Information theory for data-driven model reduction in physics and biology

Matthew S. Schmitt*,[1,2] Maciej Koch-Janusz*,[1,3,4] Michel Fruchart,[1,5]
Daniel S. Seara,[1] Michael Rust,[6,2] and Vincenzo Vitelli[1,2,7]

[1] *University of Chicago, James Franck Institute, 929 E 57th Street, Chicago, IL 60637*
[2] *University of Chicago, Department of Physics, 929 E 57th Street, Chicago, IL 60637*
[3] *Haiqu, Inc., 95 Third Street, San Francisco, CA 94103, USA*
[4] *Department of Physics, University of Zurich, 8057 Zurich, Switzerland*
[5] *ESPCI, Laboratoire Gulliver, 10 rue Vauquelin, 75231 Paris cedex 05*
[6] *University of Chicago, Department of Molecular Genetics and Cell Biology, Chicago, IL, 60637*
[7] *University of Chicago, Kadanoff Center for Theoretical Physics, 933 E 56th St, Chicago, IL 60637*

Model reduction is the construction of simple yet predictive descriptions of the dynamics of many-body systems in terms of a few relevant variables. A prerequisite to model reduction is the identification of these variables, a task for which no general method exists. Here, we develop an approach to identify relevant variables, defined as those most predictive of the future, using the so-called information bottleneck. We elucidate analytically the relation between these relevant variables and the eigenfunctions of the transfer operator describing the dynamics. In the limit of high compression, the relevant variables are directly determined by the slowest-decaying eigenfunctions. Our results provide a firm foundation to interpret deep learning tools that automatically identify reduced variables. Combined with equation learning methods this procedure yields the hidden dynamical rules governing the system's evolution in a data-driven manner. We illustrate how these tools work in diverse settings including model chaotic and quasiperiodic systems in which we also learn the underlying dynamical equations, uncurated satellite recordings of atmospheric fluid flows, and experimental videos of cyanobacteria colonies in which we discover an emergent synchronization order parameter.

The exhaustive description of a biological or physical system is usually impractical due to the sheer volume of information involved. As an example, the fluids in a steam engine may be described by a 10^{27}–dimensional state vector containing the positions and momenta of every particle in the engine. Yet, for most practical purposes, its working can be effectively described using only a small number of thermodynamic variables like pressure and temperature. Likewise, some aspects of the behavior of a nematode or a fruit fly, with 385 and ∼140,000 neurons respectively, may be effectively captured by orders of magnitude fewer degrees of freedom [1–5]. Similar reductions can be achieved for systems ranging from diffusing particles to biochemical molecules and complex networks. In all cases, certain *relevant variables* can be predicted far into the future even though individual degrees of freedom in the system are effectively unpredictable.

The process by which one goes from the complete description of a system to a simpler one is known as model reduction. Diverse procedures for model reduction exist across the natural sciences. They range from analytical methods, such as adiabatic elimination and multiple-scale analysis [6–14], to data-driven methods such as dynamic mode decomposition and other linear projection-based methods [15–20], diffusion maps [21], spectral submanifolds [22, 23], independent component analysis [24], and deep encoder-decoder neural networks [25–32].

Model reduction consists of two ingredients: a decomposition of the full system into relevant and irrelevant variables, and a self-contained dynamical equation which governs the evolution of the relevant variables. In the absence of prior knowledge and intuition (e.g. a clear separation of scales), identifying a decomposition that produces a minimal set of variables which are maximally descriptive of the full system is an open problem [9]. The answer to this question depends both on the time scale and precision with which one aims to describe the system [33, 34]. In addition, it also depends on one's ability to identify an adequate dynamical equation for the relevant variables, especially in data-driven settings, where the relevant variables and the dynamical equation are to be learned simultaneously.

To make progress, we develop an information-theoretic framework for model reduction. Very much like MP3 compression is about retaining information that matters most to the human ear [35], model reduction is about keeping information that matters most to predict the future [36, 37]. However, formalizing this intuitive statement requires some care in order to avoid trivial results, that manual approaches usually avoid using physical intuition. By carefully formulating model reduction within this information-theoretic perspective, we show that model reduction (the identification of both relevant variables and their effective dynamics) can be decomposed into a two-step process. First, the problem of variable identification can be reformulated in a way that removes any reference to the effective dynamics, and solved on its own. Second, once the relevant variables are known, an independent dynamical-equation learning step can yield the effective dynamics.

Our key step is to relate the model reduction objective to a lossy compression problem known as the information bottleneck (IB)[36, 38, 39], and to analytically show how and under what conditions the standard operator-

theoretic formalism of dynamical systems [27, 40], which underlies most methods of model reduction, naturally emerges from the optimal compression solution. This allows us to give a precise answer to the question of how to identify relevant and irrelevant variables. Crucially, our framework provides a criterion for when to stop increasing the complexity of a minimal model by directly measuring how informative relevant variables are of their own future and of the future of the entire system. Further, it provides a firm foundation to address a practical problem: the construction of deep learning tools to perform model reduction that are guaranteed to be interpretable. This practical realization of our protocol is based on recently-developed deep learning tools that allow scaling our approach to high dimensional systems [41].

The remainder of this paper is organized as follows. Section 1 formalizes model reduction and gives our information-theoretic starting point (see also SI Sections 1-3). Section 2 shows how the optimal information-bottleneck solution identifies the dominant (left) eigenfunctions of the system's transfer operator (see also SI Section 4), and delineates a procedure to extract the reduced variables. In Sections 3 and 4 we show how information measures capture features of the system's transfer operator spectrum and how this relates to the question of "when to stop" increasing the complexity of the reduced model. Section 5 (and SI Sections 5-6) shows that our framework can be used to learn relevant variables directly from data using neural networks. We illustrate our approach on several benchmark dynamical systems and demonstrate that it works even on uncurated datasets, *e.g.* satellite movies of atmospheric flows downloaded from YouTube. SI Section 7 shows applications to chaotic dynamics including the Lorenz-63 system, the Kuramoto-Sivashinsky equation and a hyperchaotic system. In Section 6 we show how our approach, combined with equation learning algorithms, forms a full model reduction pipeline yielding both the reduced variables describing the system and the equations governing its dynamics. Finally, in Section 7 (and SI Section 8) we apply our method to experimental microscopy videos of cyanobacteria colonies, discovering an emergent synchronization order parameter.

I. MODEL REDUCTION AS A COMPRESSION PROBLEM

Consider a system whose state is described by a (random) variable X_t, which might correspond to anything from the position of a single particle to an image of a fluid flow or the fluorescent molecules in a living system (Fig. 1a,b). The full state can be high dimensional, with a number of dimensions equal to the number of particles in a gas, or the number of pixels in an image. It is often difficult to predict the future state $X_{t+\Delta t}$ of such high-dimensional systems due to the enormous computational cost of evolving the system forward in time to a desired accuracy. However, in many cases there exist some functions of the full state $H_t = h(X_t)$ which are low-dimensional and very informative of the future state of the system. These features often correspond to collective variables which evolve slowly in time. While one may guess the appropriate choice of variables $h(X_t)$ in contexts where all microscopic details of the system are known, we are interested in how to find such variables without such *a priori* knowledge.

We formulate the task of identifying the reduced variables and their evolution as an optimization problem. A "good" reduced model is one which entails a compressed description of the state, and which may be evolved in time in place of the full system to make predictions. The procedure consists of three steps: first, reduction of the full system by an encoding $h_{\text{enc}} : X_t \mapsto H_t$; second, evolution of the reduced state via an evolution operator $\mathcal{S} : H_t \mapsto H_{t+\Delta t}$; third, decoding the reduced state via a function $h_{\text{dec}}^{-1} : H_{t+\Delta t} \mapsto X_t$ to predict the full system's state. These functions can be found by optimizing

$$\min_{h_{\text{enc}},\mathcal{S},h_{\text{dec}}^{-1}} \dim h_{\text{enc}}(X_t) + \beta\, d\big(h_{\text{dec}}^{-1}(\mathcal{S}h_{\text{enc}}(X_t)), X_{t+\Delta t}\big). \tag{1}$$

The first term is a compressive constraint enforcing a "reduction" of the model, with the $\dim(\cdot)$ function measuring *e.g.* the number of bits used to represent the encoded state, or the sparsity of its representation in a preferred basis. The second term captures the error in our prediction of the full state when using the reduced model, measured by a "distance" $d(\cdot,\cdot)$. The parameter β sets the trade-off between dimensionality reduction and predictive power. Note that if the encoding h_{enc} is non-invertible (as is generally the case), the decoder h_{dec}^{-1} is defined probabilistically via a conditional distribution $h_{\text{dec}}^{-1}(H_{t+\Delta t}) \sim p(\cdot|H_{t+\Delta t})$. This is analogous to a procedure producing a thermodynamic ensemble compatible with the state variables. For instance, the canonical ensemble associates to every macrostate $h_{\text{enc}} = (N, V, T)$ defined by the number of particles, the volume, and the temperature a distribution $p_{\text{can}}(\cdot|N, V, T)$ over the set of microstates from which the decoded state is drawn, $h_{\text{dec}}^{-1} \sim p_{\text{can}}(\cdot|N, V, T)$.

