# Statistical mechanics of dynamical system identification

Andrei A. Klishin [1,2,3,*] Joseph Bakarji [2,3,4] J. Nathan Kutz [2,5] and Krithika Manohar [2,3,†]
[1]Department of Mechanical Engineering, University of Hawai'i at Mānoa, Honolulu, Hawaii 96814, USA
[2]AI Institute in Dynamic Systems, University of Washington, Seattle, Washington 98195, USA
[3]Department of Mechanical Engineering, University of Washington, Seattle, Washington 98195, USA
[4]Department of Mechanical Engineering and Artificial Intelligence Hub, American University of Beirut, Beirut 1107 2020, Lebanon
[5]Departments of Applied Mathematics and Electrical and Computer Engineering, University of Washington, Seattle, Washington 98195, USA

(Received 8 May 2024; accepted 8 July 2025; published 21 August 2025)

Recovering dynamical equations from observed noisy data is the central challenge of system identification. We develop a statistical mechanics approach to analyze sparse equation discovery algorithms, which typically balance data fit and parsimony via hyperparameter tuning. In this framework, statistical mechanics offers tools to analyze the interplay between complexity and fitness similarly to that of entropy and energy in physical systems. To establish this analogy, we define the hyperparameter optimization procedure as a two-level Bayesian inference problem that separates variable selection from coefficient inference and enables the computation of the posterior parameter distribution in closed form. Our approach provides uncertainty quantification, crucial in the low-data limit that is frequently encountered in real-world applications. A key advantage of employing statistical mechanical concepts, such as free energy and the partition function, is to connect the large-data limit to thermodynamic limit and characterize the sparsity- and noise-induced phase transitions that delineate correct from incorrect identification. We thus provide a method for closed-loop inference, estimating the noise in a given model and checking if the model is tolerant to that noise amount. This perspective of sparse equation discovery is versatile and can be adapted to various other equation discovery algorithms.

DOI: 10.1103/4d98-tdlp

## I. INTRODUCTION

Identifying dynamical models from data represents a critical challenge in a world inundated with emerging time-series data which lack accurate and robust characterizations. This is the central task of modern system identification [1]. Traditional methods of constructing ordinary and partial differential equations (ODEs and PDEs) from the first principles are increasingly limited by our intuitive understanding and the growing complexity of contemporary problems involving high-dimensional data and unfamiliar nonlinearities [2]. Conversely, modern deep learning approaches, while capable of identifying highly nonlinear relationships from time-dependent data [3,4], often struggle with overfitting and lack the interpretability essential for human element in improving model generalization [5,6].

This backdrop sets the stage for modern data-driven methods that search through a large set of hypothetical differential equations to achieve an optimal fit with observational data [7]. Notable examples include sparse identification of nonlinear dynamics (SINDy) [8] and its extensions [9–12], symbolic regression [13,14], Sir Isaac [15], and equation learning [16]. These methods have seen diverse applications from sparse biochemical reaction networks [17] to atmospheric chemistry surrogate modeling, [18], uncertainty quantification [19], active matter [20], and fluid dynamics [21]. The parsimonious form of equations, apart from being directly interpretable by domain experts, also enables the key physical features of the model, namely generalization and extrapolation [22].

Despite their promise, these techniques face the added complexity of balancing parsimony with accuracy through trial-and-error hyperparameter tuning [23]. The effectiveness of these methods is contingent on the amount of available data and the level of noise present. There is currently a lack of understanding regarding these limiting cases and how the various hyperparameters interact with them. The inferred model is naturally a function of the chosen hyperparameters, usually the prior parameters or the coefficients of loss function terms such as sparsity penalty. The two most common strategies are to either fix the hyperparameters at some values that happen to work well for the example dataset [24–26] or "optimize away" the hyperparameters through cross-validation [23,27]. Both of these strategies obscure the inherent trade-off between accuracy and parsimony, which might be useful for an applied practitioner but is contrary to the scientific method.

In this paper, we directly expose the accuracy—parsimony trade-off using *exact* Bayesian inference within a statistical

*Contact author: aklishin@hawaii.edu
†Contact author: kmanohar@uw.edu

*Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.*

mechanical approach to sparse equation discovery. Statistical mechanics has long been used to analyze the average-case behavior of inference problems from classifiers to sparse sensing and network structure [28–33]. Those studies primarily focus on a large number of features that have different collective inference behaviors but are individually unimportant and anonymous; in contrast, here we apply the same ideas to identify a small set of mechanistically interpretable, nonanonymous terms of dynamical equations.

The proposed method, Z-SINDy, incorporates two key advancements: *fast closed-form computations of the full posterior distribution* and *direct $L_0$ sparsity penalty* as opposed to its approximations via $L_1$, iterative reweighing, or sequential thresholding. The key computational advantage lies in separating the accounting of the discrete and continuous degrees of freedom similar to mixed-integer optimization [34] and using a closed-form expression for multivariate Gaussian integrals. Many SINDy-inspired studies in the recent years have taken an *approximate* Bayesian view to provide error bars of the inferred models through methods such as Monte Carlo or Gibbs sampling [24,25,35–39], variational Bayes [37,40–42], or expectation-propagation expectation-maximization [41]. The methods most similar to ours are presented in Refs. [26,43], but those studies do not analyze how the confidence intervals of the inferred model parameters scale with data size or noise magnitude or how the sparsity promotion generates a family of models that trade off complexity and fitness.

While some system identification studies provide guarantees of convergence to the correct model in the low-noise limit [44], the numerical validation experiments typically do not report failure cases where inference fails in the sense of finding qualitatively wrong equation terms [12,45–48]. At the same time, even when uncertainty quantification is provided for a few particular datasets and noise levels, there is no general argument on how both the width of error bars and the systematic error in the coefficients scale with dataset parameters [49]. In other words, when identification fails, we do not know why. In this paper we aim to primarily figure out under what conditions the identified solution would match the ground truth. While computational efficiency was not the primary goal, in many practical cases Z-SINDy computation is much faster than alternative methods.

When Z-SINDy is applied to low-data scenarios, it returns the full posterior distribution over the dynamical models and thus provides uncertainty quantification at a fraction of computational cost of other approaches [24,25,35,50,51]. In high-data scenarios, we show that inference always condenses to a definite, though not necessarily correct model, and switches abruptly between models. Accordingly, as either the noise in the dataset or the sampling period are increased, we observe a detectability phase transition from correct to incorrect dynamical model, directly informing the trade-off between model fidelity and sparsity. The uncertainty quantification and noise tolerance metrics enable closed-loop inference, that is finding a model and checking if that model could be found given the amount of noise. The statistical mechanics analysis can be further integrated with other SINDy advancements or applied to other cases of sparse inference.

## II. STATISTICAL MECHANICS FOR SPARSE INFERENCE

### A. Background

System identification begins with an observed trajectory $\vec{x}(t)$ of a $d$-dimensional dynamical system. The trajectory covers the length of time $T$ sampled with period $\Delta t$, resulting in $n = T/\Delta t$ data points. For synthetic trajectory data, we presume that the integration time step is much smaller than $\Delta t$, making the integration error negligible. Trajectory measurement incurs uncorrelated additive Gaussian noise of magnitude $\eta$ in each dimension of dynamics. From the trajectory we compute the empirical derivative $\dot{\vec{x}}$ with a second-order centered finite difference method to simplify the interpretation.

The goal of sparse equation discovery is to extract a dynamical equation $\dot{\vec{x}} = f(\vec{x})$ from the observed trajectories of $\vec{x}$, where $f(\cdot)$ is a sparse analytical expression. When $f(\cdot)$ is given by a library of candidate nonlinear functions, we seek an equation of the form:

$$\dot{x}_l \approx \sum_{i=1}^{N} \Theta_i(\vec{x})^T \, \Xi_{il}, \tag{1}$$

where the index $l$ enumerates the dynamical variables, the left-hand side is the empirical derivative and right-hand side is a linear combination of $N$ nonlinear functions of dynamical variables $\Theta_i(\vec{x})$ (i.e., the library) with a matrix of coefficients $\Xi_{il}$. SINDy seeks to optimize for the *sparsest* matrix of coefficients $\Xi_{il}$ that minimizes the residual of (1). This optimization problem involving two competing objectives (sparsity and fitness) leads to a challenging nonuniqueness in the resulting models. In practice, the problem is typically solved by optimizing a loss function that combines linear regression with a sparsity penalty. While sparsity is directly measured by the $\|\Xi\|_0$ pseudonorm on the model coefficients, it is usually approximated by either the $\|\Xi\|_1$ norm or sequential thresholded least squares [8,52].

Where optimization-based SINDy provides a point estimate of the coefficients, Bayesian approaches aim to extract the maximum amount of information from the data, while providing an uncertainty quantification of the resulting equation. In the Bayesian setting, sparsity can be promoted by the choice of a sparsifying prior, such as Laplace, spike-and-slab, or horseshoe [24,35,53]. These priors aim to concentrate the posterior probability near $\Xi_{il} = 0$, while remaining differentiable to enable Monte Carlo sampling.

The interplay of sparsity with large numbers of variables and data points in a probabilistic setting attracted significant attention from the statistical mechanics community, particularly the theory of disordered systems [33]. By using the so-called replica trick, researchers averaged over random data matrices to obtain the average behavior of different classes of inference problems, revealing multiple detectability and algorithmic phase transitions [29,31,32,54]. Our approach here is different in two key aspects: First, we focus on identifying not a finite *fraction* of relevant variables but a finite *number*, thus attaching more interpretation to each term; second, instead of averaging over generic random Gaussian data matrices, we work with trajectory data, including the sampling period and numerical differentiation effects.

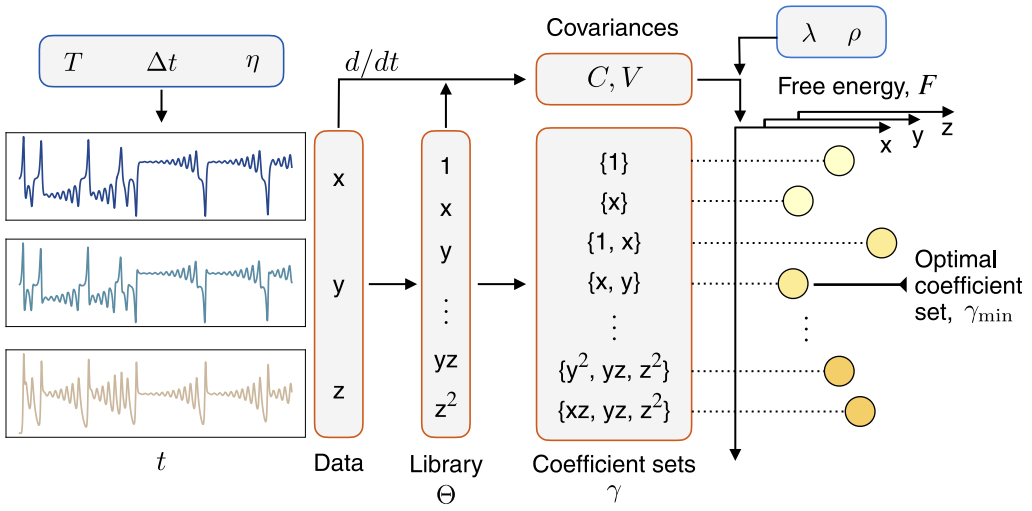



FIG. 1. Schematic representation of the questions Z-SINDy aims to answer. A dataset consists of a trajectory of time length $T$ sampled with time step $\Delta t$ and measurement noise of magnitude $\eta$ (left blue box). From the data, a library of nonlinear functions $\Theta_i(\vec{x})$ is computed. The library functions are combined with empirical derivatives of the trajectory to compute the covariance matrices $C$ and vectors $\vec{V}$. The covariances along with the algorithm hyperparameters of sparsity $\lambda$ and resolution $\rho$ (right blue box) are used to compute the free energies $F_{\gamma_l}^{(l)}$ that quantify the balance between goodness of fit and number of fitting solutions. The best fit coefficient set $\gamma_{\min}$ is the one that corresponds to the lowest free energy. In this paper we study how the probability of choosing the correct coefficient set depends on the parameters in blue boxes.

