nature reviews physics

https://doi.org/10.1038/s42254-024-00688-2

Perspective

Check for updates

Generative learning for nonlinear dynamics

William Gilpin [1,2]

Abstract

Modern generative machine learning models are able to create realistic outputs far beyond their training data, such as photorealistic artwork, accurate protein structures or conversational text. These successes suggest that generative models learn to effectively parametrize and sample arbitrarily complex distributions. Beginning half a century ago, foundational works in nonlinear dynamics used tools from information theory for a similar purpose, namely, to infer properties of chaotic attractors from real-world time series. This Perspective article aims to connect these classical works to emerging themes in large-scale generative statistical learning. It focuses specifically on two classical problems: reconstructing dynamical manifolds given partial measurements, which parallels modern latent variable methods, and inferring minimal dynamical motifs underlying complicated data sets, which mirrors interpretability probes for trained models.

Sections

[1] Department of Physics, The University of Texas at Austin, Austin, TX, USA. [2] The Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, USA. e-mail: wgilpin@utexas.edu

Introduction

The fractal geometry of a strange attractor can only be visualized by watching a chaotic system evolve for an extended duration. Chaotic systems therefore continuously produce information, which gradually reveals their structure at ever-decreasing scales[1–3]. The notion of information production by chaotic systems inspired early efforts to frame computation as a physical theory, including Richard Feynman's estimation of the information stored in an ideal gas[4] and John Archibald Wheeler's analogies between travelling salesman algorithms and molecular chaos[5]. Wheeler would later declare 'it from bit' — that physical theories ultimately encode computational primitives[6].

Contemporaneous to Wheeler's remark, work by the dynamical systems community formalized information production by chaotic systems[7–11]. Continuous-time chaotic systems encountered in the natural world, from turbulent fluid cascades to intertwined stellar orbits, act as analogue computers that manipulate and transform information encoded in their initial conditions and parameters[3,12,13]. Given a chaotic dynamical system $\mathrm{d}\mathbf{x}(t)/\mathrm{d}t = \mathbf{f}(\mathbf{x}(t))$, Pesin's formula states that its entropy production rate is proportional to the sum of its positive Lyapunov exponents[12] (Fig. 1a), which measure the rate at which nearby trajectories diverge along different directions on a chaotic attractor:

$$\mathcal{H} = \sum_{\lambda_i > 0} \lambda_i.$$

The entropy $\mathcal{H}$ represents the Kolmogorov–Sinai entropy, which can be estimated by coarse-graining the phase space of the system with infinitesimal bins and then calculating the probability of the system occupying each bin over an extended period[14]. Pesin's formula thus relates properties of the dynamics to the rate of information production of the system $\mathbf{f}(\mathbf{x})$; systems with greater chaoticity reveal more quickly the points on their attractor (Fig. 1a). The formula therefore connects dynamics, attractor geometry and information in the evolution of chaotic systems.

Ongoing developments in statistical learning motivate revisiting older results on information production by chaotic systems. Many machine learning algorithms implicitly estimate the underlying distribution of possible input values $p(\mathbf{x})$ based on a finite number of inputs $\mathbf{x}$ seen during training[15]; the information gained in the sampling process is related to the geometry of the underlying distribution. Supervised learning algorithms seek to construct the conditional probability distribution $p(\mathbf{y}|\mathbf{x})$ of a target state $\mathbf{y}$ given knowledge of an input $\mathbf{x}$. In image classification, $\mathbf{y}$ comprises a discrete label for an image $\mathbf{x}$; in forecasting, $\mathbf{y}$ represents the future system state conditioned on the past observations $\mathbf{x}$. By contrast, unsupervised learning constructs a map between the estimated $p(\mathbf{x})$ and a latent space $p(\mathbf{z}|\mathbf{x})$ in which underlying patterns in the data become apparent. Generative models, such as generative adversarial networks or variational autoencoders, seek to sample $p(\mathbf{x})$ to produce new examples $\mathbf{x}'$ resembling training data cases. These methods either directly sample a smooth estimate of $p(\mathbf{x})$ constructed from the training data, or instead sample the latent space $p(\mathbf{z}')$ and then decode the result through the inverse transformation $p(\mathbf{x}'|\mathbf{z}')$ (ref. 15).

However, because generative models are frequently used in applications in which $\mathbf{x}$ is high dimensional, drawing representative samples from $p(\mathbf{x})$ often proves difficult owing to the curse of dimensionality, leading to a high sample rejection rate. Methods suitable for high-dimensional distributions such as Markov chain Monte Carlo select the next sample $\mathbf{x}_{i+1}$ based on the previous accepted sample $\mathbf{x}_i$. A simplified such scheme (Fig. 1b) centres a Gaussian proposal distribution around the current sample, $\mathbf{x}_{i+1} \sim \mathcal{N}(\mathbf{x}_i, \boldsymbol{\Sigma})$, with the covariance matrix $\boldsymbol{\Sigma}$ aligned with the local geometry of distribution as estimated from gradient information or previous samples. A sample drawn from this proposal distribution provides an amount of local information given by

$$\mathcal{H} = \ln(2\pi e)^{N/2} + \sum_{i=1}^{N} \ln\sigma_i, \qquad (1)$$

in which σ_i denotes the standard deviation of the N-dimensional distribution along its ith principal axis. Similar to Pesin's formula, this expression relates novelty in the form of information gain to local geometric properties of the underlying manifold. Just as chaotic systems diverge along unstable manifolds associated with positive Lyapunov exponents, complex data distributions contain may flatter local directions that dominate their apparent diversity[16,17].

Consistent with this connection between dynamics and sampling, many large-scale generative models implement sampling schemes that may be viewed as dynamical systems mapping input variables $\mathbf{x}$ to latent variables $\mathbf{z}$[18,19]. When these models are applied to dynamical data sets, latent representations reveal physical properties such as attractors and invariants, enabling recent scientific applications of generative models to spiking neurons[20,21] and turbulent flow time series[22].

These applications therefore motivate revisiting it from bit and situating emerging work on statistical learning within classical works on information processing in chaotic systems. Chaos and statistics have well-established connections through ergodic theory[2,14,23], which stimulated the development of early statistical methods for identifying the dynamical systems that act as generators for experimental time series[10,24,25]. Although several reviews highlight work at the intersection of data-driven modelling and dynamical systems[26–31], this Perspective article has two specific focuses: how to represent systems given partial measurements and how discretization can reveal minimal dynamical generators of complex processes. We first consider classical attractor reconstruction, which mirrors constraints on latent representations learned by state-space models of time series. We then revisit early efforts to use symbolic approximations to compare minimal discrete generators underlying complex processes, a problem relevant to modern efforts to distil and interpret black-box statistical models. Emerging interdisciplinary works bridge nonlinear dynamics and learning theory, such as operator-theoretic methods for complex fluid flows, or detection of broken detailed balance in biological data sets. We conclude by discussing how future machine learning techniques may revisit other classical concepts from nonlinear dynamics, such as transinformation decay and complexity–entropy tradeoffs.

Representing and propagating chaotic dynamics

Large statistical learning models often parametrize complex data sets in low-dimensional latent spaces. For example, artificial image generators aim to invertibly map the space of natural images to a latent probability distribution[19]. The empirical success of such approaches is termed the manifold hypothesis: that high-dimensional data sets typically cluster near low-dimensional manifolds[32,33]. In the context of time series, successful data-driven re-parametrization implies that the dynamics arise from a low-dimensional attractor embedded within the higher-dimensional ambient space of the data. In this

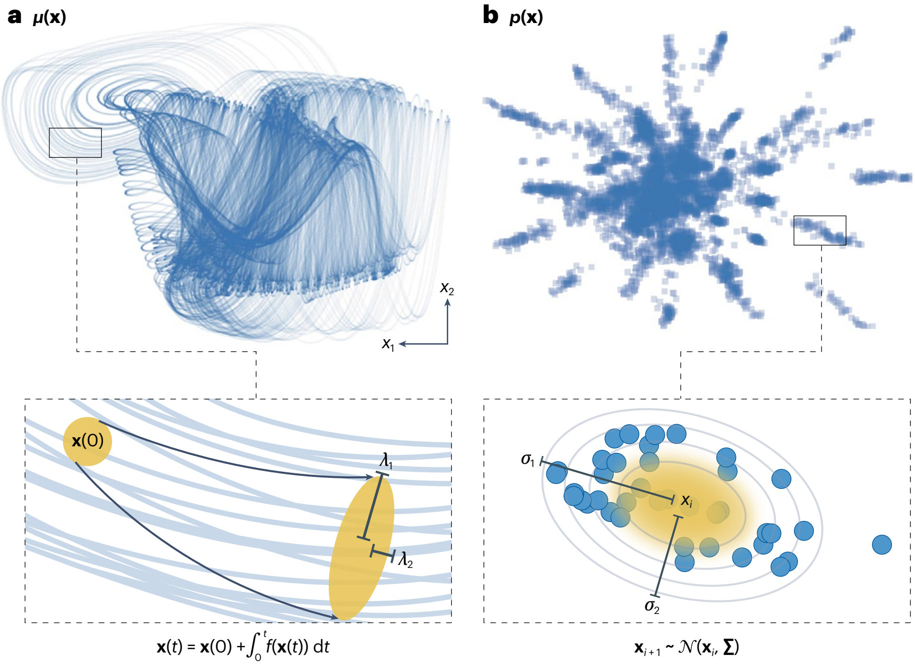
a $\mu(\mathbf{x})$
b $p(\mathbf{x})$
x_2
x_1
$\mathbf{x}(0)$
λ_1
λ_2
σ_1
x_i
σ_2
$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t f(\mathbf{x}(t))\,\mathrm{d}t$
$\mathbf{x}_{i+1} \sim \mathcal{N}(\mathbf{x}_i, \mathbf{\Sigma})$

Fig. 1 | Chaos as a generative process. a, The natural measure $\mu(\mathbf{x})$ of a strange attractor, which arises from a deterministic chaotic system $\mathbf{f}(\mathbf{x}(t))$ that evolves over time t, and a schematic of the divergence of a set of initial conditions, governed by the Lyapunov exponents λ_1 and λ_2. **b**, A probability distribution $p(\mathbf{x})$ over protein sequences learned by a variational autoencoder[181] and a simplified Markov chain Monte Carlo sampling scheme. The distribution $\mathcal{N}$ of proposed steps depends on the local covariance matrix $\mathbf{\Sigma}$; σ_1 and σ_2 denote standard deviations along principal axes. Part **b** is adapted from ref. 181, CC BY 4.0.

case, the apparent complexity of a measured time series is a matter of representation — complexity can be 'transformed out' by identifying and parameterizing this structure.