The above objective depends on the choice of distance d, which incorporates the geometry of the state space and any desired biases about what is "important" in the system; it may differ from system to system. In what follows, we measure accuracy in terms of the mutual information I shared between the reduced model prediction and the full state. This abstracts away from the geometry of the space and characterizes the similarity of variables in terms of how predictive they are of each other, rather than how close they are, in a system-independent way. This has an important consequence: as a result of the data processing inequality, we can show that our objective function simplifies to one depending only on h_{enc}:

$$\min_{h_{\text{enc}}} \quad \dim h_{\text{enc}}(X_t) - \beta I\big(h_{\text{enc}}(X_t), X_{t+\Delta t}\big) \tag{2}$$

The steps to this result are shown in detail in the Methods. This simplification is dramatic: instead of simultaneously learning the encoding function, evolution operator, and decoder, the only step we need to care about is the encoding. Consequently, we view model reduction in the rest of this paper as a task of finding the optimal reduced variables with which to represent our system. The evolution rule $\mathcal{S}$ and decoder h_{dec}^{-1} are *induced* by the choice of h_{enc}, and we may view their identification as a secondary, downstream task to be performed after the proper variables have been found.

Before we proceed, we still need to specify the dimensionality reduction. This is subtle, as in different contexts different measures may seem natural, for instance, the entropy for discrete variables, or the number of components for continuous variables. We achieve our dimensionally-reduced variables with a two-step construction. First, we use a term explicitly forcing the reduced representation to lose information about the full state, $\min I(h(X_t), X_t)$ (for notational simplicity, we refer to h_{enc} as h). The minimal objective found in this way

$$\min_h \mathcal{L}_{\text{IB}} = I(X_t, h(X_t)) - \beta I(X_{t+\Delta t}, h(X_t)), \tag{3}$$

is known as the information bottleneck (IB) [38, 39]. It (and related variants [42, 43]) has been used in a variety of contexts, from document clustering [42] to neural codes [44] to (most relevant for us) dynamical systems [36, 45, 46], as well as in deep-learning contexts [41, 47, 48].

The IB objective ((3)) is a functional over stochastic functions, which may be represented by a conditional probability distribution $h(x_t) \sim p(h|x_t)$ (we denote random variables by uppercase X_t and their values by lowercase x_t). Crucially, the parameter β controls the trade-off between compression and prediction. For small β the compression term dominates and the optimal encoder is trivial, losing all information about the system. For intermediate β the compression term does not allow X_t to be completely represented by H_t, so features of X_t must compete to pass through to the encoding variable (Fig. 1b). These selected features are reflected in the IB encoder

$$p^*(h_t|x_t) = \arg\min \mathcal{L}_{\text{IB}} \tag{4}$$

which provides the optimal trade-off between retained information and predictability [37]. In the optimal encoder, information about X_t has been removed both by reducing the dimensionality, as well as by injecting noise into h_t. Increasing β can impact both of these, either by increasing the dimensionality to learn new features, or by reducing noise to better identify the features that have already been found. In order to "perfectly" identify the features given a fixed set of components, in the second step we modify the IB encoder by discarding the part attributable to noise alone. This effectively increases the coupling of the reduced variables to the already-learned features. We shall see that this procedure isolates deterministic, interpretable reduced variables in both discrete and continuous settings.

II. THE OPTIMAL ENCODER IN TERMS OF THE TRANSFER OPERATOR

In any realistic experimental setting, the presence of noise or uncertainty means we cannot predict precisely the future state of a system but instead only a likely distribution of future states. Our prediction of the state at Δt in the future is then represented mathematically as $p(x_{t+\Delta t}|x_t)$, the probability of observing state $x_{t+\Delta t}$ given the current state x_t. For Markovian, or "memoryless" dynamics, this conditional probability distribution completely characterizes the dynamics of the system, and determines how probability distributions evolve in time. In this work we assume that the systems we consider are (or can be made) Markovian; for a detailed discussion of how to measure the memory and how it changes our below results, see SI Sections 3 and 4D. The evolution of probability distributions can then be understood as the action of a linear transfer operator $U^{\Delta t}$ on probability distributions:

$$p_{X_{t+\Delta t}}(x_{t+\Delta t}) = \int p(x_{t+\Delta t}|x_t) p_{X_t}(x_t) \mathrm{d}x_t \equiv U^{\Delta t}[p_{X_t}] \tag{5}$$

$U^{\Delta t}$ can be decomposed as follows (using standard notation from quantum mechanics and explained in the SI):

$$U^{\Delta t} = \sum_n |\rho_n\rangle \, \mathrm{e}^{\lambda_n \Delta t} \, \langle \phi_n| + U_{\text{ess}} \tag{6}$$

where $|\rho_n\rangle$ denote right eigenfunctions with eigenvalue $\Lambda_n \equiv \mathrm{e}^{\lambda_n \Delta t}$ and $\langle \phi_n|$ are the corresponding left eigenfunctions. λ_n are the eigenvalues of the infinitesimal generator of $U^{\Delta t}$, known as the Fokker-Planck operator (Fig. 1d). The operator U_{ess} corresponds to the so-called essential spectrum, and we assume that it can be neglected. This is usually possible when the system is subjected to even a small amount of noise, or when some amount of uncertainty is present in the measurements [49, 50]. The eigenfunctions $\langle \phi_n|$ in (6) are in some sense "natural" features of the dynamics, as they evolve independently in time.

Our key observation is that the optimal IB encoder in (4), which we will use as a "filter" to extract relevant features, can be expressed in terms of the eigenvalues λ_n and left eigenfunctions ϕ_n of (the generator of) U,

$$p^*_\beta(h_t|x_t) = \frac{p^*_\beta(h_t)}{\mathcal{N}(x_t)} \exp\left[\beta \sum_n \mathrm{e}^{\lambda_n \Delta t} \phi_n(x_t) f_n(h_t)\right] \tag{7}$$

where $f_n(h_t)$ are factors that do not depend on x_t. For an outline of the mathematical steps leading to this see Methods, as well as SI Section 4. The combined effect of

