Out-of-Domain Generalization in Dynamical Systems Reconstruction

Niclas Göring [1 2] **Florian Hess** [1 2] **Manuel Brenner** [1 2] **Zahra Monfared** [1] **Daniel Durstewitz** [1 2 3]

Abstract

In science we are interested in finding the governing equations, the dynamical rules, underlying empirical phenomena. While traditionally scientific models are derived through cycles of human insight and experimentation, recently deep learning (DL) techniques have been advanced to reconstruct dynamical systems (DS) directly from time series data. State-of-the-art dynamical systems reconstruction (DSR) methods show promise in capturing invariant and long-term properties of observed DS, but their ability to generalize to unobserved domains remains an open challenge. Yet, this is a crucial property we would expect from any viable scientific theory. In this work, we provide a formal framework that addresses generalization in DSR. We explain why and how out-of-domain (OOD) generalization (OODG) in DSR profoundly differs from OODG considered elsewhere in machine learning. We introduce mathematical notions based on topological concepts and ergodic theory to formalize the idea of learnability of a DSR model. We formally prove that black-box DL techniques, without adequate structural priors, generally will not be able to learn a generalizing DSR model. We also show this empirically, considering major classes of DSR algorithms proposed so far, and illustrate where and why they fail to generalize across the whole state space. Our study provides the first comprehensive mathematical treatment of OODG in DSR, and gives a deeper conceptual understanding of where the fundamental problems in OODG lie and how they could possibly be addressed in practice.

[1]Department of Theoretical Neuroscience, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany [2]Faculty of Physics and Astronomy, Heidelberg University, Heidelberg, Germany [3]Interdisciplinary Center for Scientific Computing, Heidelberg University, Heidelberg, Germany. Correspondence to: Niclas Göring <niclas.goering@gmail.com>, Daniel Durstewitz <Daniel.Durstewitz@zi-mannheim.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

The majority of complex systems we encounter in physics, biology, the social sciences, and beyond, can mathematically be described as systems of differential equations, whose behavior is the subject of dynamical systems theory (DST). Deriving accurate mathematical models of natural (or engineered) systems from observations for mechanistic insight, scientific understanding, and prediction, is the core of any scientific discipline. Recent years have seen a plethora of advances in the field of DS reconstruction (DSR) mostly based on deep learning (DL) approaches for inferring DS models directly from time series data and thus partly automatizing the scientific model building process (Brunton et al., 2016; Raissi et al., 2018; Vlachas et al., 2018; Platt et al., 2021; Brenner et al., 2022; Vlachas et al., 2022; Hess et al., 2023). Like any good scientific theory, a proper DS model inferred from data should be able to generalize to novel domains (dynamical regimes) not observed during training. Here we develop a principled mathematical framework for out-of-domain (OOD) generalization (OODG) in DSR. We mathematically and numerically demonstrate that current data-driven SOTA methods for DSR hit fundamental limits regarding OODG, and provide some directions of how these could potentially be addressed.

Current state of DSR Current DSR models attempt to either approximate the underlying system's vector field (Brunton et al., 2016), or try to directly learn the flow (solution) operator of the data-generating DS (Lu et al., 2019; Vlachas et al., 2020; Li et al., 2020; Brenner et al., 2022; Hess et al., 2023; Chen & Wu, 2023). More specifically, DSR methods have been developed based on symbolic regression (Brunton et al., 2016; d'Ascoli et al., 2023), on various forms of recurrent neural networks (RNNs) equipped with special training algorithms (Vlachas et al., 2018; Brenner et al., 2022; Hess et al., 2023), on ordinary or partial differential equation (ODE/PDE)-based DL models such as Neural ODEs (N-ODE; Chen et al. (2018); Ko et al. (2023)), on operator theory (Lu et al., 2019; Li et al., 2020; Chen & Wu, 2023) or on reservoir computing (RC; Pathak et al. (2017); Verzelli et al. (2021); Platt et al. (2022; 2023)). State-of-the-art (SOTA) methods (Brenner et al., 2022; Platt et al., 2023; Hess et al., 2023; Jiang et al., 2023) can generalize beyond the observed time horizon and capture an underlying system's dynamical

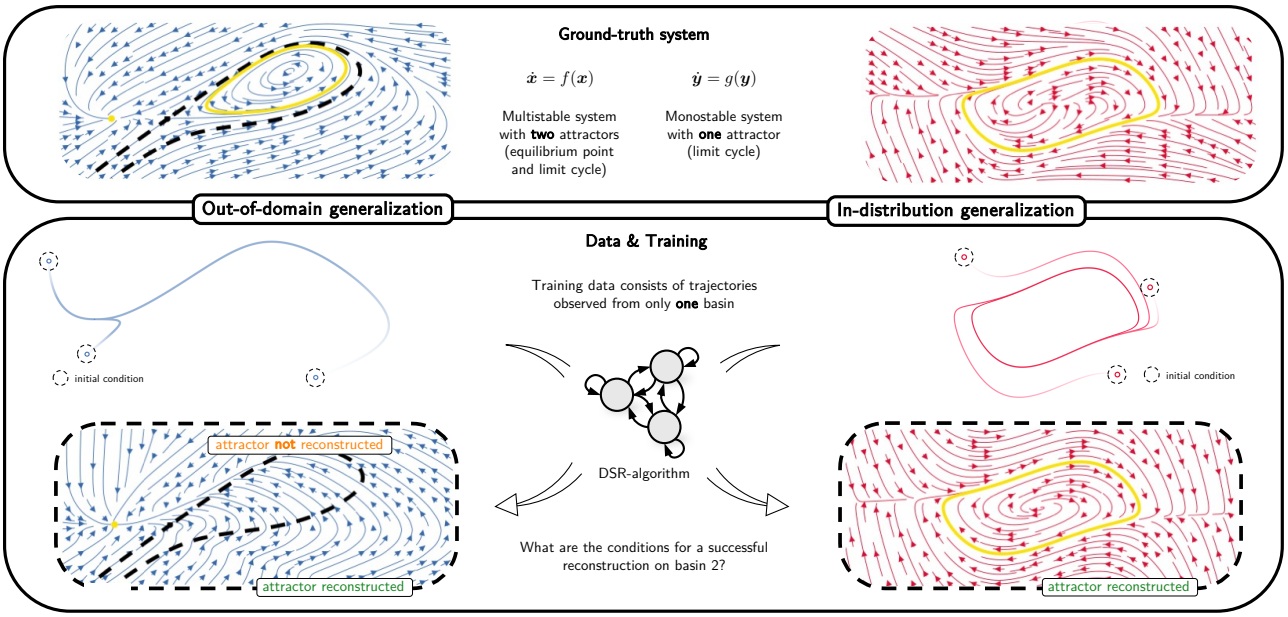

Figure 1. In-distribution generalization within one basin (right; van-der-Pol oscillator) vs. OODG across basins (left; neuron model with a limit cycle corresponding to spiking activity and an equilibrium point corresponding to the resting potential).

invariants and long-term properties, like the geometry of an attractor trajectories are converging to (Fig. A1), while providing accurate short-term forecasts on in-distribution test data. However, the current field of DSR has overwhelmingly focused on synthetic benchmark systems, e.g. given by low-order polynomial ODE and PDE systems, mostly in regimes where either only one (globally) attracting object exists in state space, as in the chaotic Lorenz system for common parameter settings (Lorenz, 1963), or at least reconstruction within just one dynamical regime was sought. Only a small number of studies considered experimental data, and even less consider systems which may harbor multiple attractor objects simultaneously, so-called *multistability* (Fig. 1, left; for a detailed overview of current benchmarks in use, see Appx. A).

An unresolved challenge in DSR: generalization to unobserved dynamical regimes While current SOTA methods for DSR may generalize to nearby initial conditions close to the domain covered by the training data, *which ultimately converge into the same limit set*, at current their ability to generalize to unobserved regions of state space is either not given or remains unexplored (Fig. 1). Generalization across the whole state space of the DS, or scientifically relevant portions of it, is, however, a feature any sound scientific theory should possess. The OODG problem is exacerbated in the presence of multistability, i.e., if multiple dynamical objects coexist in the same DS. In fact, even the simplest examples of low-dimensional nonlinear DS, like a damped-driven pendulum, often have multistable regimes. The problem is also of high practical relevance, as most complex DS encountered in nature and society are likely extensively multistable, with examples ranging from neuroscience (Schiff et al., 1994; Durstewitz et al., 2000; Izhikevich, 2007; Khona & Fiete, 2022), optics (Brambilla et al., 1991), chemistry (Ngonghala et al., 2011), biology (Dubinkina et al., 2019), ecology (Mumby et al., 2007), to financial markets (Cavalli & Naimzada, 2016) and climate science (Yoden, 1997). For such DS, crossing the boundaries between different dynamical regimes, as induced by noise or external inputs, may lead to qualitatively completely different behavior (Fig. 1, left; Fig. A2).

Current approaches toward OODG in DSR Several recent studies at least partially or implicitly address the question of OODG and multistability in DSR. For instance, in Ghadami & Epureanu (2018); Patel & Ott (2022); Bury et al. (2023), the authors attempt to anticipate tipping points in non-autonomous DS and predict the post-tipping point dynamics. A related topic is the forecasting of extreme events (Farazmand & Sapsis, 2018; Guth & Sapsis, 2019; Qi & Majda, 2020), which are events that are not, or only very sparsely, represented in the training data, thus constituting a form of generalization. Others consider learning DS across multiple environments defined by different parameter settings (Yin et al., 2022a; Bereska & Gavves, 2022). However, essentially all this work implicitly or explicitly assumed some observations from the domain on which generalization is sought to be available (i.e., reflecting more a form of transfer learning rather than true OODG; Kirch-

meyer et al. (2022); Yin et al. (2022a)). Another strategy to enable better generalization is to include physical domain knowledge into the model formulation or loss function, as in physics-informed neural networks (PINNs) (Raissi et al., 2019; Mehta et al., 2021; Subramanian & Mahadevan, 2023; Mouli et al., 2023), or directly ground-truth parameters of the DS studied (Fotiadis et al., 2023). This, again, assumes we already have prior knowledge about the domain which we would like to generalize to.

Promoting data-driven DSR models to viable scientific theories of complex systems requires a thorough understanding of whether, how, and when reconstructions generalize to the entire state space in the common empirical scenario where the measurements sample only a limited portion of that space (Fig. 1, left). This leads us to our central research question: what are the precise mathematical conditions that allow for successful reconstruction of a DS on the whole state space?

2. Dynamical Systems Background

2.1. Dynamical Systems

A DS is generally comprised of a state space $M \subseteq \mathbb{R}^n$, a set of times $\mathcal{T} \subseteq \mathbb{R}$, and an evolution law. Here we focus on continuous-time systems described by ODEs

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}), \qquad \boldsymbol{x} \in M \subseteq \mathbb{R}^n, \tag{1}$$

where $f \in \mathcal{X}^1(M)$ is a vector field (VF) from the set of functions with continuous first derivative on the (compact, metric, measurable) state space M. The VF gives rise to the evolution operator $\Phi : \mathcal{T} \times M \to M$ that maps some initial condition $\boldsymbol{x}_0$ to the state $\boldsymbol{x}_t$ at time t (Kuznetsov, 1998):

$$\boldsymbol{x}_t = \Phi(t, \boldsymbol{x}_0). \tag{2}$$

2.2. Dynamical Systems Reconstruction (DSR)

In data-driven DSR, the aim is to infer from time series observations a *generative model* of the true underlying system, approximating either its vector field $f \in \mathcal{X}^1(M)$ or evolution operator $\Phi(t, \boldsymbol{x})$ (hence its governing equations) given an inference algorithm from a hypothesis class $\mathcal{H}$. This goes beyond mere time series forecasting, in that we require the model to also capture dynamical invariants, that is long-term statistics and topological properties, of the underlying system. Thus, after training a DSR model should ideally be topologically conjugate to the true system and capable of producing trajectories with the same temporal and geometrical structure as those of the true system (Fig. A1; Platt et al. (2022; 2023); Hess et al. (2023)). Note that this subsumes various more specific goals one may have in time series modeling and DS analysis. For instance, a proper DSR model should also provide excellent time series predictions, while, vice versa, a model optimized for time series prediction would not necessarily reproduce invariant statistics and the geometry and topology of a system's attractors.

2.3. Measure Theoretic Aspects

Measure theoretical approaches investigate the long-term statistical properties of DS, the subject of ergodic theory (Eckmann & Ruelle, 1985). Specifically, there is a stable statistical property called the *(average) occupation measure*, defined as

$$\mu_{\boldsymbol{x}_0,T}(B) = \frac{1}{T}\int_0^T \mathbb{1}_B(\boldsymbol{x}(s))\, ds \tag{3}$$

where $\boldsymbol{x}(t)$ is a trajectory with $t \in [0, T]$, starting from $\boldsymbol{x}_0$, $B \subset \mathbb{R}^n$ is some Borel measurable set, and $\mathbb{1}_B$ denotes an indicator function that maps elements of the set B to one, and all other elements to zero. Intuitively, $\mu_{\boldsymbol{x}_0,T}(B)$ measures the amount of time the trajectory $\boldsymbol{x}(t)$ spends in the set B.

2.4. Topological Aspects and Attractors

Another way to capture the long-term behavior of a DS is by studying special sets, so called invariant sets, which describe the 'anatomy' of a DS. A set $U \subseteq M$ is called invariant under the flow, if $\Phi(t, U) \subseteq U \quad \forall t$. We can associate any point in state space with a special invariant set capturing its long-term behavior, the so-called ω-limit set

$$\omega(\boldsymbol{x}, \Phi) = \bigcap_{s \in \mathbb{R}} \overline{\{\Phi(t, \boldsymbol{x}) | t > s\}}, \tag{4}$$

where the overline denotes closure. If there is a set of points in state space with non-zero measure that have the same ω-limit set, it is called an attractor. More formally:

Definition 2.1. An *attractor* (Milnor, 1985; Perko, 1991) is a closed invariant set $A \subseteq M$ such that there exists an open and forward-invariant set $B(A) = \{\boldsymbol{x} \in M | \omega(\boldsymbol{x}, \Phi) \subseteq A\}$, called the *basin of attraction*, with $\omega(B(A)) = A$. We further require that A is minimal (i.e., there is no proper subset with that same property).

Stable equilibrium points, limit cycles and chaotic ('strange') sets are examples of attractors with increasingly complicated topology (see Fig. A3).

3. A General Framework for OODG in DSR

3.1. Multistability Induces Distribution Shifts

In statistical learning theory, out-of-sample generalization, and – more importantly here – OODG, is already quite well-studied (for a detailed treatment, see Appx. C.1,

Ben-David et al. (2010)). Generally, one assumes to have observed a set of training domains E on which the data is distributed according to some domain specific distribution p^e, $e \in E$. The goal is to learn a function $\hat{f} \in \mathcal{F}$ in hypothesis class $\mathcal{F}$ that yields minimum risk $R^{e_{\text{test}}}(\hat{f}) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p^{e_{\text{test}}}}[\ell(\hat{f}(\boldsymbol{x}), \boldsymbol{y})]$ on an unseen test domain $e_{\text{test}} \notin E$, where the data is distributed according to $p^{e_{\text{test}}}$. Since $p^{e_{\text{test}}}$ is unknown, one estimates the respective prediction error by taking the average loss across all empirically accessible domains as empirical risk:

$$R^E_{\text{emp}}(\hat{f}) = \frac{1}{|E|} \sum_{e\in E} \frac{1}{n_e} \sum_{i=1}^{n_e} \ell(\hat{f}(\boldsymbol{x}_i^e), \boldsymbol{y}_i^e). \tag{5}$$

Yin et al. (2022b) extend this definition to time-series data:

$$R^E_{\text{emp}}(\Phi_R) = \frac{1}{|E|} \sum_{e\in E} \frac{1}{T_e} \sum_{t=1}^{T_e} \ell(\Phi_R(t, \boldsymbol{x}_0^e), \boldsymbol{x}_t^e), \tag{6}$$

where $|E|$ is the number of domains and T_e time series length.[1] Yin et al. (2022b) mainly associate these domains with different parameter settings of the ground-truth system. For multistable DS, however, the different domains obtain a very natural interpretation.

A DS is called *multistable* if it has at least two attractors coexisting in its state space (Fig. 1, left). In this study, we generally assume that multistable systems allow for a decomposition of their state space into n disjoint basins of attraction (Milnor, 1985):

$$M = \sqcup_{e=1}^n B(A_e) \sqcup \tilde{M} \quad \text{such that} \quad \mu\big(\tilde{M}\big) = 0, \tag{7}$$

where μ is the Lebesgue measure. It is thus natural to define the domains in Eq. (6) as the different basins of attraction, each of which belongs to a different attractor. Different attractors generally give rise to different long-term dynamics with different topology (Fig. 1, left; Fig. A2), governed by different physical measures (see Appx B.1). Hence, in each basin the trajectories follow a different statistical law (regarding their long-term evolution), implying that the challenge of reconstructing multistable DS is ultimately the same as that of understanding OODG in DSR. In monostable DS, in contrast, each trajectory in the state space is governed by the same long-term statistics (Fig. 1, right). Hence, one (sufficiently long) trajectory is already enough to specify the dynamics on the attractor. Accordingly, generalization for monostable systems essentially comes down to classical in-distribution (out-of-sample) generalization. Note that this is also true for chaotic attractors. Indeed, as illustrated in Fig. A1 (and amply demonstrated in the literature, e.g. Mikhaeil et al. (2022); Brenner et al. (2022); Hess et al. (2023)), current SOTA methods fare very well on even short trajectories from complex chaotic systems, as long as these are monostable (or sufficient information from all basins of attraction is available, see Fig. A7). Yet, the DSR field overwhelmingly so far focused on just monostable hyperbolic attractor systems as benchmarks (Appx. A), making OODG for DSR an essentially unstudied problem.