Historical work by the dynamical systems community sought to reconstruct the manifolds underlying dynamical systems based on limited observations. Contemporaneously to the introduction of Pesin's formula, several works formulated methods for reconstructing dynamical attractors from partial observations[34,35]. In the simplest such approach, time-delay embedding, a given univariate measurement time series $x(t)$, is assumed to result from a non-invertible transformation of an underlying multivariate dynamical system $\mathrm{d}\mathbf{z}/\mathrm{d}t = \mathbf{f}(\mathbf{z})$ that lies on an attractor. Although the dynamical variables necessary to span this attractor are unobserved, time-delay embedding constructs proxy variables using d_E copies of the original measurements some time τ in the past, resulting in the multivariate time series $\hat{\mathbf{z}} = [x(t), x(t-\tau), \ldots, x(t-d_\mathrm{E}\tau)]$. Theoretical justification for this approach comes from Takens' theorem: if the number of time delays d_E exceeds twice the manifold dimension of the underlying attractor, the resulting time-delay embedding will be diffeomorphic to the original attractor[34] (Box 1). The surprising aspect of Takens' theorem stems from its apparent contradiction of classical observability: measurements are typically non-invertible and low-rank operators, which discard information in their nullspace[36]. Takens' theorem and its variants sidestep this issue by imposing the regularity requirement that the underlying dynamics lie on attractors with well-defined structure — an assumption that, as discussed subsequently, might be recognized as an inductive bias from the perspective of modern statistical learning algorithms. An early success of time-delay embeddings was the experimental detection of a low-dimensional strange attractor as a laboratory flow transitions to turbulence[37] — a critical prediction of the Ruelle–Takens theory of turbulent attractors[38] (Fig. 2a). Delay embeddings subsequently spurred early advances in model-free forecasting[39–41] and nonlinear control[42].

Latent representations in time series models

Classical work on attractor reconstruction bears relevance to present-day statistical forecasting models for time series, which can be seen as generative models $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots)$ that sample potential future states of a dynamical system conditioned on its past[43,44]. Popular state-space models for forecasting treat observed data as emissions from an unobserved latent process, such as an underlying attractor or probability distribution[45]. In a simplified view, these models decompose time series in three phases: encoding observations into a latent space, propagating the dynamics and decoding them back into the measurement space. These three stages are made explicit in autoencoders, which parametrize the encoder and decoder using separate artificial neural networks[16]. However, even generic statistical learning models for time series, such as recurrent neural networks and attention-based transformers, implicitly represent dynamics with hidden variables[46]. Different statistical time series models may therefore be compared in terms of how they encode and decode time series, how they propagate dynamics in the latent space and what constraints they apply to each learning stage.

The latent structure found by time series models can reveal dynamical properties not apparent in the original data set. Although attracting inertial manifolds can be shown analytically to exist for particular systems such as damped fluid flows[47], reaction-diffusion systems[48] or coupled oscillators[49], data-driven manifold learning algorithms allow tools from dynamical systems to be applied even in the absence of explicit equations. These techniques have proven successful for equation-free nonlinear control, bifurcation detection and forecasting[50–52]. A key theme of works in recent years involves training autoencoders on high-dimensional dynamical time series and analysing the dynamical attractor in the latent space (Fig. 2b). One such work decomposes dynamical manifolds through a series of local charts tiling the overlapping latent spaces of several autoencoders[47], an approach reminiscent of classical work on piecewise linear models

Box 1

Attractor reconstruction estimates dynamical measures

Classic works in nonlinear dynamics describe ‘embedology’, that is, the process of inferring properties of the attractor of a dynamical system given only low-dimensional time series observations[11]. To emphasize connections to probabilistic machine learning, we frame embedology as estimating the density of attractor points in phase space. For dissipative chaotic systems in continuous time, this probability distribution often forms a fractal set, which for ergodic systems represents the natural measure $\mu(\mathbf{z})$, or the fraction of time that a long trajectory spends in the vicinity of a given phase space point $\mathbf{z}$.

Given an observed time series $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, classical attractor reconstruction learns a proxy variable $\mathbf{z}$ associated with the natural measure $p_\theta(\mathbf{z}) \sim \mu(\mathbf{z})$. In classical time-delay embedding, the lag operator $\mathcal{L}_\tau[\mathbf{x}_i] \equiv \mathbf{x}_{i-\tau}$ lifts the system to delay coordinates $\mathbf{z}_i \equiv \mathbf{x}_i, \mathcal{L}_\tau[\mathbf{x}_i], \mathcal{L}_\tau^2[\mathbf{x}_i], \ldots, \mathcal{L}_\tau^{d_\mathrm{E}-1}[\mathbf{x}]$. This reconstruction produces a density estimate $p_\theta(\mathbf{z}) = (T - d_\mathrm{E})^{-1} \sum_{i=1}^{T-d_\mathrm{E}} \delta(\mathbf{z} - \mathbf{z}_i)$ with $\theta = \{\tau, d_\mathrm{E}\}$, which can be smoothed by centring radial basis functions at each $\mathbf{z}_i$ (refs. 59,183). Motivation for this approach stems from Takens’ embedding theorem, a consequence of the Whitney immersion theorem, that states that a time-delay embedding smoothly and invertibly deforms onto the true dynamical attractor as long as $d_\mathrm{E} > 2d_\mathrm{F}$, in which d_F describes the intrinsic dimensionality of the measure (a non-integer for fractals)[11,34,35]. However, for most time series, d_F and thus d_E are unknown a priori; instead, d_E and the lag τ may be treated as learnable parameters θ, with their values determined using heuristic methods. Many methods select the most informative τ based on local minima of an averaged pairwise similarity measure across the time series $\bar{g}(\tau) = \langle g(\mathbf{z}_i, \mathbf{z}_{i-\tau}) \rangle_i$; although autocorrelation seems a natural choice, mutual information performs more strongly in practice[25,183]. d_E is typically determined using topological considerations on the basis of neighbourhoods in embedding space. Recent works generalize Takens’ theorem to multivariate and non-stationary time series[80,184] and externally forced systems with skew-product structure[185].

However, Takens’ theorem provides no assurance that time-delay embeddings preserve density, $p_\theta(\mathbf{x}) \approx \mu(\mathbf{x})$, a key requirement to accurately sample the system. Many algorithms built upon time-delay embeddings mitigate this issue by performing calculations based on nearest neighbours, rather than absolute distances in embedding space[59,183]. Motivated by the Nash embedding theorem, extensions of Takens’ theorem demonstrate conditions under which isometric embeddings can be recovered, often through additional nonlinear transformations or time delays — introducing a tradeoff between representational dimensionality and accuracy[186,187]. Recovery of the local density has also motivated practical extensions of time-delay embeddings based on nonlinear transformations of the lagged coordinates — these include principal components analysis[77,79], kernel and diffusion map methods[51,52] and artificial neural networks[53].

Evolving a dynamical system produces correlated, not independent, samples from the underlying attractor, suggesting that the measure may be approximated by a set of trajectories rather than individual points. Unstable periodic orbit theory seeks to group the points comprising the natural measure into exact solutions of the underlying dynamical system, which act as a topological skeleton of the flow[2]. For dissipative chaotic systems, $\boldsymbol{\mu}(\mathbf{z}) \propto \sum_p \delta(\mathbf{z} - \mathbf{z}_p) |\Lambda_p(\mathbf{z})|^{-1}$, in which p indexes a set of points $\mathbf{z}_p$ that traces an unstable recurrent solution $\mathbf{z}(t+t_p) = \mathbf{z}(t)$ (refs. 188–190). Because chaotic attractors contain no stable points, this sum spans an infinite set of unstable saddles ($t_p \to 0$) and limit cycles ($t_p > 0$). However, not all solutions influence the dynamics equally, and the stability multiplier Λ_p of a given solution denotes its relative instability. Dissipative, hyperbolic chaotic systems exhibit $|\Lambda_p| > 1$ for all $\mathbf{x}_p$; for saddle points, the stability multipliers may be obtained via linear stability analysis, whereas cycles require averaging across the orbit. Solutions with values closer to one dominate the measure and thus observed dynamics, making them appealing targets for unsupervised learning. Classical methods estimate the dominant unstable periodic orbits directly from dynamical time series by detecting near-recurrences in time-delay embeddings[191]. Recently, advances in unsupervised learning and topological data analysis have yielded new methods for detecting unstable periodic orbits in high-dimensional time series[192–194], prompting new applications of cycle decomposition to complex time series such as fluid turbulence[55,195] and organismal behaviour[196].

of chaotic attractors[39]. Other studies analyse local neighbourhood incidence to determine the intrinsic dimension of the latent space of autoencoders[53,54], a concept related to classical neighbour-based methods used to calculate fractal dimensions and select the time-delay embedding dimensions d_E (ref. 25) (Box 1). Consistent with the manifold hypothesis, latent spaces learned by statistical models of high-dimensional dynamical data sets are often contracting; for example, autoencoders trained on videos of weakly turbulent flows can map the dynamics to a low-dimensional latent space associated with exact solutions[55], consistent with inertial manifolds of the underlying partial differential equations. Much like delay embeddings, the dimensionality and model capacity necessary to identify these latent representations often depend on invariants of the underlying dynamics[56,57].

The practical success of contemporary state-space models illustrates that seemingly complex dynamics may be generated by low-dimensional latent processes. In this sense, their motivation mirrors earlier efforts to explain seemingly stochastic complex time series in terms of deterministic chaos[41,58,59]. Following the development of time-delay embedding, early approaches to nonlinear forecasting fit the observed data to a dynamical model — either via analytical governing equations or via data-driven methods based on nearest neighbours on the reconstructed attractor[59,60]. Post hoc statistical analysis then determined whether including nonlinearity in a fitted model substantially improved the forecast relative to a purely linear model, signalling a deterministic nonlinear component in the dynamics[59]. Just as classical statistical model selection presumes that the residuals of a fitted model should exhibit uniform scatter, equivalent tests for forecasting models evaluate whether the time series of forecast residuals exhibits no remaining autocorrelation owing to unmodelled deterministic dynamics[61]. Early methods therefore introduced the idea that stochasticity represents intrinsically high-dimensional dynamics driven from unmodelled measurement or process noise, producing degrees of freedom that cannot collapse onto a low-dimensional latent manifold.

When some previous knowledge of a nonlinear system is available, hybrid statistical learning methods directly impose constraints on latent dynamics to ensure consistency with known physical laws. For example, one approach encodes high-dimensional time series into a latent space where they obey analytical ordinary differential equations[62]. These equations can be constrained by restricting the library of possible functions present in the differential equations, or based on known symmetry groups[63]. Alternative methods such as neural ordinary differential equations do not require the latent differential equation to have an analytic form, only that it can be numerically approximated by an artificial neural network[64]. Some works impose constraints through limitations on the architecture of the learning model; for example, Hamiltonian neural networks directly fit differentiable Hamiltonians to data, ensuring that dynamics produced by the learned model heed symplecticity[56,65,66]. When physical constraints are unavailable, other representational constraints can prove informative. In many biological data sets, such as recordings of spiking neurons, observed data may be assumed to originate from time-varying stochastic dynamics (similar to an inhomogeneous Poisson process), making the inferred latent dynamics deterministic while the observed dynamics are stochastic[20,21].