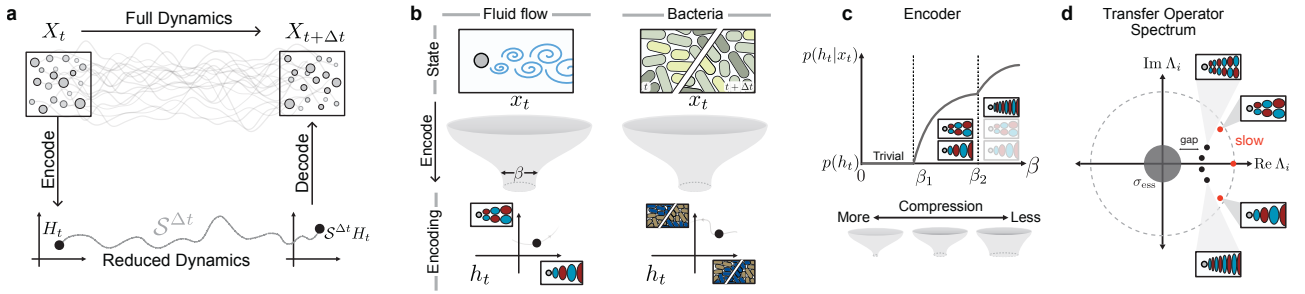

FIG. 1. **Interpretable dynamical variables in reduced models via the information bottleneck.** (a) In general model reduction aims to both learn an encoding and reduced dynamics, so that the reduced variable serves as a useful proxy that may be decoded for an estimate of the full system. Requiring the reduced variable to only be informative of the full state decouples these two steps (see SI Section 2). (b) The information bottleneck compresses high-dimensional state variables x_t, into simpler encoding variables h_t with a controllable trade-off between the degree of compression and the predictive power about the system's future. With deep neural networks, the encoding can be computed directly from data of observed fluid flows (left) or biological datasets, such as fluorescently labeled bacteria colonies (right). In general, the state of the variable x_t may comprise time-lagged variables of the intensity field, $x_t = \{I_t, I_{t+\Delta t}\}$ (right). The amount of compression is determined by the "width" of the bottleneck β (see (3)). The resulting compressed, or encoded, variables h_t represent collective variables most predictive of the system's future. (c) Schematic evolution of the encoder $p(h_t|x_t)$ for varying compression strength β. For low β (high compression), the encoder is trivial and forgets everything about the input x_t. After the first IB transition at β_1, the encoder becomes non-trivial by gaining some dependence on x_t; some features of the input are able to pass through the bottleneck. At subsequent IB transitions, additional features are learned. (d) The point spectrum of the transfer operator contains several slowly decaying modes (red). We show that the most predictive variables that IB systematically extracts correspond to the slowest eigenfunctions of the transfer operator, associated to eigenvalues Λ_i with $|\Lambda_i| \approx 1$. In fluid flows, the slowest-decaying eigenfunctions typically represent large-scale coherent patterns of the flow field, while faster-decaying eigenfunctions correspond to variations over shorter length scales. In general the system may exhibit an essential spectrum, however for noisy systems this is small (see main text).

the factors in the exponent determines what features the encoder learns about the state x_t. In general, there may be a large number of non-zero factors f_n so that the learned features are difficult to interpret in terms of individual eigenfunctions ϕ_n. However, things become simple in the limit of small β, or high compression. When β is small the encoder is trivial: $p(h_t|x_t) = p(h_t)$ so the value h_t is assigned at random with no regard to the state of the system. No feature has been learned, and all factors f_n are equal to zero. As β is increased, the encoder undergoes a series of transitions at $\beta = \beta_1 < \beta_2 < \beta_3 ...$ where new features are allowed to pass through the bottleneck (Fig. 1c) [51–54]. The first transition happens at a finite value of β_1 when the first, most predictive, feature is learned.

Surprisingly, we find that at the first IB transition the vector of f_n coefficients is dominated by a single term f_1.

$$p^*_\beta(h_t|x_t) \approx \frac{p^*_\beta(h_t)}{\mathcal{N}(x_t)} \exp\left(\beta\, e^{\lambda_1 \Delta t} \phi_1(x_t) f_1(h_t)\right) \quad (8)$$

This is our main mathematical result, which we derive by considering a perturbative expansion of the IB objective for small f_n. A proof of Eq. 8 with clearly specified technical assumptions may be found in Section 4 of the SI. One notable assumption underpinning this result, in addition to Markovianity, is that the system is in a statistical steady state, so that x_t is sampled from the steady state distribution. When this assumption is violated, eigenfunctions of the transfer operator may cease to carry information about the system's dynamics and instead its pseudospectrum may become relevant [55]; see SI Section 4D.

The above result shows that in the limit of high compression the encoder's dependence on x_t is given by the first left eigenfunction $\phi_1(x_t)$, which is the slowest-varying function of the state under dynamics given by U. Therefore, Eq. 8 makes precise the intuitive statement that slow features are the most relevant for predicting the future. Our analytical result, while applying only to the dominant eigenfunction, is valid for arbitrary, including non-Gaussian, variables.

We further observe numerically that this picture holds true more generally: at successive IB transitions, the learned features correspond to successive modes of the transfer operator. This is consistent with the exact results known for the specific case of Gaussian IB, where the encoder learns successive eigenvectors of a matrix related to the covariance of the joint $X_t, X_{t+\Delta t}$ distribution at each IB transition [53].

Due to the smallness of f_1 at the transition, the optimal IB encoder couples only weakly to the eigenfunction ϕ_1, so that the dynamics of h_t will be dominated by noise. Given the maximal number of available bits in memory, or variables used in an analytical model, the IB solution will thus not make the best use of the resources (IB is optimal with respect to the amount of information retained, but

not the size of its representation [43]). However, we can use the optimal IB encoding as a filter to identify ϕ_1, and then "distill" it by discarding the noisy part of the encoding, effectively increasing the coupling between h_t and $\phi_1(x_t)$ while keeping the number of components fixed. In practice, this is done by taking the most likely value for the reduced state, $h_t = \arg\max_h p(h|x_t)$ (see SI Section 4E). Similar procedures are performed in other probabilistic dimensionality reduction approaches, *e.g.* probabilistic principal component analysis [56]. There one finds a stochastic mapping from the full state x to the reduced representation h, and then discards the noise to obtain the deterministic reduction.

Together this lays the foundation of an operational prescription for model reduction which is built on IB for identifying relevant, slow, features. As we show later, this insight can be leveraged to systematically learn these slow variables directly from data with neural networks [41] which in turn may be used as inputs to an equation learning pipeline.

III. INFORMATION DECAY AND THE SPECTRUM OF THE TRANSFER OPERATOR

To develop intuition for information in a dynamical system, we turn to the simple example of a Brownian particle in a double-well potential. This may represent, for example, a molecule with a single degree of freedom that transitions between two metastable configurations [57]. In the overdamped limit the state of the particle is completely determined by its position $X_t \in \mathbb{R}$, with dynamics given by the Langevin equation

$$\dot{x}_t = -\partial_x V(x_t) + \sigma\eta_t. \tag{9a}$$

$$V(x) = \frac{1}{4}(\mu - x^2)^2 \tag{9b}$$

Here, η_t is unit-variance white noise, σ controls its strength, and μ controls the shape of the potential $V(x)$.

The deterministic dynamical system undergoes a bifurcation at $\mu = 0$ (Fig. 2a). To quantify the amount of information about the future state $X_{t+\Delta t}$ contained in the initial state X_t we compute their mutual information (see SI for details). The dynamics of X_t are Markovian, so that for any sequence of times $t_0 < t_1 < t_2$, $p(X_{t_2}|X_{t_1}, X_{t_0}) = p(X_{t_2}|X_{t_1})$. From the data processing inequality, one has [58]

$$I(X_{t_2}, X_{t_0}) \leq I(X_{t_1}, X_{t_0}),$$

which implies that information can only decrease in time.

What governs the rate at which information decays? Here we can already see the role of the spectrum of the dynamics' transfer operator. By exploiting the spectral expansion of the conditional distribution $p(x_{t+\Delta t}|x_t)$ one finds that for long times the information decays as

$$I(X_t, X_{t+\Delta t}) = e^{2\lambda_1 \Delta t}\langle\phi_1^2\rangle\langle\rho_1^2/\rho_0^2\rangle + \mathcal{O}(e^{2\lambda_2\Delta t}) \tag{10}$$

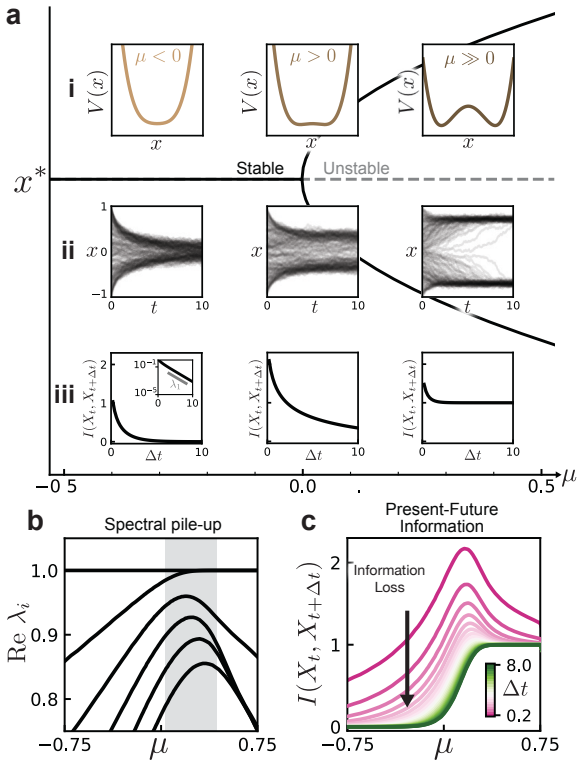

FIG. 2. **Information loss of a Brownian particle in a double well potential.** (a) Fixed point (FP) diagram of the dynamics given by Eq. 9 for zero noise. There is a bifurcation at $\mu = 0$ where the stable FP at $x = 0$ becomes unstable and two new stable FPs appear at $\pm\sqrt{\mu}$. (a,i) The corresponding potential $V(x)$ for varying μ, with the emergence of a double-well structure for $\mu > 0$. (a,ii) Dynamics of the system Eq. 9 for varying values of μ corresponding to the potentials above, with uniformly-distributed initial conditions. (a,iii) Loss of information between the initial condition and the future state. Inset shows scaling given by the first eigenvalue of the transfer operator. (b) Spectrum of the transfer operator U, showing a pile-up of eigenvalues for $\mu \gtrsim 0$. These are related to the eigenvalues of its infinitesimal generator by $\Lambda_i = e^{\lambda_i \Delta t}$. (c) Mutual information between the present and future state for varying time delay Δt and bifurcation parameter μ.