### B. Z-SINDy

In the present paper we separate the Bayesian inference problem of coefficients $\Xi$ into two layers: The discrete layer describes which set of coefficients $\gamma$ is active (nonzero), e.g., $\gamma = \varnothing$, {x,yz}, {1,y,z,xy,xz}, while the continuous layer describes the values of coefficients $\Xi_\gamma$ in the active set. The inference framework is schematically represented in Fig. 1.

In particular, this new setting does not require the prior to be differentiable, allowing us to use a much simpler Bernoulli-Gaussian functional form [32]:

$$p(\Xi) = \prod_i \frac{1}{1+e^{-\Lambda}}(\delta(\Xi_i) + w(\Xi_i)e^{-\Lambda}), \tag{2}$$

$$= \frac{1}{(1+e^{-\Lambda})^N} \sum_\gamma \left(\prod_{i\notin\gamma} \delta(\Xi_i)\right)\left(\prod_{i\in\gamma} w(\Xi_i)\right) e^{-\Lambda|\gamma|}, \tag{3}$$

where we omitted the index $l$ for simplicity of notation, since the inference is independent in each dimension with a shared set of library terms $\Theta_i(\vec{x})$. Within the Bernoulli-Gaussian prior, the hyperparameter $\Lambda$ regulates sparsity; $\delta(\cdot)$ is Dirac delta function that can be thought of as an infinitely narrow Gaussian; and $w(\cdot)$ is a prior function that quantifies the parameter uncertainty in the absence of data, which we take to be in the limit of an infinitely wide Gaussian (see Supplemental Material (SM) [55] for additional discussion of the prior).

The prior form of (2) is symmetric under the permutation of indices $i$, highlighting that the inference procedure does not have a preference of any variable combination before seeing any data. The symmetry is made even more explicit by multiplying out the brackets to get to (3), which is given as a combinatorial sum over the *coefficient sets* $\gamma$. For $N$ library terms there are $2^N$ possible discrete coefficient sets that form the power set of library indices $\{\gamma_l\} = \wp(\{i\})$. The prior probability of any coefficient set depends only on its size $|\gamma|$ but not its identity, and increasing the value of $\Lambda$ shifts progressively more probability weight from larger to smaller sets, while maintaining unbroken symmetry under index permutation.

Along with the prior, we define the forward model, i.e., the probability of observing the data given the coefficients (likelihood):

$$p(x|\Xi) \propto \exp\left(-\frac{1}{2\rho^2}\sum_t (\dot{x} - \vec{\Theta}^T\vec{\Xi})^2\right), \tag{4}$$

where we suppressed the time indexing on the empirical derivative and the library terms. The resolution hyperparameter $\rho$ regulates how closely the nonlinear function library aims to approximate the empirical dynamics derivative.

Using the forward model and the prior, we derive the posterior via Bayes rule:

$$p(\Xi|x) = \frac{1}{\mathcal{Z}} \exp\left(-\frac{1}{2\rho^2}\sum_t (\dot{x} - \vec{\Theta}^T\vec{\Xi})^2 - \Lambda\|\vec{\Xi}\|_0\right), \tag{5}$$

where the expression inside $\exp(\cdot)$ is the $L_0$ regularized SINDy functional [8]. In statistical mechanics, this would be akin to Boltzmann's distribution with $\rho^2$ playing the role of temperature, $\Lambda$ playing the role of chemical potential (cost of adding new particles to the system), and $\mathcal{Z}$ the partition function (normalization) of the distribution. The mode of this distribution, or the maximum *a posteriori* (MAP) estimate of the coefficients, would be equivalent to the original SINDy problem statement, but the $L_0$ regularization is challenging to optimize in practice.

Instead of performing optimization, we disentangle the mixture of continuous and discrete degrees of freedom in (5) by using the factorized form of the prior (3) to rewrite the posterior into a hierarchical form as a choice of active

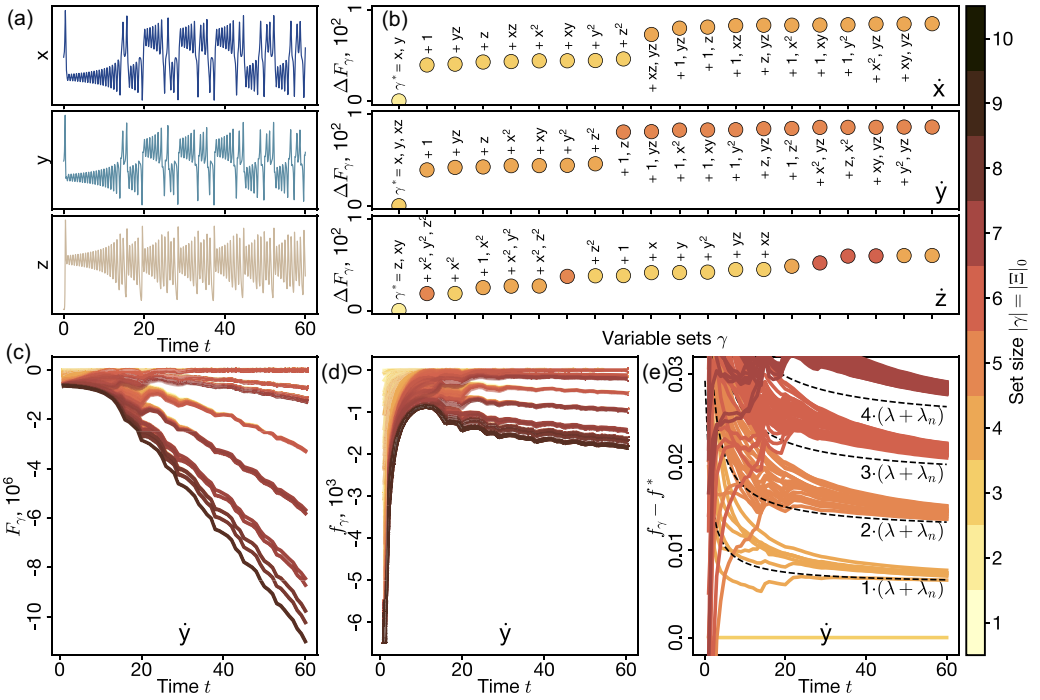



FIG. 2. Free-energy trends in Z-SINDy inference. (a) Sample trajectories of the Lorenz attractor sampled with time sampling period of $\Delta t = 0.01$ and no noise. (b) Free energies of different coefficient sets for each dimension of dynamics relative to the best fit set $\gamma^*$ computed at $\rho = 1.0$. Other coefficient sets are labeled by the extra library terms with respect to $\gamma^*$. (c) Free energy $F_\gamma$ of each of the $2^N = 1024$ coefficient sets over time. (d) Free energy *per data point* $f_\gamma$ of each coefficient set. (e) Relative free energy per data point $f_\gamma - f^*$ of each coefficient set, stratified by the sparsity penalty $\lambda = 0.006$ and the natural sparsity penalty $\lambda_n$. The color of markers and curves corresponds to the number of coefficients in each set $|\gamma|$.

coefficient set followed by a choice of coefficient values. In this factorized form the posterior takes form of a linear mixture of multivariate Gaussians:

$$p(\Xi|x) = \frac{\sum_\gamma p(\vec{\Xi}_\gamma|\gamma, x)\mathcal{Z}_\gamma}{\sum_\gamma \mathcal{Z}_\gamma}, \tag{6}$$

$$\mathcal{Z}_\gamma \equiv e^{-F_\gamma} = e^{-F'_\gamma} e^{-\Lambda|\gamma|}, \tag{7}$$

$$F'_\gamma = -\ln \mathcal{Z}_0 - \frac{|\gamma|}{2}\ln(2\pi\rho^2) + \frac{1}{2}\ln\det C_\gamma - \frac{1}{2\rho^2}\vec{V}^T_{l,\gamma} C^{-1}_\gamma \vec{V}_{l,\gamma l}, \tag{8}$$

$$F_\gamma \equiv F'_\gamma + \Lambda|\gamma| = F'_\gamma + \lambda\cdot n|\gamma|, \tag{9}$$

where the *statistical weights* $\mathcal{Z}_\gamma$ quantify the relative importance of each coefficient set (evidence) and are thus the central objects of the method, inspiring the name Z-SINDy. Since the statistical weights vary over many orders of magnitude, it is more convenient to represent them on logarithmic scale as *free energies* $F_\gamma$. In Bayesian terminology, $\mathcal{Z}_\gamma$ is the evidence of each coefficient set and $F_\gamma$ is the negative log-evidence [56]. The free energy is computed directly from subsets of the precomputed empirical correlation matrix $C$ and vectors $\vec{V}_l$ of library functions with each other and with empirical derivatives (see Fig. 1 for schematic illustration and SM [55] for derivation).

Much like in statistical physics the free energy quantifies the balance of energy and entropy of a coarse-grained state [57], here the free energy represents the goodness of fit of the empirical derivatives $\dot{x}_l$ with respect to all possible continuous values of coefficients within the same active set. The derivation and the final functional form of (8) are closely similar to the Akaike and Bayesian information criteria (AIC and BIC) that combines the likelihood of a model with a penalty based on the number of parameters [58,59]. The free-energy expression (9) selects for sparse solutions in two ways: While the AIC- and BIC-like penalty we term "natural sparsity" scales with the trajectory length $n$ as $\mathcal{O}(1) + \mathcal{O}(\ln n)$ (see SM [55] for derivation) [60], the prior driven penalty scales as $\Lambda = \lambda \cdot n = \mathcal{O}(n)$. The scaling of the sparsity penalty with data size has been previously discussed for $L_1$ sparsity promotion [61]. The following sections use the free-energy computation to compute the full posterior and its marginals to analyze model inference in different regimes.

## III. RESULTS

We choose to focus our in-depth analysis in the main text on one three-dimensional chaotic dynamical system, the classic Lorenz attractor first integrated numerically by Ellen Fetter [62,63] [Fig. 2(a)], since it is a fairly typical chaotic system with polynomial nonlinearity [64,65], and the performance of SINDy-family algorithms does not strongly correlate with different features of chaotic systems [66]. Polynomial representations cover many dynamical systems with a wide range of dynamical behaviors [64,65]. This is in large part due to the manifestation of dominant balance physics which is often

expressed as leading polynomial terms in a Taylor series expansion of more complex behavior. Dominant balance physics is commonly used to express the mathematical similarity between highly disparate physical systems [67]. However, there are certainly many empirical systems are not well described by polynomial forms [7]. While the Z-SINDy framework, like most other SINDy forms, can handle arbitrary nonlinear library terms, we restrict this study to polynomial forms. We thus consider the library terms in form of all monomials in variables $x, y, z$ up to second order, resulting in $N = 10$ library terms and $2^N = 1024$ possible coefficient sets. The computational infrastructure of evaluating the library terms is based on the PySINDy package [68]. Throughout this chapter we study different aspects of the Z-SINDy fit and introduce multiple graphical formats of analysis, focusing on the Lorenz system. We repeated a similar analysis flow for six other nonlinear systems, including limit cycles, decaying oscillations, a higher-order nonlinear chaotic system, and a hyper chaotic system (see SM [55]) [26,69–71].