Modern time series methods therefore navigate a general dichotomy between directly imposing structure (such as latent symmetries, symplecticity and distribution families) or inferring these properties from the observed data. The former represents the use of inductive biases that shrink the space of possible trained models to reduce data intensivity and errors, at the expense of generalizability. This use of external knowledge about a physical system to tune models along the bias–variance tradeoff echoes classical tradeoffs in nonlinear time series models. Early data assimilation algorithms for chaotic time series, such as nonlinear extensions of the Kalman filter, directly fit the parameters of either known ordinary differential equations or their linearizations[67,68]. The bias–variance tradeoff in these systems appears as rank conditions on the resulting nonlinear fits[69]. This tradeoff has physical interpretation in early works that use the quality of data-driven models to differentiate low-dimensional chaos from noisy linear dynamics. These works diagnose nonlinearity by comparing linear and nonlinear models fitted on a given data set[40]; if a nonlinear model provides a better fit, then the system deviates from the expected behaviour of a stationary process[70]. These results lead to scaling laws relating the amount of available data, degree of nonlinearity, attractor dimensionality and the number of time delays required for accurate state-space reconstruction[39,41,71]. Empirical scaling laws relating data volume and model complexity are the focus of many recent studies by statistical learning practitioners[72], suggesting that large generative models of dynamical data sets may eventually obey practical scaling laws backed by theoretical constraints on dynamical systems.

Lifting linearizes complex dynamics

The intuition behind time-delay embedding – that reparameterization of observed data provides information about unobserved variables – underlies emerging methods at the interface of dynamical systems and machine learning. Originally developed to analyse velocity field measurements in complex fluid flows[73], dynamic mode decomposition seeks to identify the linear transformation mapping a time-delay embedding of a time series onto itself at a later time. For spatially indexed data such as fluid flows, the spectral properties of the resulting linear reveal spatiotemporal motifs such as oscillations, large-scale currents and other coherent structures[29,74]. Operator approaches to dynamical systems trace their roots to the Koopman theory of ergodic systems in the first half of the twentieth century[75]; however, they gained greater prominence nearly a century later, in works showing that linear propagators for complex dynamics have spectra that factor into isolated (nearly periodic) and continuous (chaotic) components[76]. Nonlinear dynamics often become more linear with additional time-delayed variables[77–79] – a concept echoing earlier efforts to unravel

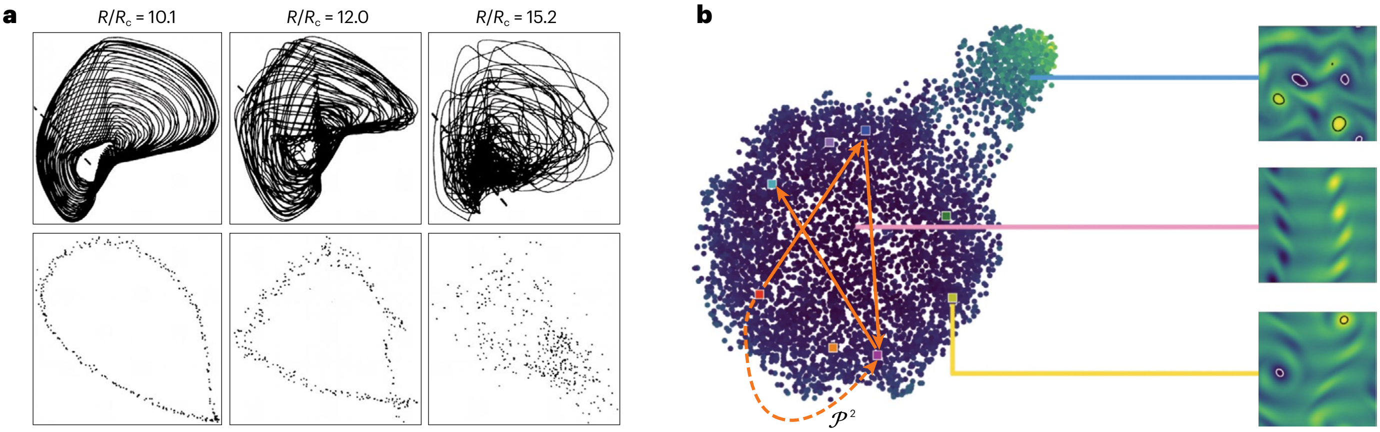

Fig. 2 | Latent dynamics revisit classical attractor reconstruction. **a**, Time-delay embeddings of a univariate time series representing the radial velocity of a flow, at three different Reynolds numbers (R) leading to turbulence at the critical value R_c. Poincaré sections are shown below each embedding. **b**, The latent space of an autoencoder neural network trained on weak turbulence ($R = 40$). The latent states are further embedded in 2D using t-distributed stochastic neighbour embedding. Shading indicates power dissipation, and connected states indicate equivalent flow configurations due to underlying symmetries. Part **a** reprinted with permission from ref. 37, APS. Part **b** reprinted with permission from ref. 55, APS.

non-stationary systems by 'overembedding' beyond the prescription of Takens' theorem[80]. Given a complex system such as a turbulent flow or spiking neuronal array, we can forgo modelling a system in terms of measured variables (velocity fields or individual neuron voltages) and instead 'lift' the system to a higher dimensionality than is strictly necessary to fully describe the dynamics. Although chaotic systems cannot be linearized with a finite number of lifting transformations, in many cases an infinite dimensional transformation exists for which a linear Koopman operator propagates the dynamics[81]. In practice, even finite-dimensional approximations of this operator unravel complex dynamical systems, by making them appear quasilinear for extended durations[30,77].

However, given a data set without known governing equations, it is difficult to determine in advance the particular lifting transformations that best approximate the Koopman operator in finite dimensions. Besides time delays, potential Koopman observables include: fixed nonlinear transformations based on known physical symmetries (such as spatial Fourier coefficients)[82]; generic nonlinear features such as polynomial kernels[83,84]; custom transformations learned directly from the data using autoencoders or custom kernels[85–87]; and transformations identified via equation discovery methods[88]. Because the optimal observables to approximate the Koopman operator are usually unknown a priori, data-driven methods require regularization and cross-validation to prevent overfitting[89,90].

Beyond Koopman methods, other emerging operator-theoretic techniques explore the interplay between lifted representations and dynamical complexity. These include data-driven discovery of quadratic forms[91], neural operators for partial differential equations[27,92] and works that combine nonlinear transformations of data with symbolic regression of analytical governing equations[54,62]. These frameworks share the theme that dynamical complexity can be unravelled at the expense of increased representational intricacy — echoing the tradeoffs between dimensionality and accuracy that underlie Takens' theorem[71]. Beyond a general competition between cost and accuracy[93,94], an inefficient choice of lifting transformations undermines interpretability while needlessly increasing computational demands. Similar tradeoffs have been noted in other emerging methods; for example, neural ordinary differential equations use artificial neural networks to construct numerical surrogates for the right-hand sides of differential equations[64]. Original formulations of these methods struggled to model complex trajectories near kinetic barriers, but later works introduced auxiliary dynamical variables that untangle the trajectories in a lifted space, allowing learned flows to bypass transport obstacles[95]. Machine learning practitioners are therefore beginning to confront basic questions regarding transport in dynamical systems with coherent structures inhibiting flow — a return to the original impetus for the development of dynamic mode decomposition and a demonstration of how dynamical systems theory may inform ongoing practical developments in statistical learning for time series.

Outlook for future learning architectures

Historical work on chaotic dynamics may continue to provide inductive biases guiding future statistical learning algorithms for time series. Although dynamic mode decomposition and Koopman methods have been adapted to broad scientific problems[28–30], other insights from dynamical systems may prove informative for general time series approaches even beyond the natural sciences. For example, the 'Hamiltonian manifold hypothesis' argues that because, in principle, all natural videos implicitly illustrate the consequences of physical laws, models trained on sufficiently large data sets will eventually converge to learning latent Hamiltonian dynamics[66].

On a practical level, constraints drawn from dynamical systems theory have begun to reveal whether the practical success of deep learning arises from unrecognized inductive biases. In deep neural networks, representations of inputs propagate across many layers as they are transformed into output labels or latent representations. Early works formulated multilayer artificial neural networks as continuous-time dynamical systems across layers[96–98], a connection that has gained renewed relevance in methods such as neural ordinary differential equations, continuous normalizing flows and diffusion models[19,64,99]. The dynamics of input representations propagating across layers can even exhibit transient chaos before settling into fixed points associated with output labels[100,101], and the stretching and folding of input representations across layers gives rise to measures of model expressivity that resemble topological entropy in complex flows[102]. Gradient-based training methods for large models implicitly reverse the layerwise dynamics, motivating theoretical works that describe large models as infinite-dimensional linear dynamical systems[103], with attendant implications for their ability to learn complex functions[26].

Compressibility and minimal dynamical generators

Latent space representations imply that the apparent complexity of a dynamical process may depend on the choice of measurement coordinates. Yet, Pesin's formula associates entropy production with chaotic dynamics, implying that some aspects of chaos are irreducible: intuitively, no invertible reparametrization can map a strange attractor to a limit cycle. Moreover, entropy production often has observable effects, such as heat production, that are independent of representation[104,105]. Classical works explore the irreducibility of chaos in the context of symbolic dynamics, which consider the computational properties of continuous systems under discrete coarse-graining[7,106]. A continuous-valued dynamical time series may be converted into a symbolic series by partitioning phase space and then analysing the properties of the symbol sequence produced by recording the partition label whenever the deterministic dynamics cross a boundary. Pesin's formula implies that this sequence is non-repeating for any nontrivial partitioning of a continuous-time chaotic system. However, because no deterministic finite-state automaton can exhibit non-recurrent dynamics, only a stochastic automaton can describe the symbol sequence produced by partitioning a deterministic chaotic system[9,107]. Symbolic dynamics thus link the properties of analogue dynamical systems to digital computers.

An early motivation for symbolic dynamics was identifying computational equivalence among systems[23,108]: setting aside differences in representation, are certain dynamical systems functionally identical? For example, the celebrated period-doubling cascade provides a universal description of bifurcations leading to chaos across diverse systems, from turbulent flows to ecological population fluctuations[1,109]. The dynamics preceding a given period-doubling bifurcation can be mapped to those following the bifurcation via a renormalization operation[110], the repeated application of which drives the system towards an accumulation point where the period diverges and chaos emerges. Symbolic dynamics provides an alternative view of this process, in which periodic dynamics correspond to a two-state automaton implementing a discrete shift on a symbolic register[109]. Each bifurcation doubles the number of unique states (and thus memory requirements) of the automaton, but renormalization decimates the register and renders the automaton invariant. At the accumulation point, the periodicity and thus memory requirements diverge, and the asymptotic entropy production becomes non-zero. Similar analyses based on symbolization were used by early practitioners to identify universal structure in diverse systems such as kicked rotors and chaotic scattering[111,112], as well as to map experimental data sets onto characteristic minimal systems[113]. Other contemporaneous works even sought to enumerate and categorize minimal symbolic dynamical systems based on their ability to support computation[114,115].

Models that learn to distil dynamics

The motivation behind symbolic dynamics – reducing systems down to their essential components – seemingly contradicts current trends of scaling general-purpose learning models to ever-larger data sets. However, in recent years, works on model compression and distillation have revisited the original ambitions of symbolic dynamics. When fitting a many-parameter learning model to a given data set, patterns within the original data (such as symmetries or stereotypy) may be revealed through the analysis of the trained model. For example, many state-space time series models directly map continuous-time observations to discrete modes of the underlying system (Fig. 3). In particular, hidden Markov models treat continuous observations as emissions from probability distributions conditioned on sequences of discrete internal states[45]. Likewise, switching linear models fit piecewise linear operators to subsets of the full phase space of a system, thereby approximating the global dynamics through a series of switches among local linear maps[39,116–118] (Fig. 4a). These approaches have proven particularly successful for data sets such as organismal behaviour and speech patterns, for which high measurement dimensionality meets low dynamical dimensionality owing to biomechanical constraints[119,120]. In such cases, latent discretization provides interpretability[119,121]; for example, in continuous-time recordings of organismal behaviour or neuronal activity, the latent variable sequence indicates distinct cognitive imperatives[118,122].