where expectations are taken over the steady state distribution (see SI Section 3). Asymptotically, the information decay is set by the value of λ_1, the rate of decay of the slowest-varying function $\phi_1(x)$ under the dynamics of U. In the limit of infinite time, for any value of μ even weak noise will cause the mutual information to become zero as there is a non-zero (perhaps exponentially small) probability of hopping between the wells [49].

The loss of information in time depends on the bifurcation parameter μ as summarized in Fig. 2c. Note the peak in $I(X_t, X_{t+\Delta t})$ for small, positive μ. This corresponds to dynamics where observation of X_t strongly informs the future state; recall that the mutual information is maximized when the conditional entropy $\mathcal{H}(X_{t+\Delta t}|X_t) \approx 0$

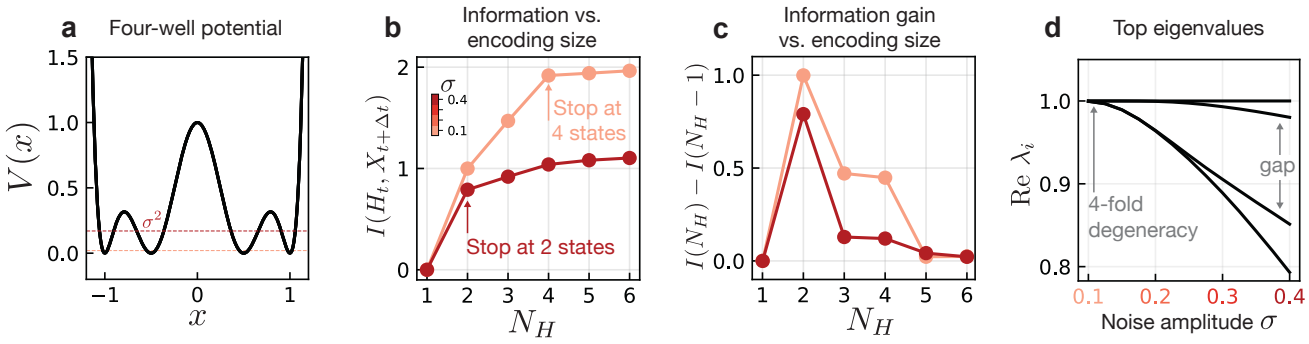

FIG. 3. **"Knowing when to stop".** The spectral properties of the transfer operator determine the necessary complexity (i.e. "when to stop" [33]) of the reduced model, which we show is also visible in information theoretic metrics. (a) Four-well potential in which a Brownian particle fluctuates. The magnitude σ of the fluctuating noise is related to an energy scale $E_\sigma = \sigma^2$. (b) Information contained in the encoding variable H_t about the future state $X_{t+\Delta t}$ for varying levels of noise and alphabet sizes N_H. (c) Information gain achieved by increasing the alphabet size by a single variable. This is the discrete derivative of the curve in (b). (d) Spectrum of the transfer operator for changing values of noise amplitude.

(see SI Section 1). In contrast, for large positive or negative μ, X_t is not as informative of $X_{t+\Delta t}$ even for small times: the initial state is quickly forgotten as the particle approaches the bottom of the single (for $\mu < 0$) or double (for $\mu > 0$) well.

This phenomenon is reminiscent of critical slowing down, which occurs in the noise-free system as μ passes through the bifurcation at $\mu = 0$. For the deterministic dynamics, the slowing down is reflected in the spectrum as a "pile up" of eigenvalues to form a continuous spectrum [49]. In the presence of noise, although the continuous spectrum becomes discrete [49, 50] there is still a pile-up of eigenvalues characterized by several eigenvalues becoming close to 1 (Fig. 2b). This pile-up gives rise to the information peak seen in Fig. 2c. The peak is not solely due to the closing spectral gap $\lambda_1 - \lambda_2$, but is also impacted by the subdominant eigenvalues which accumulate at $\mu \approx 0.2$ (SI Fig. S2).

IV. KNOWING WHEN TO STOP

For discrete encoding variables h, the information bottleneck partitions state space and reduces the dynamics on x to a discrete dynamics on h. Such reductions of complex systems to symbolic sequences via partitioning of state space has attracted attention for more than half a century in both theoretical and data-driven contexts [4, 59–64]. Several works have approached this partition problem from a dynamical systems perspective, linking optimal partitions to eigenfunctions of the (adjoint) transfer operator [33, 65]. In this setting, a central question is "when to stop" [4, 33, 34, 63]: how many states does h need in order to capture statistical properties of the original dynamics?

We consider this question by finding the optimal IB encoder in the limit of low compression, $\beta \gg 1$, but fixed encoding capacity N_H (where $H_t \in \{0, ..., N_H - 1\}$), i.e. the encoder is only restricted by the number of symbols it can use. An analogous setup was used in the context of renormalization group (RG) transformations in [66–68], which results in effective model reduction due to the "sloppiness", or irrelevance, of certain system variables [69, 70]. In this regime, the encoder learned by IB is deterministic; we are learning an optimal hard partition of state space. This can be seen by noting that $I(H_t; X_{t+\Delta t}) = \mathcal{H}(H_t) - \mathcal{H}(H_t|X_{t+\Delta t})$ is maximized when the latter term is zero, which happens when x_t unambiguously determines h_t, i.e. when $p(h_t|x_t) \in \{0, 1\}$ for all x_t. The details of how the encoder is computed are discussed in the next section.

Consider a fluctuating Brownian particle as in the double well above, where now each of the wells is split into two smaller wells, giving a total of four potential minima (Fig. 3a). As the system is in steady state, the variance of the fluctuations defines an energy scale $E_\sigma = \sigma^2 = 2k_\mathrm{B}T$. For small E_σ, the system rarely transitions between the four potential minima. In this case, knowledge of the initial minimum is very informative of the future state of the particle. In contrast, for large fluctuations the particle can spontaneously jump between shallow minima in each large well, so that the system immediately forgets about the precise potential minimum it was in. Information about the shallow minima has been "washed out", and only the information about the larger double-well structure remains.