[1] For chaotic systems, both stationary and non-stationary probability distributions are possible (Parthasarathy & Rajasekar, 1998).

3.2. OODG Error

The most commonly employed loss function ℓ is the mean-squared-error (MSE), derived from the maximum likelihood principle assuming i.i.d. Gaussian model residuals. While this is still the default when it comes to training RNNs, N-ODEs or RCs, it cannot be used to *assess* the reconstruction for three reasons: 1) The MSE breaks down as a suitable test loss in Eq. (6) because of exponential trajectory divergence in chaotic DS (Wood, 2010; Koppe et al., 2019). Even for a perfectly reconstructed DS, numerical uncertainties in the initial conditions or small amounts of noise quickly lead to large prediction errors. This is accounted for in training methods like generalized teacher forcing (Hess et al., 2023); 2) The MSE does not capture any long-term, invariant or topological properties of the DS and its reconstruction. As discussed in Sect. 2.3 & 2.4, these are the central mathematical tools to study DS; 3) The MSE is not guaranteed to be sensitive to multistability, yet this property of a measure is much needed in light of Sect. 3.1. We thus propose a novel way of assessing generalization across state space by defining a statistical and a topological error that are provably sensitive to multistability (Theorem 3.3).

Statistical error As the MSE is not a useful quantity for comparing (chaotic) trajectories, we define the statistical error through the sliced Wasserstein-1 distance (SW_1; Bonneel et al. (2015)) between the occupation measures $\mu^{\Phi}_{\boldsymbol{x},T}$ of the ground-truth DS Φ and $\mu^{\Phi_R}_{\boldsymbol{x},T}$ of the DSR model Φ_R:

$$\text{SW}_1(\mu^{\Phi}_{\boldsymbol{x},T}, \mu^{\Phi_R}_{\boldsymbol{x},T}) = \mathbb{E}_{\xi\sim\mathcal{U}(\mathbb{S}^{n-1})}\left[W_1(g_{\boldsymbol{\xi}}\sharp\mu^{\Phi}_{\boldsymbol{x},T}, g_{\boldsymbol{\xi}}\sharp\mu^{\Phi_R}_{\boldsymbol{x},T})\right], \tag{8}$$

where $\mathbb{S}^{n-1} := \{\boldsymbol{\xi} \in \mathbb{R}^n \mid \|\boldsymbol{\xi}\|_2^2 = 1\}$ is the unit hyper-sphere, $g_{\boldsymbol{\xi}}\sharp\mu$ denotes the pushforward of μ, $g_{\boldsymbol{\xi}}(\boldsymbol{x}) = \boldsymbol{\xi}^T\boldsymbol{x}$ is the one-dimensional slice projection, and W_1 the Wasserstein-1 distance (Villani et al., 2009). Since the expectation in Eq. (8) is intractable, it is commonly approximated by Monte Carlo sampling (see Appx. C.2 for computational details).

Definition 3.1. The statistical error $\mathcal{E}_{\text{stat}}$ is defined as

$$\mathcal{E}^U_{\text{stat}}(\Phi_R) := \int_{U\subseteq M} \text{SW}_1(\mu^{\Phi}_{\boldsymbol{x},T}, \mu^{\Phi_R}_{\boldsymbol{x},T})\, d\boldsymbol{x}, \tag{9}$$

which integrates across initial conditions from a subset U of state space.

Topological error An important concept to assess if two DS agree in their topology is *topological equivalence*. Two DS are called topologically equivalent if there exists a homeomorphism between the two system's orbits preserving the direction of flow (see Appx. B.3 for more details). However, this homeomorphism is usually not known a priori and hard to access numerically. Hence, we will replace the condition of topological equivalence with three weaker conditions based on the Lyapunov spectrum, which contains topological and stability information about limit sets of a DS. Let us denote the ordered, largest n Lyapunov exponents of limit sets $\omega(\boldsymbol{x}, \Phi)$ and $\omega(\boldsymbol{x}, \Phi_R)$ by $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $\lambda_1^R \leq \lambda_2^R \leq \cdots \leq \lambda_n^R$, respectively. First, we require that all Lyapunov exponents agree in their signs, $sgn(\lambda_i) = sgn(\lambda_i^R)\, \forall i$. Second, we demand that the maximum Lyapunov exponent is close in relative error, $|\lambda_n - \lambda_n^R| \,/\, |\lambda_n| < \varepsilon_{\lambda_n}$, where ε_{λ_n} is a tolerance. Lastly, the limit sets need to be close in state space, $d_H(\omega(\boldsymbol{x}, \Phi_R), \omega(\boldsymbol{x}, \Phi)) < \varepsilon_{d_H}$, assessed through the Hausdorff distance (see Appx. C.3 for more details). We then define an indicator function on M, $\mathbb{1}_{\Phi_R}(\boldsymbol{x})$, which is equal to 1 for a given point $\boldsymbol{x} \in U \subseteq M$ iff the associated limit sets fulfill all of the three conditions above.

Definition 3.2. We define the topological generalization error on $U \subseteq M$ as

$$\mathcal{E}_{\text{top}}^{U}(\Phi_R) = 1 - \frac{1}{\text{vol}(U)} \int_{U \subseteq M} \mathbb{1}_{\Phi_R}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{10}$$

In the following, we will use $\mathcal{E}_{\text{gen}}$ as a placeholder for both $\mathcal{E}_{\text{top}}$ and $\mathcal{E}_{\text{stat}}$, and statements involving $\mathcal{E}_{\text{gen}}$ must hold for *both* errors.[2] These errors are highly sensitive to a failure to reconstruct multistable systems:

Theorem 3.3. *Assume Φ is multistable with decomposition as in Eq. (7) and connected basins, and there exists one attractor A_k, $k \leq n$, not reconstructed by Φ_R. Then, the generalization error of Φ_R is proportional to the volume of the basin of this non-reconstructed attractor:*

$$\mathcal{E}_{\text{gen}}^{M_{\text{test}}}(\Phi_R) \propto \text{vol}(B(A_k)). \tag{11}$$

This statement naturally generalizes to the case of multiple non-reconstructed attractors (with different proportionality constants).

Proof. See Appx. E.1. □

3.3. OOD Learnability in DSR

Learnability is a fundamental concept in statistical learning theory (Vapnik, 2000; Shalev-Shwartz et al., 2010), with many different definitions advanced (Valle-Pérez & Louis, 2020). In its simplest form, a hypothesis class is called learnable if, for any distribution of training data, the error between the learned and ground-truth function decreases monotonically with sample size and converges to zero in the limit of infinitely many data points. This concept has been extended to OODG settings in Fang et al. (2023). To apply these definitions to DSR, assume the state space segregates into n basins (domains), Eq. (7), $|E| < n$ of which form the training domains $M_{\text{train}} = \cup_{e \in E} B(A_e)$ and all others the test domains M_{test}. For simplicity, we assume we have access to the data generating process Φ on M_{train}, such that the training data can be expressed as $\mathcal{D} \subseteq \cup_{\boldsymbol{x}_0 \in M_{\text{train}}} \Phi(T, \boldsymbol{x}_0)$ where $[0,\ T]$ is the time interval in which the trajectories are observed. In line with statistical DL theory, we further assume that $\mathcal{H}$ includes hypotheses consistent with both the training and test data (Valle-Pérez et al., 2019; Belkin, 2021). In other words, there exist models within $\mathcal{H}$ that, in theory, achieve zero reconstruction error on both the training and test domains (but in practice will depend on uncertainties introduced by the DSR algorithm). Denote by $\Theta_0 = \{\boldsymbol{\theta} \in \Theta | \mathcal{E}_{\text{gen}}^{M_{\text{train}}}(\Phi_\theta) = 0\}$ the set of parameters associated with models having (near) zero reconstruction error on the training domain, and by $\mathcal{H}_0$ the corresponding set of DS. Then, learnability in DSR boils down to:

Definition 3.4. The OODG problem $(\mathcal{H}, \mathcal{D})$ defined by hypothesis class $\mathcal{H}$ and dataset $\mathcal{D}$ is *strictly learnable* if

$$\forall\, \Phi_R \in \mathcal{H}_0 : \quad \mathcal{E}_{\text{gen}}^{M_{\text{test}}}(\Phi_R) = 0. \tag{12}$$

Hence, the OODG-problem is *strictly* learnable, if zero reconstruction error on the training domain leads to zero reconstruction error on the test domain.

For highly expressive hypothesis classes there can be multiple, if not infinitely many, models in $\mathcal{H}_0$ with different generalization errors on M_{test}. If we assume we are dealing with a *parameterized* function class $\mathcal{H}_\theta = \{\Phi_\theta | \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^P\}$, as practically the case in all DL & DSR settings, the quantity of interest becomes the distribution of generalization errors of models in $\mathcal{H}_0$:

Definition 3.5. We define the *learnability*-distribution of the OODG problem $(\mathcal{H}_\theta, \mathcal{D})$ as

$$p(\varepsilon_{\text{gen}} | \mathcal{D}) = \frac{1}{\text{vol}(\Theta_0)} \int_{\Theta_0} \mathbb{1}[\mathcal{E}_{\text{gen}}^{M_{\text{test}}}(\Phi_{\boldsymbol{\theta}}) = \varepsilon_{\text{gen}}] d\boldsymbol{\theta}, \tag{13}$$

the probability of a model with zero reconstruction error on the training domain having a generalization error of ε_{gen} on M_{test}, where $\mathbb{1}[\cdot]$ returns 1 if the condition in square brackets holds and 0 otherwise. The more mass the distribution has at zero, the better the problem is learnable. In

[2] Note that $\mathcal{E}_{\text{top}}$ and $\mathcal{E}_{\text{stat}}$ are solely theoretical constructs we introduce to formalize the OODG in DSR problem, not loss functions to be used in training.

the limit of $p(\varepsilon_{\text{gen}}|\mathcal{D})$ being fully concentrated at $\varepsilon_{\text{gen}} = 0$, the OODG problem becomes strictly learnable. Table A17 summarizes the most important differences between OODG from the perspectives of standard statistical learning theory vs. of DST as advanced here.

4. Results

The learnability of an OODG problem depends both on the hypothesis class $\mathcal{H}$ as well as the chosen prior in $\mathcal{H}$ through the training algorithm. Therefore, we will examine the following two scenarios:

- **Strong prior** (Sect. 4.1): Library-based algorithms such as SINDy (Brunton et al., 2016) introduce a strong inductive bias by explicitly providing a function class for the underlying VF.
- **No prior** (Sect. 4.2): Approaches based on universal approximators of DS, like RNNs (Funahashi & Nakamura, 1993; Kimura & Nakano, 1998; Hanson & Raginsky, 2020), do not incorporate any explicit prior (but may still introduce implicit priors through the choice of training algorithm and parameter initialization).

4.1. Strong Prior: Methods Based on Predefined Function Libraries

Following the classical statistical approach of basis expansions (Hastie et al., 2009; Durstewitz, 2017), some popular DSR methods rest on a predefined library of basis functions in the observables (Brunton et al., 2016; Reinbold et al., 2020), most prominently SINDy and its further developments (Brunton et al., 2016; Loiseau & Brunton, 2018; Kaiser et al., 2018; Cortiella et al., 2021; Messenger & Bortz, 2021; Kaheman et al., 2022). These models usually are linear in the parameters, thus easing statistical inference. Since the library of functions needs to be specified a priori, these methods induce a strong inductive bias. A strong sparsity prior on the parameters, and – correspondingly – sparse regression methods like LASSO or sequential thresholding (Brunton et al., 2016), ensure that only a small subset of functions from the library is selected for modeling the vector field (for details on SINDy, see Appx. D.2). More formally, this defines the class of finite-dimensional linearly parameterized functions with m differentiable basis functions $\psi_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \dots, m$,

$$\mathcal{B}_L = \left\{ f_j(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{m} \theta_{i,j}\psi_i(x) \,\middle|\, \forall j, \boldsymbol{\theta} \in \mathbb{R}^{m\times n} \right\}, \quad (14)$$

where these basis functions may be arbitrarily chosen. In this hypothesis class, one trajectory is sufficient to fully specify the DS, unless a certain uniqueness condition is violated:

Theorem 4.1. *Let $f \in \mathcal{B}_L$ be a multistable VF, and assume SINDy (or related) is used to learn Φ_R, including the right terms from $\mathcal{B}_L$ in its library. If there exists a trajectory $\Gamma_{\boldsymbol{x}_0} \subset \mathcal{D}$ not solving an algebraic equation in the parameters of Eq. (14), then the OODG problem given by $(\mathcal{B}_L, \mathcal{D})$ is strictly learnable.*

Proof. See Appx. E.2. □

This implies that a single trajectory from one basin is enough for the DSR model to capture the dynamics on all other basins, as long as $\mathcal{D}$ contains a trajectory not solving an algebraic equation in the parameters of Eq. (14)[3] and the correct function library is provided[4]. These observations are illustrated in Fig. 2a. Appx. D.4 provides an efficient formal procedure for checking whether the conditions on a given trajectory are met, and hence a generalizing solution could be found. In situations where the trajectory solves an algebraic equation, we can further restrict the library to find a unique solution (Corollary E.7). We remark that these conditions are usually established by LASSO.

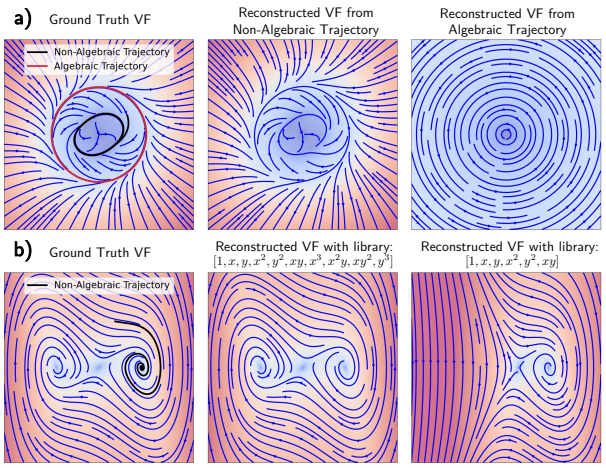

Figure 2. a) Example reconstructions using SINDy (details in Appx. D.2). The underlying VF has two cycle solutions. One solves an algebraic equation (red), while the other does not (black). The VF is only correctly identified from a trajectory containing the inner cycle (center), but not for the outer cycle (right). b) SINDy needs the proper function library to correctly infer a system across the whole state space (center). If the 3rd order term present in the Duffing equations is lacking (right), the inferred VF may only be locally correct (or not at all for more complex systems).

[3]There are in fact particular systems where *each* trajectory may solve an algebraic equation, e.g. vector fields with a rational first integral like algebraic Hamiltonians (see Appx. D.3 for more details).

[4]If this is not the case, sometimes SINDy may still be able to find a good approximation, depending on the degree of mismatch and the expressiveness of the library

As laid out in Appx. A, this has important implications, since many benchmark systems and scientific models in physics (Ramberg & Osgood, 1943), chemistry (Fernández-Ramos et al., 2006), ecology (Goel et al., 1971), or epidemiology (Kermack & McKendrick, 1927), are expressed in terms of polynomials. However, for many complex real-world systems, like climate or the brain, which we observe through time series measurements, this assumption is likely to be violated, with polynomials at best a convenient simplification. As Figs. 2b and A10 make clear, library approaches like SINDy will generally fail if the library does not contain the right terms describing the GT model.[5] Hence, in scientific ML we often turn to more flexible and expressive models.

4.2. No Priors: Universal Approximators

Next, we examine the most common data-driven approach of choosing a 'black-box' model, like an RNN, to approximate the flow of the underlying system, i.e. without assuming any prior knowledge about the to-be-modeled system. If we assume that these models operate in the universal approximation limit, we can show that there is an infinity of models in the hypothesis class having zero reconstruction error on the training domain but a very high error on the whole state space or for OOD data from M_{test}. This is in stark contrast to SINDy, where – *given the assumptions of Theorem 4.1 are met* – every model with zero reconstruction error on the training domain also has zero generalization error.