Natural images and other real-world data sets often span low-dimensional manifolds relative to their feature set[32]. As a result, many large-scale generative learning models are designed to map between complex data sets and simplified latent representations; for example, variational autoencoders use artificial neural networks to map training data to a tractable probability distribution, which can then be sampled to generate new surrogate data (Box 2). This latent space can therefore illuminate the inner workings of the learning model, even when the learned transformation between ambient and latent spaces remains opaque itself. For example, several architectures apply constraints during learning that cause variational autoencoders to learn a quantized latent space[123–126], in which patterns in the input data correspond to discrete entries within a latent codebook. These discrete modes reveal clusters of related training examples, making certain models a generalization of classical self-organizing maps[127,128] (Fig. 4b). Beyond interpretability, imposing discretization helps large generative models avoid posterior collapse, a limitation of autoregressive generation in which the model begins ignoring the latent state space and instead relies only on the decoder to determine its output – thereby reducing the complexity of the generated samples[15,123]. To identify this and other failure modes, it has been proposed to use entropy production to identify miscalibration in generative models[129].

A limitation of modern overparameterized learning models stems from degeneracy: many different trained models may exhibit equivalent performance on a given task, thereby precluding systematic comparison of models across tasks or model architectures. However, simplified latent representations of learned dynamics can reveal commonalities across learning models, echoing the computational

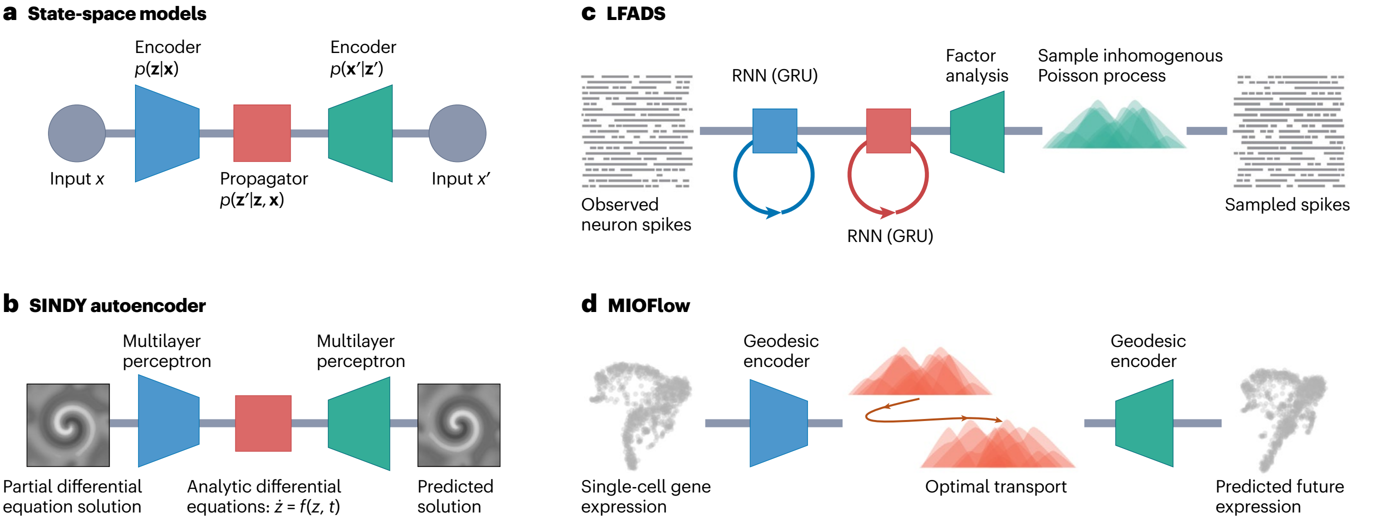

Fig. 3 | State-space models generate complex dynamics. a, Components of a generic state-space model. **b**, In an autoencoder employing Sparse Identification of Nonlinear Dynamics (SINDY)[62], multilayer perceptrons deterministically transform high-dimensional observations to a low-dimensional latent space, in which the dynamics are propagated using analytical differential equations learned via sparse regression from a library of known functions. **c**, In Latent Factor Analysis via Dynamical Systems (LFADS)[20], neuron spiking time series are deterministically encoded into latent initial conditions, which are evolved using a second recurrent neural network, and then decoded into latent factor time series. These latent factors parameterize the stochastic firing rate of an inhomogeneous Poisson process. **d**, In Manifold Interpolating Optimal-Transport Flows (MIOFlow)[99], high-dimensional gene expression measurements are encoded to a latent distribution that preserves the manifold diffusion distance. The latent measure is then propagated with optimal transport. Part **b** is adapted from ref. 62, CC BY 4.0. RNN, recurrent neural network. GRU, gated recurrent unit.

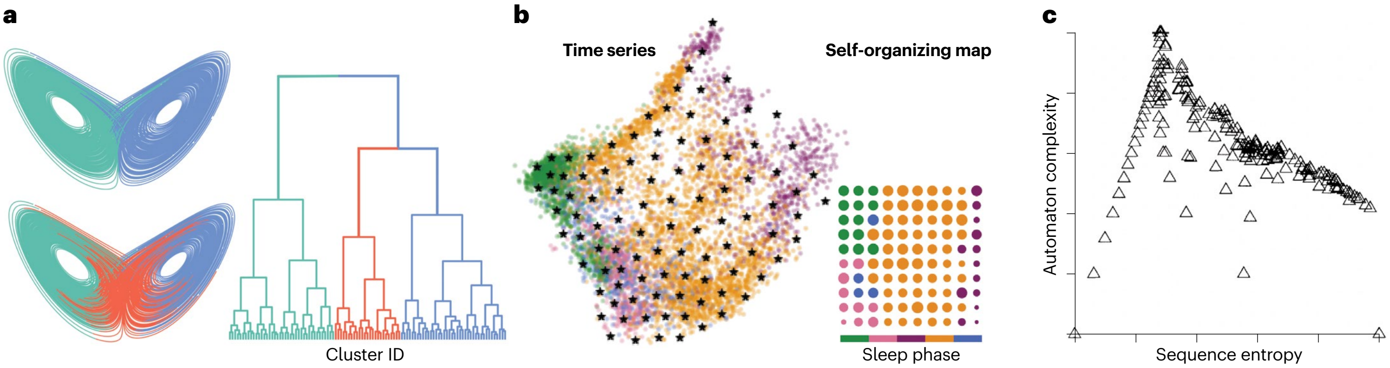

Fig. 4 | Latent discretization and interpretability. a, Successive stages of an adaptive approximation algorithm that fits locally linear dynamics to parts of the phase space of a chaotic system. Colours indicate discrete clusters at different levels of approximation. Data from ref. 120. **b**, A continuous-valued learning model that creates a discrete, latent self-organizing map (right) from a continuous time series of sleep recordings (left). Stars in the continuous space correspond to centroids of similar datapoints, each associated with discrete nodes in the map. **c**, The topological complexity of probabilistic automata fitted to a dynamical map across a range of chaotic and periodic regimes, plotted against the entropy of the time series. The most structurally complex automaton occurs when the dynamics exhibit intermediate entropy. Part **b** adapted with permission from ref. 182, PMLR. Part **c** adapted with permission from ref. 9, APS.

primitives sought by symbolic dynamics. Quantized latent states can reveal the internal logic of black-box neural networks, by allowing their internal grammar to be probed with post hoc analysis[118,126,130,131]. For example, traditional recurrent neural networks are provably capable of encoding arbitrary continuous-time dynamical attractors[132] and even discrete logic[133,134], given sufficient training data and model scale. Trained black-box recurrent learning models can be analysed by fitting probabilistic automata to their latent dynamics[135], and newer classes of generative learning models, such as attention-based transformers, may outperform earlier architectures because they can internally represent more sophisticated grammars[136–138]. Rather than analysing continuous-valued learning models post-training, some methods instead impose discrete dynamics directly through architectural constraints on the model. Emerging neuro-symbolic approaches combine the trainability of continuous-valued learning models with the representational guarantees of exact symbolic procedures, such as digital logic or arithmetic, by incorporating the latter within separate modules[134,139–142].

A drawback to combining discrete operations with continuous-valued model parameters stems from the difficulty of computing gradients of discrete states, which complicates the training of large models with gradient descent. This limitation mirrors a common shortcoming of symbolic dynamical systems, for which non-differentiability precludes the application of mathematical tools such as linear stability analysis – leading some early practitioners to view symbolic dynamics as intrinsically computational objects that can only be understood through direct simulation[114,143,144]. End-to-end trainable learning models containing discrete modules frequently use straight-through estimators, in which symbolic components are treated as identity functions when computing gradients of the model error with respect to its parameters[123]. Other works bypass gradient-based training entirely, instead directly modifying the latent code within continuous-valued neural networks, to programme the dynamics to perform discrete computations[145,146].

Taken together, these works illustrate how implicit or imposed symbolization within trainable learning models allows the generation of more complex and interpretable dynamics. Future works may use symbolic methods to automatically identify universality in generators of time series found by different trained learning models. For example, systems biology often requires comparison of gene regulatory dynamics across multiple organisms. Although these measurements vary widely based on differences in imaging modalities and fluorescent reporters, symbolic distillation could identify shared latent dynamical motifs arising from orthologous regulatory structures[147]. A key concept that may inform future developments in generative modelling is unifilarity[9]. Although a given time series can be mapped to multiple possible state-space models, a unifilar representation comprises the minimal maximally predictive generator for the dynamics – thus facilitating comparison of generators across systems[148–150]. Improvements in inference methods offer a key step towards extracting unifilar representations consistently across data sets and trained learning methods[151–153].

Measuring entropy production from data

Classical work on symbolic dynamics bears relevance to emerging interdisciplinary problems for which entropy production has physical implications. Recent studies have sought to identify macroscopic signatures of microscale non-equilibrium processes directly from experimental data[154–157]. Equilibrium thermodynamic systems exhibit detailed balance, in which the net flux between any pair of microstates equals zero. By contrast, active or living systems dissipate energy at the microscale, thus producing apparent violations of detailed balance at larger scales. Finite-resolution time series measurements of such systems, such as those produced by video microscopy, exhibit signatures of these microscale effects when quantized in the spatial or frequency domains[154]. Non-equilibrium behaviours manifest as net circulation in the phase space of the coarse-grained data – in contrast to detailed balance, in which probability currents vanish. These methods have successfully identified mesoscopic non-equilibrium states in diverse systems, ranging from the locomotory states of swimming cells[158,159] to oxygen levels in the brain during cognitive exertion[160].

