simple model
Simple model, effective
Full model
Full model fit, effective
Number of data points N_D

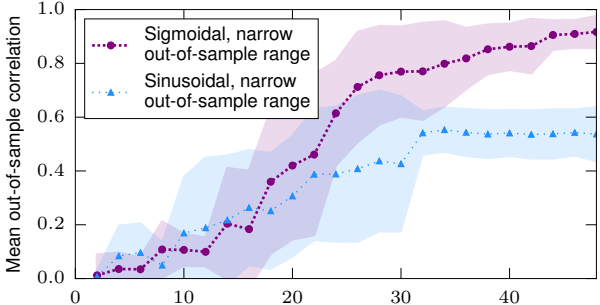

2.4
1.2
0.0
Total phosphorylation
10^3
10^0
10^{-3}
Input V
0
1
2
3
4
5
Input V
Total phosphorylation output
Full model (best fit)
Sigmoidal best fit
S-systems best fit
Full model true parameters

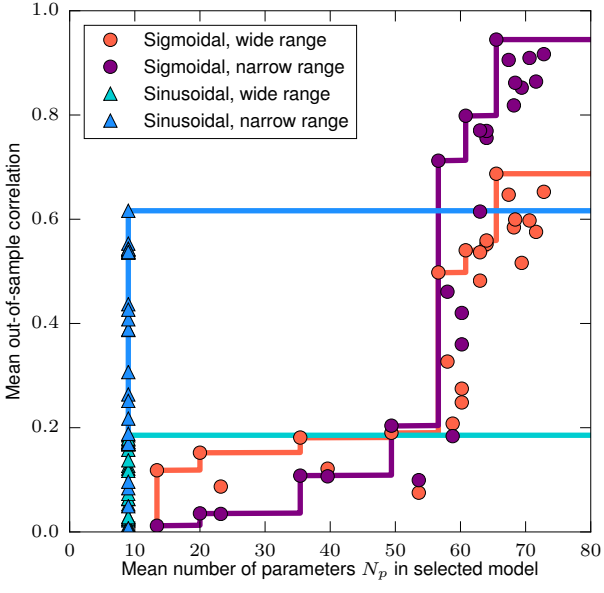

Full model max. likelihood
Sigmoidal adaptive model
Simple model
Out-of-sample mean squared error
10^{-1}
10^{-2}
10^{-3}
2
5
10
20
50
Mean number of parameters N_p in selected model