### A. Free-energy trends

In order to establish intuition for free-energy scaling, we use the expression (8) to compute the free energies of all variable sets for the Lorenz model. For a constant and moderate trajectory length $T$ and sparsity penalty $\lambda$ the free energies of different variable sets form a hierarchy shown in Fig. 2(b) for each dimension of dynamics: The correct variable set $\gamma^*$ has the lowest free energy, while the next several sets with higher free energies all have extra terms. If the trajectory data are only considered up to a variable upper limit of time $t$, the free energies of each variable set have asymptotically linear trajectories of different slopes [Fig. 2(c)]. In order to compare the slopes, we compute the *intensive* free energy per data point $f_\gamma = F_\gamma/n$ that is asymptotically constant for each variable set [Fig. 2(d)]. We further disentangle the different sets by computing the intensive free energy relative to its lowest value $\Delta f_\gamma = f_\gamma - f^*$ [Fig. 2(e)]. By construction the relative free energy of the correct variable set is zero, and free energies of other sets are asymptotically stratified by the constant intensive sparsity penalty $\lambda$. While the sparsity penalty is an externally chosen hyperparameter of Z-SINDy, at low sample sizes the Bayesian inference procedure itself introduces an additional "natural" sparsity $\lambda_n = \ln(n/2\pi\rho^2)/2n$ that enhances the selection for sparse coefficient sets similar to the AIC and BIC (see the SM [55] for derivation).

The asymptotically constant free energy per data point in this inference problem is similar to the thermodynamic free energy of interacting particle systems. For particle systems, nonlinear scaling of free energy with system size usually implies either long-range particle interactions or strong boundary effects. Indeed, the inference free energy has a nonlinear scaling at early times [$t < 10$ on Fig. 2(c)], when the Lorenz dynamical system has only explored one lobe of the attractor and has not yet demonstrated the switching behavior. At longer times, the chaotic dynamics forget the initial condition and the effective sample size scales linearly with the amount of data, leading to a condensation of inference, as we explore in the following section.

### B. Inference condensation

In order to connect the scaling of free energy to the outcomes of the inference procedure, we consider the limiting case of the posterior distribution (6). The probability of selecting the lowest free–energy set is driven by the free-energy gap between it and the next set:

$$p(\gamma^*) = \frac{1}{1 + \sum_{\gamma \neq \gamma^*} e^{-(F_\gamma - F_{\gamma^*})}} \approx 1 - e^{-n\Delta f}, \tag{10}$$

where $\Delta f$ is the asymptotic difference of free energy per data point between the best and the second-best fitting coefficient sets. This expression implies that the probability of selecting *any other* variable set decays exponentially with trajectory length across a wide range of sampling frequencies [Fig. 3(a)]. While the statistical weights $\mathcal{Z}_\gamma$ can get exponentially large or small, risking numerical overflow or underflow problems, the values of free energies do not face that problem. The exponential suppression of suboptimal coefficient sets implies that it is sufficient to look for the lowest free-energy coefficient set at given dataset parameters and inference hyperparameters.

Given the condensation of the discrete part of inference, what happens to the continuous part? Per (6), the Gaussian mixture reduces to a single multivariate Gaussian distribution with the covariance and mean parameters driven by the empirical correlations $C_{\gamma^*}$ and $\vec{V}_{l,\gamma^*}$. The posterior covariance matrix is given by $\Sigma_{\gamma^*} = \rho^2 C_{\gamma^*}^{-1}$, scaling with the resolution parameter $\rho$ but decaying with increasing trajectory length, which can be combined into an effective timescale $n_e = t/\Delta t\,\rho^2$. The standard deviations of the posterior along each coefficient direction have different magnitudes but identical scaling of $n_e^{-1/2}$ as in the central limit theorem [Fig. 3(b)].

The posterior Gaussian mean is given by $\vec{\mu} = C_\gamma^{-1}\vec{V}_{l,\gamma}$, which quickly converges to a constant value, which is not necessarily equal to the ground truth $\vec{\mu}_{\text{true}}$. Since the linear regression against the nonlinear library terms aims to explain the empirical derivative, it inherits the systematic error of the numerical differentiation procedure, well known in the studies of numerical integration [72] but rarely highlighted in system identification. For the second-order finite difference derivative employed here the systematic error scales as $\mathcal{O}(\Delta t^2)$, and this scaling propagates to the error of the mean [Fig. 3(c)]. For large-enough trajectory length the standard deviation becomes smaller than error of the mean $\sigma_i < |\mu_i - \mu_{\text{true}}|$, and thus the inference converges to a small region that does not include the ground-truth values.

The scaling reported here is due to collecting an ever-longer single trajectory. An alternative way to collect more data is to stitch together many short trajectories [47,73]. Once the numerical differentiation of each trajectory is performed, the data take the form of pairs $(\dot{x}, \vec{\Theta})$, the order of which is not important. The covariances $\vec{V}, C$ are computed by summation over all available data points, whether they originally came from the same continuous trajectory or different ones. We analyze the inference condensation for many short trajectories of just 10 time steps each and see qualitatively similar results, though the standard deviations $\sigma_i$ get somewhat tighter if data are collected from transient trajectories that have not yet

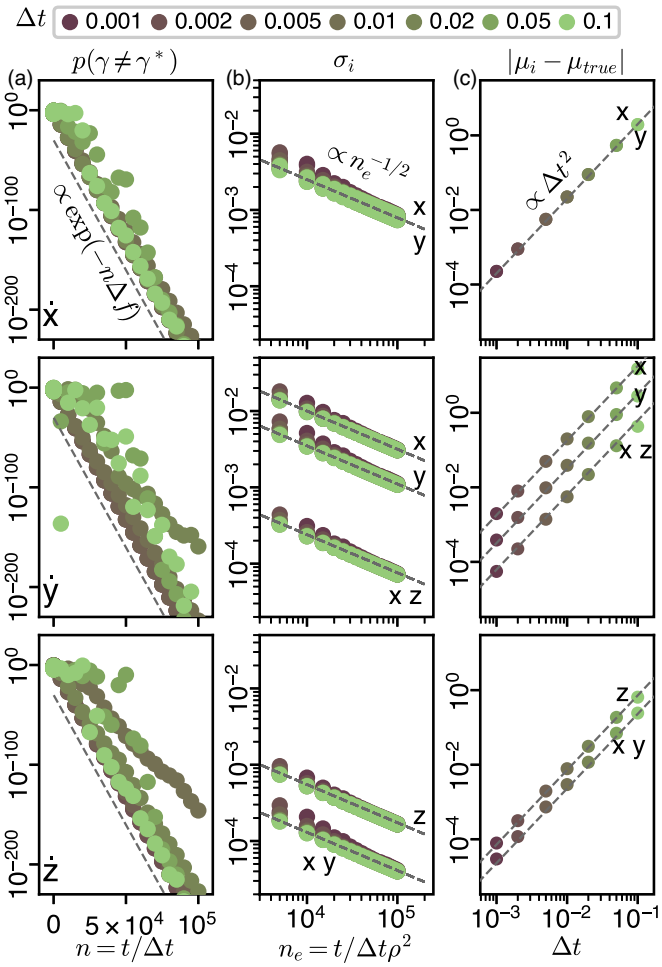



FIG. 3. Condensation of Z-SINDy inference for trajectory datasets of variable runtime $t$, sampling time period $\Delta t$, resolution parameter $\rho \in \{0.1, 1.0, 10\}$. (a) The probability of choosing any variable set other than the one with lowest free energy decays exponentially with trajectory length. (b) The posterior standard deviation of each inferred coefficient decays as inverse square root of the trajectory length. (c) The absolute error of the posterior mean coefficient is limited by the systematic error of the finite difference derivative of the trajectory. Trend lines guide the eye to illustrate the functional form of the scalings. The line labels in (b) and (c) indicate which coefficient the standard deviation and the mean correspond to. Rows correspond to the three dimensions of dynamics. The standard deviation and error of the mean for the two variables explaining $\dot{x}$ coincide.

reached the Lorenz attractor (see the SM [55] for figure and details).

Increasing the trajectory length leads to a condensation of inference to a particular set of coefficients with a narrow range of values. However, if the amount of data asymptotically does not distinguish between inferred models, then what does? In thermodynamic systems, free energy per particle is typically a function of external thermodynamic variables, such as temperature, magnetic field, or chemical potential. Small changes of the external parameter can shift the global free energy to a different state, leading to a thermodynamic phase transition. In a similar way, small changes in the inference hyperparameter $\lambda$ or noise in the data $\eta$ can lead to *abrupt* changes in the inferred model, as we show in the following sections.

### C. Sparsity transitions

The goal of SINDy-family approaches is to balance the data fit with the parsimony of the inferred models, operationalized by sparsity. Instead of prescribing a particular number of equation terms, the algorithm is supposed to find it adaptively, but how exactly does the sparsity penalty parameter $\lambda$ lead to sparse solutions?

We have established in the previous section that with enough data, Z-SINDy would always select the model with lowest free energy. Given a constant dataset, the penalized free energies are linear functions of the penalty [Eq. (9)], with the intercept given by data fit at zero penalty and the slope given by the number of terms. Graphically, the ensemble of all the linear functions looks like a fan plot with all-integer slopes $1, 2, 3, \ldots$ [Fig. 4(a)]. As the external sparsity penalty $\lambda$ increases, the lowest free-energy line changes in a series of abrupt transitions from lower-intercept higher-slope to higher-intercept lower-slope (shown in vertical dashed lines), similarly to the plots produced by least angle regression (LARS) [74]. However, because of the natural sparsity effect, the inference selects for sparse solutions even at $\lambda = 0$. In order to include both external and natural sparsity, we therefore vary both $\lambda$ and the trajectory length $t$.

At moderate to high sparsity penalty $\lambda$ the selected variable sets are practically independent of trajectory length [Fig. 4(b)], but at low $\lambda$ the selection changes in two ways. Very short trajectories do not explore the entirety of available phase space: For the Lorenz system, the trajectory stays within a single lobe of the attractor for $t < 15$ [Fig. 2(a)], and thus Z-SINDy identifies less-sparse models [bottom-left corner of Fig. 4(b) panels]. For moderate trajectory length the correct coefficient set is recovered for a wide range of $\lambda$ because of the natural sparsity $\lambda_n$. For long trajectories the natural sparsity disappears, leading to identification of less sparse sets [top-left corner of Fig. 4(b) panels]. We conclude that Z-SINDy correctly identifies the sparse set of coefficients within a window of several orders of magnitude of sparsity penalty $\lambda$, but the boundaries of the region are quite abrupt. Aggregation of many short trajectories results in a similar sequence of inferred models, with larger amount of jitter for low $\lambda$ (see the SM [55] for figure and discussion).

### D. Noise transition

Along with sparsity, another important limitation to the performance of SINDy algorithms is the noise in the data. The robustness of SINDy is usually measured by how fast the error in the inferred coefficients grows with noise magnitude and thus how much noise can the inference tolerate. While the denoising approaches [75,76] and the weak form SINDy [11] improve noise tolerance significantly, they do not explain how the noise-induced breakdown happens, whether collecting a longer trajectory helps, and why denoising improves performance so much.