To see this reflected in the information, we again consider an encoding of the initial state into a discrete variable $H_t \in \{0, ..., N_H - 1\}$. In both the small and large noise scenarios, a variable with $N_H = 2$ encodes approximately one bit of information (Fig. 3b), corresponding to an H_t which distinguishes the two large wells for $x \lessgtr 0$. For large noise this is essentially all the information that can be learned; increasing the capacity of the encoding variable beyond this provides only marginally more information about the future state (Fig. 3c). In the small

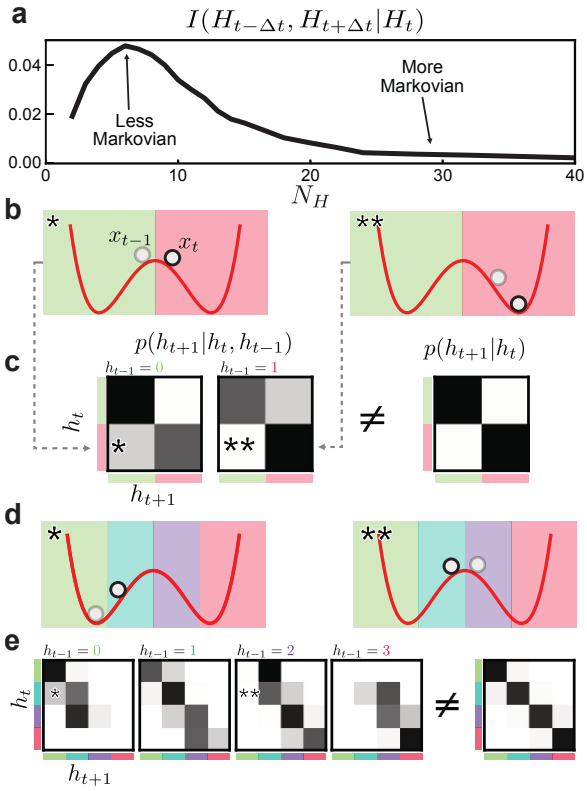

FIG. 4. **Non-Markovianity for a Brownian particle in a double well potential.** (a) Conditional mutual information $I(H_{t-\Delta t}, H_{t+\Delta t}|H_t)$ as a function of N_H. (b) Two possible scenarios (left and right) with the same value of h_t, but different h_{t-1}. Here, the circles represent the value of the full state x_t while the colored regions represent the value of the reduced state $h_t \in \{1, 2\}$ which are mapped to for each x_t. (c) Illustration of the transition probability which takes into account the previous state h_{t-1} (left), compared to the Markovian transition probability obtained by marginalizing out h_{t-1} (right). In a Markovian system, these transition probabilities would coincide. (d-e) Same as in (h), but for $N_H = 4$ states.

noise case, the information between the encoding and the future state continues to increase to approximately two bits at $N_H = 4$, after which it plateaus. The encoding has learned to distinguish each of the four potential wells. These observations are reflected in the transfer operator spectrum shown in Fig. 3d. For small noise, the eigenvalue $\lambda = 1$ is nearly four-fold degenerate, indicating the existence of four regions that can evolve independently under U, giving rise to four steady state distributions satisfying $U\rho = \rho$. These regions correspond to the potential minima. Hops between the separate minima are exceedingly rare, so that the dynamics essentially take place in the four minima independently. With larger σ the degeneracy is lifted, resulting in one dominant subleading eigenvalue followed by a gap. The corresponding eigenfunction is one which is positive (negative) on the right (left) side of the large potential barrier at $x = 0$: the only relevant piece of information is which of the large wells the initial condition is contained in, and all other information is lost exponentially quickly.

In many cases, our goal is not only to find a reduced representation H_t that is predictive of the future state, but one which may be evolved forward for long times as a surrogate for the full system. This is a practical but much more stringent requirement of our reduced state, and it may result in a different answer to the question of "when to stop."

In general the dynamics of the reduced variable will be non-Markovian, meaning that its evolution will be determined not only by its current state, but also its past. With a fixed encoder $p(h_t|x_t)$, transition probabilities $p(h_{t+1}|h_t)$ are induced by decoding h_t to x_t, evolving x_t forward in time to x_{t+1}, and then encoding x_{t+1} to h_{t+1} (here we assume discrete time dynamics with $\Delta t = 1$ for simplicity). Information theory provides a natural quantification of the (non-)Markovianity of the dynamics as measured by the conditional mutual information $I(H_{t-1}, H_{t+1}|H_t)$ (see SI Section 3 and alternative approaches in Refs. [63, 71, 72]). This quantity tells us how much additional information the past state H_{t-1} contains about the future H_{t+1}, on top of what is contained in the present state H_t.

To illustrate this, we again consider a particle in a double well, which we compress into a discrete variable $H_t \in \{0, ..., N_H - 1\}$. Because the full system is Markovian by design, we expect that for large N_H the conditional mutual information $I(H_{t-1}, H_{t+1}|H_t)$ should go to zero as H_t begins to more closely approximate the full state X_t (Fig. 4a). Interestingly, the approach to Markovianity is highly non-monotonic, and the system is more non-Markovian with $N_H = 4$ than with $N_H = 2$.

The role of memory is illustrated in the two cases shown in Fig. 4b. In both cases, the current state is in the right well, so that $h_t = 1$. However, more precise information about the state (and hence its future) is revealed by knowing the system's past. If $h_{t-1} = 0$ (so the state was in the left well), it is likely that the full state is near the boundary between wells, so that $h_{t+1} = 0$ is more likely than if $h_{t-1} = 1$, which would suggest that the state is deep in the right well. This behavior can be quantified by computing the transfer probabilities $p(h_{t+1}|h_t, h_{t-1})$ (Fig. 4c). These differ significantly from $p(h_{t+1}|h_t)$, which would not be the case for Markovian dynamics. From these, we see that the primary contribution to the non-Markovian dynamics occurs when $(h_{t-1}, h_t) = (0, 1)$ or $(1, 0)$, which can only happen when the state is very near to the boundary between the wells. As N_H increases, the likelihood that the state is near the boundary of two regions with different h_t labels will increase, which will increase the non-Markovianity of the system (Fig. 4d-e). However, at the same time the different h_t regions will shrink in size so that h_{t-1} will no longer provide any meaningful information about the system's location x_t. The trade-off between these two effects leads to the observed non-monotonic behavior of $I(H_{t-1}, H_{t+1}|H_t)$.

V. DATA-DRIVEN DISCOVERY OF SLOW VARIABLES

IB finds a reduced state variable by optimizing an information theoretic-objective that makes no reference to physics or dynamics. This suggests it may be used for the discovery of slow variables in situations where one lacks physical intuition. In the regime of small β, or high compression, features of the state x_t are forced to compete to make it through the bottleneck h_t, which allows us to extract relevant features of the system.

These variables coincide with left eigenfunctions of the transfer operator, as discussed in Section II. We verify this theoretical result numerically in cases where the IB objective (3) can be solved exactly by using an iterative scheme known as the Blahut-Arimoto algorithm [38, 58] (see SI Section 3). By taking the logarithm of the encoder as it begins to deviate from the trivial uniform encoder above β_1 we can confirm that it depends on x only through $\phi_1(x)$ (SI Fig. S3). This can be independently verified by studying the stability of the trivial encoder with respect to perturbations in the parameters $f_n(h_t)$ in Eq. 7, where we see that the encoder becomes unstable to perturbations in f_1, as predicted by our theoretical results (SI Section 4).

The utility of exact IB for variable discovery is limited because it requires knowledge of the exact conditional distribution $p(x_{t+\Delta t}|x_t)$ which is difficult to estimate in many real-world scenarios. Fortunately however, the IB optimization problem can be replaced by an approximate variational objective introduced in Ref. [41] that can be solved with neural networks and which we refer to as variational IB. This is achieved by introducing a tractable Ansatz for the form of $p(h_t|x_t)$; we show in SI Section 4 that this does not change the encoder's dependence on transfer operator eigenfunctions.

Our core result linking the behavior of the learned encoder with transfer operator eigenfunctions remains valid even for variational IB applied to high-dimensional systems, which we show by considering a simulated data of fluid flow past a disk [74]. The state of the system is given by a two-dimensional velocity field $\mathbf{v}(\mathbf{x}) \in \mathbb{R}^{2\times N_{\text{pixels}}}$, where $N_{\text{pixels}} \sim \mathcal{O}(10^5)$ (Fig. 5a). Fluid flows in from the left boundary with a constant velocity $v_0\hat{e}_x$ past a disk of unit diameter. At Reynolds number Re $\gtrsim$ 150, the fluid undergoes periodic vortex shedding behind the disk, forming what is known as a von Kármán street.

The eigenfunctions for this system are in general complex, and come in conjugate pairs: $\phi_2(x) = \phi_1^*(x)$. In this situation any linear combination of ϕ_1 and ϕ_2 will decay at the same rate, and hence we expect to learn some arbitrary combination of the two dominant eigenfunctions, or equivalently a combination of the real and imaginary parts of ϕ_1. We therefore take a two-dimensional encoding variable $[h_0, h_1]$, so that it can represent the full complex eigenfunction.