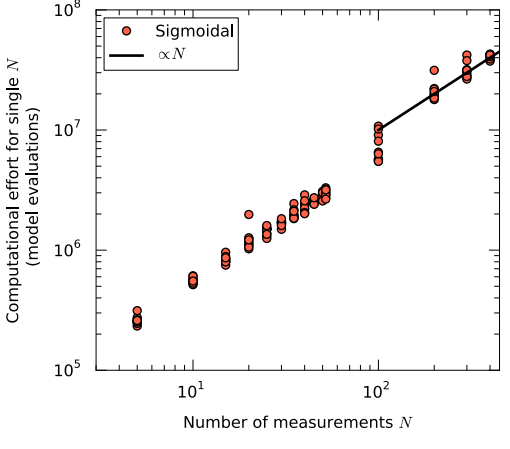

S3
S2
S4
S1
N2
S4ex
A3

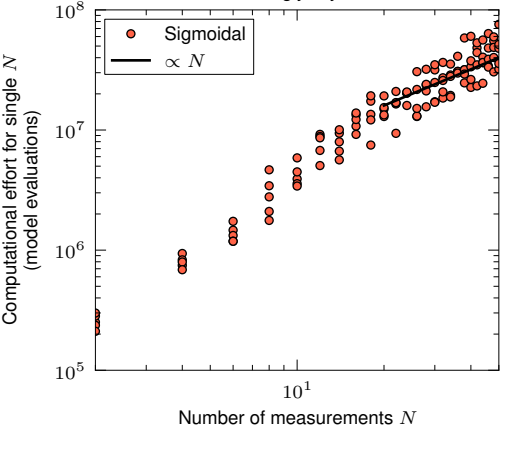

S3
S2
X4
S1
X5
X6

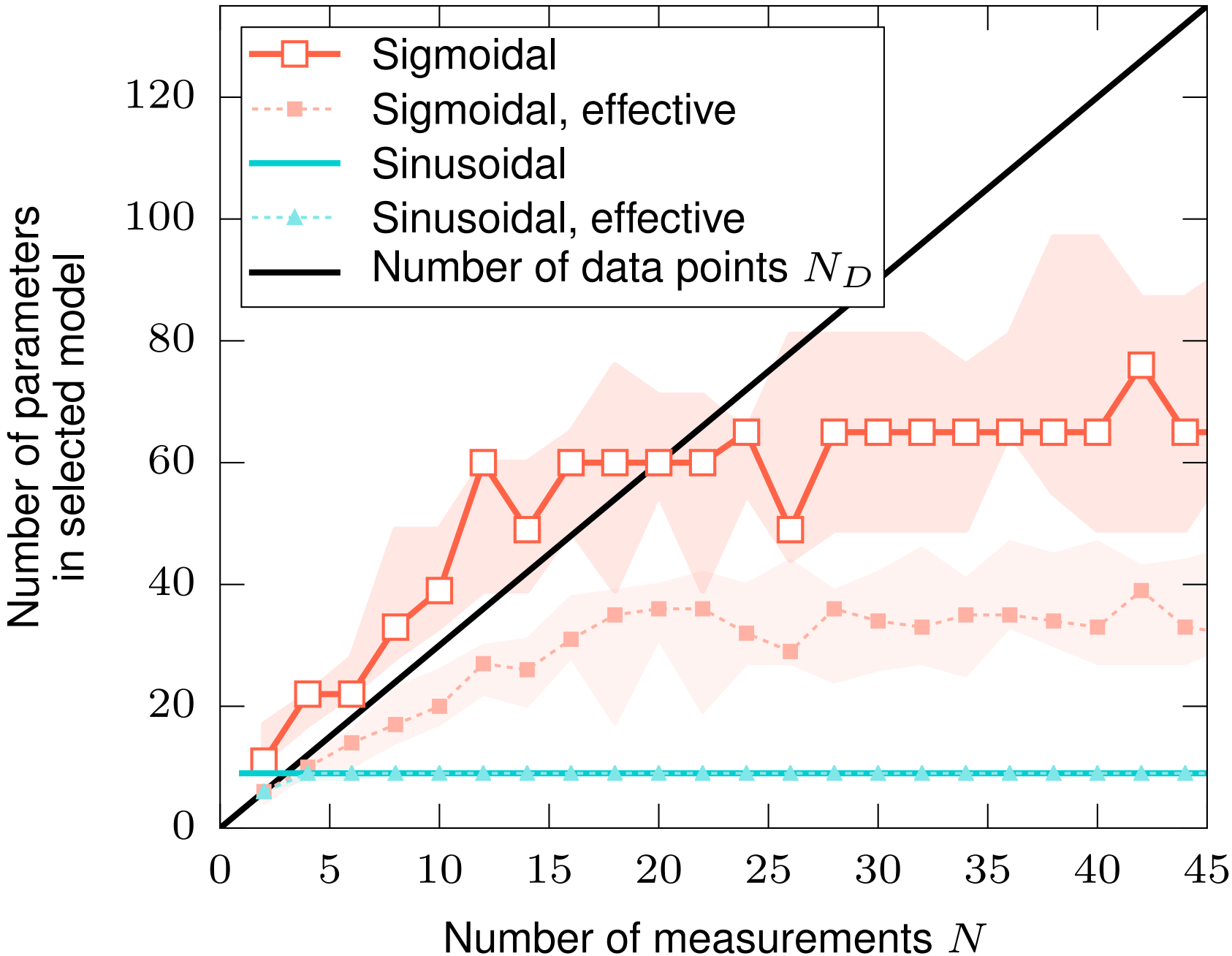

Sigmoidal, wide
out-of-sample range
Sinusoidal, wide
out-of-sample range
Mean out-of-sample correlation
Number of measurements N


Sigmoidal, narrow
out-of-sample range
Sinusoidal, narrow
out-of-sample range
Mean out-of-sample correlation
Number of measurements N


Sigmoidal, wide range
Sigmoidal, narrow range
Sinusoidal, wide range
Sinusoidal, narrow range
Mean out-of-sample correlation
Mean number of parameters N_p in selected model


Multi-site phosphorlyation
Sigmoidal
$\propto N$
Computational effort for single N
(model evaluations)
Number of measurements N
10^5
10^6
10^7
10^8
10^1
10^2


Yeast glycolysis
Sigmoidal
$\propto N$
Computational effort for single N
(model evaluations)
Number of measurements N
10^5
10^6
10^7
10^8
10^1


Sigmoidal
Sigmoidal, effective
Sinusoidal
Sinusoidal, effective
Number of data points N_D
Number of parameters in selected model
Number of measurements N
0
20
40
60
80
100
120
0
5
10
15
20
25
30
35
40
45