Theorem 4.2. *Let Φ be a multitstable flow that is not topologically transitive (cf. Appx. B.4) on M_{test}, generated by a VF $f \in \mathcal{X}^1$. Then, the OODG problem $(\mathcal{X}^1, \mathcal{D})$ is not strictly learnable. In fact, there exists an infinite family of $f \in \mathcal{X}^1$ and an $\varepsilon > 0$ such that the corresponding flows fulfill*

$$\mathcal{E}_{\text{gen}}^{M_{\text{train}}}(\Phi) = 0 \quad \text{and} \quad \mathcal{E}_{\text{gen}}^{M_{\text{test}}}(\Phi) \geq \varepsilon. \tag{15}$$

Proof. See Appx. E.3. □

Note that this result is independent from the loss function used in training. Fig. 3, where data were just sampled from one basin of attraction of the multistable Duffing system (Eq. (29)), illustrates this idea for three of the most commonly used DSR models (in stark contrast to DSR performance on monostable systems, cf. Fig. A1). We emphasize that this is not a sampling issue: Regardless of how much data are drawn from one basin, generalization fails, while increasing sample size quickly helps to identify the whole state space if data from both basins are available (Fig. A9). SINDy on the other hand, *provided the correct function library*, generalizes (Fig. 2b).

[5] In fact, SINDy fails on many empirical datasets from complex systems (Brenner et al., 2022; Hess et al., 2023).

It is important to note that while on multistable settings like the one above, if trajectories are drawn from just one basin OODG will generally fail, the very same architectures can be trained to approximately zero training error on the full state space M (see Fig. A5). This implies there are indeed regions in the loss landscape that would generalize, raising the question of why these are hardly ever discovered by the optimization algorithm.

4.3. Why OODG Fails

We will shed light on this failure, focusing on RNNs trained with SGD. Given data $\mathcal{D}$, the probability that an RNN after training has a generalization error ε_{gen} is formally given by

$$\begin{aligned} p_{\text{SGD}}(\varepsilon_{\text{stat}} \mid \mathcal{D}) = \int_{\Theta} \mathbb{1}[\mathcal{E}_{\text{gen}}^{M}(\Phi_{\boldsymbol{\theta}_f}) = \varepsilon_{\text{gen}}] \\ p_{\text{opt}}\left(\boldsymbol{\theta}_f \mid \boldsymbol{\theta}_i, \mathcal{D}\right) p_{\text{ini}}\left(\boldsymbol{\theta}_i\right) d\boldsymbol{\theta}_i d\boldsymbol{\theta}_f, \end{aligned} \tag{16}$$

where p_{ini} characterizes the initialization scheme and p_{opt} formalizes the training process, quantifying the probability of obtaining a final set of parameters $\boldsymbol{\theta}_f$ given an initial set $\boldsymbol{\theta}_i$. Under certain assumptions (cf. Appx. D.5 for details), Eq. (16) exactly aligns with the learnability distribution (Eq. (13)). We now illustrate how the implicit biases in p_{ini} or p_{opt} will impede OODG.

Simplicity bias in p_{ini} In recent studies of standard NNs (Valle-Pérez et al., 2019; Mingard et al., 2023) and transformers (Bhattamishra et al., 2023) it was shown that the parameter-function map $\mathcal{M}$ (Appx. D.6) is biased towards 'simple' functions, which in turn may explain the good generalization capability of these models on i.i.d. data (of course, time series data are not i.i.d. to begin with). Here we show that RNNs also exhibit a bias towards simplicity, which, in this case, unfortunately, manifests as a bias towards monostable DS.

To this end we initialized a shPLRNN Φ_θ (Hess et al. (2023), Appx. D.2) using the Glorot uniform and Glorot normal (Glorot & Bengio, 2010) scheme where we systematically varied the gain scaling the variance. We then uniformly drew N_I initial conditions and evolved them with this randomly initialized DSR model Φ_θ until the resulting trajectories had converged to a limit set. The distribution of these limit set points across state space was then quantified through the Shannon entropy, which gives a measure for the complexity of the attractor structure at initialization (Fig. 4a). In Fig. 4b, the mean entropy is plotted as a function of gain (variance), revealing a clear trend (see also Fig. A12 for a higher-dimensional RNN example). Increasing the parameter variance hence leads to more complex dynamics at initialization. However, we empirically observe that models initialized with high gains become almost impossible to train by SGD, as commonly observed for vanilla

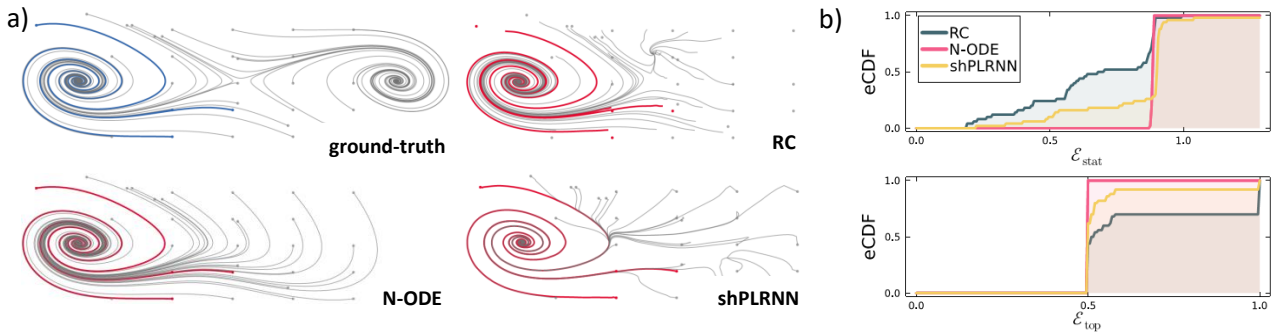

Figure 3. Learnability of three SOTA DSR algorithms evaluated on the Duffing system in a multistable regime. a) Reconstructions of DSR models trained on four ground-truth trajectories (blue) from one basin. Red trajectories are freely generated using initial conditions of the training data and the respective DSR model. Grey trajectories comprise example ground-truth test trajectories and generated ones from both the training basin and OOD basin. While training data trajectories align with the ground-truth, all models fail to properly generalize to the unobserved attractor/basin. b) Empirical cumulative distribution function (eCDF) of both $\mathcal{E}_{\text{stat}}$ and $\mathcal{E}_{\text{top}}$ based on $N = 50$ independent trainings of each DSR model evaluated over a grid of initial conditions covering both basins (see Fig. A4).

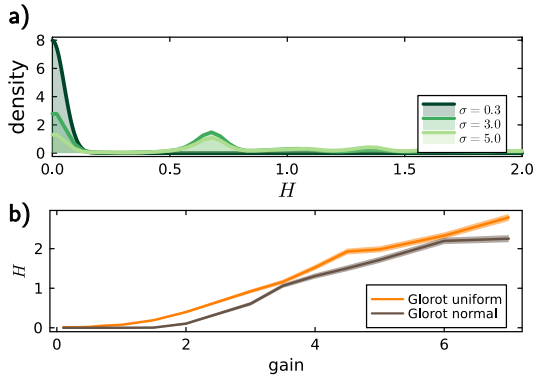

Figure 4. a) Distribution of Shannon entropies (in Nat) for the limit sets of shPLRNNs ($M = 2, H = 100$) initialized with different gains (parameter variances) using the Glorot uniform scheme. For a low gain ($\sigma = 0.3$), as predominantly used in DSR, the attractors of all models at initialization had $H = 0$, which means that these consisted only of a single equilibrium point. For higher gains, further peaks at $H > 0$ started to appear, implying that either more and/or higher-order objects (like cycles) exist upon initialization. b) Mean Shannon entropy for the same data plotted against gain, using the Glorot uniform and Glorot normal initialization scheme.

feed-forward NNs (Glorot & Bengio, 2010). This conflict effectively biases all trainable (shPL)RNNs toward monostability, and merely increasing the gain by itself is therefore not a viable option for enhancing OODG.

Generalizing solutions are saddle points While the implicit bias introduced by p_{ini} plays a role in OODG failure, uncertainties in the optimization, as quantified through p_{opt}, turned out to be even more crucial. To illustrate this, we consider the bistable Duffing oscillator (see Appx. A13 for a chaotic multistable example) and denote by $\boldsymbol{\theta}_{\text{gen}}$ the parameters of a model generalizing across M, i.e., with close to zero training loss $\ell_M = \ell_{B(A_1)} + \ell_{B(A_2)}$ and reconstruction error $\mathcal{E}_{\text{gen}}$ on both basins. We then retrain a model initialized with $\boldsymbol{\theta}_{\text{gen}}$ on trajectories from just one of the two basins, i.e. employing $\ell_{B(A_1)}$ as a loss function. In Fig. 5a we present the distribution of statistical errors for various generalizing models $\Phi_{\theta_{\text{gen}}}$ and retrained models $\Phi_{\theta_{re}}$ across the two basins, $B(A_1)$ and $B(A_2)$. We observe an about 20-fold increase in the reconstruction error of retrained compared to initialized models on $B(A_2)$, even though the error on $B(A_1)$ remained largely the same. Hence, the process of retraining effectively leads the models to *unlearn* the dynamics on the second basin. While here we illustrated

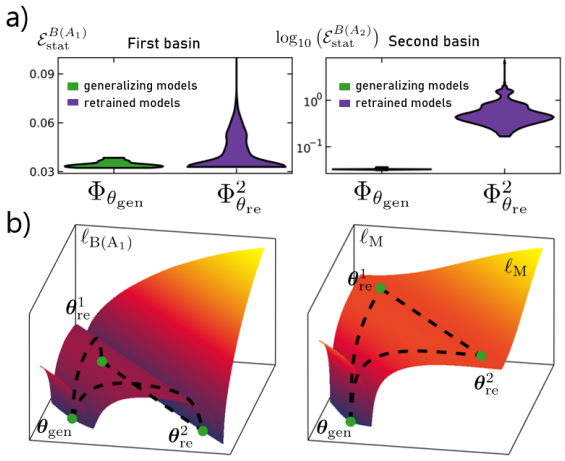

Figure 5. a) Statistical error distribution on basins $B(A_1)$ and $B(A_2)$ for 20 generalizing models (green) and 20×20 models retrained (purple) using only $B(A_1)$ data. b) Illustration of loss landscapes using data from just one (left) or both (right) basin(s) of attraction, with parameters corresponding to generalizing solution ($\boldsymbol{\theta}_{\text{gen}}$), and to models retrained for $125k$ ($\boldsymbol{\theta}^1_{re}$) and $250k$ ($\boldsymbol{\theta}^2_{re}$) parameter updates, respectively. Note that ℓ_{M} does not exhibit the spurious loss valley present in $\ell_{\text{B}(A_1)}$.

that this issue arises even in fairly simple systems like the bistable Duffing oscillator, Fig. A14 shows it is equally present in higher-dimensional, more complex systems like the generalized spatially extended chaotic Lorenz-96 (Pelzer & Sterk, 2020) model of atmospheric convection.

Since $\boldsymbol{\theta}_{\text{gen}}$ corresponds to a model that already agrees well with trajectories from *both basins* (low ℓ_M & $\mathcal{E}_{\text{stat}}^M$), this raises the question of why the optimizer leaves this regime during the retraining phase in the first place. To further understand this, we studied the Hessian of the loss functions evaluated on trajectories from just one ($\ell_{B(A_1)}$) or both (ℓ_M) basins w.r.t. $\boldsymbol{\theta}_{\text{gen}}$ (see Tab. A5). First, we noticed that $\boldsymbol{\theta}_{\text{gen}}$ is not a minimum but a saddle in both loss landscapes. Further, the Hessian of $\ell_{B(A_1)}$ has much fewer positive eigenvalues than that of ℓ_M, implying that the saddle is more stable (with less directions to escape) when trajectories from both basins are provided. Hence, as soon as data from one basin are removed from the training set, the optimizer will run into new directions with zero or small negative eigenvalue, thus forgetting the second equilibrium point. Fig. 5b further shows that the removal of data from the second basin leads to the emergence of spurious extrema. Current training routines may thus not be built to learn multistable systems, as they unlearn the multistable property even upon perfect initialization.

Generalizing minima are sharp In 'standard' DL, the width of minima correlates with generalization, with wider minima generalizing better than narrow ones (Hochreiter & Schmidhuber, 1997). While certain studies have contested this correlation (Dinh et al., 2017), large-scale studies, such as (Jiang et al., 2019), validate this association. Here, we adopt a specific notion of width based on the minima volumes or radii as outlined in Huang et al. (2020), where Appx. D.7 explains how this concept also applies to saddle regions. To further examine this idea, we trained shPLRNNs – as above – with identical architecture and hyperparameters once on a trajectory from just one basin and once from both basins of attraction of the Duffing system (see Fig. A15 for the same analysis for a chaotic multistable system). We made sure that both models have approximately the same training error when evaluated only on a single trajectory from the first basin. We then examined the width (radius) $r(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\min}\|_2$ of the minimum $\boldsymbol{\theta}_{\min}$ corresponding to the loss evaluated only on the one trajectory common to both (the mono- and the multistable) training setups, at a height 5% above the minimum value (other heights gave similar results, see Appx. D.7). On average, the minimum valleys corresponding to generalizing models, i.e. those trained on the whole state space, have a *smaller* radius (Fig. 6), in contrast to the more common observation in DL that generalizing minima are usually wider. This, in addition to the fact that SGD is more likely to converge to wider minima (Chaudhari et al., 2019; Foret et al., 2020; Xie et al., 2021), may further explain why generalizing minima are avoided in DSR.

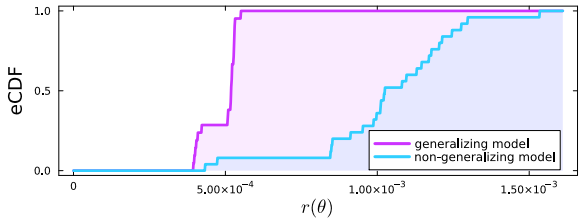

Figure 6. eCDF of minima radii for generalizing and non-generalizing models.

5. Discussion

Here we provide the first systematic mathematical treatment of OODG in DSR. We aimed to lay a theoretical foundation which could serve to guide the field toward future solutions of the OODG problem in DSR, by providing a new set of theoretically guided measures, providing theorems which clearly state what is, and what is not, possible, and by delineating where the hard problems lie and exactly why current SOTA algorithms struggle with them. The core problem is that most naturally observed DS will harbor many co-existing dynamical regimes, characterized by different VF topologies and long-term statistics, but usually we have observed data only from one or few of them. If we already know the correct function class, we can infer models (like SINDy) which generalize across the whole state space. But for the likely much more common empirical scenario where this is not the case, unique identification of a generalizing solution is no longer possible. In fact, if a chosen library does not even work on the training domain, this is already a strong hint that crucial terms are missing. Unfortunately, intentionally choosing a very expressive, too-large function library is not a remedy either (let alone for computational reasons), as it makes the problem underspecified.

Practically, one DS-agnostic way to potentially address OODG may be by targeting implicit biases in the initialization and, more importantly, the training processes (cf. Sect. 4.3), for instance by promoting solutions that explicitly encourage multistability. Often, however, we may still need to guide the training process by a more profound physical or biological understanding of the DS in question, and evaluate trained models by explicitly (experimentally) testing novel predictions. More generally, future work may want to put the focus on training algorithms that encourage and preserve multistability and avoid overfitting the training basin.

All code used here is available at `https://github.com/DurstewitzLab/OODG-in-DSR`.

Acknowledgements

This work was funded by the German Research Foundation (DFG) within Germany's Excellence Strategy EXC 2181/1 – 390900948 (STRUCTURES) and by DFG grant Du354/15-1 to DD.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, May 2021. ISSN 0962-4929, 1474-0508. doi: 10.1017/S0962492921000039. Publisher: Cambridge University Press.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1): 151–175, May 2010. ISSN 1573-0565. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Bereska, L. and Gavves, E. Continual Learning of Dynamical Systems With Competitive Federated Reservoir Computing. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, pp. 335–350. PMLR, November 2022. URL https://proceedings.mlr.press/v199/bereska22a.html. ISSN: 2640-3498.

Bhattamishra, S., Patel, A., Kanade, V., and Blunsom, P. Simplicity bias in transformers and their ability to learn sparse boolean functions, 2023.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Brambilla, M., Lugiato, L. A., Penna, V., Prati, F., Tamm, C., and Weiss, C. O. Transverse laser patterns. ii. variational principle for pattern selection, spatial multistability, and laser hydrodynamics. *Phys. Rev. A*, 43:5114–5120, May 1991. doi: 10.1103/PhysRevA.43.5114. URL https://link.aps.org/doi/10.1103/PhysRevA.43.5114.

Brenner, M., Hess, F., Mikhaeil, J. M., Bereska, L. F., Monfared, Z., Kuo, P.-C., and Durstewitz, D. Tractable Dendritic RNNs for Reconstructing Nonlinear Dynamical Systems. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 2292–2320. PMLR, June 2022. URL https://proceedings.mlr.press/v162/brenner22a.html. ISSN: 2640-3498.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data: Sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, April 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1517384113. URL http://arxiv.org/abs/1509.03580. arXiv: 1509.03580.