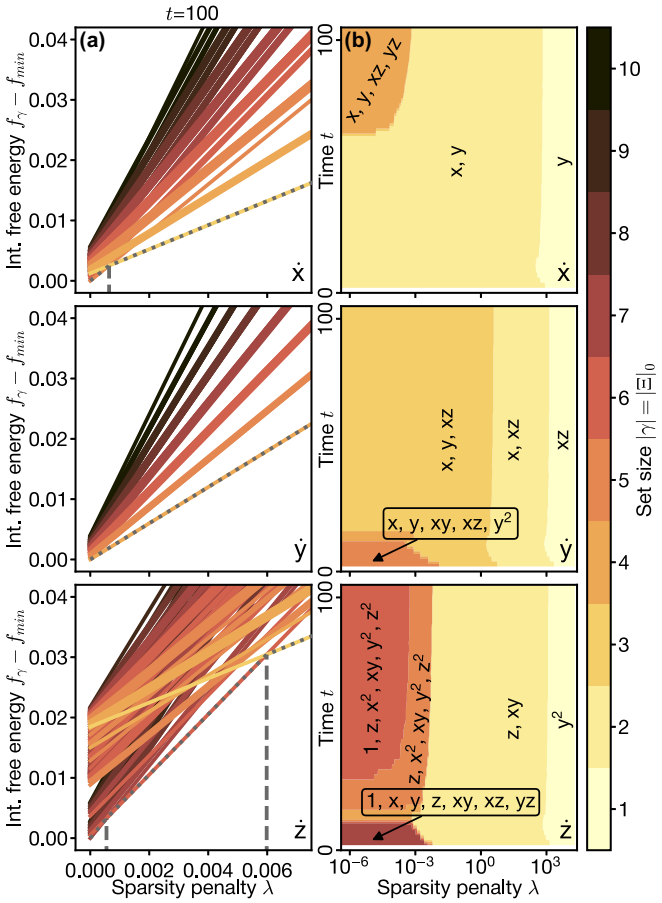



FIG. 4. Sparsity penalty selects sparse solutions. (a) For a fixed trajectory of length $t = 100$, the penalized intensive free energies of each coefficient set are linear functions of the sparsity penalty $\lambda$ with the slope given by the number of terms and the intercept by the goodness of fit. At any value of $\lambda$, the posterior condenses on the lowest free-energy set (piecewise linear dotted lines), switching abruptly between solutions of different sparsity (vertical dashed lines). (b) As the sparsity penalty varies over many orders of magnitude, the best fit coefficient set changes abruptly several times from large to sparse, with a finite range of $\lambda$ recovering the correct set. The color of lines and backgrounds corresponds to the number of coefficient in each set.

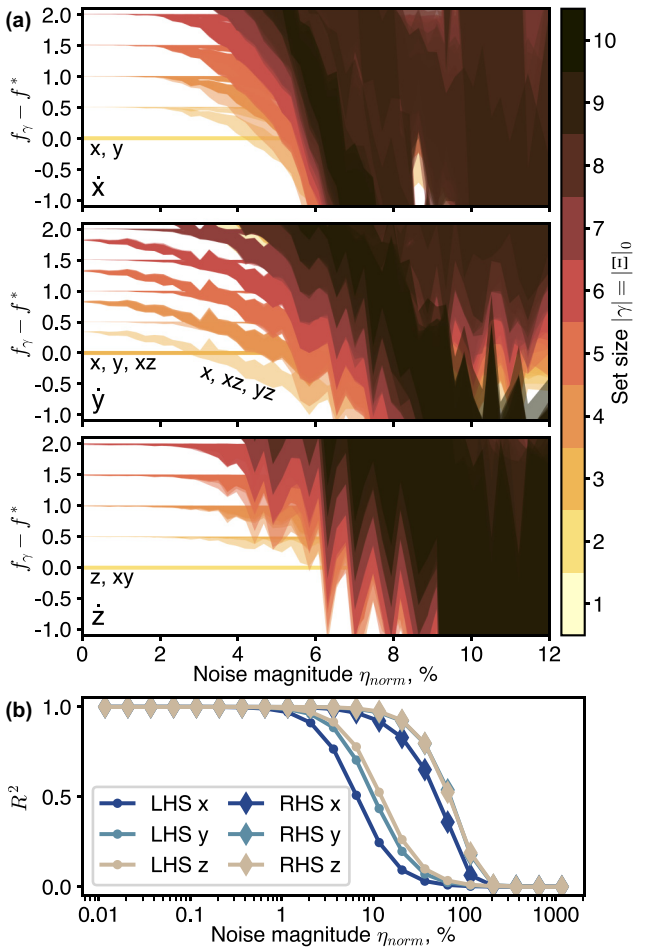



FIG. 5. Additive noise prevents the identification of the correct coefficient set. (a) Each panel shows the free energies per data point with respect to the correct set for a trajectory of length $t = 60$ at sampling period $\Delta t = 0.01$. For each coefficient set, the shaded region shows the range of free energies across 10 realizations of random noise of given relative magnitude $\eta/\bar{\sigma}_{\text{data}}$, stratified by sparsity penalty $\lambda = 0.5$. The color of the shaded regions corresponds to the number of coefficients in each set. (b) Coefficient of determination $R^2$ of the noisy left- and right-hand sides of the equation of motion with the correct right-hand side for each dimension of dynamics, averaged over five noise realizations.

We seek to explain the noise-induced breakdown in free-energy terms. The free energy of any variable set $\gamma$ depends not only on the trajectory length $t$, sparsity penalty $\lambda$, and noise *magnitude* $\eta$ but also on the noise *realization*. Within each realization, we compute the deterministic free energy per data point relative to the correct set $\gamma^*$ in each dimension and then collect free-energy statistics across multiple noise realizations. The relevant free-energy statistic is not its mean but its range of fluctuations, since the inference condenses to the coefficient set with the lowest free energy.

The noise-induced transition graphically looks like an overlap between the horizontal line of the correct set and the free-energy range of one of the competing sets [shaded regions in Fig. 5(a)]. We normalize the noise magnitude by the standard deviation of the original trajectory averaged across the dimensions $\eta_{\text{norm}} = \eta/\sigma_{\text{data}}$. At low noise magnitude $\eta_{\text{norm}}$, the correct coefficient set $\gamma_l^*$ has the lowest free energy, clearly stratified from all other ones by the sparsity penalty, and is thus selected in inference. As noise magnitude increases, at about $\eta_{\text{norm}} \approx 4\%$ the model for $\dot{y}$ changes as the set $\gamma^* = \{\text{x, y, xz}\}$ first overlaps with the range of free energies of $\gamma = \{\text{x, xz, yz}\}$ and then lies entirely above the range. Within the free-energy range overlap, the inference condenses to a single coefficient set that depends on the *realization* of the random noise, and thus the inference is unstable. At higher levels of noise above $\eta_{\text{norm}} \approx 6\%$ the horizontal line of the correct set lies fully above multiple overlapping free-energy ranges, and thus the inference procedure would confidently select a coefficient set from many possible alternatives, none of which are correct. This inference scenario is qualitatively similar to the detection of weak communities in complex networks, where the single eigenvalue that carries community

information gets buried within a continuous band of random eigenvalues [30].

Can this inference collapse be avoided with larger amounts of data? As trajectory length $t$ increases, the width of the free-energy range shrinks proportional to $t^{-1/2}$, thus reducing the range of noise magnitudes $\eta$ where the inference outcome is realization dependent (see the SM [55] for discussion). However, since the mean free energy per data point converges to a constant $\eta$-dependent value, the takeover of the correct coefficient set $\gamma^*$ by one or more competing ones is inevitable.

What part of the data-processing pipeline drives the inference collapse? SINDy approaches aim to balance the noisy versions of the left- and right-hand sides of (1) $\mathrm{LHS}_{\mathrm{noisy}}$ and $\mathrm{RHS}_{\mathrm{noisy}}$ (LHS is derivatives and RHS is library terms) *approximately*, but how good is that approximation? We can perform that comparison since for a synthetic dataset we have access to both clean and noisy trajectories. The clean trajectory is an exact solution to (1) up to integration error, for which the two sides of the equation match $\mathrm{LHS}_{\mathrm{true}} = \mathrm{RHS}_{\mathrm{true}}$ and can be computed exactly by using the ground-truth coefficients $\Xi_{il,\mathrm{true}}$. We thus have three time series for each dimension, the linear correlation between which is easily measured by the coefficient of determination $R^2$.

The $R^2$ smoothly decrease from 1 at low noise to 0 at high noise, where neither LHS nor RHS of the noisy equation carry any resemblance to the truth [Fig. 5(b)]. While there is a slight variation of the decay point of curves between dimensions, LHS decays at almost an order of magnitude lower noise level and is thus primarily responsible for SINDy breakdown. Since the core of SINDy is linear regression, the curves such as Fig. 5(b) can be plotted for any denoising numerical derivative method and sampling time period and used as a diagnostic method to identify the limiting factor of performance and thus the most promising algorithmic improvement.

The functional form of the LHS curves in Fig. 5(b) can be estimated as follows. Let $\sigma^2_{l,\mathrm{data}}$ be the variance of the true derivative in a given dimension of dynamics and $\sigma^2_{\mathrm{diff}}$ the variance of measurement noise propagated through the differentiation procedure. Since SINDy cannot fit noise, the most variance that can be explained by a model is $\sigma^2_{l,\mathrm{data}}$. $R^2$ then expresses the fraction of variance explained and $R^2 \approx \sigma^2_{l,\mathrm{data}}/(\sigma^2_{l,\mathrm{data}} + \sigma^2_{\mathrm{diff}}) = 1/(1 + \sigma^2_{\mathrm{diff}}/\sigma^2_{l,\mathrm{data}})$, where the fraction in the denominator expresses the noise-to-signal ratio in a given dimension of dynamics. This expression would produce sigmoidal curves with different inflection point akin to those visible in Fig. 5(b). Similar analysis can be repeated for the RHS side, but the measurement noise needs to be propagated through the nonlinear library terms rather than the differentiation procedure.

### E. Uncertainty quantification

The analysis of the noisy LHS and RHS leads to the uncertainty quantification of the inferred coefficients in the limit of low data. The posterior distributions of all coefficients are Gaussian with parameters derived from the correlations $C, \vec{V}_l$ but require knowing the resolution parameter $\rho$. The resolution parameter describes the distribution of the the residual between the noisy LHS and RHS of the dynamical equation and can be estimated from the empirical distribution by using the MAP values of the coefficients $\vec{\mu}$:

$$\rho_{\mathrm{est}} = \sqrt{\frac{1}{n \cdot d} \sum_{t,l} (\mathrm{LHS}_{l,\mathrm{noisy}} - \mathrm{RHS}_{l,\mathrm{noisy}})^2}. \tag{11}$$

While for each dimension the residual itself $\mathrm{LHS}_{\mathrm{noisy}} - \mathrm{RHS}_{\mathrm{noisy}}$ is expected to have a Gaussian distribution, we can check the shape of the distribution empirically [Figs. 6(a) and 6(b)]. At low noise level the distribution is different for each dimension and has a complex multimodal shape with long tails [Fig. 6(a)], while at moderate noise the distributions for all three dimensions have a consistent Gaussian shape [Fig. 6(b)]. As the amount of noise on the trajectory varies over several orders of magnitude, the estimated resolution parameter $\rho_{\mathrm{est}}$ switches from a flat value driven by the systematic error of the finite sampling period to the noise-driven value $\rho = \eta/\Delta t\sqrt{2}$ [Fig. 6(c); see the SM [55] for derivation].

The resolution parameter is the last missing piece in computing the error bar on the inferred coefficients and comparing them to the ground truth. At low noise the error bar is incredibly small, so that the systematic error in the coefficients is immediately visible for all trajectory lengths [Fig. 6(d)]. At moderate noise the error bar increases and overlaps with the ground truth so that the systematic error is not visible at short trajectory length. As the trajectory length increases, the error bar shrinks proportional to $t^{-1/2}$ as discussed before, resulting in the same value of the statistically significant systematic error.