Our learned latent variables are oscillatory with the correct frequency as shown in Fig. 5b. However, this

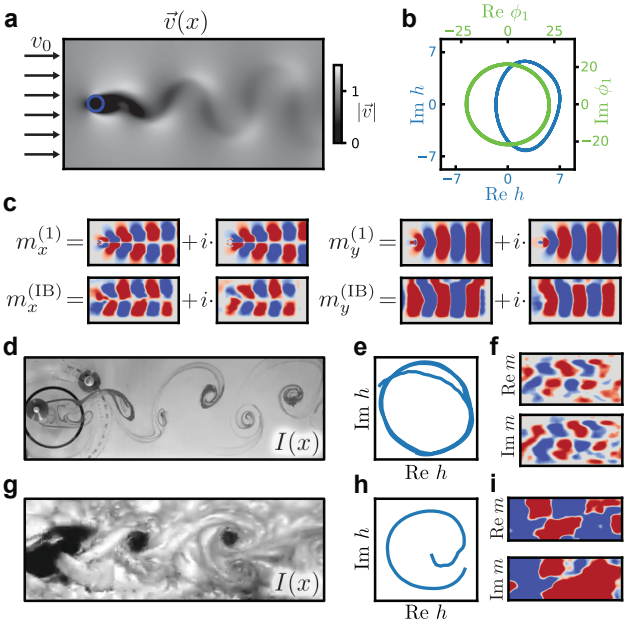

FIG. 5. **Variational IB for high-dimensional simulated fluid flow.** (a) Fluid flow past a disk with uniform in-flow velocity v_0 exhibits vortex shedding behind the object in a so-called von Kármán street. The state of the system is given by a spatially varying two-component vector field $\mathbf{v}(\mathbf{x})$. (b) Dynamics in latent space (blue) traverse a nearly circular trajectory (see Supplementary Movie 1). For comparison we show the evolution of the mode amplitudes obtained by projecting the velocity field onto the first DMD mode (green). (c) Comparison of the first Koopman mode obtained from DMD ($\mathbf{m}^{(1)}$) and from VIB ($\mathbf{m}^{(\text{IB})}$). Koopman modes from VIB are computed as gradients of the latent encoding variables as described in the main text. Red corresponds to positive values and blue to negative; the magnitudes of the modes are not directly comparable. (d) Intensity field $I(x)$ from a video of water flowing past an obstacle; video from Ref. [73]. (e) Trajectory of the system in latent space. (f) Gradients of the latent encodings with respect to the input field $I(x)$. (g-i) Same as (d-f), but using satellite images of clouds near the island of Guadalupe as our state $I(x)$.

alone does not indicate whether we are learning the correct eigenfunctions. Because the system is well approximated by linear dynamics, eigenfunctions of the adjoint transfer operator are linear functions of the state variable, and may hence be represented as $\phi_n[\mathbf{v}] = \langle \mathbf{v}(\mathbf{x}), \mathbf{m}^{(n)}(\mathbf{x})\rangle$, where angled brackets denote an average over space and $\mathbf{m}^{(n)}(\mathbf{x})$ are so-called Koopman modes. The true $\mathbf{m}^{(n)}$ (and hence the eigenfunctions) can be thus computed via dynamic mode decomposition (DMD) [15, 16] as described in the SI. For the functions $h[\mathbf{v}]$ learned using variational IB, we can infer the corresponding $\mathbf{m}^{(\text{IB})}$ using gradients of the neural network with respect to the input field. Fig. 5c shows these inferred modes $\mathbf{m}^{(\text{IB})}$ compared to the true mode $\mathbf{m}^{(1)}$ (see SI Section 5 for details). This shows that variational IB not only recovers the essential oscillatory nature of the dynamics, but does so by learning the correct

slowly varying functions of the state variable given by the adjoint transfer operator eigenfunctions.

Our framework provides interpretability for the learned latent variables even in messy real-world fluid flow datasets scraped directly from videos on Youtube [73, 75] (Supplementary Movie 1). The first shows a von Kármán street which forms as water passes by a cylindrical obstacle at Reynolds number 171, with flow visualized by a dye injected at the site of the obstacle [73]. We take a background-subtracted grayscale image of the flow field as our input (Fig. 5d) and task VIB with learning a two-dimensional latent variable as above. Also here, variational IB learns oscillatory dynamics of the latent variables (Fig. 5e). We visualize the function learned by the encoder by considering gradients of the latent variables, which show the same structure as those obtained for the x component of the simulated data (Fig. 5f). This is expected, as the x-component of the velocity field has similar glide reflection symmetry as the intensity image.

We also apply variational IB to a von Kármán street arising due to air flow around Guadalupe Island, which was imaged by a National Oceanic and Atmospheric Administration (NOAA) satellite [75] (Fig. 5g). Although the video consists of only 62 frames and less than one full oscillation, the variational IB neural network learns latent variables which capture this oscillation and have the expected dependence on the input variables (Fig. 5h-i). As in the first experimental example, the gradients of the encoding variables show the glide symmetry of m_x due to the symmetry of the intensity pattern in Fig. 5g.

VI. LEARNING THE DYNAMICS OF LATENT VARIABLES

In Section I we showed how using an information-theoretic measure of predictability reduces the problem of model reduction to the problem of identifying relevant variables. In this setting, the dynamical rule governing the evolution of H_t is not explicitly learned, but is instead induced by the choice of encoding variable. We now show how variational IB can be incorporated into a complete model reduction pipeline by combining it with sparse equation learning methods such as SINDy [76, 77]. We consider a system described by an underdamped, driven, one-dimensional PDE known as the sine-Gordon equation [78] (Fig. 6a; see SI Section 9). For a particular choice of driving frequency and strength, this system undergoes quasiperiodic dynamics, and hence lives on a toroidal manifold.

To identify the reduced variables for the system, we must make a choice about the dimensionality of H_t. As discussed above, the mutual information between H_t and $X_{t+\Delta t}$ provides an indicator of "when to stop" (Fig. 6b). This metric plateaus already at $\dim H_t = 2$. The fractal dimension of the latent manifold, however, does not plateau until $\dim H_t = 3$ (Fig. 6c). Interestingly, when using the encoding to learn a deterministic equation for the evolution of H_t using SINDy, the prediction error is only minimized at $\dim H_t = 4$ (Fig. 6d).

Each of these indicators of "when to stop" tells us something different about the system. The encoding learned for $\dim H_t = 2$ is shown in Figure 6e, with a portion of the trajectory highlighted. Although the full dynamics takes place on a torus, the encoding neural network is able to "cut" the torus to embed it in two dimensions. The dynamics of the latent variables are completely deterministic: at any point on the latent manifold it is known where the system will evolve next, so that the information about the future is complete. However, the dynamics will be highly irregular, featuring large jumps in latent space (Fig. 6f).

The "fraying" evident at the boundary of the torus embedded in two dimensions leads to a fractal dimension which is less than two, although this difference may disappear in the limit of infinite data. At three dimensions, the latent manifold is two dimensions everywhere (Fig. 6c). However, while the torus may be embedded in three dimensions, it does not admit simple, linear dynamics until $\dim H_t = 4$. In this dimension, SINDy is able to identify the correct linear equations of motion

$$\begin{aligned} \dot{h}_0 &= -\omega_1 h_1 & \dot{h}_2 &= \omega_2 h_3 \\ \dot{h}_1 &= \omega_1 h_0 & \dot{h}_3 &= -\omega_2 h_2 \end{aligned} \tag{11}$$

with $\omega_1 = 1.5$ and $\omega_2 = 8.7$(Fig. 6g-h).

The above example illustrates the subtle difference between a model being the simplest in the sense of providing a minimal predictive set of variables, and the simplest in terms of the governing equations. The latter may benefit from the inclusion of informationally redundant variables. The IB objective, being purely information based, is indifferent to the geometric or analytic properties of the encodings. Such model properties, often desirable, as *e.g.* smoothness or linearity must be introduced as additional constraints.

VII. RELEVANT VARIABLE DISCOVERY IN CYANOBACTERIAL POPULATIONS

We now demonstrate how variational IB may be used as an aid for collective variable discovery in situations where physical intuition may not be a useful guide – collective behavior of biological organisms (Supplementary Movie 2). Here, we ask what the most predictive variables are for predicting the evolution of populations of cyanobacteria (*Synechococcus elongatus*). The dynamics of the colonies are driven by several factors: growth and division of individual bacteria, translational motion of groups of bacteria as they are pushed by their neighbors, as well as the circadian oscillations within each bacterium (Fig. 7a). These oscillations are controlled by three Kai proteins [79] and depend in particular on the ratios of the copy number of these proteins which can be tuned experimentally [80].