Bury, T. M., Dylewsky, D., Bauch, C. T., Anand, M., Glass, L., Shrier, A., and Bub, G. Predicting discrete-time bifurcations with deep learning. *Nature Communications*, 14 (1):6331, October 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-42020-z. URL https://www.nature.com/articles/s41467-023-42020-z. Number: 1 Publisher: Nature Publishing Group.

Cavalli, F. and Naimzada, A. Complex dynamics and multistability with increasing rationality in market games. *Chaos, Solitons & Fractals*, 93:151–161, 2016. ISSN 0960-0779. doi: https://doi.org/10.1016/j.chaos.2016.10.014. URL https://www.sciencedirect.com/science/article/pii/S0960077916303095.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-SGD: biasing gradient descent into wide valleys*. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, December 2019. ISSN 1742-5468. doi: 10.1088/1742-5468/ab39d9. URL https://dx.doi.org/10.1088/1742-5468/ab39d9. Publisher: IOP Publishing and SISSA.

Chen, J. and Wu, K. Deep-osg: Deep learning of operators in semigroup. *Journal of Computational Physics*, 493: 112498, 2023.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Climenhaga, V., Luzzatto, S., and Pesin, Y. The Geometric Approach for Constructing Sinai–Ruelle–Bowen Measures. *Journal of Statistical Physics*, 166(3):467–493, February 2017. ISSN 1572-9613. doi: 10.1007/s10955-016-1608-7. URL https://doi.org/10.1007/s10955-016-1608-7.

Cortiella, A., Park, K.-C., and Doostan, A. Sparse identification of nonlinear dynamical systems via reweighted l1-regularized least squares. *Computer Methods in Applied Mechanics and Engineering*, 376:113620, April 2021. ISSN 0045-7825. doi: 10.1016/j.cma.2020.113620. URL https://www.sciencedirect.com/science/article/pii/S0045782520308057.

d'Ascoli, S., Becker, S., Mathis, A., Schwaller, P., and Kilbertus, N. Odeformer: Symbolic regression of dynamical systems with transformers. *arXiv preprint arXiv:2310.05573*, 2023.

Dawkins, J. Same occupation measure → same trajectory. MathOverflow, 2024. URL https://mathoverflow.net/q/463088. URL:https://mathoverflow.net/q/463088 (version: 2024-01-29), USER: (https://mathoverflow.net/users/42851/john-dawkins).

de Silva, B. M., Champion, K., Quade, M., Loiseau, J.-C., Kutz, J. N., and Brunton, S. L. PySINDy: A Python package for the Sparse Identification of Nonlinear Dynamics from Data. *arXiv preprint arXiv:2004.08424*, 2020. URL http://arxiv.org/abs/2004.08424.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp Minima Can Generalize For Deep Nets. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1019–1028. PMLR, July 2017. URL https://proceedings.mlr.press/v70/dinh17b.html. ISSN: 2640-3498.

Dubinkina, V., Fridman, Y., Pandey, P. P., and Maslov, S. Multistability and regime shifts in microbial communities explained by competition for essential nutrients. *eLife*, 8:e49720, nov 2019. ISSN 2050-084X. doi: 10.7554/eLife.49720. URL https://doi.org/10.7554/eLife.49720.

Duffing, G. *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung*. Number 41-42. Vieweg, 1918.

Durstewitz, D. *Advanced Data Analysis in Neuroscience*. Bernstein Series in Computational Neuroscience. Springer International Publishing, Cham, 2017. ISBN 978-3-319-59974-8 978-3-319-59976-2. doi: 10.1007/978-3-319-59976-2. URL http://link.springer.com/10.1007/978-3-319-59976-2.

Durstewitz, D., Seamans, J. K., and Sejnowski, T. J. Neurocomputational models of working memory. *Nature Neuroscience*, 3(11):1184–1191, November 2000. ISSN 1546-1726. doi: 10.1038/81460. URL https://www.nature.com/articles/nn1100_1184. Number: 11 Publisher: Nature Publishing Group.

Eckmann, J. P. and Ruelle, D. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, 57 (3):617–656, July 1985. doi: 10.1103/RevModPhys.57.617. URL https://link.aps.org/doi/10.1103/RevModPhys.57.617. Publisher: American Physical Society.

Fang, Z., Li, Y., Lu, J., Dong, J., Han, B., and Liu, F. Is Out-of-Distribution Detection Learnable?, February 2023. URL http://arxiv.org/abs/2210.14707. arXiv:2210.14707 [cs, stat].

Farazmand, M. and Sapsis, T. P. Extreme Events: Mechanisms and Prediction, March 2018. URL http://arxiv.org/abs/1803.06277. arXiv:1803.06277 [nlin, physics:physics].

Fernández-Ramos, A., Miller, J. A., Klippenstein, S. J., and Truhlar, D. G. Modeling the Kinetics of Bimolecular Reactions. *Chemical Reviews*, 106(11):4518–4584, November 2006. ISSN 0009-2665. doi: 10.1021/cr050205w. URL https://doi.org/10.1021/cr050205w. Publisher: American Chemical Society.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware Minimization for Efficiently Improving Generalization. October 2020. URL https://openreview.net/forum?id=6Tm1mposlrM.

Fotiadis, S., Valencia, M. L., Hu, S., Garasto, S., Cantwell, C. D., and Bharath, A. A. Disentangled Generative Models for Robust Prediction of System Dynamics. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10222–10248. PMLR, July 2023. URL https://proceedings.mlr.press/v202/fotiadis23a.html. ISSN: 2640-3498.

Funahashi, K.-i. and Nakamura, Y. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801–806, January 1993. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80125-X. URL https://www.sciencedirect.com/science/article/pii/S089360800580125X.

Geist, K., Parlitz, U., and Lauterborn, W. Comparison of different methods for computing lyapunov exponents. *Progress of theoretical physics*, 83(5):875–893, 1990.

Ghadami, A. and Epureanu, B. I. Forecasting critical points and post-critical limit cycles in nonlinear oscillatory systems using pre-critical transient responses. *International Journal of Non-Linear Mechanics*, 101:146–156, May 2018. ISSN 0020-7462. doi: 10.1016/j.ijnonlinmec.2018.02.008. URL https://www.sciencedirect.com/science/article/pii/S0020746217306716.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, March 2010. URL https://proceedings.mlr.press/v9/glorot10a.html. ISSN: 1938-7228.

Goel, N. S., MAITRA, S. C., and MONTROLL, E. W. On the Volterra and Other Nonlinear Models of Interacting Populations. *Reviews of Modern Physics*, 43 (2):231–276, April 1971. doi: 10.1103/RevModPhys.43.231. URL https://link.aps.org/doi/10.1103/RevModPhys.43.231. Publisher: American Physical Society.

Guth, S. and Sapsis, T. P. Machine Learning Predictors of Extreme Events Occurring in Complex Dynamical Systems. *Entropy*, 21(10):925, October 2019. ISSN 1099-4300. doi: 10.3390/e21100925. URL https://www.mdpi.com/1099-4300/21/10/925. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

Hanson, J. and Raginsky, M. Universal simulation of stable dynamical systems by recurrent neural nets. In Bayen, A. M., Jadbabaie, A., Pappas, G., Parrilo, P. A., Recht, B., Tomlin, C., and Zeilinger, M. (eds.), *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 384–392. PMLR, 10–11 Jun 2020. URL https://proceedings.mlr.press/v120/hanson20a.html.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7. URL http://link.springer.com/10.1007/978-0-387-84858-7.

Hess, F., Monfared, Z., Brenner, M., and Durstewitz, D. Generalized Teacher Forcing for Learning Chaotic Dynamics. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 13017–13049. PMLR, July 2023. URL https://proceedings.mlr.press/v202/hess23a.html. ISSN: 2640-3498.

Hochreiter, S. and Schmidhuber, J. Flat Minima. *Neural Computation*, 9(1):1–42, January 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.1. URL https://doi.org/10.1162/neco.1997.9.1.1.

Huang, W. R., Emam, Z., Goldblum, M., Fowl, L., Terry, J. K., Huang, F., and Goldstein, T. Understanding Generalization Through Visualizations. pp. 87–97. PMLR, February 2020. URL https://proceedings.mlr.press/v137/huang20a.html. ISSN: 2640-3498.

Innes, M., Saba, E., Fischer, K., Gandhi, D., Rudilosso, M. C., Joy, N. M., Karmali, T., Pal, A., and Shah, V. Fashionable modelling with flux. *CoRR*, abs/1811.01457, 2018. URL https://arxiv.org/abs/1811.01457.

Izhikevich, E. M. *Dynamical systems in neuroscience: the geometry of excitability and bursting*. Computational neuroscience. MIT Press, Cambridge, Mass, 2007. ISBN 978-0-262-09043-8. OCLC: ocm65400606.

Jiang, R., Lu, P. Y., Orlova, E., and Willett, R. Training neural operators to preserve invariant measures of chaotic attractors, 2023.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic Generalization Measures and Where to Find Them, December 2019. URL http://arxiv.org/abs/1912.02178. arXiv:1912.02178 [cs, stat].

Kaheman, K., Brunton, S. L., and Kutz, J. N. Automatic differentiation to simultaneously identify nonlinear dynamics and extract noise probability distributions from data. *Machine Learning: Science and Technology*, 3(1):015031, March 2022. ISSN 2632-2153. doi: 10.1088/2632-2153/ac567a. URL https://dx.doi.org/10.1088/2632-2153/ac567a. Publisher: IOP Publishing.

Kaiser, E., Kutz, J. N., and Brunton, S. L. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2219):20180335, November 2018. doi: 10.1098/rspa.2018.0335. URL https://royalsocietypublishing.org/doi/full/10.1098/rspa.2018.0335. Publisher: Royal Society.

Kermack, W. O. and McKendrick, A. G. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772): 700–721, 1927. ISSN 0950-1207. URL https://www.jstor.org/stable/94815. Publisher: The Royal Society.

Khona, M. and Fiete, I. R. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12):744–766, December 2022. ISSN 1471-0048. doi: 10.1038/s41583-022-00642-0. URL https://www.nature.com/articles/s41583-022-00642-0. Number: 12 Publisher: Nature Publishing Group.

Kimura, M. and Nakano, R. Learning dynamical systems by recurrent neural networks from orbits. *Neural Networks*, 11(9):1589–1599, 1998.

Kirchmeyer, M., Yin, Y., Donà, J., Baskiotis, N., Rakotomamonjy, A., and Gallinari, P. Generalizing to New Physical Systems via Context-Informed Dynamics Model, June 2022. URL http://arxiv.org/abs/2202.01889. arXiv:2202.01889 [cs, stat].

Ko, J.-H., Koh, H., Park, N., and Jhe, W. Homotopy-based training of NeuralODEs for accurate dynamics discovery, May 2023. URL http://arxiv.org/abs/2210.01407. arXiv:2210.01407 [physics].

Kolouri, S., Zou, Y., and Rohde, G. K. Sliced Wasserstein Kernels for Probability Distributions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5258–5267, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.568. URL http://ieeexplore.ieee.org/document/7780937/.

Koppe, G., Toutounji, H., Kirsch, P., Lis, S., and Durstewitz, D. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. *PLOS Computational Biology*, 15(8):e1007263, August 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007263. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007263. Publisher: Public Library of Science.

Kunze, H. Solving inverse problems for odes using the picard contraction mapping. 05 1999.

Kuznetsov, Y. A. *Elements of Applied Bifurcation Theory (2nd Ed.)*. Springer-Verlag, Berlin, Heidelberg, 1998. ISBN 0387983821.

Lee, J. M. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2012. ISBN 978-1-4419-9981-8 978-1-4419-9982-5. doi: 10.1007/978-1-4419-9982-5. URL https://link.springer.com/10.1007/978-1-4419-9982-5.

Li, Z., Kovachki, N. B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., Anandkumar, A., et al. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2020.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, April 2020.

Loiseau, J.-C. and Brunton, S. L. Constrained sparse Galerkin regression. *Journal of Fluid Mechanics*, 838: 42–67, March 2018. ISSN 0022-1120, 1469-7645. doi: 10.1017/jfm.2017.823. Publisher: Cambridge University Press.

Lorenz, E. N. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141, 1963.

Lu, L., Jin, P., and Karniadakis, G. E. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

Lü, J., Chen, G., and Cheng, D. A new chaotic system and beyond: the generalized lorenz-like system. *International Journal of Bifurcation and Chaos*, 14(05):1507–1537, May 2004. ISSN 0218-1274. doi: 10.1142/S021812740401014X. URL https://www.worldscientific.com/doi/abs/10.1142/S021812740401014X. Publisher: World Scientific Publishing Co.

Mehta, V., Char, I., Neiswanger, W., Chung, Y., Nelson, A., Boyer, M., Kolemen, E., and Schneider, J. Neural Dynamical Systems: Balancing Structure and Flexibility in Physical Prediction. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 3735–3742, December 2021. doi: 10.1109/CDC45484.2021.9682807. ISSN: 2576-2370.

Messenger, D. A. and Bortz, D. M. Weak SINDy: Galerkin-Based Data-Driven Model Selection. *Multiscale Modeling & Simulation*, 19(3):1474–1497, January 2021. ISSN 1540-3459. doi: 10.1137/20M1343166. URL https://epubs.siam.org/doi/10.1137/20M1343166. Publisher: Society for Industrial and Applied Mathematics.

Mikhaeil, J. M., Monfared, Z., and Durstewitz, D. On the difficulty of learning chaotic dynamics with RNNs. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, October 2022. URL https://openreview.net/forum?id=-_AMpmyV0Ll.

Milnor, J. On the concept of attractor. *Communications in Mathematical Physics*, 99(2):177–195, June 1985. ISSN 1432-0916. doi: 10.1007/BF01212280. URL https://doi.org/10.1007/BF01212280.

Mingard, C., Rees, H., Valle-Pérez, G., and Louis, A. A. Do deep neural networks have an inbuilt occam's razor?, 2023.

Mouli, S. C., Alam, M. A., and Ribeiro, B. MetaPhysiCa: OOD Robustness in Physics-informed Machine Learning, March 2023. URL http://arxiv.org/abs/2303.03181. arXiv:2303.03181 [cs, stat].

Mumby, P. J., Hastings, A., and Edwards, H. J. Thresholds and the resilience of caribbean coral reefs. *Nature*, 450 (7166):98–101, 2007. doi: 10.1038/nature06252. URL https://doi.org/10.1038/nature06252.

Ngonghala, C. N., Feudel, U., and Showalter, K. Extreme multistability in a chemical model system. *Phys. Rev.*

E, 83:056206, May 2011. doi: 10.1103/PhysRevE.83.056206. URL https://link.aps.org/doi/10.1103/PhysRevE.83.056206.

Odani, K. The Limit Cycle of the van der Pol Equation Is Not Algebraic. *Journal of Differential Equations*, 115(1):146–152, January 1995. ISSN 0022-0396. doi: 10.1006/jdeq.1995.1008. URL https://www.sciencedirect.com/science/article/pii/S002203968571008X.

Parthasarathy, S. and Rajasekar, S. Probability distribution characteristics of chaos in a simple population model and the Bonhoeffer–van der Pol oscillator. *Physical Review E*, 58(5):6839–6842, November 1998. doi: 10.1103/PhysRevE.58.6839. URL https://link.aps.org/doi/10.1103/PhysRevE.58.6839. Publisher: American Physical Society.

Patel, D. and Ott, E. Using Machine Learning to Anticipate Tipping Points and Extrapolate to Post-Tipping Dynamics of Non-Stationary Dynamical Systems, July 2022. URL http://arxiv.org/abs/2207.00521. arXiv:2207.00521 [physics].

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., and Ott, E. Using Machine Learning to Replicate Chaotic Attractors and Calculate Lyapunov Exponents from Data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(12): 121102, December 2017. ISSN 1054-1500, 1089-7682. doi: 10.1063/1.5010300. URL http://arxiv.org/abs/1710.07313. arXiv: 1710.07313.

Pelzer, A. F. G. and Sterk, A. E. Finite Cascades of Pitchfork Bifurcations and Multistability in Generalized Lorenz-96 Models. *Mathematical and Computational Applications*, 25(4):78, December 2020. ISSN 2297-8747. doi: 10.3390/mca25040078. URL https://www.mdpi.com/2297-8747/25/4/78. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Perko, L. *Differential Equations and Dynamical Systems*. Texts in applied mathematics. World Publishing Company, 1991. ISBN 9780387974439.

Platt, J. A., Wong, A., Clark, R., Penny, S. G., and Abarbanel, H. D. Robust forecasting using predictive generalized synchronization in reservoir computing. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(12), 2021.

Platt, J. A., Penny, S. G., Smith, T. A., Chen, T.-C., and Abarbanel, H. D. I. A Systematic Exploration of Reservoir Computing for Forecasting Complex Spatiotemporal Dynamics, January 2022. URL http://arxiv.org/abs/2201.08910. arXiv:2201.08910 [cs].

Platt, J. A., Penny, S. G., Smith, T. A., Chen, T.-C., and Abarbanel, H. D. Constraining chaos: Enforcing dynamical invariants in the training of reservoir computers. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33 (10), 2023.