### F. Inference phase diagram

Having characterized the inference breakdown with sparsity and noise separately, we can now answer the questions about the joint effect of dataset parameters and inference hyperparameters on the viability of system identification. We have shown that inference rapidly condenses to a single coefficient set with growing trajectory length $t$ that is determined by the lowest free energy *per data point*. Moderate values of $t$ incur the natural sparsity effect, but it vanishes for $t \to \infty$, so that the selected coefficient set becomes independent of data quantity. The value of the resolution hyperparameter $\rho$ modulates the inference condensation and affects the size of the confidence interval for the coefficients. This value can be chosen to match the statistics of the residual $\dot{x}_l - \vec{\Theta}^T \cdot \vec{\Xi}_l$ and provide accurate uncertainty quantification.

The remaining parameters interact in a more complex way as shown in the phase diagrams of Fig. 7. The noise magnitude $\eta_{\mathrm{norm}}$ and sampling period $\Delta t$ impose separate, orthogonal limitations [Fig. 7(a)]. At large sampling period, the samples can no longer resolve the smallest timescales of system dynamics and result in a large *systematic* error in the empirical derivative. At the same time, a small sampling period combined with presence of noise results in a large *statistical* error in the empirical derivative. However, while the error of each derivative sample grows as $1/\Delta t$, the number of samples per unit time grows at precisely the same rate $1/\Delta t$ and the two effects almost cancel each other out, resulting in a nearly horizontal upper bound of correct identification (see the SM [55] for additional discussion).

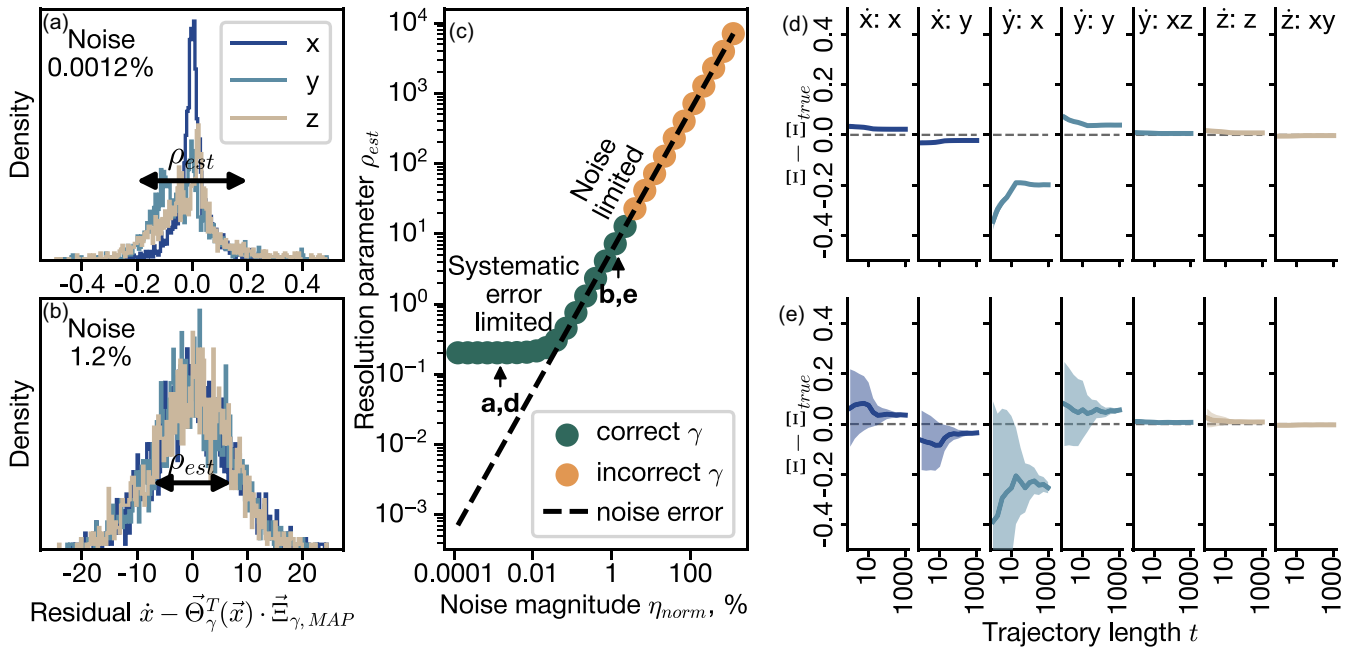



FIG. 6. Uncertainty quantification from empirical residual. [(a) and (b)] Histograms of the derivative residual across the length of the trajectory for each of three dimensions of dynamics. The black arrow indicates the $\pm 1 \cdot \rho_{\text{est}}$ range, equal to the sample standard deviation of the residual. (c) Inferred resolution parameter $\rho_{\text{est}}$ at each noise level. The marker color indicates whether the correct set of coefficients has been identified at given noise level. At low noise the residual is limited by the systematic error and thus the resolution parameter is constant. The dashed line indicates the analytic prediction of noise-driven error in the derivative, which dominates at high noise. [(d) and (e)] Marginal posterior distributions for each of the seven active terms in the Lorenz system. The solid curves show the posterior mean relative to ground truth, the shaded region indicates $\pm 1$ standard deviation. The horizontal axes cover the range of trajectory length $t \in [1, 1000]$ on logarithmic scale. Panels (a) and (d) are computed at low noise and panels (b) and (e) at moderate noise.

The interaction of noise magnitude $\eta_{\text{norm}}$ with sparsity penalty $\lambda$ is even more complex [Fig. 7(b)]. The noiseless regime of identification has been explored in Fig. 4, revealing a finite range of several orders in $\lambda$ where the identified coefficient set has not too many and not too few terms, with a narrower range for $\dot{y}$ that has more terms in the correct coefficient set. While the sparsity penalty creates a gap between the free energies of sets of different size, growing noise level gradually reduces this gap until a free-energy crossover [Fig. 5(a)]. Qualitatively, a larger initial gap $\lambda$ would require more noise $\eta$ to close, resulting in a positive slope of the limiting curve on the phase diagram. The trade-off has the approximate shape of a power law $\eta \propto \lambda^{1/3}$ across six orders of magnitude of $\lambda$, but the limited resolution of the phase diagram and the complexity of the free-energy landscape prevent a simple explanation of this scaling, leaving an important opening for further work.

## IV. CLOSED-LOOP INFERENCE

The graphical analyses presented in the previous sections, specifically the residual distribution [Figs. 6(a) and 6(b)] and the phase diagrams (Fig. 7), are powerful diagnostic tools that quantify the limits of performance of a particular data-driven algorithm. In particular, they can be combined into a closed-loop inference system: For a given dataset one can fit and integrate a SINDy model, compute its noise tolerance, and estimate the empirical noise magnitude. If the empirical noise is above the tolerance, then the model is misleading and should be rejected.

Here we demonstrate a proof-of-concept closed-loop analysis on the famous dataset of hare and lynx pelts collected each year by the Hudson Bay Company. The original theoretical works by Lotka [77] and Volterra [78] were *inspired* by observations of fluctuations in species abundance. The resulting Lotka-Volterra equations are widely celebrated but not directly derived from empirical data. It is important to clarify that there is not a single hare-lynx dataset but multiple versions of it. Two famous historical data compilations are the Leigh dataset (1847–1906) [79] and the Odum dataset (1845–1935) [80], [81] More recently, a version of the Hudson Bay Company dataset became widely used for pedagogical reasons in multiple courses and textbooks [82–86] and used in SINDy studies [26,35], which we refer to as the Mahaffy dataset (1900–1920) [83]. While the time resolution of $\Delta t =$ 1 year and the qualitative shape of oscillations are consistent across the three datasets, the covered date ranges overlap only partially and the reported numerical values are different [Figs. 8(a), 8(c), and 8(e)].

In order to compare the three empirical datasets, we construct the closed-loop inference as follows. First, we fit a polynomial Z-SINDy model to the entirety of each dataset while varying the the sparsity penalty $\lambda$ over several orders of magnitude, resulting in a sequence of models of different sparsity. We then estimate the variance of the residual $\rho$ following the method in Figs. 6(a) and 6(b) and convert it into an estimate of noise $\eta_{\text{est}} = \rho \cdot \sqrt{2}$ ($\Delta t = 1$) [green stepped curve in Figs. 8(b), 8(d), and 8(f)]. We then take the maximum *a posteriori* model at each noise level as ground truth and

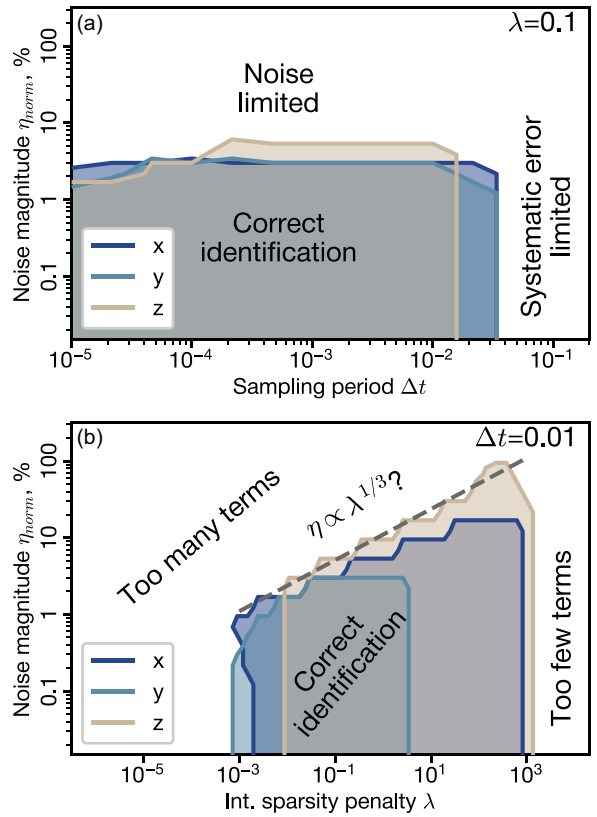



FIG. 7. Inference of the correct dynamical equations is limited by the data noise $\eta$, sampling period $\Delta t$, and sparsity penalty $\lambda$. (a) Inference fails when either the noise level crosses a threshold (noise limited), or the sampling period becomes too large to resolve the smallest timescale of the dynamical system (systematic error limited), with the boundaries consistent across the three dimensions of dynamics. (b) The region of correct inference in noise and sparsity parameters is limited on three sides: too few terms at high $\lambda$ and too many terms at low $\lambda$ or high $\eta$. The dashed line guides the eye to illustrate a speculative power law on critical noise level $\eta \propto \lambda^{1/3}$.

evaluate the inference phase diagram for it following Fig. 7(b) across a range of synthetic noise magnitudes to identify the tolerance $\eta_{\text{tol.}}$ [shaded region in Figs. 8(b), 8(d), and 8(f)].

We find that for the Mahaffy dataset at moderate sparsity $\lambda$ we can identify the familiar Lotka-Volterra functional form $\dot{x} \propto x + xy; \dot{y} \propto y + xy$ (coefficients omitted) with the estimated noise below the noise tolerance threshold [Fig. 8(b)]. Note that adding the second term to each of the two equations causes a significant reduction in the residual variance, but adding more terms does not change variance much. For the Odum and Leigh datasets the identified functional forms are all below noise threshold and are thus spurious identifications that are idiosyncratic to the particular noise *realization* in the empirical dataset [Fig. 8(d) and 8(f)]. Addition of more terms to the model does not decrease variance significantly, and some of the identified models are unstable, defined as $x^2 + y^2 > 10^6$ within simulation time and marked with red font on the figure.