Because executing computations in finite time and noisy environments requires energy dissipation, the physical entropy produced by non-equilibrium thermodynamic flows can be related to the information-theoretic entropy production associated with symbolic dynamics. This concept underlies studies that discretize

Box 2

Generative models of complex dynamics

A generative model $p_\theta(\mathbf{x}_{1:T})$ for a time series $\mathbf{x}_{1:T} \equiv \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ has the form[15,46,197]:

$$p_\theta(\mathbf{x}_{1:T}) = \int p_\theta(\mathbf{x}_{1:T}|\mathbf{z}_{1:T}) p_\theta(\mathbf{z}_{1:T})\, d\mathbf{z}_{1:T}, \qquad (2)$$

in which $p(\mathbf{z}_{1:T})$ denotes the previous distribution of latent state series of the model, which we assume matches the temporal resolution of the measured time series; $p_\theta(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})$ denotes the likelihood of an observed time series given a latent sequence; and θ denotes the trainable parameters. Many time series exhibit a characteristic timescale $\tau < T$ over which future values become decorrelated from past; for deterministic chaotic time series, this timescale approximately comprises the Lyapunov time $\tau \approx \lambda_{\max}^{-1}$. In this case, the likelihood exhibits conditional independence, in which all relevant dependence between past and future values of $\mathbf{x}$ is captured in the τ most recent timepoints:

$$p_\theta(\mathbf{x}_{1:T}|\mathbf{z}_{1:T}) = \prod_{t=1}^{\tau} p_\theta(\mathbf{x}_t|\mathbf{z}_{1:t}) \prod_{t=\tau+1}^{T} p_\theta(\mathbf{x}_t|\mathbf{z}_{t-\tau+1:t}).$$

Moreover, the latent prior becomes autoregressive:

$$p_\theta(\mathbf{z}_{1:T}) = p_\theta(\mathbf{z}_{1:\tau}) \prod_{t=\tau+1}^{T} p_\theta(\mathbf{z}_t|\mathbf{z}_{t-\tau:t-1}).$$

Inserting this expression into equation (2) and rearranging terms yields a general expression for a generative time series model:

$$p_\theta(\mathbf{x}_{1:T}) = \int \left[p_\theta(\mathbf{z}_{1:\tau}) \prod_{t=1}^{\tau} p_\theta(\mathbf{x}_t|\mathbf{z}_{1:t}) \right] \left(\prod_{t=\tau+1}^{T} p_\theta(\mathbf{x}_t|\mathbf{z}_{t-\tau+1:t}) p_\theta(\mathbf{z}_t|\mathbf{z}_{t-\tau:t-1}) \right) d\mathbf{z}_{1:T}, \qquad (3)$$

in which the first bracketed term represents the initial conditions at the start of the time series. The second term comprises two parts: an emission model for $\mathbf{x}$ based on the past τ values of $\mathbf{z}$ and $\mathbf{x}$ and a transition model implementing the latent dynamics. In most practical applications, τ is unknown a priori and instead represents an adjustable hyperparameter of the model — in forecasting, τ is the lookback window, and in other machine learning applications τ is the context length, akin to a working memory. In applications such as data smoothing and assimilation, the latent series $\mathbf{z}_{1:T}$ may represent the variable of practical interest, whereas forecasting seeks to directly sample future states of $\mathbf{x}$. For a traditional hidden Markov model, the transmission kernel simplifies to $p_\theta(\mathbf{z}_t|\mathbf{x}_{t-\tau:t-1}, \mathbf{z}_{t-\tau:t-1}) = p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1})$. For models that implement deterministic dynamics (for example, latent ordinary differential equations), the transmission becomes $p_\theta(\mathbf{z}_t|\mathbf{x}_{t-1}, \mathbf{z}_{t-1}) = \delta(\mathbf{z}_t - \mathbf{F}(\mathbf{z}_{t-1}, \mathbf{x}_{t-1}))$, in which the flow map $\mathbf{F}$ propagates $\mathbf{z}_{t-1}$ to $\mathbf{z}_t$, potentially under external forcing by $\mathbf{x}$.

Training probabilistic models requires maximizing the marginal log-likelihood $\log p_\theta(\mathbf{x}_{1:T})$ of the training data under the learned parameters θ. If the underlying dynamical system is linear and all conditional probabilities and process noise are Gaussian, then equation (3) reduces to a form similar to equation (1), and the resulting model represents the Kalman filter[198,199]. The likelihood of other classical state-space models such as hidden Markov models may be trained iteratively with expectation-maximization procedures[15,45]. However, for many complicated time series, $p_\theta(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ becomes difficult to sample, and so many works approximate the posterior distribution using artificial neural networks. Supervised training of forecasting models, such as recurrent neural networks or attention-based transformers, requires comparing the generated predictions of the model against ground truth future values — minimizing the forecast error therefore maximizes the training data likelihood. In unsupervised settings, training can instead proceed by comparing a given training example with one sampled from a nearby latent space location. In generative adversarial networks, training feedback is derived by passing the generator outputs to a separate discriminator that attempts to distinguish true versus sampled points[200].

The marginal likelihood $p_\theta(\mathbf{x}_{1:T})$ is typically difficult to compute, and so approaches such as variational autoencoders instead optimize a lower bound on the marginal log-likelihood called the evidence lower bound[201]. These approaches introduce a second trainable model that approximates the posterior $q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) \approx p_\theta(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$. Training thus requires minimizing a variational bound on the negative log-likelihood,

$$-\log p_\theta(\mathbf{x}_{1:T}) = -\log \int p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})\, d\mathbf{z}_{1:T} \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})}{q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})} \right],$$

in which the latter term arises from Jensen's inequality. We equate this expression with a loss function and rearrange terms to reveal an entropy-like expression:

$$\begin{aligned} \mathcal{L}_{\theta,\phi}(\mathbf{x}_{1:T}) = &-\mathbb{E}_q[\log p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})] \\ &+ \int q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) \log q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\, d\mathbf{z}_{1:T}. \end{aligned}$$

We next apply the same conditional independence assumptions used to derive equation (3). We neglect the boundary term by assuming that the loss depends negligibly on the first $\tau \ll T$ timepoints:

$$\begin{aligned} &\mathcal{L}_{\theta,\phi}(\mathbf{x}_{1:T}) \\ &\quad \approx -\sum_{t=\tau+1}^{T} \mathbb{E}_q\left[\log p_\theta(\mathbf{x}_t|\mathbf{z}_{t-\tau+1:t}) - \log\left(\frac{q_\phi(\mathbf{z}_t|\mathbf{z}_{t-\tau:t-1}, \mathbf{x}_{t-\tau+1:t})}{p_\theta(\mathbf{z}_t|\mathbf{z}_{t-\tau:t-1})} \right) \right]. \end{aligned} \qquad (4)$$

The loss therefore splits into a series of separate contributions from each τ-timepoint window, resembling classical state-space factorization of chaotic time series[25]. During training, the emission term $p_\theta(\mathbf{x}|\mathbf{z})$ and approximate transition term $q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1}\ldots)$ may be parameterized with models such as attention-based transformers or recurrent neural networks[202]. After training, forecasts may be generated autoregressively using equation (3). The first term in equation (4) corresponds to a maximum likelihood term that encourages accurate reconstruction of the training data, whereas the second term minimizes the Kullback–Leibler divergence between

(continued from previous page)
the true latent transition term p_θ and its trainable surrogate q_ϕ. A similar information-theoretic expression appears in classical measures of synchronization in chaotic systems[203–205], and generalized synchronization between true and latent dynamics has been proposed as a potential learning mechanism for recurrent neural networks[206]. This effect may explain recent empirical works demonstrating that modern recurrent neural networks can successfully forecast chaotic systems ~10 Lyapunov times into the future[142,207,208].

continuous-valued measurements of active systems and then use existing digital compression algorithms to extract descriptive order parameters – such as the difference in probability between forward and backward sequences, a measure of irreversibility[161–163]. Non-equilibrium steady states can thus reveal information about minimal dynamical systems that underlie biological motifs. Emerging works on data-driven detection of non-equilibrium flows have begun to infer computational primitives of the underlying living systems[156,164], particularly in systems known to execute computations such as neuronal ensembles[160,165] or cellular decision-making[147,159,166,167].

These works provide biophysical motivation to revisit classical questions regarding the physical nature of information in dynamical systems that support computation[4,107]. In computers operating at finite temperatures, small-scale thermal fluctuations represent a precision floor. If computation is implemented with chaotic dynamics, errors cascade from small to large scales at a rate that depends only on the Lyapunov exponents of the deterministic dynamics – and not on the temperature itself[7]. The mutual information between the initial and final states of a chaotic system, deemed the transinformation in early works, therefore decays over time, a form of information erasure that underlies the effective irreversibility of chaos[8]. Work predating the widespread study of chaos established a minimal thermodynamic cost for information erasure[105], a connection that influenced later efforts to understand computation in chaotic systems[168]. More recently, these concepts have been revisited in works that formalize the non-equilibrium thermodynamics of information processing systems[104,164,169].

Non-equilibrium thermodynamics has influenced the recent development of practical generative models for complex data sets[170,171]. A leading approach to natural image generation is diffusion models, which learn to iteratively invert a diffusive flow connecting a tractable latent distribution to the observed distribution of natural images[19]. Training a diffusion model on natural images consists of gradually combining high-dimensional noise with each input image, while simultaneously training a set of denoising learning models to invert each incremental noise addition. After training, synthetic images may be generated by sampling an image of random noise and then applying the sequence of trained denoising models[172]. Early works on diffusion models noted that this sequence of intermediate operations comprises a non-equilibrium flow[18] and that generation of new images requires weighted non-equilibrium sampling of rarer trajectories that lead to realistic natural images. Empirical studies assessing the quality of large, autoregressive generative models observe long-time decay in mutual information between input states and generated outputs[129], a phenomenon comparable to classical transinformation erasure in chaotic systems[7]. By analogy to the data processing inequality, transinformation decay implies that, given knowledge of a dynamical system at finite precision, no statistical learning model can recover predictive information about the initial conditions, once a sufficient number of Lyapunov times have elapsed[8,164].

Outlook

Much as early computers and the resulting visualizations of fractals inspired excitement in applying chaos to other fields[173], advances in statistical learning have sparked renewed interest in classical ideas from nonlinear dynamics. Connections between these fields range from physics-based inductive biases on latent representations in generative learning models, to the identification of minimal dynamical generators underlying complex time series.

Future works may implicate fundamental relationships between the observability and representability of complex dynamics. Early efforts to relate chaos to computational principles related the apparent entropy of a system to the complexity of its underlying representation[4,9]. A system settling to a fixed point or limit cycle eventually ceases to produce new information because its attractor has been fully observed after a finite observation period[148,174]. Conversely, a completely stochastic system such as a random number generator seemingly produces information, but without any underlying structure. The complexity of the generator of a system plotted against the entropy of its outputs therefore exhibits non-monotonicity with an intermediate peak – suggestively termed the 'edge of chaos' by some practitioners – that represents systems that can, at different times, switch between fully ordered and seemingly random outputs (Fig. 4c). Early works considered whether this edge represents those systems capable of supporting information processing and intelligence[114,174,175], a concept revisited in studies that analyse the capacity of modern statistical learning models[176–180].

As the scale and quality of generative learning models improve, structural complexity compared with data randomness may emerge as an observable relationship between problem difficulty and model selection. A complexity–entropy relation could describe the intricacy of latent representations learned by large models in unsupervised settings, or the complexity of the underlying architectures necessary to achieve a given accuracy on supervised learning problems. This dynamical refinement of the bias–variance tradeoff could inform future developments, bridging Wheeler's physical bits with the practicalities of modern large-scale learning systems.