Pérez-Hernández, J. A. and Benet, L. PerezHz/TaylorIntegration.jl: TaylorIntegration v0.4.1, February 2019. URL https://doi.org/10.5281/zenodo.2562353.

Qi, D. and Majda, A. J. Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences*, 117(1):52–59, January 2020. doi: 10.1073/pnas.1917285117. URL https://www.pnas.org/doi/full/10.1073/pnas.1917285117. Publisher: Proceedings of the National Academy of Sciences.

Rackauckas, C. and Nie, Q. DifferentialEquations.jl–a performant and feature-rich ecosystem for solving differential equations in Julia. *Journal of Open Research Software*, 5(1), 2017.

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., and Ramadhan, A. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Multistep Neural Networks for Data-driven Discovery of Nonlinear Dynamical Systems. *arXiv:1801.01236 [nlin, physics:physics, stat]*, January 2018. URL http://arxiv.org/abs/1801.01236. arXiv: 1801.01236.

Raissi, M., Perdikaris, P., and Karniadakis, G. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019. ISSN 00219991. doi: 10.1016/j.jcp.2018.10.045. URL https://linkinghub.elsevier.com/retrieve/pii/S0021999118307125.

Ramberg, W. and Osgood, W. R. Description of stress-strain curves by three parameters, July 1943. URL https://ntrs.nasa.gov/citations/19930081614. NTRS Author Affiliations: National Bureau of Standards NTRS Report/Patent Number: NACA-TN-902 NTRS Document ID: 19930081614 NTRS Research Center: Legacy CDMS (CDMS).

Reinbold, P. A. K., Gurevich, D. R., and Grigoriev, R. O. Using Noisy or Incomplete Data to Discover Models of Spatiotemporal Dynamics. *Physical Review E*, 101(1): 010203, January 2020. ISSN 2470-0045, 2470-0053. doi:

10.1103/PhysRevE.101.010203. URL http://arxiv.org/abs/1911.03365. arXiv:1911.03365 [physics].

Schiff, S. J., Jerger, K., Duong, D. H., Chang, T., Spano, M. L., and Ditto, W. L. Controlling chaos in the brain. *Nature*, 370(6491):615–620, August 1994. ISSN 1476-4687. doi: 10.1038/370615a0. URL https://www.nature.com/articles/370615a0. Number: 6491 Publisher: Nature Publishing Group.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, Stability and Uniform Convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010. ISSN 1533-7928. URL http://jmlr.org/papers/v11/shalev-shwartz10a.html.

Subramanian, A. and Mahadevan, S. Probabilistic physics-informed machine learning for dynamic systems. *Reliability Engineering & System Safety*, 230:108899, February 2023. ISSN 0951-8320. doi: 10.1016/j.ress.2022.108899. URL https://www.sciencedirect.com/science/article/pii/S0951832022005142.

Valle-Pérez, G. and Louis, A. A. Generalization bounds for deep learning, December 2020. URL http://arxiv.org/abs/2012.04115. arXiv:2012.04115 [cs, stat].

Valle-Pérez, G., Camargo, C. Q., and Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions, 2019.

Vapnik, V. N. *The Nature of Statistical Learning Theory*. Springer, New York, NY, 2000. ISBN 978-1-4419-3160-3 978-1-4757-3264-1. doi: 10.1007/978-1-4757-3264-1. URL http://link.springer.com/10.1007/978-1-4757-3264-1.

Verzelli, P., Alippi, C., and Livi, L. Learn to synchronize, synchronize to learn. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(8), 2021.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Vlachas, P. R., Byeon, W., Wan, Z. Y., Sapsis, T. P., and Koumoutsakos, P. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2213): 20170844, May 2018. doi: 10.1098/rspa.2017.0844. URL https://royalsocietypublishing.org/doi/10.1098/rspa.2017.0844. Publisher: Royal Society.

Vlachas, P. R., Pathak, J., Hunt, B. R., Sapsis, T. P., Girvan, M., Ott, E., and Koumoutsakos, P. Backpropagation algorithms and Reservoir Computing in Recurrent Neural Networks for the forecasting of complex spatiotemporal dynamics. *Neural Networks*, 126:191–217, June 2020. ISSN 0893-6080. doi: 10.1016/j.neunet.2020.02.016. URL https://www.sciencedirect.com/science/article/pii/S0893608020300708.

Vlachas, P. R., Arampatzis, G., Uhler, C., and Koumoutsakos, P. Multiscale simulations of complex systems by learning their effective dynamics. *Nature Machine Intelligence*, 4(4):359–366, 2022.

Vogt, R., Puelma Touzel, M., Shlizerman, E., and Lajoie, G. On lyapunov exponents for RNNs: Understanding information propagation using dynamical systems tools. *Frontiers in Applied Mathematics and Statistics*, 8, 2022. ISSN 2297-4687. URL https://www.frontiersin.org/articles/10.3389/fams.2022.818799.

Wood, S. N. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, August 2010. ISSN 1476-4687. doi: 10.1038/nature09319. URL https://www.nature.com/articles/nature09319. Number: 7310 Publisher: Nature Publishing Group.

Xie, Z., Sato, I., and Sugiyama, M. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. 2021.

Yin, Y., Ayed, I., de Bézenac, E., Baskiotis, N., and Gallinari, P. Leads: Learning dynamical systems that generalize across environments, 2022a.

Yin, Y., Ayed, I., de Bézenac, E., Baskiotis, N., and Gallinari, P. LEADS: Learning Dynamical Systems that Generalize Across Environments, February 2022b. URL http://arxiv.org/abs/2106.04546. arXiv:2106.04546 [cs, stat].

Yoden, S. Classification of simple low-order models in geophysical fluid dynamics and climate dynamics. *Nonlinear Analysis: Theory, Methods & Applications*, 30(7):4607–4618, 1997. ISSN 0362-546X. doi: https://doi.org/10.1016/S0362-546X(97)00306-4. URL https://www.sciencedirect.com/science/article/pii/S0362546X97003064. Proceedings of the Second World Congress of Nonlinear Analysts.

A. Survey of Benchmark Systems

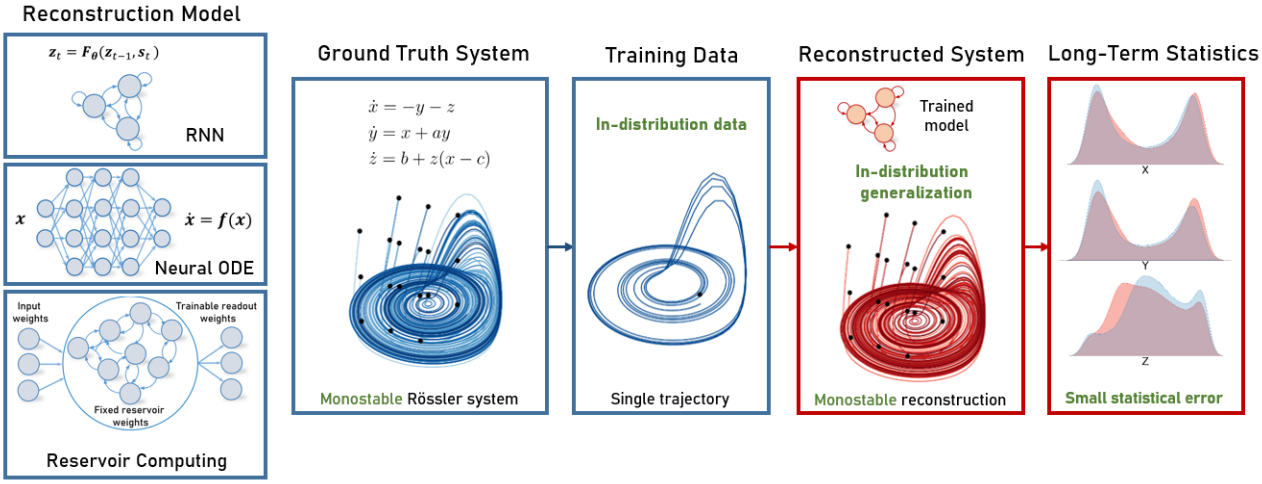

Figure A1. In-distribution generalization in DSR.

We surveyed 59 papers in the field of DSR, containing a wide range of methods and applications, with respect to the benchmark systems or datasets considered (Farmer & Sidorowich (1987); Wang & Lin (1998); Voss et al. (2004); Brunton et al. (2016); Trischler & D'Eleuterio (2016); Sussillo et al. (2016); Linderman et al. (2016); Tran & Ward (2017); Pathak et al. (2017); Raissi et al. (2018); Mohajerin & Waslander (2018) Vlachas et al. (2018); Lu et al. (2018); Lusch et al. (2018); Karlsson & Svanström (2019); Otto & Rowley (2019); Raissi et al. (2019); Duncker et al. (2019); Ayed et al. (2019); Nguyen et al. (2019); Qin et al. (2019); Fu et al. (2019) Champion et al. (2019); Singh et al. (2019); Lee & Carlberg (2020); Shalova & Oseledets (2020); Vlachas et al. (2020); Zhao et al. (2020); Hernandez et al. (2020); Azencot et al. (2020); Strauss (2020); Gilpin (2020); Nguyen et al. (2021) (Kraemer et al., 2021; Li & Duan, 2021; Lu et al., 2021; Schmidt et al., 2021; Jordana et al., 2021; Kim et al., 2021; Lai et al., 2021; Gauthier et al., 2021; Goyal & Benner, 2021; Liu & Jin, 2021; Schlaginhaufen et al., 2021) Mehta et al. (2021); Zhang et al. (2022); Uribarri & Mindlin (2022); Gilpin (2022); Yin et al. (2022b); Rusch et al. (2022); Brenner et al. (2022); Lejarza & Baldea (2022); Chen et al. (2022); Geneva & Zabaras (2022); Mikhaeil et al. (2022) Yang et al. (2023); Linot et al. (2023); Tripura & Chakraborty (2023); Hess et al. (2023)). This survey motivated the classification in Table A1, where three types of systems dominate the literature:

- Simple, low-dimensional linear or nonlinear systems like the Fitz-Hugh-Nagumo equations, Lotka-Volterra system, or coupled or damped harmonic oscillators/ pendulums like the van-der-Pol oscillator.
- Simple monostable 3d chaotic attractors, predominantly the Lorenz-63, Rössler or Duffing systems.
- Nonlinear PDEs as models of fluid dynamics and convection (e.g. Burgers equation, Navier Stokes equation, Lorenz-96 or Kuramoto–Sivashinsky equations).

Experimental data or explicitly multistable systems were rarely considered (or at least not explored in their multiple stable regimes).

Table A1. Classification of benchmark systems in the field of dynamical systems reconstruction.

Category	**Counts**
Linear Models/Oscillators	24
Chaotic 3D Models	29
Fluid Dynamics/PDEs	13
Experimental Data	6
Multistable	3

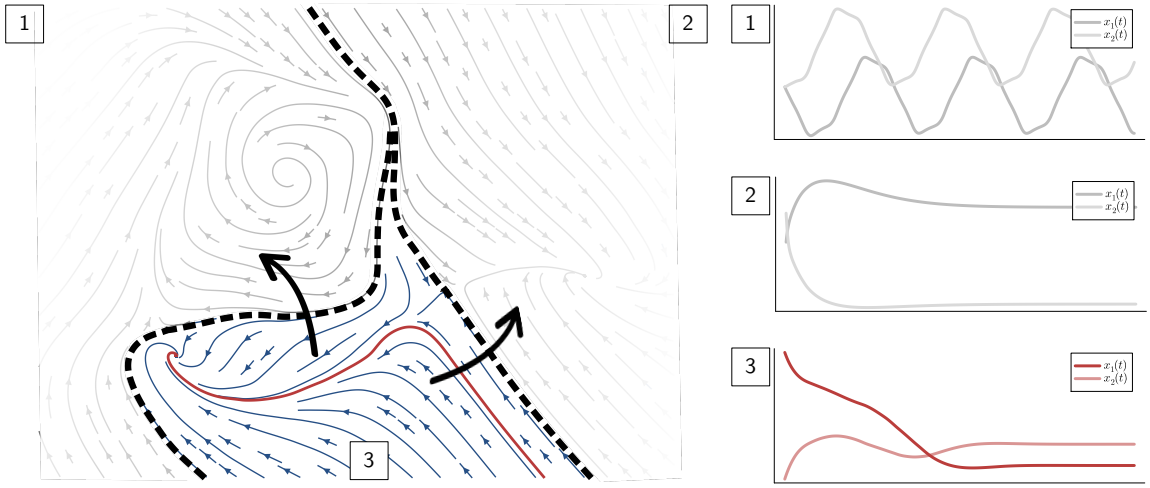

Figure A2. Illustration of multistability. Different basins of attraction can lead to completely different dynamical regimes with different topologies.

B. Further Details on Section 2

B.1. Ergodic Theory and Topology

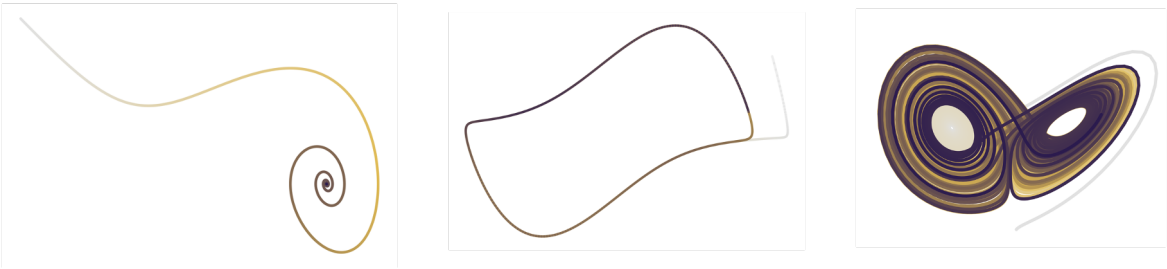

Figure A3. Trajectories of systems with an equilibrium point (Duffing oscillator), cycle (van der Pol oscillator) and chaotic attractor (Lorenz system).

Physical measure

Definition B.1. We call μ^* a *physical measure*, related to Sinai-Ruelle-Bowen (SRB) measures (Climenhaga et al., 2017), if for some set U with positive Lebesgue measure ($\mathcal{L}^n(U) > 0$) and $\boldsymbol{x}_0 \in U$,

$$\lim_{T\to\infty} \mu_{\boldsymbol{x}_0,T} \;=\; \mu^*. \tag{17}$$

In essence, this means that the measure is physically realisable. However, not every attractor (or even the DS itself) has to have a physical measure.

Hausdorff Distance

Definition B.2. Let X, Y be two non-empty subsets of a metric space (M, d). The Hausdorff-distance is defined by

$$d_{\mathrm{H}}(X,Y) = \max\left\{\sup_{x\in X} d(x,Y), \sup_{y\in Y} d(X,y)\right\} \tag{18}$$

where $d(a,B) = \inf_{b\in B} d(a,b)$ with $a \in X$ and $B \subseteq X$.

The choice of Hausdorff distance in this context is motivated by its robustness and sensitivity to outliers between the two sets, which in the context of the topological error makes it a suitable choice for assessing the closeness of the limit sets.

Topological Equivalence

Definition B.3. Let $F_1, F_2 \in \mathcal{C}^1(U)$ with flow maps $\phi_t^{F_1}, \phi_t^{F_2}$. The two vector fields (VFs) are topologically equivalent (Perko, 1991), denoted by $F_1 \simeq F_2$, if there exists a homeomorphism $h : U \to U$ mapping orbits of the first system onto orbits of the second system, i.e.

$$\forall t \in \mathbb{R},\ \forall x \in U : \quad \phi_t^{F_1}(x) = h^{-1} \circ \phi_\tau^{F_2} \circ h(x), \tag{19}$$

with $\tau : U \times \mathbb{R} \to \mathbb{R}, \quad \frac{\partial \tau(x,t)}{\partial t} > 0 \quad \forall x \in U$. This means the time direction of the orbits is preserved.

Loosely speaking, two VFs are topologically equivalent if we can continuously deform one VF into the other, i.e. such that each orbit is only deformed in a continuous manner without 'ripping it apart'. This implies that equilibrium points are mapped onto equilibrium points, and closed orbits to closed orbits. An open orbit will not be closed through h, and vice versa.

Topological Transitivity

Definition B.4. Let $F \in \mathcal{C}^1(U)$ be a vector field on a topological space U, with its associated flow map ϕ_t^F. The dynamical system induced by F is said to be topologically transitive if for any two non-empty open sets $A, B \subseteq U$, there exists a time $t \in \mathbb{R}$ such that the flow map at time t, ϕ_t^F, maps some part of A into B; that is, $\phi_t^F(A) \cap B \neq \emptyset$.

This implies that trajectories of the vector field F, starting from an arbitrary region in the space U, will eventually enter any other region, given that these regions are open and non-empty.