We thus found that the Lotka-Volterra equations are robustly recovered from the Mahaffy dataset with both our method and many other methods [26,35], explaining the popularity of this dataset in pedagogical literature and data science benchmarks. It is, however, unclear where exactly the Mahaffy dataset comes from as it is arbitrarily restricted to the narrow time range of 1900–1920 and presumably back-converted into tabular form from a graph [86], [87] The older Odum and Leigh datasets are thus not useful for learning the Lotka-Volterra equations, at least not in the differential form. These datasets have a number of other controversies such as the seemingly wrong phase order of the abundance peaks, known as the hares eat lynx paradox [88]. Reference [89] offers a detailed history of these datasets, an extended model that accounts for the dynamics of fur trappers, and an algorithm to fit the empirical trajectories in an integral form without numerical differentiation.

## V. DISCUSSION

In this paper we introduce Z-SINDy, an *exact* Bayesian version of a simple form of SINDy that we analyze through the prism of statistical mechanics to understand how it works—and, importantly, how it breaks. We explicitly separate the discrete and continuous parts of equation inference and derive a closed-form posterior probability distribution of the coefficients. The probability of a particular discrete set condenses exponentially with growing trajectory length to a single coefficient set, though the set is not guaranteed to be correct. Even if the identified library terms are qualitatively correct, the value of the coefficients is subject to both systematic and noise-driven errors. A combination of natural and externally imposed sparsity penalties induce a set of discontinuous transitions from a large coefficient set to a sparse one. The inference procedure correctly identifies the sparse set when the trajectory has moderate noise but fails at large noise, where the inferred equation depends strongly not only on noise magnitude but also its specific *realization*. Before the noise-induced transition, matching the residual statistics provides UQ for the inferred model. The combination of these results establishes the boundaries of applicability of SINDy, and the warning mechanisms for its breakdown. The UQ and noise tolerance can be combined to have close-form inference and diagnostic of parsimonious dynamical models.

In more practical terms, equation learning is often framed as an inverse problem, the solutions to which are generically nonunique. Uniqueness can be mathematically assured by choosing a regularization (such as the sparsity penalty), but the regularization strength becomes a hyperparameter. A common way to treat a hyperparameter in machine learning is to "optimize away," usually through automated cross-validation. While cross-validation can improve model generalization, it hides the modeling choices from the user and ignores the domain-specific context of the data source. When SINDy or its versions are treated as an off-the-shelf method, the reasons for their failure or even success on a given dataset are obscure. In contrast to optimizing away, frameworks such as Z-SINDy make the exploration and discovery explicit by treating sparsity as a free parameter controlled by the modeler's subjective choices. Much like traditional modeling and problem solving is an art built on a mathematical foundation, data-driven equation learning can be seen as a modern version of that art.

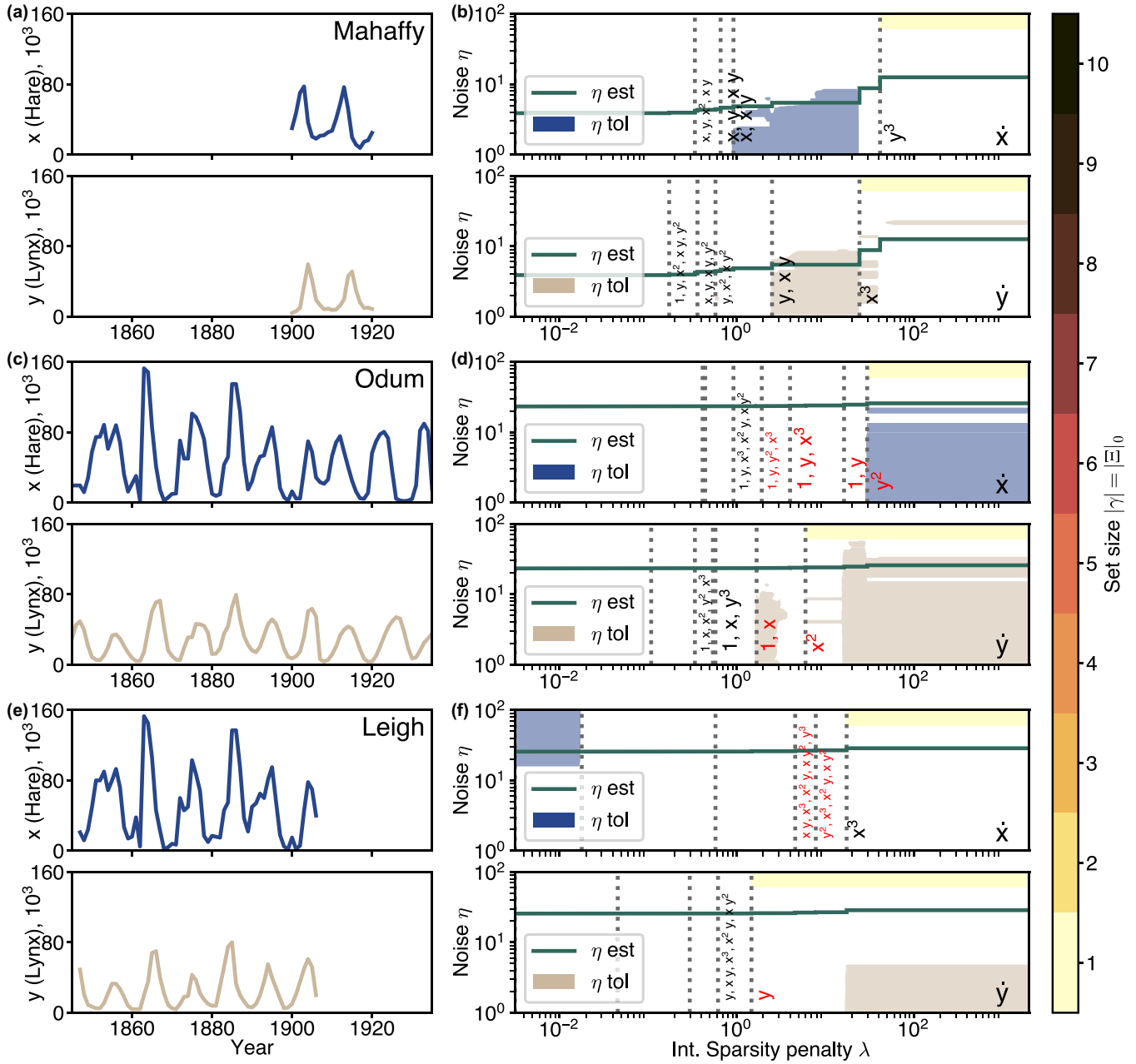



FIG. 8. Closed loop inference of predator-prey interactions. [(a), (c), and (e)] Raw empirical data on the number of hare and lynx pelts reported in each year within the data range for the three datasets, in units of $10^3$. [(b), (d), and (e)] Noise estimate (green step curves) and tolerance diagrams (shaded regions). The vertical gray dashed lines show the sparsity-induced transitions between the models. The coefficient sets are labeled to the right of the transitions, with unstable models marked in red. The color-coded shaded stripe at the top of each panel marks the size of each coefficient set.

### A. Error analysis

Our study is not unique in application of Bayesian inference to the SINDy framework. While previous recent work relied on costly Monte Carlo sampling from the posterior [24,35–39] or variational Bayes techniques [37,40,41], several studies conducted concurrently but independently from ours used the closed-form computation of model evidence [26,43]. They show that using Bayesian inference with Gaussian priors gives rise to the effect we term "natural sparsity" but do not consider an explicitly sparsity-promoting prior such as Bernoulli-Gaussian. As a result, they focus either on the probability of identifying the correct solution form [26] or the $L_2$ norm of the residual [43] at constant noise and trajectory length. In contrast, we provide both the boundaries of correct identification of the model form, as well as the asymptotic scaling of coefficient errors and uncertainties with dataset size and sampling period.

Early system identification papers acknowledged the error in the determined coefficients [90,91] but were unable to separate the statistical (noise-induced) from systematic

(differentiation-induced) errors [9]. Early forms of system identification focused on the Padé approximants traced the error in the coefficient to noise in both the matrix and vector components of regression ($C$ and $\vec{V}$ in our notation) [90]. Our study proposes a simple visual diagnostic of the origins of error in Fig. 5(b), which shows that for many practical cases the error in the vector due to the derivative estimation $\dot{\vec{x}}$) is vastly higher than the error in the matrix due to evaluation of nonlinear library terms $\Theta(\vec{x})$ at noisy arguments. At the same time, the derivative estimation is contaminated by both bias and variance, thus leading to both systematic and statistical error (see SM [55] for derivation).

The different origin and scaling of the two error terms, as well as their mitigation were only recognized recently in both strong and weak forms of SINDy [92]. The numerical differentiation error can typically be reduced by using higher-order schemes [93], whereas noise is most commonly treated with the weak form of SINDy [11,44]. At the same time, the weak formulation acts as an effective low-pass filter, thus distorting the high-frequency components of the time series that might be important to the dynamics [26]. Reference [94] assumes that the model form is known and focuses on *estimation* rather than *identification* of dynamics and provides the most comprehensive error analysis for the weak form and thus can in the future serve as a basis for a weak formulation of an exact Bayesian method. Although all of these studies were focused on the scaling and bounding of a point estimator error, the work presented here is Bayesian, thus expressing the statistical error in the scaling of the error bar width, and the systematic error in the offset between the center of the error bar and the ground truth (Fig. 6). In case of the Lorenz system presented in the main text, the systematic error in coefficients inferred from a long trajectory is roughly equivalent in low-noise and moderate-noise cases [Fig. 6(d) vs 6(e)], whereas for other systems addition of noise can induce a significant systematic error (see SM [55]). The systematic errors depend on the subtle interplay between the sampling time step $\Delta t$, the noise magnitude $\eta$, and the natural timescales of the studied dynamical system, which opens avenues for further investigation.

### B. Statistical mechanics analogies

Instead of searching for the best fit alone, we characterize the whole landscape of models on a logarithmic scale by computing their free energies. This computation integrates out the continuous coefficient values and focuses on the coefficient sets, building a connection with coarse-grained statistical mechanics and thermodynamics of interacting particle systems [57]. The consideration of long time trajectories is akin to the thermodynamic limit and motivates the evaluation of intensive free energy per data point similar to the free energy per particle. The thermodynamic cost of adding another "particle" type into the model, i.e., another SINDy term, is given by the sparsity penalty $\lambda$ that functions as (negative) chemical potential. The chemical potential of all "particles" is identical, but the interactions between them are driven by the trajectory data and thus select a particular "particle" set, i.e., a set of active SINDy library terms. As the parameters of the dataset or the hyperparameters of the inference procedure are adjusted, the inferred model can change discontinuously, akin to a phase transition or a dynamical system bifurcation. The pattern of free energies of different coefficient sets exhibits rich structure, opening avenues for further study, including connection to the entropy metrics of dynamical system trajectories [95].

The connections between statistical mechanics, statistical inference, and machine learning have a rich history, focusing primarily on the average-case behavior of prediction risk [28]. Statistical mechanics helped identify and describe multiple inference regimes and phase transitions between them, from network structure inference to constraint satisfaction to compressed sensing [29–31,33]. The so-called replica method has been used to characterize the regularized least-squares regression that also lies at the core of SINDy [54], identifying the regimes where local greedy algorithms can efficiently identify the optimal set of predictor variables [32]. However, statistical physics studies often consider the large-data case $n \to \infty$ in which, in our notation, $|\gamma|/n = \text{const}$, i.e., the set of selected coefficients is sparse but growing. This paper enriches the discussion by analyzing the case of dynamical systems in which the number of differential equation terms staying constant regardless of trajectory length $|\gamma| = \text{const}$, paying attention to the specific interpretable nature of individual terms and painting a detailed picture of inference breakdown.