We were provided with videos of 10 cyanobacteria colonies that were grown under various conditions that im-

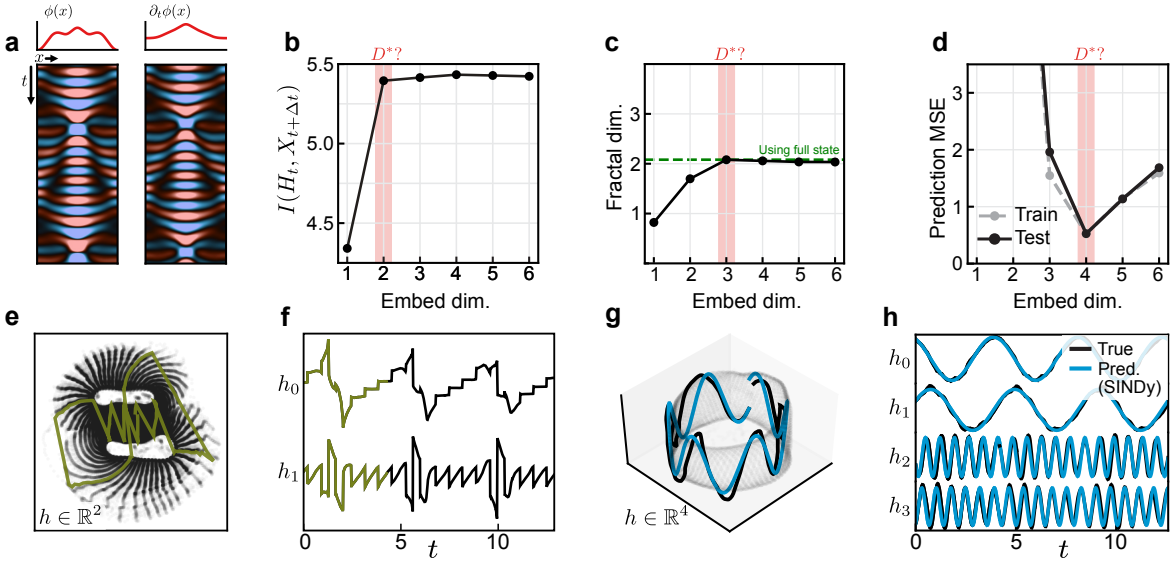

FIG. 6. **IB as part of an equation learning pipeline.** (a) Kymographs of the sine-Gordon system, discussed in the text. Left shows the field $\phi(x)$ while the right shows time-derivative of the field. Initial conditions are shown above as curves. (b) Information content contained in the encoding for varying dimensionality of the embedding space. While the information appears to plateau at $D^* = 2$, this does not necessarily mean a "proper" embedding dimension has been reached. (c) Estimated fractal dimension of the embedded dynamics for varying embedding dimension. The full fractal dimension of ≈ 2 is only attained with an encoding dimension $D^* = 3$. (d) Error of the predictions for the latent variable evolution produced by the dynamical system learned by SINDy. The error reaches a minimum at $D^* = 4$. Interestingly the error begins to rise for $d > 4$, likely due to the fact that the IB encoding begins to store other features of the full state variable which are irrelevant for the dynamics. (e) For $d = 2$, the encoding manages to project the full torus into two dimensions by cutting and unfolding it (left). This introduces discontinuous jumps in the evolution of h. (f) Time evolution of both components of h, with highlighted part corresponding to the trajectory highlighted in (e). (g) Trajectory for a four-dimensional latent variable, visualized by projected onto the first three principal components. True trajectory shown in black, trajectory predicted by SINDy shown in blue with the equations given in the main text. (h) Predicted trajectory for all the degrees of freedom of h.

pact their dynamics. However, as a test of our method, we were blinded to these conditions until we had performed our analysis. The videos are sequences of fluorescent images, taken once per hour, which show the clock state of each individual bacteria visualized with a fluorescent marker EYFP driven by the *kaiBC* promoter. Here, we focus on collective variables which are predictive of the state of the interior of the colony and not the growth in area of the colonies. We therefore crop the images to the interiors of each colony (SI Fig. S13). This allows us to isolate the motion of individual bacteria and fluorescence oscillations.

Our input to the variational IB neural network are these cropped images augmented with a time-lagged image of the same region (Fig. 7b-c). The purpose of this time lag is to make the dynamics Markovian: the intensity field oscillating, the observation at a single time point does not allow to determine whether the intensity is increasing or decreasing. In principle more time-lags may be necessary, and their number could be estimated from the behavior of the mutual information (see SI Section 3) or the transition entropy [63] . As these quantities are difficult to measure from images with limited computational resources, we keep only one lag: the state is given by pairs $X_t = \{I(\mathbf{x}, t), I(\mathbf{x}, t + \tau)\}$, where τ is the lag duration. Here we take $\tau = 3$ hr and a prediction horizon $\Delta t = 8$ hr, but find that different Δt or τ does not change our results (SI Fig. S13).

With variational IB we compress the state X_t into a latent variable h of variable dimension. We train the neural network on the entire dataset of all 10 colonies simultaneously. The dynamics in latent space undergo clear oscillations, indicating that the relevant variables encode primarily the intensity fluctuations rather than, for example, the spatial locations of the bacteria. Notably, the trajectories are essentially two-dimensional, even when the encoding space is higher dimensional. This is reflected in the information retained about the future state, $I(X_{t+\Delta t}, H_t)$: increasing the dimension of the embedding space beyond two only marginally increases $I(X_{t+\Delta t}, H_t)$; this tells us "when to stop" (Fig. 7d). We independently verify this by using principal component analysis to characterize the geometry of embedded trajectories, and find that even in higher dimensions the trajectories occupy a two dimensional subspace (SI Fig. S10). In the following, we thus restrict our focus to $\dim H = 2$.

We noticed that there were notable differences in the radius of latent space oscillations from colony to colony,