C. Further Details on Section 3

C.1. OODG in Statistical Learning Theory

Consider a regression or classification setting, where $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^k$ are the input and output spaces, and $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in X \times Y\}_{i=1}^n$ denotes a dataset sampled from a distribution $p(\boldsymbol{x}, \boldsymbol{y})$ defined on the domain $X \times Y$. Further assume there is a set E of *training* domains with cardinality $|E|$. Let the dataset of size n_e from a single environment $e \in E$ be $\mathcal{S}^e = \{(\boldsymbol{x}_i^e, \boldsymbol{y}_i^e) \in X^e \times Y^e\}_{i=1}^{n_e}$, where samples are drawn i.i.d. from the unknown, data-generating distribution $p^e(\boldsymbol{x}, \boldsymbol{y})$. Consider the class $\mathcal{F}$ of functions $f : X \to Y$ and a loss function $\ell : Y \times Y \to \mathbb{R}^+ \cup \{0\}$ measuring the goodness of fit. In OODG, the goal is to learn a generalizing, predictive function $\hat{f} \in \mathcal{F}$ from the $|E|$ training domains to obtain a minimum prediction error on an unseen test domain e_{test} with distribution $p^{e_{\text{test}}}(\boldsymbol{x}, \boldsymbol{y})$, i.e.

$$\min_{\hat{f} \in \mathcal{F}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}) \sim p^{e_{\text{test}}}} \left[\ell(\hat{f}(\boldsymbol{x}), \boldsymbol{y}) \right], \tag{20}$$

where

$$R^{e_{\text{test}}}(\hat{f}) \ := \ \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}) \sim p^{e_{\text{test}}}} \left[\ell(\hat{f}(\boldsymbol{x}), \boldsymbol{y}) \right] \ = \ \int \ell(\hat{f}(\boldsymbol{x}), \boldsymbol{y})\, dp^{e_{\text{test}}}(\boldsymbol{x}, \boldsymbol{y}) \tag{21}$$

is the expected loss for $\hat{f}$, called the test risk. However, we cannot compute $R^{e_{\text{test}}}(\hat{f})$ because the distribution $p^{e_{\text{test}}}(\boldsymbol{x}, \boldsymbol{y})$ is unknown. Hence, we estimate the expectation by the sample mean across all the training domains, called *empirical risk*, defined as

$$R_{\text{emp}}^E(\hat{f}) \ = \ \frac{1}{|E|} \sum_{e \in E} \frac{1}{n_e} \sum_{i=1}^{n_e} \ell(\hat{f}(\boldsymbol{x}_i^e), \boldsymbol{y}_i^e). \tag{22}$$

Based on this, the *OODG error* is defined as the gap between the test risk and the empirical risk,

$$\left| R^{e_{\text{test}}}(\hat{f}) - R_{\text{emp}}^E(\hat{f}) \right|. \tag{23}$$

C.2. Statistical Error

To compute Eq. (8), we use a common Monte-Carlo approximation of the expectation

$$\text{SW}_1(\mu_{\boldsymbol{x},T}^{\Phi},\ \mu_{\boldsymbol{x},T}^{\Phi_R}) \approx \frac{1}{L} \sum_{l=1}^{L} W_1(g_{\boldsymbol{\xi}^{(l)}} \sharp \mu_{\boldsymbol{x},T}^{\Phi},\ g_{\boldsymbol{\xi}^{(l)}} \sharp \mu_{\boldsymbol{x},T}^{\Phi_R}), \tag{24}$$

where projection vectors $\boldsymbol{\xi}^{(l)} \sim \mathcal{U}(\mathbb{S}^{n-1})$ are drawn uniformly across the unit hypersphere embedded in $\mathbb{R}^n$. The Wasserstein-1 distance is computed across trajectories (empirical distributions) of the ground-truth flow Φ and the reconstructed flow Φ_R. Trajectories are drawn by evolving the respective system for T time units from initial conditions $\boldsymbol{x} \in \mathbb{R}^n$. Between two one-dimensional distributions, the Wasserstein-1 distance can then efficiently be computed as

$$W_1(\mu, \nu) = \int_0^1 \left| F_\mu^{-1}(q) - F_\nu^{-1}(q) \right| dq, \tag{25}$$

where $F_\bullet^{-1}$ denotes the quantile function (inverse CDF). In practice, we approximate the integral in Eq. (25) by evaluating the quantile functions at a resolution of $\Delta q = 10^{-3}$. We use $L = 1000$ samples in Eq. (24). For the final error $\mathcal{E}_{stat}^U$, Eq. (9), we sample K initial conditions from a uniformly spaced grid $\text{Gr}(U) = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)}\}$ over $U \subset M \subset \mathbb{R}^n$. The integral in Eq. (9) is then approximated by

$$\mathcal{E}_{stat}^U(\Phi_R) \approx \frac{1}{K} \sum_{x \in \text{Gr}(U)} \text{SW}_1(\mu_{\boldsymbol{x},T}^{\Phi}, \mu_{\boldsymbol{x},T}^{\Phi_R}). \tag{26}$$

C.3. Topological Error

For the topological error, the Lyapunov spectra of orbits in limit sets $\omega(\boldsymbol{x}, \Phi)$ and $\omega(\boldsymbol{x}, \Phi_R)$ need to be computed. To compute the Lyapunov spectrum of continuous-time systems, i.e. ground-truth systems and Neural ODEs, we use the Julia library `TaylorIntegration.jl` (Pérez-Hernández & Benet, 2019). For RNNs and RCs we use our own implementation of an algorithm described in (Geist et al., 1990; Vogt et al., 2022), which computes the Lyapunov spectrum by evaluating the Jacobian product along orbits of length T:

$$\lambda_i = \lim_{T \to \infty} \frac{1}{T} \log \sigma_i \left(\prod_{t=0}^{T-1} \boldsymbol{J}_{T-t} \right), \tag{27}$$

where σ_i is the i-th singular value. For numerical stability, the product of Jacobians is repeatedly re-orthogonalized using a QR decomposition. To ensure convergence to the limit set spectrum, transients are discarded from the computation of Eq. (27). For the Duffing system (Appx. D.1), we discard the first $T_{trans} = 3000$ time steps and compute the Lyapunov spectrum across an additional $T = 3000$ time steps, while re-orthogonalizing every 50 time steps. For the multistable Lorenz-like system (Eq. (30)), we use $T_{trans} = 5000$ and $T = 10,000$. For the tolerance of the relative error between the maxmimum Lyapunov exponents λ_n and λ_n^R of the ground-truth and reconstructed system, respectively, we choose $\varepsilon_{\lambda_n} = 0.25$. For evaluating the agreement of limit sets, $d_H(\omega(\boldsymbol{x}, \Phi_R), \omega(\boldsymbol{x}, \Phi)) < \varepsilon_{d_H}$, we used the same setup as for computation of the Lyapunov spectra, but only use the $T' = 500$ and $T' = 5000$ last time steps. We set $\varepsilon_{d_H} = V/L$, where V is the volume of U, and L the number of initial conditions contained in the grid $\text{Gr}(U)$, which is the same as used for computation of $\mathcal{E}_{stat}^U(\Phi_R)$. For the Duffing system, this comes down to $\varepsilon_{d_H} = 40.0/100 = 0.4$. The integral in Eq. (10) is approximated and computed across the very same grid of initial conditions $Gr(U)$ as used for the statistical error $\mathcal{E}_{stat}^U$:

$$\mathcal{E}_{\text{top}}^U(\Phi_R) \approx 1 - \frac{1}{K} \sum_{\boldsymbol{x} \in \text{Gr}(U)} \mathbb{1}_{\Phi_R}(\boldsymbol{x}) \tag{28}$$

See Fig. A4 for a visualization of the grid of initial conditions used for the Duffing system.

D. Further Details on Section 4

D.1. Ground-truth Models

Duffing system The unforced Duffing system (Duffing, 1918) is given by a set of coupled ODEs:

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= ay - x\left(b + cx^2\right) \end{aligned} \tag{29}$$

where $[a, b, c] = [-\frac{1}{2}, -1, \frac{1}{10}]$ places the system into a multistable regime with two coexisting equilibrium points. To generate datasets, we numerically integrate Eq. (29) for $t_{int} = 40.0$ time units with a read-out interval of $\Delta t = 0.01$ using the adaptive step size integrator `Tsit5` provided within the Julia library `DifferentialEquations.jl` (Rackauckas & Nie, 2017). Using K initial conditions, this results in an array of shape $4000 \times 2 \times K$. To facilitate training, we standardize our datasets by the overall mean and standard deviation across all trajectories.

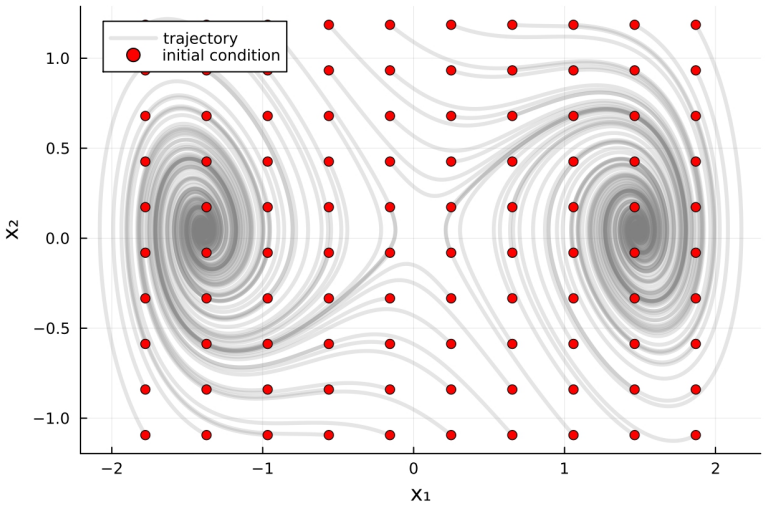

Figure A4. Grid $\mathrm{Gr}(U)$ used to compute $\mathcal{E}^U_{stat}$ as well as $\mathcal{E}^U_{top}$ for the Duffing system.

Multistable Lorenz-like system As an example system of multistable chaotic attractors, we use the multistable Lorenz-like system introduced in Lü et al. (2004):

$$\begin{aligned} \dot{x} &= -\tfrac{ab}{a+b}x - yz + c \\ \dot{y} &= ay + xz \\ \dot{z} &= bz + xy\,, \end{aligned} \tag{30}$$

where we chose parameters $[a, b, c] = [-10, -4, 18.1]$ such that the system exhibits two chaotic 1-scroll attractors in state space. We numerically integrate Eq. (30) for $t_{int} = 80.0$ time units with a read-out interval of $\Delta t = 0.005$ using the `Tsit5` integrator. Using K initial conditions, this results in an array of shape $16000 \times 3 \times K$. As for the Duffing datasets, we standardize using the overall mean and standard deviation.

D.2. DSR Models and Training Routines

SINDy Sparse Identification of Nonlinear Dynamics (SINDy) (Brunton et al., 2016) aims to identify a sparse representation of the governing dynamical equations from data. Given a set of measurements of the state $\mathbf{x}(t) \in \mathbb{R}^n$, where n is the number of system variables and $t = \{t_1 \dots t_m\}$ represents observation times, application of SINDy first requires approximating the flow $\frac{d\mathbf{x}}{dt} = \dot{\mathbf{x}}$ numerically, e.g. by finite difference methods. Following the notation in Brunton et al. (2016), the derivatives are arranged into matrix form:

$$\dot{\mathbf{X}} = \begin{bmatrix} \dot{\mathbf{x}}^\top(t_1) \\ \dot{\mathbf{x}}^\top(t_2) \\ \vdots \\ \dot{\mathbf{x}}^\top(t_m) \end{bmatrix} \begin{bmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \cdots & \dot{x}_n(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \cdots & \dot{x}_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{x}_1(t_m) & \dot{x}_2(t_m) & \cdots & \dot{x}_n(t_m) \end{bmatrix}. \tag{31}$$

SINDy optimization then tries to determine a sparse matrix of regression coefficients $\boldsymbol{\Xi}$ such that:

$$\dot{\mathbf{X}} = \Theta(\mathbf{x})\boldsymbol{\Xi}, \tag{32}$$

Here, $\Theta(\mathbf{x})$ is a library of candidate functions on the state variables $\mathbf{x}$ that is defined beforehand, e.g.:

$$\Theta(\mathbf{X}) = \begin{bmatrix} | & | & | & | & | & | & \\ \mathbf{1} & \mathbf{X} & \mathbf{X}^2 & \mathbf{X}^3 & \mathbf{X}^4 & \cos(\mathbf{X}) & \dots \\ | & | & | & | & | & | & \end{bmatrix}, \tag{33}$$

The regression coefficients are found by applying a sparsity-promoting optimization technique, such as the least absolute shrinkage and selection operator (LASSO regression) or the Sequentially Thresholded Least Squares (STLSQ) algorithm, to solve for $\boldsymbol{\Xi}$.

The VF used for Figure 2 is defined as:

$$\dot{x} = x + x(x^2 + y^2 - 1)(4x^2 - 4xy + 4y^2) + (x^2 + y^2)(-2x + 2y + x^3 + xy^2), \tag{34}$$
$$\dot{y} = y + y(x^2 + y^2 - 1)(4x^2 - 4xy + 4y^2) + (x^2 + y^2)(-2x - 2y + y^3 + x^2y).$$

For the inner cycle, a trajectory was drawn from a randomly chosen initial condition $(x_0, y_0) = (0.6, 0.4)$. A long trajectory was then sampled with $T = 100$ and $\Delta t = 0.01$. To infer the VF with SINDy, we used the Python implementation (`PySINDy`, de Silva et al. (2020)) with STLSQ optimizer and threshold 0.01, and a `PolynomialLibrary` up to degree 6. As the outer cycle is an unstable solution, small perturbations lead the system to diverge away from it, and so for the reconstructions in Fig. 2 (center) we made sure to initialize exactly on that cycle. For the results in Figure A8, we provided a polynomial library of second order for the multistable Lorenz-like system and a library of third order for the Duffing system, with other settings the same as used for Fig. 2.

RNNs We trained clipped shallow piecewise-linear RNNs (shPLRNNs; Hess et al. (2023)) using Backpropagation through time (BPTT) with sparse teacher forcing (STF, Mikhaeil et al. (2022); Brenner et al. (2022)) and identity teacher forcing (id-STF; Brenner et al. (2022)). The clipped shPLRNN has a simple 1-hidden-layer architecture

$$\boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{W}_1\left[\phi(\boldsymbol{W}_2\boldsymbol{z}_{t-1} + \boldsymbol{h}_2) - \phi\left(\boldsymbol{W}_2\boldsymbol{z}_{t-1}\right)\right] + \boldsymbol{h}_1, \tag{35}$$

with latent state $\boldsymbol{z}_t \in \mathbb{R}^M$, diagonal matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$, rectangular connectivity matrices $\boldsymbol{W}_1 \in \mathbb{R}^{M\times H}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{H\times M}$, and thresholds $\boldsymbol{h}_2 \in \mathbb{R}^H$ and $\boldsymbol{h}_1 \in \mathbb{R}^M$. The nonlinear activation function ϕ is given by the $\text{ReLU}(\bullet) = \max(\bullet, 0)$. The idea behind id-STF is to replace latent states with states inferred from the observations at optimally chosen intervals τ, such as to pull model-generated trajectories back on track and to avoid strong gradient divergence for chaotic dynamics (Mikhaeil et al., 2022). Teacher forcing also has the effect of smoothening the loss landscape (Hess et al., 2023). As in Brenner et al. (2022), we take an "identity-mapping" for the observation model, $\hat{\boldsymbol{x}}_t = \mathcal{I}\mathbf{z}_t$, with $\mathcal{I} \in \mathbb{R}^{N\times M}$ and $\mathcal{I}_{kk} = 1$ taken to be the identity for the k read-out neurons, $k \leq N$, and zeros for all other elements. STF is only used in training the model, but not when deploying and testing it. The loss function to be minimized is the MSE:

$$\ell_{MSE}(\hat{\boldsymbol{X}}, \boldsymbol{X}) = \frac{1}{NT_s}\sum_{t=1}^{T_s} \|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t\|_2^2, \tag{36}$$

where $\hat{\boldsymbol{X}}$ are model predictions and $\boldsymbol{X}$ is the a training sequence of length T_s. For performing SGD updates, we employ the RAdam (Liu et al., 2020) optimizer paired with an exponential decay learning rate schedule. The shPLRNN and training routine are implemented using the `Flux.jl` DL stack (Innes et al., 2018). Detailed hyperparameter settings are collected in Table A2.

Hyperparameter	Duffing	Lorenz-like
M	5	30
H	100	500
τ	15	15
T_s	100	50
batch size	32	32
η_{start}	10^{-3}	10^{-3}
η_{end}	10^{-6}	10^{-5}
# trainable parameters	$1,116$	$30,641$
# SGD steps	$250,000$	$250,000$

Table A2. Hyperparameter settings of shPLRNNs trained on the Duffing and Lorenz-like systems.