### C. Integration with other SINDy techniques

The system identification scenario considered here aims to extract a parsimonious model [22] but focuses only on one aspect of parsimony—sparsity—over other aspects such as discovery of coordinates and parametric dependencies, both of which have been included in other data-driven methods. The coordinate discovery has been addressed by combining dynamics discovery with an autoencoder neural network that automatically discovers the sparse coordinates either in the optimization framework [96,97] or the Bayesian framework [24]. The parametric dependence can be inferred by including a parameter library along with the dynamical equation library [98]. The parsimony requirements can be supplemented by other desired features of dynamical systems such as global stability [99,100]. Integrating fast posterior computations from the present paper with nonlinear coordinate transformation discovery, parametric inference, and dynamic stability remain important avenues for further work. Within the analysis of trajectories, additional improvements can be achieved by a finer-scale analysis of free-energy fluctuations with respect to the general trend $F \sim f \cdot n$, as well as active learning to proactively sample the unexplored parts of phase space akin to the technique suggested in Refs. [26,50].

All SINDy approaches rely on regression and thus require a reliable estimation of the dependent variable, the trajectory derivative. While here we employ the simplest derivative method, finite difference, most other contemporary SINDy algorithms rely on some version of denoising derivative, such as total variation [101], spectral derivative [102], weak form [11], or basis expansion [12,25], which drastically improve the benchmark noise tolerance. However, such performance improvement might be misleading in quantifying the parameter uncertainty since the data uncertainty is discarded [25].

Moreover, the denoising process is itself parametric and trades random noise for a systematic error in the derivative and the library terms [72,101], requiring hyperparameter optimization [23,103]. The weak formulation of SINDy [11,44] is a promising direction for improving noise tolerance, has been used in the ensemble resampling version of SINDy [50], and can be incorporated into the exact Bayesian framework of Z-SINDy in the future. However, the weak formulation can act as an effective low-pass filter, suppressing the rapid dynamics that might be important [26]. For an experimental system where the ground-truth equations are unknown, it is thus unclear whether small variations of the dynamical variables are due to measurement noise or genuine fine-scale dynamics. Z-SINDy does not make a choice between those options by keeping $\rho$ as a free parameter of how closely the linear combination of library terms should approximate the derivative. Since both the derivative and the library terms suffer from noise contamination, it is challenging to pick $\rho$ *a priori* but it can be estimated after the SINDy fit from the remaining unexplained variance in the derivative $\dot{x} - \Theta^T \cdot \Xi$ (see Sec. III E and the SM [55] for additional discussion).

The free-energy analysis presented here makes a prediction of the identification phase diagram for a known ground-truth dynamical model. On one side, the residual distribution [Fig. 6(a) and 6(b)], the phase diagrams (Fig. 7), and the $R^2$ plots [Fig. 5(b)] are powerful diagnostic tools that quantify the limits of performance of a particular numerical algorithm and thus suggest how the boundary of detectability can be pushed, with the most immediate gains available through denoising. On the other side, the analysis establishes the noise tolerance that can be part of a closed-loop inference system: For a given dataset one can fit and integrate a SINDy model, compute its noise tolerance, and estimate the empirical noise magnitude (Fig. 8).

### D. Computational considerations

The main computational advantage of Z-SINDy is the closed form evaluation of the posterior distribution within a few seconds on a laptop computer, avoiding the costly techniques of bootstrap resampling [50], Markov-chain Monte Carlo [24], and repeated ODE integration [35] that are required in other UQ system identification methods (see the SM [55] for a detailed comparison table). On the other side, an important computational limitation to our method is the combinatorial enumeration of coefficient sets. While the free energy (8) of any particular set $\gamma$ is very cheap to compute, the number of all possible sets grows exponentially with the number of library terms $|\{\gamma\}| = 2^N$; in other words, Z-SINDy has aggressive scaling with the number of library terms but a very small prefactor that makes enumeration still viable in many practical cases.

In this paper we chose to evaluate free energies exhaustively for all sets to illustrate their scaling behavior. At the same time, the free energies of different sets are heavily stratified [e.g., Fig. 2(c)], and only a small fraction of sets are involved in sparsity- or noise-induced breakdowns. In order to make Z-SINDy more computationally efficient and thus tractable for realistic systems, future work should aim to understand the patterns of free energies better to reduce the number of sets to evaluate, for instance by using greedy methods [26], subset selection [104], least angle regression [74], mixed-integer optimization [34], or branch-and-bound methods [105]. Other avenues for computational speedup include using multiple threads to evaluate the embarrassingly parallel set of free energies, taking advantage of the boundary method to iteratively compute the inversions of progressively larger matrix inverses $C_\gamma^{-1}$, or using Monte Carlo sampling of sets. At the same time, we hope that the physical intuition of coarse-graining, chemical potentials, and free energies per data point would inform the further development of statistical methods for dynamical systems.

Python code for our implementation of Z-SINDy is available at [106].

## ACKNOWLEDGMENTS

The authors thank U. Fasel, L. Fung, L. M. Gao, M. Juniper, P. Langley, J. Michel, R. Roy, and S. Still for helpful discussions; B. Deng for providing the tabulated predator-prey data from Ref. [89]; anonymous referees for extensive reference suggestions; and L. D. Lederer for administrative support. This work uses Scientific colour Maps for visualization [107]. The authors acknowledge support from the National Science Foundation AI Institute in Dynamic Systems (Grant No. 2112085). J.B. acknowledges the support of the Center for Advanced Mathematical Sciences, and the Artificial Intelligence, Data Science and Computing Hub at the American University of Beirut.

## DATA AVAILABILITY

The data that support the findings of this article are openly available. The benchmark ODEs and parameters for synthetic data generation are listed in the SM [55]. The predator-prey datasets are available in [83,89].

[1] L. Ljung, Perspectives on system identification, Annu. Rev. Control **34**, 1 (2010).

[2] J. S. North, C. K. Wikle, and E. M. Schliep, A review of data-driven discovery for dynamic systems, Int. Stat. Rev. **91**, 464 (2023).

[3] Z. Li, N. B. Kovachki, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, A. Anandkumar *et al.*, Fourier neural operator for parametric partial differential equations, in *International Conference on Learning Representations* (ICLR, Vienna, Austria, 2021).

[4] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, Stat. Surv. **16**, 1 (2022).

[5] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, Queue **16**, 31 (2018).

[6] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. **1**, 206 (2019).

[7] J. P. Crutchfield and B. S. McNamara, Equations of motion from a data series, Complex Syst. **1**, 417 (1987).

[8] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA **113**, 3932 (2016).

[9] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Data-driven discovery of partial differential equations, Sci. Adv. **3**, e1602614 (2017).

[10] E. Kaiser, J. N. Kutz, and S. L. Brunton, Sparse identification of nonlinear dynamics for model predictive control in the low-data limit, Proc. R. Soc. A **474**, 20180335 (2018).

[11] D. A. Messenger and D. M. Bortz, Weak SINDy: Galerkin-based data-driven model selection, Multiscale Model. Simul. **19**, 1474 (2021).

[12] J. M. Hokanson, G. Iaccarino, and A. Doostan, Simultaneous identification and denoising of dynamical systems, SIAM J. Sci. Comput. **45**, A1413 (2023).

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

[14] S.-M. Udrescu, A. Tan, J. Feng, O. Neto, T. Wu, and M. Tegmark, AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity, Adv. Neural Inf. Process. Syst. **33**, 4860 (2020).

[15] B. C. Daniels and I. Nemenman, Automated adaptive inference of phenomenological dynamical models, Nat. Commun. **6**, 8133 (2015).

[16] S. Sahoo, C. Lampert, and G. Martius, Learning equations for extrapolation and control, in *International Conference on Machine Learning* (PMLR, 2018), pp. 4442–4450.

[17] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Inferring biological networks by sparse identification of nonlinear dynamics, IEEE Trans. Mol. Biol. Multi-Scale Commun. **2**, 52 (2016).

[18] X. Yang, L. Guo, Z. Zheng, N. Riemer, and C. W. Tessum, Atmospheric chemistry surrogate modeling with sparse identification of nonlinear dynamics, J. Geophys. Res.: Mach. Learn. Comput. **1**, e2024JH000132 (2024).

[19] J. Bakarji and D. M. Tartakovsky, Data-driven discovery of coarse-grained equations, J. Comput. Phys. **434**, 110219 (2021).

[20] R. Supekar, B. Song, A. Hastewell, G. P. Choi, A. Mietke, and J. Dunkel, Learning hydrodynamic equations for active matter from particle simulations and experiments, Proc. Natl. Acad. Sci. USA **120**, e2206994120 (2023).

[21] J.-C. Loiseau and S. L. Brunton, Constrained sparse galerkin regression, J. Fluid Mech. **838**, 42 (2018).

[22] J. N. Kutz and S. L. Brunton, Parsimony as the ultimate regularizer for physics-informed machine learning, Nonlin. Dyn. **107**, 1801 (2022).

[23] F. Van Breugel, J. N. Kutz, and B. W. Brunton, Numerical differentiation of noisy data: A unifying multi-objective optimization framework, IEEE Access **8**, 196865 (2020).

[24] M. L. Gao and J. N. Kutz, Bayesian autoencoders for data-driven discovery of coordinates, governing equations and fundamental constants, Proc. R. Soc. A **480**, 20230506 (2024).

[25] J. S. North, C. K. Wikle, and E. M. Schliep, A Bayesian approach for data-driven dynamic equation discovery, J. Agric. Biol. Environ. Stat. **27**, 728 (2022).

[26] L. Fung, U. Fasel, and M. Juniper, Rapid Bayesian identification of sparse nonlinear dynamics from scarce and noisy data, Proc. R. Soc. A. **481**, 20240200 (2025).

[27] J. Wentz and A. Doostan, Derivative-based SINDy (DSINDy): Addressing the challenge of discovering governing equations from noisy data, Comput. Methods Appl. Mech. Eng. **413**, 116096 (2023).

[28] P. Langley and S. Sage, Tractable average-case analysis of naive Bayesian classifiers, in *ICML*, Vol. 99 (Citeseer, Bled, Slovenia, 1999), pp. 220–228.

[29] S. Ganguli and H. Sompolinsky, Statistical mechanics of compressed sensing, Phys. Rev. Lett. **104**, 188701 (2010).

[30] R. R. Nadakuditi and M. E. Newman, Graph spectra and the detectability of community structure in networks, Phys. Rev. Lett. **108**, 188701 (2012).

[31] L. Zdeborová and F. Krzakala, Statistical physics of inference: Thresholds and algorithms, Adv. Phys. **65**, 453 (2016).

[32] T. Obuchi, Y. Nakanishi-Ohno, M. Okada, and Y. Kabashima, Statistical mechanical analysis of sparse linear regression as a variable selection problem, J. Stat. Mech. (2018) 103401.

[33] F. Krzakala and L. Zdeborová, *Statistical physics methods in optimization and machine learning: An introduction to replica, cavity & message-passing techniques* (2022).

[34] D. Bertsimas and W. Gurnee, Learning sparse nonlinear dynamics via mixed-integer optimization, Nonlin. Dyn. **111**, 6585 (2023).

[35] S. M. Hirsh, D. A. Barajas-Solano, and J. N. Kutz, Sparsifying priors for Bayesian uncertainty quantification in model discovery, R. Soc. Open Sci. **9**, 211823 (2022).

[36] Z. Wang, B. Wu, K. Garikipati, and X. Huan, A perspective on regression and Bayesian approaches for system identification of pattern formation dynamics, Theor. Appl. Mech. Lett. **10**, 188 (2020).