Published online: 12 February 2024

References

1. Crutchfield, J. & Packard, N. Symbolic dynamics of one-dimensional maps: entropies, finite precision, and noise. *Int. J. Theor. Phys.* **21**, 433–466 (1982).
2. Cvitanovic, P. et al. in *Chaos: Classical and Quantum* Vol. 69, 25 (2005).
3. Farmer, J. D. Information dimension and the probabilistic structure of chaos. *Z. Naturforsch. A* **37**, 1304–1326 (1982).
4. Feynman, R. P. *Feynman Lectures on Computation* (CRC, 2018).
5. Wheeler, J. A. "On recognizing 'law without law'," Oersted medal response at the joint APS–AAPT Meeting, New York, 25 January 1983. *Am. J. Phys.* **51**, 398–404 (1983).
6. Wheeler, J. A. Recent thinking about the nature of the physical world: it from bit a. *Ann. N. Y. Acad. Sci.* **655**, 349–364 (1992).
7. Shaw, R. Strange attractors, chaotic behavior, and information flow. *Z. Naturforsch. A* **36**, 80–112 (1981).
8. Pompe, B., Kruscha, J. & Leven, R. State predictability and information flow in simple chaotic systems. *Z. Naturforsch. A* **41**, 801–818 (1986).

9. Crutchfield, J. P. & Young, K. Inferring statistical complexity. *Phys. Rev. Lett.* **63**, 105 (1989).
10. Grassberger, P. Information and complexity measures in dynamical systems. In *Proc. NATO Advanced Study Institute on Information Dynamics* 15–33 (Springer, 1991).
11. Sauer, T., Yorke, J. A. & Casdagli, M. Embedology. *J. Stat. Phys.* **65**, 579–616 (1991).
12. Pesin, Y. B. Characteristic Lyapunov exponents and smooth ergodic theory. *Russ. Math. Surv.* **32**, 55 (1977).
13. Gilpin, W. Cryptographic hashing using chaotic hydrodynamics. *Proc. Natl Acad. Sci. USA* **115**, 4869–4874 (2018).
14. Sinai, Y. G. Gibbs measures in ergodic theory. *Russ. Math. Surv.* **27**, 21 (1972).
15. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
16. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
17. Edelman, A., Arias, T. A. & Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**, 303–353 (1998).
18. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning* 2256–2265 (PMLR, 2015).
19. Song, Y. & Ermon, S. Generative modeling by estimating gradients of the data distribution. In *33rd Conference on Neural Information Processing Systems* (NeurIPS, 2019).
20. Pandarinath, C. et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
21. Koppe, G., Toutounji, H., Kirsch, P., Lis, S. & Durstewitz, D. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. *PLoS Comput. Biol.* **15**, e1007263 (2019).
22. Yousif, M. Z., Yu, L. & Lim, H.-C. High-fidelity reconstruction of turbulent flow from spatially limited data using enhanced super-resolution generative adversarial network. *Phys. Fluids* **33**, 125119 (2021).
23. Bowen, R. & Ruelle, D. The ergodic theory of axiom a flows. *Invent. Math.* **29**, 181–202 (1975).
24. Gershenfeld, N. An experimentalist's introduction to the observation of dynamical systems. In *Directions in Chaos* Vol. *2*, 310–353 (World Scientific, 1988).
25. Abarbanel, H. D., Brown, R., Sidorowich, J. J. & Tsimring, L. S. The analysis of observed chaotic data in physical systems. *Rev. Mod. Phys.* **65**, 1331 (1993).
26. Bahri, Y. et al. Statistical mechanics of deep learning. *Annu. Rev. Condens. Matter Phys.* **11**, 501–528 (2020).
27. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
28. Brunton, S. L., Budisi´c, M., Kaiser, E. & Kutz, J. N. Modern Koopman theory for dynamical systems. *SIAM Rev.* **64**, 229–340 (2022).
29. Mezić, I. Analysis of fluid flows via spectral properties of the Koopman operator. *Annu. Rev. Fluid Mech.* **45**, 357–378 (2013).
30. Otto, S. E. & Rowley, C. W. Koopman operators for estimation and control of dynamical systems. *Annu. Rev. Control Robot. Auton. Syst.* **4**, 59–87 (2021).
31. Ghadami, A. & Epureanu, B. I. Data-driven prediction in dynamical systems: recent developments. *Philos. Trans. Royal Soc. A* **380**, 20210213 (2022).
32. Fefferman, C., Mitter, S. & Narayanan, H. Testing the manifold hypothesis. *J. Am. Math. Soc.* **29**, 983–1049 (2016).
33. Boumal, N. *An Introduction to Optimization on Smooth Manifolds* (Cambridge Univ. Press, 2023).
34. Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980: Proceedings of a Symposium Held at the University of Warwick* 1979/80, 366–381 (Springer, 1980).
35. Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. Geometry from a time series. *Phys. Rev. Lett.* **45**, 712 (1980).
36. Bechhoefer, J. *Control Theory for Physicists* (Cambridge Univ. Press, 2021).
37. Brandstäter, A. et al. Low-dimensional chaos in a hydrodynamic system. *Phys. Rev. Lett.* **51**, 1442 (1983).
38. Ruelle, D. & Takens, F. On the nature of turbulence. *Commun. Math. Phys* **20**, 167–192 (1971).
39. Casdagli, M. Nonlinear prediction of chaotic time series. *Phys. D* **35**, 335–356 (1989).
40. Sugihara, G. & May, R. M. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* **344**, 734–741 (1990).
41. Tsonis, A. & Elsner, J. Nonlinear prediction as a way of distinguishing chaos from random fractal sequences. *Nature* **358**, 217–220 (1992).
42. Ott, E., Grebogi, C. & Yorke, J. A. Controlling chaos. *Phys. Rev. Lett.* **64**, 1196 (1990).
43. Petropoulos, F. et al. Forecasting: theory and practice. *Int. J. Forecast.* **38**, 705–871 (2022).
44. Gershenfeld, N., Schoner, B. & Metois, E. Cluster-weighted modelling for time-series analysis. *Nature* **397**, 329–332 (1999).
45. Durbin, J. & Koopman, S. J. *Time Series Analysis by State Space Methods* Vol. 38 (Oxford Univ Press, 2012).
46. Girin, L. et al. Dynamical variational autoencoders: a comprehensive review. *Found. Trends Mach. Learn.* **15**, 1–175 (2021).
47. Floryan, D. & Graham, M. D. Data-driven discovery of intrinsic dynamics. *Nat. Mach. Intell.* **4**, 1113–1120 (2022).
48. Doering, C. R. & Gibbon, J. D. *Applied Analysis of the Navier–Stokes Equations* Vol. 12 (Cambridge Univ. Press, 1995).
49. Ott, E. & Antonsen, T. M. Low dimensional behavior of large systems of globally coupled oscillators. *Chaos* **18**, 037113 (2008).
50. Blanchard, A. & Sapsis, T. P. Learning the tangent space of dynamical instabilities from data. *Chaos* **29**, 113120 (2019).
51. Cenedese, M., Axås, J., Bäuerlein, B., Avila, K. & Haller, G. Data-driven modeling and prediction of non-linearizable dynamics via spectral submanifolds. *Nat. Commun.* **13**, 872 (2022).
52. Berry, T., Giannakis, D. & Harlim, J. Nonparametric forecasting of low-dimensional dynamical systems. *Phys. Rev. E* **91**, 032915 (2015).
53. Gilpin, W. Deep reconstruction of strange attractors from time series. In *Advances in Neural Information Processing Systems* Vol. 33 (NeurIPS, 2020).
54. Chen, B. et al. Automated discovery of fundamental variables hidden in experimental data. *Nat. Comput. Sci.* **2**, 433–442 (2022).
55. Page, J., Brenner, M. P. & Kerswell, R. R. Revealing the state space of turbulence using machine learning. *Phys. Rev. Fluids* **6**, 034402 (2021).
56. Greydanus, S., Dzamba, M. & Yosinski, J. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems* Vol. 32 (NeurIPS 2019).
57. Linot, A. J. & Graham, M. D. Deep learning to discover and predict dynamics on an inertial manifold. *Phys. Rev. E* **101**, 062209 (2020).
58. Lefebvre, J., Goodings, D., Kamath, M. & Fallen, E. Predictability of normal heart rhythms and deterministic chaos. *Chaos* **3**, 267–276 (1993).
59. Sugihara, G. Nonlinear forecasting for the classification of natural time series. *Philos. Trans. Royal Soc. A Phys. Eng. Sci.* **348**, 477–495 (1994).
60. Casdagli, M. Chaos and deterministic versus stochastic non-linear modelling. *J. R. Stat. Soc. Ser. B* **54**, 303–328 (1992).
61. Broock, W. A., Scheinkman, J. A., Dechert, W. D. & LeBaron, B. A test for independence based on the correlation dimension. *Econom. Rev.* **15**, 197–235 (1996).
62. Champion, K., Lusch, B., Kutz, J. N. & Brunton, S. L. Data-driven discovery of coordinates and governing equations. *Proc. Natl Acad. Sci. USA* **116**, 22445–22451 (2019).
63. Udrescu, S.-M. et al. AI Feynman 2.0: pareto-optimal symbolic regression exploiting graph modularity. *Adv. Neural Inform. Process. Syst.* **33**, 4860–4871 (2020).
64. Chen, R. T., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. Neural ordinary differential equations. In *32nd Conference on Neural Information Processing Systems* (NeurIPS, 2018).
65. Choudhary, A. et al. Physics-enhanced neural networks learn order and chaos. *Phys. Rev. E* **101**, 062207 (2020).
66. Toth, P. et al. Hamiltonian generative networks. In *International Conference on Learning Representations* (2019).
67. Brown, R., Rulkov, N. F. & Tracy, E. R. Modeling and synchronizing chaotic systems from time-series data. *Phys. Rev. E* **49**, 3784 (1994).
68. Julier, S. J. & Uhlmann, J. K. Unscented filtering and nonlinear estimation. *Proc. IEEE* **92**, 401–422 (2004).
69. Reif, K., Gunther, S., Yaz, E. & Unbehauen, R. Stochastic stability of the discrete-time extended Kalman filter. *IEEE Trans. Autom. Control.* **44**, 714–728 (1999).
70. Kaplan, D. T. Model-independent technique for determining the embedding dimension. in *Chaos in Communications*, Vol. 2038, 236–240 (SPIE, 1993).
71. Gershenfeld, N. A. Dimension measurement on high-dimensional systems. *Phys. D* **55**, 135–154 (1992).
72. Hoffmann, J. et al. Training compute-optimal large language models. Preprint at https://arXiv.org/abs/2203.15556 (2022).
73. Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *J. Fluid Mech.* **656**, 5–28 (2010).
74. Haller, G. Lagrangian coherent structures. *Annu. Rev. Fluid Mech.* **47**, 137–162 (2015).
75. Koopman, B. O. & Neumann, J. V. Dynamical systems of continuous spectra. *Proc. Natl Acad. Sci. USA* **18**, 255–263 (1932).
76. Mezić, I. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dyn.* **41**, 309–325 (2005).
77. Brunton, S. L., Brunton, B. W., Proctor, J. L., Kaiser, E. & Kutz, J. N. Chaos as an intermittently forced linear system. *Nat. Commun.* **8**, 19 (2017).
78. Arbabi, H. & Mezic, I. Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator. *SIAM J. Appl. Dyn. Syst.* **16**, 2096–2126 (2017).
79. Kamb, M., Kaiser, E., Brunton, S. L. & Kutz, J. N. Time-delay observables for Koopman: theory and applications. *SIAM J. Appl. Dyn. Syst.* **19**, 886–917 (2020).
80. Hegger, R., Kantz, H., Matassini, L. & Schreiber, T. Coping with nonstationarity by overembedding. *Phys. Rev. Lett.* **84**, 4092 (2000).
81. Budisić, M., Mohr, R. & Mezić, I. Applied koopmanism. *Chaos* **22**, 047510 (2012).
82. Nathan Kutz, J., Proctor, J. L. & Brunton, S. L. Applied Koopman theory for partial differential equations and data-driven modeling of spatio-temporal systems. *Complexity* **2018**, 1–16 (2018).
83. Williams, M. O., Kevrekidis, I. G. & Rowley, C. W. A data-driven approximation of the Koopman operator: extending dynamic mode decomposition. *J. Nonlinear Sci.* **25**, 1307–1346 (2015).
84. Nuske, F., Keller, B. G., Pérez-Hernández, G., Mey, A. S. & Noé, F. Variational approach to molecular kinetics. *J. Chem. Theory Comput.* **10**, 1739–1752 (2014).
85. Takeishi, N., Kawahara, Y. & Yairi, T. Learning Koopman invariant subspaces for dynamic mode decomposition. In *31st Conference on Neural Information Processing Systems* (NIPS, 2017).
86. Lusch, B., Kutz, J. N. & Brunton, S. L. Deep learning for universal linear embeddings of nonlinear dynamics. *Nat. Commun.* **9**, 4950 (2018).
87. Wehmeyer, C. & Noé, F. Time-lagged autoencoders: deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **148**, 241703 (2018).
88. Kaiser, E., Kutz, J. N. & Brunton, S. L. Data-driven discovery of Koopman eigenfunctions for control. *Mach. Learn. Sci. Technol.* **2**, 035023 (2021).