Reservoir Computing (RC) We used a formulation of the RC architecture often employed in work on DSR (Patel & Ott, 2022):

$$\boldsymbol{r}_t = \alpha\boldsymbol{r}_{t-1} + (1-\alpha)\tanh\left(\boldsymbol{W}\boldsymbol{r}_{t-1} + \boldsymbol{W}_{in}\boldsymbol{u}_t + \boldsymbol{b}\right) \tag{37}$$
$$\hat{\boldsymbol{x}}_t = \boldsymbol{W}_{out}\boldsymbol{r}_t, \tag{38}$$

where $\boldsymbol{r}_t \in \mathbb{R}^M$ is the reservoir state, $\alpha \in \mathbb{R}$ the leakage parameter, $\boldsymbol{W} \in \mathbb{R}^{M\times M}$ the reservoir connectivity matrix, $\boldsymbol{W}_{in} \in \mathbb{R}^{M\times N}$ the input-to-reservoir matrix weighing inputs $\boldsymbol{u}_t \in \mathbb{R}^N$, $\boldsymbol{b} \in \mathbb{R}^M$ a bias vector, and $\boldsymbol{W}_{out} \in \mathbb{R}^{N\times M}$ the matrix mapping reservoir states to the observed data. In RCs, the dynamical reservoir parameters $\theta_r = \{\alpha, \boldsymbol{W}, \boldsymbol{W}_{in}, \boldsymbol{b}\}$ are fixed after initialization. Here we initialized $\boldsymbol{W}$ to be fully connected with entries sampled from a standard normal distribution, and then scaled to have a predefined spectral radius specified by a hyperparameter ρ. Input-to-reservoir matrix $\boldsymbol{W}_{in}$ and bias $\boldsymbol{b}$ are also drawn from Gaussian distributions with variances σ^2 and β^2, respectively. In RCs, only the reservoir-to-output matrix $\boldsymbol{W}_{out}$ is learned. The RC is trained by first driving the reservoir using ground-truth data $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_T] \in \mathbb{R}^{N\times T}$ supplied through $\boldsymbol{u}_t = \boldsymbol{x}_t$. This results in a trajectory of reservoir states $\boldsymbol{R} = [\boldsymbol{r}_1, \dots, \boldsymbol{r}_T] \in \mathbb{R}^{M\times T}$. The only trainable parameters $\boldsymbol{W}_{out}$ are then determined by minimizing the least-squares error $\|\boldsymbol{X} - \boldsymbol{W}_{out}\boldsymbol{R}\|_2^2$, a straightforward linear regression problem with closed form solution

$$\boldsymbol{W}_{out} = \boldsymbol{X}\boldsymbol{R}^T \left(\boldsymbol{R}\boldsymbol{R}^T\right)^{-1}. \tag{39}$$

After training, the reservoir state is initialized with zeros and the RC is only provided a short sequence of ground-truth data $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_{T_W}\}$ to 'warm-up' the dynamics of the reservoir state $\boldsymbol{r}_t$, where T_W denotes the warm-up time. Afterwards, the RC runs closed-loop (autonomously) by feeding predictions $\hat{\boldsymbol{x}}_t$ back to the reservoir through the input-to-reservoir connection. To keep the comparison between DSR models fair in Fig. 3, we only provide a single initial condition, i.e. $T_W = 1$. For visual clarity, however, we still drop the first few time steps of RC-generated trajectories (e.g. in Fig. 3), which the zero-initialized reservoir state needs to converge to the correct dynamics. Detailed hyperparameter settings are in Table A3.

Hyperparameter	Duffing	Lorenz-like
M	500	2000
ρ	1.0	0.75
α	0.7	0.4
σ	0.2	0.3
β	0.5	0.7
# trainable parameters	$1,000$	$6,000$

Table A3. Hyperparameter settings of RCs trained on the Duffing and Lorenz-like systems.

N-ODE We train N-ODEs (Chen et al., 2018) using the Julia library `DiffEqFlux.jl` (Rackauckas et al., 2020). We use a simple multi-layer perceptron (MLP) architecture where parameters are optimized using the adjoint method. The loss function is the MSE, Eq. (36). As for RNNs we perform SGD updates using RAdam paired with an exponential decay learning rate schedule. Detailed hyperparameter settings are in Table A4.

Hyperparameter	Duffing	Lorenz-like
# hidden layer	2	3
hidden layer sizes	$[40, 40]$	$[100, 100, 100]$
activation	tanh	ReLU
T_s	30	30
batch size	32	32
ODE solver	Tsit5	Tsit5
η_{start}	10^{-3}	10^{-3}
η_{end}	10^{-5}	10^{-5}
# trainable parameters	$1,842$	$20,903$
# SGD steps	$100,000$	$100,000$

Table A4. Hyperparameter settings of N-ODEs trained on the Duffing and Lorenz-like systems.

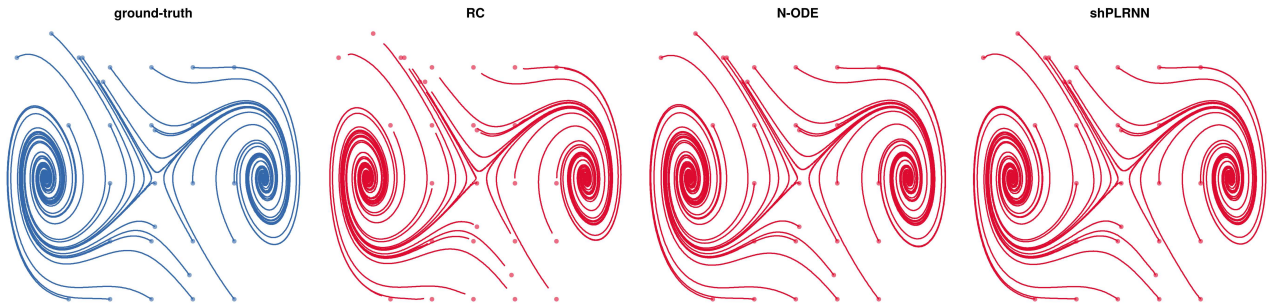

Figure A5. Reconstructions of the Duffing system as in Fig. 3, but with models trained on ground-truth data (blue trajectories) from both basins. The models are capable of learning the multistable dynamics (red trajectories) of the system when supplied with data from both basins. This is also reflected in the drastically smaller statistical error $\mathcal{E}_{stat}$ when applied to the same test data as used for Fig. 3: RC $\approx 2.7 \cdot 10^{-3}$, N-ODE $\approx 2.1 \cdot 10^{-3}$ and shPLRNN $\approx 1.4 \cdot 10^{-3}$.

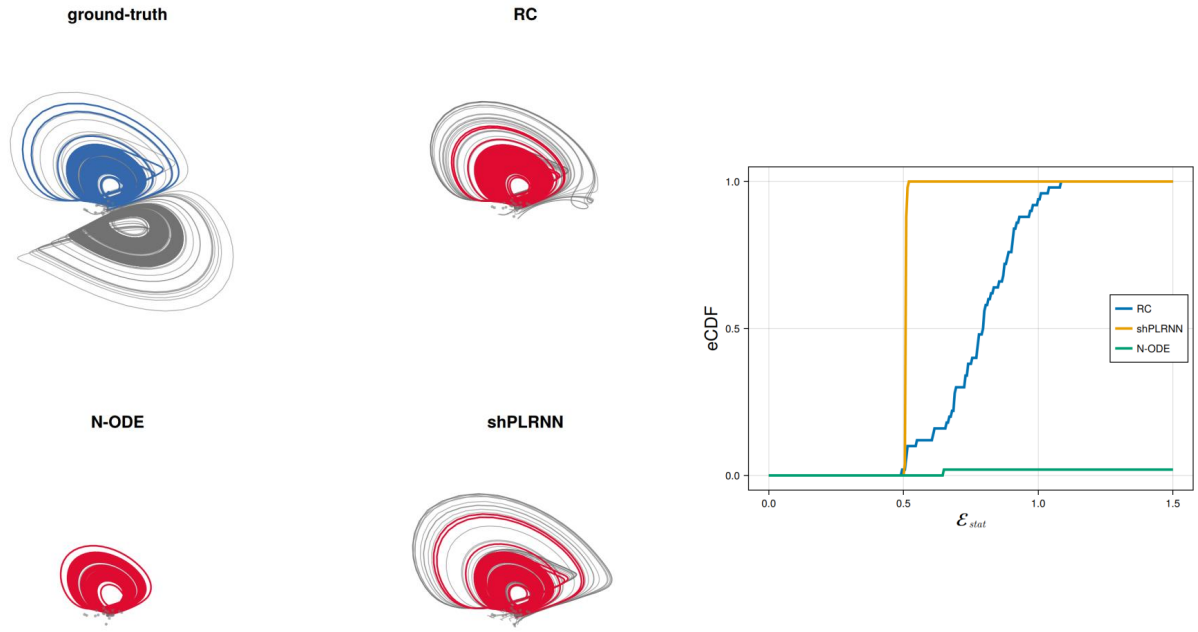

Figure A6. Learnability evaluated on the multistable Lorenz-like system. Left: Example reconstructions of DSR models trained on 4 ground-truth trajectories (blue) from one basin. Red trajectories are freely generated using initial conditions of the training data and the respective DSR model. Grey trajectories are example ground-truth test trajectories and model-generated trajectories, respectively, from both the training basin and the OOD basin. Again, all models fail to properly generalize to the unobserved attractor/basin. Right: eCDF of $\mathcal{E}_{stat}$ with a sample size of $N = 50$ independent trainings of each DSR model evaluated over a grid of 125 initial conditions covering both basins. Note that the dynamics of the N-ODE models consistently diverged for many initial conditions from the grid. For N-ODE $\mathcal{E}_{stat}$ values, this means that most mass is concentrated at much higher error values, which were cut off in the eCDF plot.

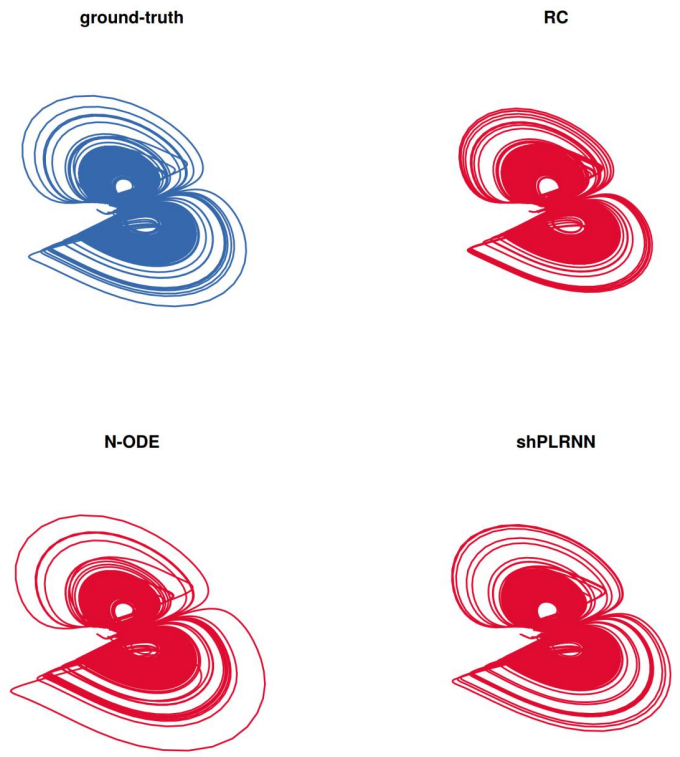

Figure A7. Reconstructions of the multistable Lorenz-like system using the same DSR models as in Fig. A6, but trained on ground-truth training data (blue) from both basins. As for the Duffing system, the models are capable of learning the multistable dynamics (red trajectories) of the Lorenz-like system when supplied with data from both basins. This is also reflected in lower statistical errors $\mathcal{E}_{stat}$ when compared to the monostable training data, evaluating to ≈ 0.131 for RC, ≈ 0.133 for N-ODE and ≈ 0.064 for the shPLRNN.

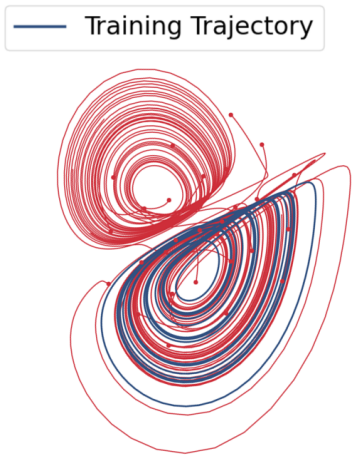

Figure A8. Reconstruction of the multistable Lorenz-like system from a trajectory from just one basin, using PySINDy (de Silva et al., 2020). Since in this case, the correct polynomial function library was provided, both basins are correctly identified (see also Fig. 2b).

D.3. Specifications on Theorem 4.1

There are also examples of VFs with a dense set of trajectories which solve an algebraic equation. Any VF with a rational first integral (e.g. algebraic Hamiltonian) is not learnable (since then every solution solves an algebraic equation in terms of the basis functions). A simple example would be the standard harmonic oscillator in the regime where it has a center, i.e. a

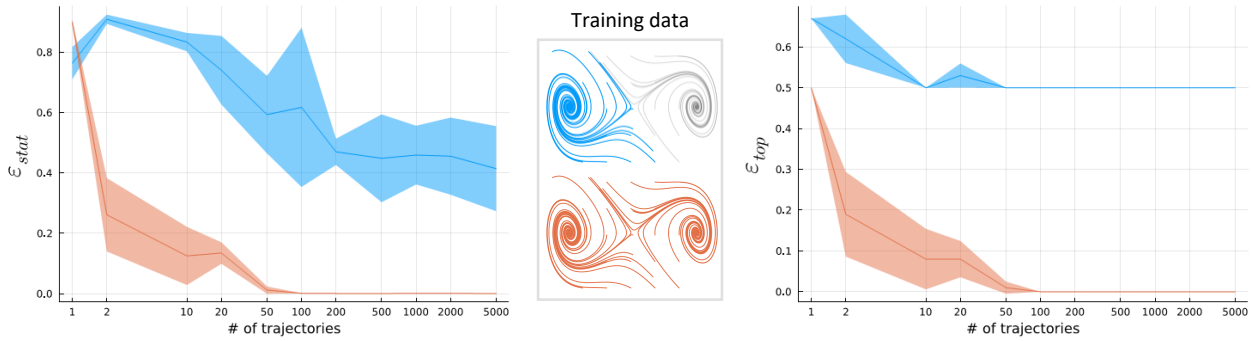

Figure A9. Generalization errors $\mathcal{E}^M_{\text{stat}}$ (left) and $\mathcal{E}^M_{\text{top}}$ (right) for shPLRNNs ($M = 5$, $H = 100$) as a function of sample size (amount of training data), drawn from one basin (blue) or both basins (orange) of the Duffing system (center). Each data point is median $\pm$ median absolute deviation (ribbon) across 10 models. Increasing training data from one basin makes no difference for the generalization to the second basin (blue), i.e. both errors plateau at high values.

dense set of closed orbits (each of them solving an algebraic equation). More generally, systems with a center manifold (as in many Hamiltonian systems and biological systems like the FitzHugh-Nagumo equation) may have this property, with all trajectories lying on the center manifold solving an algebraic equation.

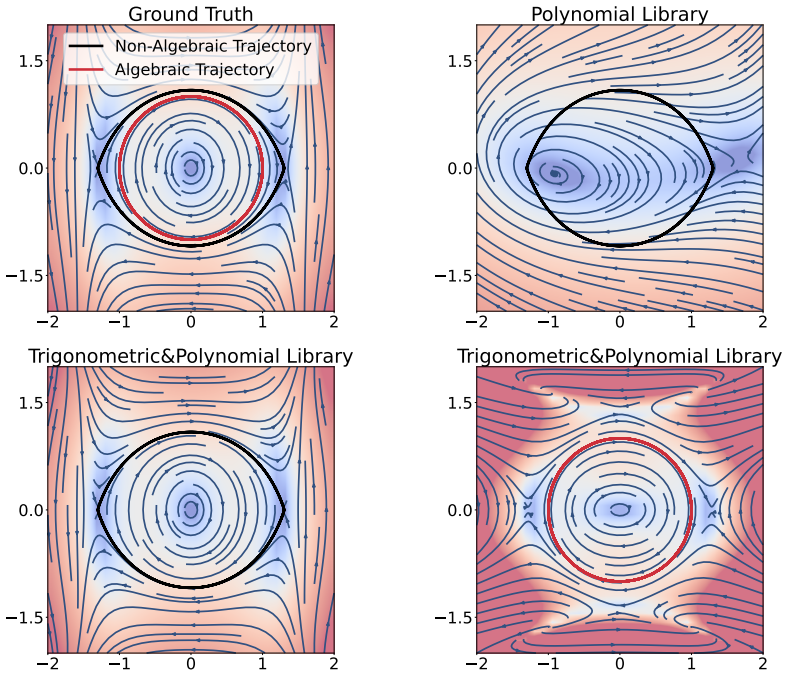

Figure A10. Example reconstructions using SINDy (details in Appx. D.2). a) The ground truth (GT) VF has two cycle solutions. One solves an algebraic equation (red), while the other does not (black). b) Providing SINDy with the black trajectory and the correct library (including both trigonometric and polynomial functions) leads to a correctly inferred VF. c) Providing as data the curve solving an algebraic trajectory leads to an incorrectly inferred VF, as stated in theorem 4.1, despite the correct library. d) Providing SINDy with a non-algebraic trajectory but a library that lacks the trigonometric terms also fails to reproduce the GT VF.