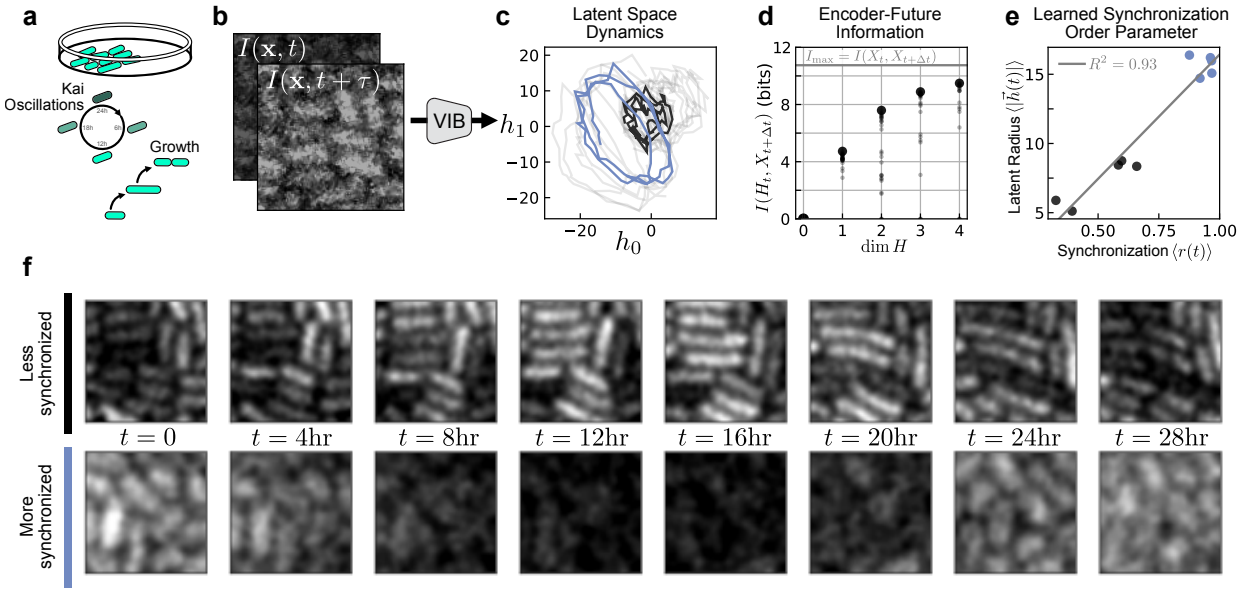

FIG. 7. **Discovering slow collective variables in cyanobacteria populations** (a) Fluorescent images of cyanobacteria colonies labeled with EYFP driven by the *kaiBC* promoter, allowing the visualization of Kai protein transcription. The colonies are imaged as they undergo cell growth and oscillations in Kai expression associated with the circadian rhythm. (b) The "state" used for variational IB is the time-lagged intensity field with lag time τ (see SI Fig. S13 for details). (c) Variational IB embeddings of time-lagged images into two dimensions (see Supplementary Movie 2). Every line corresponds to one colony's evolution in time. Note the apparent oscillations of different radii. Three-dimensional embedding is shown in SI Fig. S15. One large-radius (blue) and one small-radius (black) trajectory are highlighted. (d) Mutual information between the future state and the encoding, given by $I(X_{t+\Delta t}, H_t)$, for varying dimension of the latent space. Each small point represents one training instance of the variational IB model, while the large point shows the maximum estimated value. $I_{\max} = I(X_t, X_{t+\Delta t})$ is the mutual information estimated for the full dynamics. (e) Mean radius versus synchronization parameter for each colony (see SI Section 8). VIB identifies clusters of cells characterized by high (blue) or low (black) synchronization. This clustering corresponds to differing theophylline concentrations across experiments. Within each experimental movie (each of which contains 2-3 colonies) the radii are mostly constant. (f) Sample time series of weakly (black) and highly (blue) synchronized colonies. We apply a slight Gaussian blur to better visualize the bacteria boundaries.

two of which are highlighted in (Fig. 7c). To understand this difference, we examined the original microscopy time series corresponding to both large and small latent radius (Fig. 7e-f) and found that while the large-radius sample showed clear, nearly uniform oscillations in intensity, the small-radius samples appeared much more heterogeneous.

To quantify this we consider each pixel to be an independent oscillator, akin to a spatial Kuramoto model [81–83], and compute a global synchronization order parameter $r(t)$ (see SI Section 7). For each colony we calculate the time-averaged synchronization $\langle r(t)\rangle_t$ and find that two clusters emerge corresponding to high and low synchronization (Fig. 7e). These clusters are precisely those representing trajectories of large and small latent radius, suggesting that variational IB learns to encode the synchronization of the colony in the latent variable radius. As a check, we perform IB on a simulated locally-coupled Kuramoto model as a system sharing many features of the experimental system. Here we also learn an encoding in which the latent radius corresponds to the synchronization order parameter (SI Fig. S11).

In the SI we compare the performance of variational IB to several other model reduction methods and find that IB delivers more interpretable and well-behaved features. This is likely due to the fact that many standard methods for data-driven model reduction rely on assumptions about the dynamics which may not be appropriate in the case at hand, such as linearity. Even among deep learning methods free of such assumptions, such as time-lagged autoencoders, the variables learned by IB appear more interpretable. This increased interpretability is likely due to the compression term which effectively regularizes the latent space by encouraging the network to learn slow transfer operator eigenfunctions. While there are many specific variants of DMD [17, 84–87] or autoencoders for dynamics [28–30, 88] that may outperform variational IB in some cases, we find that in this real-world example it yields the smoothest and most interpretable latent variables without a tailored pre-processing (SI Fig. S17).

By using variational IB we could reduce a complex system with multiple dynamical components – cell growth, division, and gene expression fluctuations – into a low dimensional form that retains only the most relevant information for the future. In addition to the insight that the

dynamics are dominated by oscillations in two dimensions, the latent variables clearly distinguished trajectories into two groups that were not apparent *a priori*. We were provided this data as a "blind" test with no knowledge of the underlying system. After we performed our analysis, it was revealed to us that these bacterial colonies have been engineered to control the translational efficiency of the Kai proteins by varying theophylline concentration [80]. The synchronization order parameter discovered by variational IB corresponds to differing experimental concentrations of theophylline, which is in agreement with the findings in Ref. [80]. IB can thus serve as a way to connect experimental control parameters to effective changes in dynamics.

VIII. CONCLUSION

We have related information-theoretical properties of dynamical systems to the spectrum of the transfer operator. We illustrate our findings on several simple and analytically tractable systems, and turn them into a practical tool using variational IB, which learns an encoding variable with a neural network. The latent variables of these networks can be interpreted as transfer operator eigenfunctions even though the network was not explicitly constructed to learn these: it optimizes a purely information-theoretic objective that contains no knowledge of a transfer operator or dynamics. This allows one to harness the power of neural networks to learn physically-relevant latent variables and, combined with methods such as SINDY, even complete reduced models with their governing equations. Biological systems are an ideal setting for such methods: despite their apparent complexity, they can often be captured by low-dimensional descriptions which are difficult to identify by physical considerations alone [3, 62, 89, 90]. We have shown that variational IB is a potentially powerful tool for these cases, and can discover slow variables even directly from image data without significant preprocessing. In the SI we showcase further applications, including to chaotic systems, where they may yield computationally efficient methods to calculate *e.g.* fractal dimensions.

IX. METHODS

Information processing inequality

Here we show how to obtain (2) from (1). The variables we are considering form the Markov chain

$$X_{t+\Delta t} - X_t - h_{\text{enc}}(X_t) - \mathcal{S}^{\Delta t} h_{\text{enc}}(X_t) - h_{\text{dec}}^{-1}(\mathcal{S}^{\Delta t} h_{\text{enc}}(X_t))$$

where $X - Y - Z$ means that Z is conditionally independent of X, given Y: $p(z, x|y) = p(z|y)p(x|y)$. As a consequence of the data processing inequality, for any transformation $Z = g(Y)$ one has

$$I(X, g(Y)) \leq I(X, Y).$$

where equality is obtained if g is a bijection [58]. Consequently, we have

$$\begin{aligned} I(h_{\text{dec}}^{-1}(\mathcal{S}^{\Delta t} h_{\text{enc}}(X_t)), X_{t+\Delta t}) &\leq I(\mathcal{S}^{\Delta t} h_{\text{enc}}(X_t), X_{t+\Delta t}) \\ &\leq I(h_{\text{enc}}(X_t), X_{t+\Delta t}), \end{aligned}$$

where the inequalities are maximized if h_{dec}^{-1} and $\mathcal{S}^{\Delta t}$ are both bijective functions. Notably, $\mathcal{S}^{\Delta t} =$ identity is a solution. However, this should not be taken to mean that H_t has no dynamics, as it will still have dynamics induced by the evolution of X_t.

The information bottleneck

The information bottleneck [38] is an example of a rate-distortion problem which seeks to find an optimal compression which minimizes some distortion measure with the original signal [58]. Concretely, we call X the source signal, and let H denote the compressed signal. In IB, rather than using an *a priori* unknown distortion function, one seeks to ensure that the compression retains information about an additional relevance variable Y. As noted in the main text, the IB optimization objective is given by the Lagrangian

$$\mathcal{L}_{\text{IB}}[p(h|x)] = I(X, H) - \beta I(Y, H), \tag{12}$$

where in our case the source signal X is the state of the system X_t at time t, and the relevance variable is the state of the system $X_{t+\Delta t}$ at a future time $t + \Delta t$. The encoder which optimizes this objective can be solved for exactly and is given by [38]

$$p(h|x) = \frac{p(h)}{N(x)} \exp\left[-\beta D_{KL}(p(y|x)\|p(y|h))\right]. \tag{13}$$

Encoder in terms of transfer operator eigenfunctions

To connect the optimal encoder to the transfer operator, we first rewrite (13) in terms of the transition probabilities,

$$p(h_t|x_t) = \frac{\tilde{p}(h_t)}{\tilde{N}(x_t)} \exp\left[\beta \int \mathrm{d}x_{t+\Delta t} p(x_{t+\Delta t}|x_t) \log p(x_{t+\Delta t}|h_t))\right]. \tag{14}$$

where we have absorbed terms in the exponent which depend only on h_t or x_t into the normalization factors. Into the above equation, we replace the transition probability with the spectral decomposition

$$p(x_{t+\Delta t}|x_t) = \sum_n \mathrm{e}^{\lambda_n \Delta t} \rho_n(x_{t+\Delta t}) \phi_n(x_t). \tag{15}$$