[37] Y. C. Mathpati, T. Tripura, R. Nayek, and S. Chakraborty, Discovering stochastic partial differential equations from limited data using variational Bayes inference, Comput. Methods Appl. Mech. Eng. **418**, 116512 (2024).

[38] T. Tripura and S. Chakraborty, A sparse Bayesian framework for discovering interpretable nonlinear stochastic dynamical systems with Gaussian white noise, Mech. Syst. Signal Process. **187**, 109939 (2023).

[39] R. Nayek, R. Fuentes, K. Worden, and E. J. Cross, On spike-and-slab priors for Bayesian equation discovery of nonlinear dynamical systems via sparse linear regression, Mech. Syst. Signal Proc. **161**, 107986 (2021).

[40] K. S. More, T. Tripura, R. Nayek, and S. Chakraborty, A Bayesian framework for learning governing partial differential equation from data, Physica D **456**, 133927 (2023).

[41] D. Long, W. Xing, A. Krishnapriyan, R. Kirby, S. Zhe, and M. W. Mahoney, Equation discovery with Bayesian spike-and-slab priors and efficient kernels, in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2024), pp. 2413–2421.

[42] R. K. Niven, A. Mohammad-Djafari, L. Cordier, M. Abel, and M. Quade, Bayesian identification of dynamical systems, in *Proceedings*, Vol. 33 (MDPI, Garching, Germany, 2020), p. 33.

[43] R. K. Niven, L. Cordier, A. Mohammad-Djafari, M. Abel, and M. Quade, Dynamical system identification, model selection, and model uncertainty quantification by Bayesian inference, Chaos **34**, 083140 (2024).

[44] D. A. Messenger and D. M. Bortz, Asymptotic consistency of the wsindy algorithm in the limit of continuum data, IMA J. Numer. Anal. drae086 (2024).

[45] L. Zhang and H. Schaeffer, On the convergence of the sindy algorithm, Multiscale Model. Simul. **17**, 948 (2019).

[46] G. Tran and R. Ward, Exact recovery of chaotic systems from highly corrupted data, Multiscale Model. Simul. **15**, 1108 (2017).

[47] H. Schaeffer, G. Tran, and R. Ward, Extracting sparse high-dimensional dynamics from limited data, SIAM J. Appl. Math. **78**, 3279 (2018).

[48] H. Schaeffer, G. Tran, R. Ward, and L. Zhang, Extracting structured dynamical systems using sparse optimization with very few samples, Multiscale Model. Simul. **18**, 1435 (2020).

[49] S. Zhang and G. Lin, Robust data-driven discovery of governing physical laws with error bars, Proc. R. Soc. A **474**, 20180305 (2018).

[50] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton, Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control, Proc. R. Soc. A **478**, 20210904 (2022).

[51] M. L. Gao, U. Fasel, S. L. Brunton, and J. N. Kutz, Convergence of uncertainty estimates in ensemble and Bayesian sparse model discovery, arXiv:2301.12649.

[52] D. L. Donoho, For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution, Commun. Pure Appl. Math. **59**, 797 (2006).

[53] C. M. Carvalho, N. G. Polson, and J. G. Scott, Handling sparsity via the horseshoe, in *Artificial Intelligence and Statistics* (PMLR, 2009), pp. 73–80.

[54] A. Bereyhi, Statistical mechanics of regularized least squares, Ph.D. thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2020.

[55] See Supplemental Materialat http://link.aps.org/supplemental/10.1103/4d98-tdlp for detailed derivations, supplementary results, and computational time benchmarks.

[56] D. J. C. Mackay, *Bayesian Methods for Adaptive Models* (California Institute of Technology, Pasadena, 1992).

[57] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Addison-Wesley, Reading MA, 1992).

[58] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control **19**, 716 (1974).

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

[60] We thank L. Fung and M. Juniper for directing our attention to the natural sparsity effect which they term "Occam factor" in their paper [26].

[61] P. Zhao and B. Yu, On model selection consistency of lasso, J. Mach. Learn. Res. **7**, 2541 (2006).

[62] E. N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. **20**, 130 (1963).

[63] J. Sokol, The hidden heroines of chaos, Quanta Mag. (2019).

[64] W. Gilpin, Chaos as an interpretable benchmark for forecasting and data-driven modelling, in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021).

[65] V. I. Zubov, *Methods of AM Lyapunov and Their Application*, Vol. 4439 (US Atomic Energy Commission, Germantown, MD, 1961).

[66] A. A. Kaptanoglu, L. Zhang, Z. G. Nicolaou, U. Fasel, and S. L. Brunton, Benchmarking sparse system identification with low-dimensional chaos, Nonlin. Dynam. **11**, 13143 (2023).

[67] M. C. Cross and P. C. Hohenberg, Pattern formation outside of equilibrium, Rev. Mod. Phys. **65**, 851 (1993).

[68] B. de Silva, K. Champion, M. Quade, J.-C. Loiseau, J. Kutz, and S. Brunton, PySINDy: A Python package for the sparse identification of nonlinear dynamical systems from data, J. Open Source Softw. **5**, 2104 (2020).

[69] O. Rossler, An equation for hyperchaos, Phys. Lett. A **71**, 155 (1979).

[70] C. Letellier and O. E. Rossler, Hyperchaos, Scholarpedia **2**, 1936 (2007).

[71] A. Abooee, H. Yaghini-Bonabi, and M. R. Jahed-Motlagh, Analysis and circuitry realization of a novel three-dimensional chaotic system, Commun. Nonlin. Sci. Numer. Simul. **18**, 1235 (2013).

[72] X. Hu, V. Tritschler, S. Pirozzoli, and N. Adams, Dispersion-dissipation condition for finite difference schemes, arXiv:1204.5088.

[73] K. Wu and D. Xiu, Numerical aspects for approximating governing equations using data, J. Comput. Phys. **384**, 200 (2019).

[74] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression, Ann. Stat. **32**, 407 (2004).

[75] C. B. Delahunt and J. N. Kutz, A toolkit for data-driven discovery of governing equations in high-noise regimes, IEEE Access **10**, 31210 (2022).

[76] A. Cortiella, K.-C. Park, and A. Doostan, A priori denoising strategies for sparse identification of nonlinear dynamical systems: A comparative study, J. Comput. Inf. Sci. Eng. **23**, 011004 (2023).

[77] A. Lotka, *Elements of Physical Biology* (Williams and Wilkins, Philadelphia, PA (1925).

[78] V. Volterra, *Variazioni e Fluttuazioni Del Numero D'individui in Specie Animali Conviventi* (Società anonima tipografica "Leonardo da Vinci,", Città di Castello, 1926).

[79] E. Leigh, The ecological role of Volterra's equations, in *Some Mathematical Problems in Biology*, edited by M. Gerstenhaber (American Mathematical Society, Washington, DC, 1968), pp. 1–61.

[80] E. P. Odum, *Fundamentals of Ecology*, 3rd ed. (W B Saunders, Philadelphia, PA, 1971).

[81] We use the tabulation of the Leigh and Odom datasets provided via personal communication by B. Deng [89].

[82] P. Howard, M442: Mathematical modeling (2022), https://people.tamu.edu/∼phoward//M442.html.

[83] J. M. Mahaffy, Lotka-Volterra models (2010), https://jmahaffy.sdsu.edu/courses/f09/math636/lectures/lotka/qualde2.html.

[84] B. Carpenter, Predator-prey population dynamics: the Lotka-Volterra model in stan, https://mc-stan.org/users/documentation/case-studies/lotka-volterra-predator-prey.html (2018).

[85] U. Wilensky and K. Reisman, Thinking like a wolf, a sheep, or a firefly: Learning biology through constructing and testing computational theories—an embodied modeling approach, Cogn. Instruct. **24**, 171 (2006).

[86] W. K. Purves, G. H. Orians, and H. C. Heller, *Life, the Science of Biology* (Sinauer Associates, Inc., Sunderland, 1992).

[87] Carpenter [84] traces the origin of the numerical table as follows: "Howard (personal communication) traces the numerical data to Joseph M. Mahaffy's (2010) lecture notes on mathematical modeling for his San Diego State University course (Math 636) [83], which in turn cites Purves, Sadava, Orians, and Heller's book, Life: The Science of Biology [86]" (references ours).

[88] M. E. Gilpin, Do hares eat lynx? Am. Nat. **107**, 727 (1973).

[89] B. Deng, An inverse problem: Trappers drove hares to eat lynx, Acta Biotheoret. **66**, 213 (2018).

[90] P. A. Regalia, An unbiased equation error identifier and reduced-order approximations, IEEE Trans. Signal Process. **42**, 1397 (1994).

[91] S. Rudy, A. Alla, S. L. Brunton, and J. N. Kutz, Data-driven identification of parametric partial differential equations, SIAM J. Appl. Dyn. Syst. **18**, 643 (2019).

[92] M. Tang, W. Liao, R. Kuske, and S. H. Kang, Weakident: Weak formulation for identifying differential equation using narrow-fit and trimming, J. Comput. Phys. **483**, 112069 (2023).

[93] H. Chen, Data-driven sparse identification of nonlinear dynamical systems using linear multistep methods, Calcolo **60**, 11 (2023).

[94] D. M. Bortz, D. A. Messenger, and V. Dukic, Direct estimation of parameters in ode models using wendy: Weak-form estimation of nonlinear dynamics, Bull. Math. Biol. **85**, 110 (2023).

[95] P. Gaspard and X.-J. Wang, Noise, chaos, and $(\varepsilon, \tau)$-entropy per unit time, Phys. Rep. **235**, 291 (1993).

[96] K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, Data-driven discovery of coordinates and governing equations, Proc. Natl. Acad. Sci. USA **116**, 22445 (2019).

[97] J. Bakarji, K. Champion, J. Nathan Kutz, and S. L. Brunton, Discovering governing equations from partial measurements with deep delay autoencoders, Proc. R. Soc. A **479**, 20230422 (2023).

[98] Z. G. Nicolaou, G. Huo, Y. Chen, S. L. Brunton, and J. N. Kutz, Data-driven discovery and extrapolation of parameterized pattern-forming dynamics, Phys. Rev. Res. **5**, L042017 (2023).

[99] A. A. Kaptanoglu, J. L. Callaham, A. Aravkin, C. J. Hansen, and S. L. Brunton, Promoting global stability in data-driven models of quadratic nonlinear dynamics, Phys. Rev. Fluids **6**, 094401 (2021).

[100] M. Peng, A. Kaptanoglu, C. Hansen, K. Manohar, and S. Brunton, Local stability guarantees for data-driven quadratically nonlinear models, arXiv:2403.00324.

[101] R. Chartrand, Numerical differentiation of noisy, nonsmooth data, Int. Scholar. Res. Notices **2011** (2011).

[102] R. W. Schafer, What is a Savitzky-Golay filter?[lecture notes], IEEE Signal Proc. Mag. **28**, 111 (2011).

[103] F. Van Breugel, Y. Liu, B. W. Brunton, and J. N. Kutz, PyNumDiff: A python package for numerical differentiation of noisy time-series data, J. Open Source Softw. **7**, 4078 (2022).

[104] D. P. Foster and E. I. George, The risk inflation criterion for multiple regression, Ann. Stat. **22**, 1947 (1994).

[105] D. Bertsimas and B. V. Parys, Sparse high-dimensional regression: Exact scalable algorithms and phase transitions, Ann. Stat. **48**, 300 (2020).

[106] J. Bakarji and A. A. Klishin, ZSINDy (2024), https://github.com/josephbakarji/zsindy.

[107] F. Crameri, Scientific colour maps (2018), https://doi.org/10.5281/zenodo.1243862.