89. Bollt, E. Regularized kernel machine learning for data driven forecasting of chaos. *Annu. Rev. Chaos Theor. Bifurcat. Dyn. Syst.* **9**, 1–26 (2020).
90. Li, Q., Dietrich, F., Bollt, E. M. & Kevrekidis, I. G. Extended dynamic mode decomposition with dictionary learning: a data-driven adaptive spectral decomposition of the Koopman operator. *Chaos* **27**, 103111 (2017).
91. Qian, E., Kramer, B., Peherstorfer, B. & Willcox, K. Lift & learn: physics-informed machine learning for large-scale nonlinear dynamical systems. *Phys. D* **406**, 132401 (2020).
92. Li, Z. et al. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations* (2020).
93. De Hoop, M., Huang, D. Z., Qian, E. & Stuart, A. M. The cost-accuracy trade-off in operator learning with neural networks. Preprint at https://arxiv.org/abs/2203.13181 (2022).
94. Lu, L., Jin, P., Pang, G., Zhang, Z. & Karniadakis, G. E. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.* **3**, 218–229 (2021).
95. Dupont, E., Doucet, A. & Teh, Y. W. Augmented neural ODEs. In *33rd Conference on Neural Information Processing Systems* (NeurIPS, 2019).
96. Pineda, F. J. Generalization of back-propagation to recurrent neural networks. *Phys. Rev. Lett.* **59**, 2229 (1987).
97. Chua, L. O. & Yang, L. Cellular neural networks: theory. *IEEE Trans. Circuits Syst.* **35**, 1257–1272 (1988).
98. Saad, D. & Solla, S. A. On-line learning in soft committee machines. *Phys. Rev. E* **52**, 4225 (1995).
99. Huguet, G. et al. Manifold interpolating optimal-transport flows for trajectory inference. *Adv. Neural Inf. Process. Syst.* **35**, 29705–29718 (2022).
100. Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J. & Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems* Vol. 29 (NIPS, 2016).
101. Schoenholz, S. S., Gilmer, J., Ganguli, S. & Sohl-Dickstein, J. Deep information propagation. Preprint at https://arxiv.org/abs/1611.01232 (2016).
102. Montufar, G. F., Pascanu, R., Cho, K. & Bengio, Y. On the number of linear regions of deep neural networks. In *Proc. 27th International Conference on Neural Information Processing Systems* (NeurIPS, 2014).
103. Jacot, A., Gabriel, F. & Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. In *32nd Conference on Neural Information Processing Systems* (NeurIPS 2018).
104. Conte, T. et al. Thermodynamic computing. Preprint at https://arxiv.org/abs/1911.01968 (2019).
105. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **5**, 183–191 (1961).
106. Morse, M. & Hedlund, G. A. Symbolic dynamics. *Am. J. Math.* **60**, 815–866 (1938).
107. Moore, C. Unpredictability and undecidability in dynamical systems. *Phys. Rev. Lett.* **64**, 2354 (1990).
108. Metropolis, N., Stein, M. & Stein, P. On finite limit sets for transformations on the unit interval. *J. Comb. Theory Ser. A.* **15**, 25–44 (1973).
109. Hao, B.-l. Symbolic dynamics and characterization of complexity. *Phys. D Nonlinear Phenom.* **51**, 161–176 (1991).
110. Feigenbaum, M. J. The universal metric properties of nonlinear transformations. *J. Stat. Phys.* **21**, 669–706 (1979).
111. Lewis, J. E. & Glass, L. Nonlinear dynamics and symbolic dynamics of neural networks. *Neural Comput.* **4**, 621–642 (1992).
112. Hao, B.-L. *Elementary Symbolic Dynamics and Chaos in Dissipative Systems* (World Scientific, 1989).
113. Daw, C. S., Finney, C. E. A. & Tracy, E. R. A review of symbolic analysis of experimental data. *Rev. Sci. Instrum.* **74**, 915–930 (2003).
114. Langton, C. G. Computation at the edge of chaos: phase transitions and emergent computation. *Phys. D* **42**, 12–37 (1990).
115. Wolfram, S. Universality and complexity in cellular automata. *Phys. D* **10**, 1–35 (1984).
116. Ghahramani, Z. & Hinton, G. E. Variational learning for switching state-space models. *Neural Comput.* **12**, 831–864 (2000).
117. Fox, E., Sudderth, E., Jordan, M. & Willsky, A. Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems* Vol. 21 (NIPS, 2008).
118. Smith, J., Linderman, S. & Sussillo, D. Reverse engineering recurrent neural networks with Jacobian switching linear dynamical systems. *Adv. Neural Inf. Process. Syst.* **34**, 16700–16713 (2021).
119. Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P. & Datta, S. R. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems* Vol. 29 (NIPS, 2016).
120. Costa, A. C., Ahamed, T. & Stephens, G. J. Adaptive, locally linear models of complex dynamics. *Proc. Natl Acad. Sci. USA* **116**, 1501–1510 (2019).
121. Krakovna, V. & Doshi-Velez, F. Increasing the interpretability of recurrent neural networks using hidden Markov models. Preprint at https://arxiv.org/abs/1606.05320 (2016).
122. Mudrik, N., Chen, Y., Yezerets, E., Rozell, C. J. & Charles, A. S. Decomposed linear dynamical systems (dLDS) for learning the latent components of neural dynamics. Preprint at https://arxiv.org/abs/2206.02972 (2022).
123. Van Den Oord, A. et al. Neural discrete representation learning. In *31st Conference on Neural Information Processing Systems* (NIPS, 2017).
124. Devaraj, C. et al. From symbols to signals: symbolic variational autoencoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3317–3321 (IEEE, 2020).
125. Rasul, K., Park, Y.-J., Ramström, M. N. & Kim, K.-M. VQ-AR: vector quantized autoregressive probabilistic time series forecasting. Preprint at https://arxiv.org/abs/2205.15894 (2022).
126. Falck, F. et al. Multi-facet clustering variational autoencoders. *Adv. Neural Inf. Process. Syst.* **34**, 8676–8690 (2021).
127. Fortuin, V., Hüser, M., Locatello, F., Strathmann, H. & Rätsch, G. SOM-VAE: interpretable discrete representation learning on time series. In *International Conference on Learning Representations* (2018).
128. Kohonen, T. The self-organizing map. *Proc. IEEE* **78**, 1464–1480 (1990).
129. Braverman, M. et al. Calibration, entropy rates, and memory in language models. In *International Conference on Machine Learning*, 1089–1099 (PMLR, 2020).
130. Tschannen, M., Bachem, O. & Lucic, M. Recent advances in autoencoder-based representation learning. Preprint at https://arxiv.org/abs/1812.05069 (2018).
131. Jang, E., Gu, S. & Poole, B. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations* (2017).
132. Funahashi, K.-I. & Nakamura, Y. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Netw.* **6**, 801–806 (1993).
133. Neto, J. P., Siegelmann, H. T., Costa, J. F. & Araujo, C. S. Turing universality of neural nets (revisited). In *Computer Aided Systems Theory — EUROCAST'97: A Selection of Papers from the 6th International Workshop on Computer Aided Systems Theory Las Palmas de Gran Canaria, Spain, February* 24–28, 1997 *Proceedings* 6, 361–366 (Springer, 1997).
134. Kaiser, Ł. & Sutskever, I. Neural GPUs learn algorithms. Preprint at https://arxiv.org/abs/1511.08228 (2015).
135. Weiss, G., Goldberg, Y. & Yahav, E. Learning deterministic weighted automata with queries and counterexamples. In *Advances in Neural Information Processing Systems* Vol. 32 (NeurIPS, 2019).
136. Michalenko, J. J. et al. Representing formal languages: a comparison between finite automata and recurrent neural networks. In *International Conference on Learning Representations* (2019).
137. Resnick, C., Gupta, A., Foerster, J., Dai, A. M. & Cho, K. Capacity, bandwidth, and compositionality in emergent language learning. Preprint at https://arxiv.org/abs/1910.11424 (2019).
138. Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A. & Zhang, C. Transformers learn shortcuts to automata. Preprint at https://arxiv.org/abs/2210.10749 (2022).
139. Tsamoura, E., Hospedales, T. & Michael, L. Neural-symbolic integration: a compositional perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 35, 5051–5060 (2021).
140. Daniele, A., Campari, T., Malhotra, S. & Serafini, L. Deep symbolic learning: discovering symbols and rules from perceptions. Preprint at https://arxiv.org/abs/2208.11561 (2022).
141. Trask, A. et al. Neural arithmetic logic units. In *32nd Conference on Neural Information Processing Systems* (NeurIPS, 2018).
142. Yik, J. et al. Neurobench: advancing neuromorphic computing through collaborative, fair and representative benchmarking. Preprint at https://arxiv.org/abs/2304.04640 (2023).
143. Neumann, J. V. *Theory of Self-Reproducing Automata* (ed. Burks, A. W.) (Univ. Illinois Press, 1966).
144. Wolfram, S. Statistical mechanics of cellular automata. *Rev. Mod. Phys.* **55**, 601 (1983).
145. Gilpin, W. Cellular automata as convolutional neural networks. *Phys. Rev. E* **100**, 032402 (2019).
146. Kim, J. Z. & Bassett, D. S. A neural machine code and programming framework for the reservoir computer. *Nat. Mach. Intell.* **5**, 1–9 (2023).
147. Wong, F. & Gunawardena, J. Gene regulation in and out of equilibrium. *Annu. Rev. Biophys.* **49**, 199–226 (2020).
148. Crutchfield, J. P. Between order and chaos. *Nat. Phys.* **8**, 17–24 (2012).
149. Ephraim, Y. & Merhav, N. Hidden Markov processes. *IEEE Trans. Inf. Theory* **48**, 1518–1569 (2002).
150. Marzen, S. E. & Crutchfield, J. P. Nearly maximally predictive features and their dimensions. *Phys. Rev. E* **95**, 051301 (2017).
151. Strelioff, C. C. & Crutchfield, J. P. Bayesian structural inference for hidden processes. *Phys. Rev. E* **89**, 042119 (2014).
152. Marzen, S. E. & Crutchfield, J. P. Structure and randomness of continuous-time, discrete-event processes. *J. Stat. Phys.* **169**, 303–315 (2017).
153. Pfau, D., Bartlett, N. & Wood, F. Probabilistic deterministic infinite automata. In *Advances in Neural Information Processing Systems* Vol. 23 (NIPS, 2010).
154. Battle, C. et al. Broken detailed balance at mesoscopic scales in active biological systems. *Science* **352**, 604–607 (2016).
155. Lucente, D., Baldassarri, A., Puglisi, A., Vulpiani, A. & Viale, M. Inference of time irreversibility from incomplete information: linear systems and its pitfalls. *Phys. Rev. Res.* **4**, 043103 (2022).
156. Frishman, A. & Ronceray, P. Learning force fields from stochastic trajectories. *Phys. Rev. X* **10**, 021009 (2020).
157. Skinner, D. J. & Dunkel, J. Improved bounds on entropy production in living systems. *Proc. Natl Acad. Sci. USA* **118**, e2024300118 (2021).
158. Wan, K. Y. & Goldstein, R. E. Time irreversibility and criticality in the motility of a flagellate microorganism. *Phys. Rev. Lett.* **121**, 058103 (2018).
159. Larson, B. T., Garbus, J., Pollack, J. B. & Marshall, W. F. A unicellular walker controlled by a microtubule-based finite-state machine. *Curr. Biol.* **32**, 3745–3757 (2022).

160. Lynn, C. W., Cornblath, E. J., Papadopoulos, L., Bertolero, M. A. & Bassett, D. S. Broken detailed balance and entropy production in the human brain. *Proc. Natl Acad. Sci. USA* **118**, e2109889118 (2021).
161. Martiniani, S., Lemberg, Y., Chaikin, P. M. & Levine, D. Correlation lengths in the language of computable information. *Phys. Rev. Lett.* **125**, 170601 (2020).
162. Ro, S. et al. Model-free measurement of local entropy production and extractable work in active matter. *Phys. Rev. Lett.* **129**, 220601 (2022).
163. Nardini, C. et al. Entropy production in field theories without time-reversal symmetry: quantifying the non-equilibrium character of active matter. *Phys. Rev. X* **7**, 021007 (2017).
164. Tkacik, G. & Bialek, W. Information processing in living systems. *Annu. Rev. Condens. Matter Phys.* **7**, 89–117 (2016).
165. Lynn, C. W., Holmes, C. M., Bialek, W. & Schwab, D. J. Decomposing the local arrow of time in interacting systems. *Phys. Rev. Lett.* **129**, 118101 (2022).
166. Bauer, M., Petkova, M. D., Gregor, T., Wieschaus, E. F. & Bialek, W. Trading bits in the readout from a genetic network. *Proc. Natl Acad. Sci. USA* **118**, e2109011118 (2021).
167. Mattingly, H., Kamino, K., Machta, B. & Emonet, T. *Escherichia coli* chemotaxis is information limited. *Nat. Phys.* **17**, 1426–1431 (2021).
168. Landauer, R. Computation: a fundamental physical view. *Phys. Scr.* **35**, 88 (1987).
169. Still, S., Sivak, D. A., Bell, A. J. & Crooks, G. E. Thermodynamics of prediction. *Phys. Rev. Lett.* **109**, 120604 (2012).
170. Adhikari, S., Kabakçıoğlu, A., Strang, A., Yuret, D. & Hinczewski, M. Machine learning in and out of equilibrium. Preprint at https://arxiv.org/abs/2306.03521 (2023).
171. Li, J., Liu, C.-W. J., Szurek, M. & Fakhri, N. Measuring irreversibility from learned representations of biological patterns. Preprint at https://arxiv.org/abs/2305.19983 (2023).
172. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
173. Campbell, D., Farmer, D., Crutchfield, J. & Jen, E. Experimental mathematics: the role of computation in nonlinear science. *Commun. ACM* **28**, 374–384 (1985).
174. Feldman, D. P., McTague, C. S. & Crutchfield, J. P. The organization of intrinsic computation: complexity–entropy diagrams and the diversity of natural information processing. *Chaos* **18**, 043106 (2008).
175. Mitchell, M., Crutchfield, J. P. & Hraber, P. T. Dynamics, computation, and the 'edge of chaos': a re-examination. In *Santa Fe Institute Studies in the Sciences of Complexity* Vol. 19, 497–497 (Addison-Wesley Publishing Co, 1994).
176. Carroll, T. L. Do reservoir computers work best at the edge of chaos? *Chaos* **30**, 121109 (2020).
177. Fajardo-Fontiveros, O. et al. Fundamental limits to learning closed-form mathematical models from data. *Nat. Commun.* **14**, 1043 (2023).
178. Krishnamurthy, K., Can, T. & Schwab, D. J. Theory of gating in recurrent neural networks. *Phys. Rev. X* **12**, 011011 (2022).
179. Mikhaeil, J., Monfared, Z. & Durstewitz, D. On the difficulty of learning chaotic dynamics with RNNs. *Adv. Neural Inf. Process. Syst.* **35**, 11297–11312 (2022).
180. Marzen, S. E., Riechers, P. M. & Crutchfield, J. P. Complexity-calibrated benchmarks for machine learning reveal when next-generation reservoir computer predictions succeed and mislead. Preprint at https://arxiv.org/abs/2303.14553 (2023).
181. Ding, X., Zou, Z. & Brooks III, C. L. Deciphering protein evolution and fitness landscapes with latent space models. *Nat. Commun.* **10**, 5644 (2019).
182. Huijben, I. A., Nijdam, A. A., Overeem, S., Van Gilst, M. M. & Van Sloun, R. SOM-CPC: unsupervised contrastive learning with self-organizing maps for structured representations of high-rate time series. In *International Conference on Machine Learning* 14132–14152 (PMLR, 2023).
183. Kantz, H. & Schreiber, T. *Nonlinear Time Series Analysis* Vol. 7 (Cambridge Univ. Press, 2004).
184. Deyle, E. R. & Sugihara, G. Generalized theorems for nonlinear state space reconstruction. *PLoS ONE* **6**, e18295 (2011).
185. Stark, J. Delay embeddings for forced systems. i. Deterministic forcing. *J. Nonlinear Sci.* **9**, 255–332 (1999).
186. Nash, J. The imbedding problem for Riemannian manifolds. *Ann. Math.* **63**, 20–63 (1956).
187. Eftekhari, A., Yap, H. L., Wakin, M. B. & Rozell, C. J. Stabilizing embedology: geometry-preserving delay-coordinate maps. *Phys. Rev. E* **97**, 022222 (2018).
188. Grebogi, C., Ott, E. & Yorke, J. A. Unstable periodic orbits and the dimensions of multifractal chaotic attractors. *Phys. Rev. A* **37**, 1711 (1988).
189. Cvitanović, P. Invariant measurement of strange sets in terms of cycles. *Phys. Rev. Lett.* **61**, 2729 (1988).
190. Lai, Y.-C., Nagai, Y. & Grebogi, C. Characterization of the natural measure by unstable periodic orbits in chaotic attractors. *Phys. Rev. Lett.* **79**, 649 (1997).
191. Lathrop, D. P. & Kostelich, E. J. Characterization of an experimental strange attractor by periodic orbits. *Phys. Rev. A* **40**, 4028 (1989).
192. Yalnız, G., Hof, B. & Budanur, N. B. Coarse graining the state space of a turbulent flow using periodic orbits. *Phys. Rev. Lett.* **126**, 244502 (2021).
193. Graham, M. D. & Floryan, D. Exact coherent states and the nonlinear dynamics of wall-bounded turbulent flows. *Annu. Rev. Fluid Mech.* **53**, 227–253 (2021).
194. Bramburger, J. J. & Fantuzzi, G. Data-driven discovery of invariant measures. Preprint at https://arxiv.org/abs/2308.15318 (2023).
195. Crowley, C. J. et al. Turbulence tracks recurrent solutions. *Proc. Natl Acad. Sci. USA* **119**, e2120665119 (2022).
196. Ahamed, T., Costa, A. C. & Stephens, G. J. Capturing the continuous complexity of behaviour in caenorhabditis elegans. *Nat. Phys.* **17**, 275–283 (2021).
197. Foti, N., Xu, J., Laird, D. & Fox, E. Stochastic variational inference for hidden Markov models. In *Advances in Neural Information Processing Systems* Vol. 27 (NIPS 2014).
198. Kalman, R. E. A new approach to linear filtering and prediction problems. *Trans. ASME D* **82**, 35–45 (1960).
199. Roweis, S. & Ghahramani, Z. A unifying review of linear Gaussian models. *Neural Comput.* **11**, 305–345 (1999).
200. Goodfellow, I. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems* Vol. 27 (NIPS 2014).
201. Kingma, D. P., Mohamed, S., Jimenez Rezende, D. & Welling, M. Semi-supervised learning with deep generative models. In *Proc. 27th International Conference on Neural Information Processing Systems* (NeurIPS, 2014).
202. Tang, B. & Matteson, D. S. Probabilistic transformer for time series analysis. *Adv. Neural Inf. Process. Syst.* **34**, 23592–23608 (2021).
203. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **85**, 461 (2000).
204. Bollt, E. M. Review of chaos communication by feedback control of symbolic dynamics. *Int. J. Bifurcat. Chaos* **13**, 269–285 (2003).
205. Baptista, M. & Kurths, J. Chaotic channel. *Phys. Rev. E* **72**, 045202 (2005).
206. Lu, Z. & Bassett, D. S. Invertible generalized synchronization: a putative mechanism for implicit learning in neural systems. *Chaos* **30**, 063133 (2020).
207. Pathak, J., Hunt, B., Girvan, M., Lu, Z. & Ott, E. Model-free prediction of large spatiotemporally chaotic systems from data: a reservoir computing approach. *Phys. Rev. Lett.* **120**, 024102 (2018).
208. Gilpin, W. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *35th Conference on Neural Information Processing Systems* (NeurIPS, 2021).

Acknowledgements

The author thanks H. Abarbanel and H. Swinney for informative discussions. This project has been made possible in part by grant number DAF2023-329596 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

Competing interests

The author declares no competing interests.

Additional information

Peer review information *Nature Reviews Physics* thanks the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.