The VF used for Figure A10 is defined as:

$$\begin{aligned} \dot{x} &= 2y\cos(x), \\ \dot{y} &= x^2\sin(x) - 2x\cos(x) + y^2\sin(x) - \sin(x). \end{aligned} \tag{40}$$

D.4. Specifications on Theorem 4.1 - Evaluating Identifiability Conditions

Assume we are given a dataset (trajectory)

$$\mathcal{D} = \{\boldsymbol{x}(t_0), \boldsymbol{x}(t_1), ..., \boldsymbol{x}(t_N)\}, \tag{41}$$

for which we would like to check whether it satisfies an algebraic equation in the basis functions, i.e.,

$$\nexists \boldsymbol{\theta} \in \mathbb{R}^m \backslash \{\mathbf{0}\} : \ \Lambda(\boldsymbol{x}(t)) = \sum_{i=0}^{m} \theta_i \psi_i(\boldsymbol{x}(t)) = 0 \quad \forall t \in I^f_{\boldsymbol{x}_0}. \tag{42}$$

Multiplying Eq. (42) by $\psi_j(\boldsymbol{x}(t))$ yields

$$-\forall j = 0, ..., m : \quad \psi_j(\boldsymbol{x}(t)) \sum_{i=0}^{m} \theta_i \psi_i(\boldsymbol{x}(t)) = 0. \tag{43}$$

As this equation is always zero, it is also zero when evaluated at the data points

$$\forall j = 0, ..., m : \quad \sum_{k=0}^{N} \psi_j(\boldsymbol{x}(t_k)) \sum_{i=0}^{m} \theta_i \psi_i(\boldsymbol{x}(t_k)) = 0. \tag{44}$$

Equation (44) consists of $m+1$ linear equations which can be written in matrix form:

$$\underbrace{\sum_{k=0}^{N} \begin{bmatrix} \psi_0(\boldsymbol{x}(t_k)) \cdot \psi_0(\boldsymbol{x}(t_k)) & \psi_0(\boldsymbol{x}(t_k)) \cdot \psi_1(\boldsymbol{x}(t_k)) & \cdots & \psi_0(\boldsymbol{x}(t_k)) \cdot \psi_m(\boldsymbol{x}(t_k)) \\ \psi_1(\boldsymbol{x}(t_k)) \cdot \psi_0(\boldsymbol{x}(t_k)) & \psi_1(\boldsymbol{x}(t_k)) \cdot \psi_1(\boldsymbol{x}(t_k)) & & \psi_1(\boldsymbol{x}(t_k)) \cdot \psi_m(\boldsymbol{x}(t_k)) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_m(\boldsymbol{x}(t_k)) \cdot \psi_0(\boldsymbol{x}(t_k)) & \psi_m(\boldsymbol{x}(t_k)) \cdot \psi_1(\boldsymbol{x}(t_k)) & \dots & \psi_m(\boldsymbol{x}(t_k)) \cdot \psi_m(\boldsymbol{x}(t_k)) \end{bmatrix}}_{\underline{\Psi}} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} = 0. \tag{45}$$

Hence, if $\ker(\underline{\Psi})$ is non-trivial (contains not only the zero-vector), then the data points solve an algebraic equation in the basis functions. For instance, let us revisit the scenario described in the main text, where our training dataset comprises points on the circle $\Gamma_{\boldsymbol{x}_0} = \big\{(\cos(t), \sin(t)) \mid t \in [0, 2\pi)\big\}$. With this information, we are able to determine the null space of $\underline{\Psi}$ by solving Eq. (45) using a library that includes polynomials up to third order. The null space is three-dimensional and consists of the three algebraic curves shown in Fig. A11. (Note that, by definition, also all linear combinations of these invariant algebraic curves are in the nullspace.) In contrast, the nullspace of $\underline{\Psi}$ for a trajectory that does not solve an algebraic equation in the basis functions contains only the zero-vector.

Thus, checking whether the null space of $\underline{\Psi}$ contains only the zero-vector is a quick and efficient method for verifying whether any trajectory satisfies an algebraic equation in the basis functions. If the nullspace contains only the zero-vector and we have provided a proper (correct) library, SINDy (or related library methods) will generalize across the state space M. Conversely, if the null space is non-trivial, either a different trajectory must be selected or the hypothesis class must be limited, as outlined in Corollary E.7, to enable proper generalization.

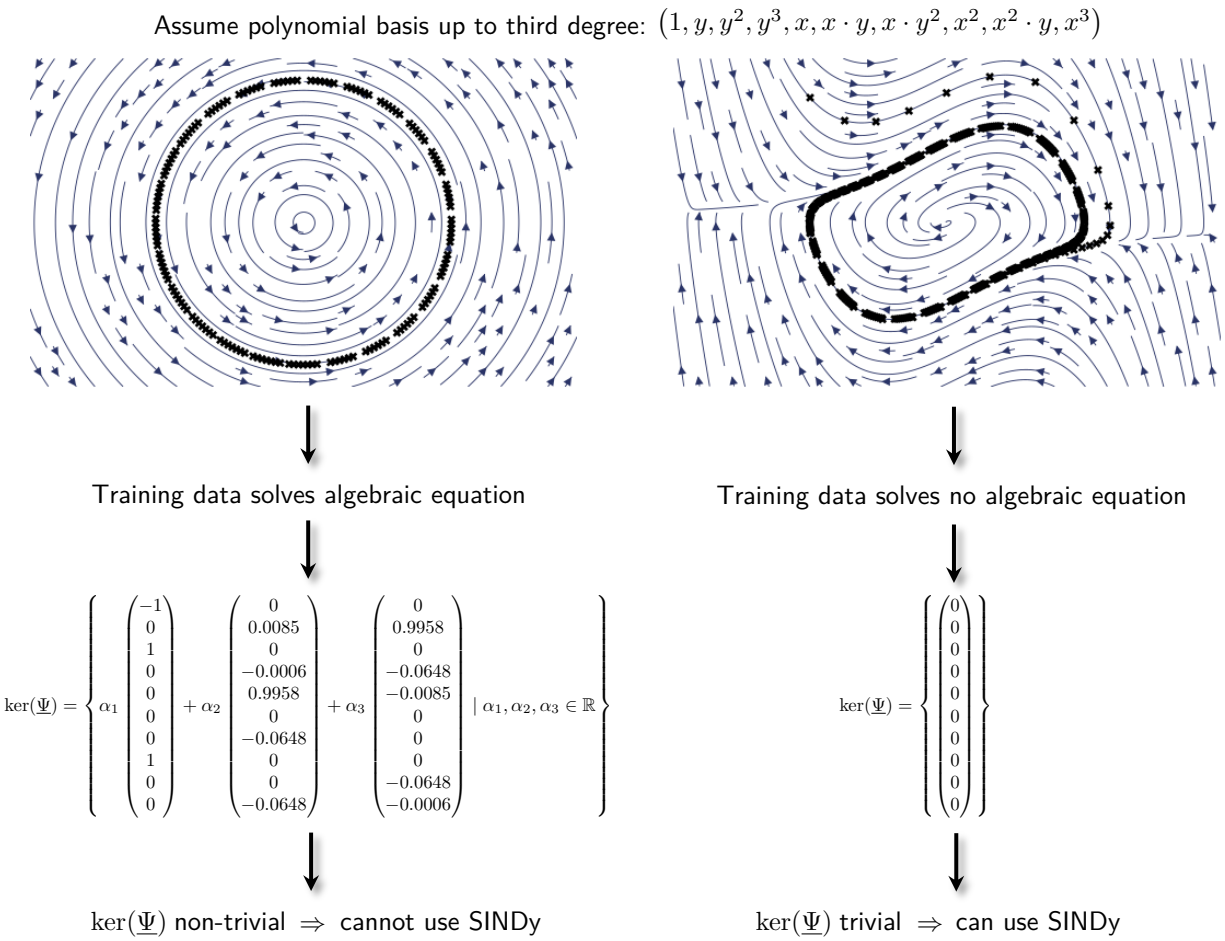

Figure A11. Illustration of how to check whether training data solve, or do not solve, an algebraic equation in a practical setting, for the harmonic oscillator (left) and van-der-Pol oscillator (right). The limit cycle of the van-der-Pol oscillator is non-algebraic (Odani, 1995).

D.5. Why OODG Fails

The probability of a model trained with SGD having error $\varepsilon_{\text{stat}}$ is defined by

$$p_{\text{SGD}}(\varepsilon_{\text{stat}} \mid \mathcal{D}) = \int_{\Theta} \mathbb{1}[\mathcal{E}_{\text{gen}}^{M}(\Phi_{\boldsymbol{\theta}_f}) = \varepsilon_{\text{gen}}] p_{\text{opt}}\left(\boldsymbol{\theta}_f \mid \boldsymbol{\theta}_i, \mathcal{D}\right) p_{\text{ini}}\left(\boldsymbol{\theta}_i\right) d\boldsymbol{\theta}_i d\boldsymbol{\theta}_f. \tag{46}$$

This coincides with the learnability distribution

$$p(\varepsilon_{\text{gen}}|\mathcal{D}) = \frac{1}{\text{vol}(\Theta_0)} \int_{\Theta_0} \mathbb{1}[\mathcal{E}_{\text{gen}}^{M_{\text{test}}}(\Phi_{\boldsymbol{\theta}}) = \varepsilon_{\text{gen}}] d\boldsymbol{\theta}, \tag{47}$$

under two conditions, the failure of either one introduces an implicit bias. First, it might be that the optimizer does not converge to a model with zero reconstruction error on M but a slightly larger error. The more significant implicit bias arises from the combination of p_{ini} and p_{stat}. Assuming the optimizer always converges to a model with zero generalization error, we have

$$p_{\text{SGD}}(\varepsilon_{\text{gen}}|\mathcal{D}) = \frac{1}{\text{vol}(\Theta_0)} \int_{\Theta_0} \mathbb{1}[\mathcal{E}_{\text{gen}}^{M_{\text{test}}}(\Phi_{\boldsymbol{\theta}}) = \varepsilon_{\text{gen}}] \cdot p_{\text{bias}}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{48}$$

i.e., p_{SGD} aligns with the learnability distribution only when an implicit bias term p_{bias} is accounted for. In general, it is likely that a specific combination of p_{ini} and p_{stat} preferentially converges to certain parameters $\boldsymbol{\theta}_f$, as observed for SGD. This preference is termed the implicit bias of the learning algorithm, and skews the learnability distribution toward the domain implicitly preferred by the learning algorithm.

D.6. Simplicity Bias

We adopt the following definition for the parameter-function map from (Valle-Pérez et al., 2019):

Definition D.1. The parameter-function map $\mathcal{M}$ associates a given parameter set with the flow, expressed as:

$$\mathcal{M} : \Theta \to \mathcal{H}_\theta, \quad \boldsymbol{\theta} \mapsto \Phi_\theta. \tag{49}$$

Depending on the choice of model, multiple parameter vectors can map onto the same flow. For instance, for the shPLRNN (Sect. D.2), scaling $\boldsymbol{W}_1$ by a positive scalar factor $c \in \mathbb{R}_+$ and adjusting the weights in $\boldsymbol{W}_2$ by $\frac{1}{c}$ results in the same dynamical model. The initial parameter distribution p_{ini} can lead to flows with different characteristics not directly apparent from the initial distribution. In the context of the simplicity bias, we are interested in the distribution over the complexity of the flows $K(\Phi_{p_{\text{ini}}})$ induced by a choice of distribution over initial parameters. To assess this complexity, we select the Shannon entropy over the limit sets of the resulting flows (Eckmann & Ruelle, 1985). We evaluate this by drawing long trajectories from a grid of initial conditions and compute the entropy over the histogram of final states, using the Rényi algorithm from `ComplexityMeasures.jl`. This entropy has a natural interpretation for flows with different topologies, where e.g. global equilibrium points have low entropy and chaotic attractors have high entropy (see Fig. 4b).

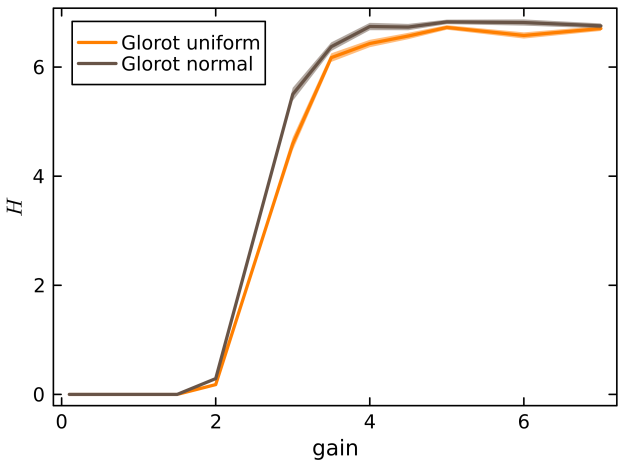

Figure A12. Simplicity bias for shPLRNNs for $M = 10$ and $H = 250$.

Table A5. Number of positive, zero and negative eigendirections for the Hessian of the loss function evaluated on trajectories from just one ($\ell_{B(A_1)}$) or both (ℓ_M) basins w.r.t. $\boldsymbol{\theta}_{\text{gen}}$. (For this analysis only 5 of the 20 generalizing models plotted in Fig. 5 where used.)

	ℓ_M	$\ell_{B(A_1)}$
$\#\lambda_+$	171.93 ± 0.62	103.08 ± 2.10
$\#\lambda_0$	182.38 ± 0.75	310.54 ± 3.76
$\#\lambda_-$	149.69 ± 0.24	90.38 ± 1.66
λ_{max}	241.09 ± 1.25	364.09 ± 3.29
λ_{min}	-0.00126 ± 0.00004	-0.00186 ± 0.00001

D.7. Assessing Sharpness of Minima

In order to assess the volume of minima, we employ a sampling-based approach. Following Huang et al. (2020), we randomly select a vector $\boldsymbol{\theta}'$ from the parameter space Θ with dimensionality d within the hyper-sphere $\mathcal{S}_r^{d-1}$ with radius r. Subsequently, we evaluate the loss of models with parameters along a straight line in parameter space given by

$$\boldsymbol{\theta}_{\min} + a(\boldsymbol{\theta}' - \boldsymbol{\theta}_{\min}). \tag{50}$$

Here, a takes values in the interval $[0, 1]$. We define the threshold value $a = a_{th}$ as the one at which the loss of the corresponding model exceeds the predefined threshold $\ell > \ell(1 + p_{th})$. We tested a threshold of $p_{th} = 1\%$ and $p_{th} = 5\%$ for our experiments. However, as the results in Fig. 6 and A15 indicate, the results obtained are not overly sensitive to this hyperparameter. The resulting threshold value a_{th} yields an estimate of the minimum radius in the direction of $\boldsymbol{\theta}'$, given by $r \cdot a_{th}$. This radius serves as a lower bound for the minimum volume $V = \frac{\pi^{n/2}}{\Gamma(1+n/2)} \mathbb{E}_{\boldsymbol{\theta}} \left[r^d(\boldsymbol{\theta}) \right]$, given by:

$$
\begin{aligned}
\log V &= \frac{d}{2} \log \pi - \log \Gamma(\frac{d}{2} + 1) + \log \mathbb{E}_{\boldsymbol{\phi} \sim \mathcal{U}}[r^d(\boldsymbol{\theta})] \\
&\geq \frac{d}{2} \log \pi - \log \Gamma(\frac{d}{2} + 1) + \mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{U}}[d \log r(\boldsymbol{\theta})] \\
&\approx \frac{d}{2} \log \pi - \log \Gamma(\frac{d}{2} + 1) + \frac{d}{N} \sum_{i=1}^{N} \log r(\boldsymbol{\theta}_i),
\end{aligned}
$$

where we used Jensen's inequality to pull the logarithm into the expectation, and $\Gamma(\frac{d}{2} + 1)$ is Euler's gamma function.

Radius estimates for saddle points It is widely acknowledged that many critical points to which stochastic gradient descent (SGD) converges in high-dimensional spaces are saddle points (Chaudhari et al., 2019). Mathematically, saddle points lack a well-defined radius since there exist directions in which the loss decreases. Despite this, estimation techniques for the radius, such as the one proposed here, are commonly applied (Huang et al., 2020). Our convergence analysis supports the viability of our method, as demonstrated in Fig. A16a, where the estimation of the radius converges after approximately 3000 randomly drawn samples. A possible explanation for this convergence could be the prevalence of positive directions. As illustrated in Table A5, the Hessian has more positive eigendirections than zero or negative ones. This could lead to a much larger volume of parameters around the minima having an ascending loss. Further support for this hypothesis is provided in Fig. A16b, where random sampling of parameter vectors around minima, as described earlier, results more frequently in ascending than flat curves.

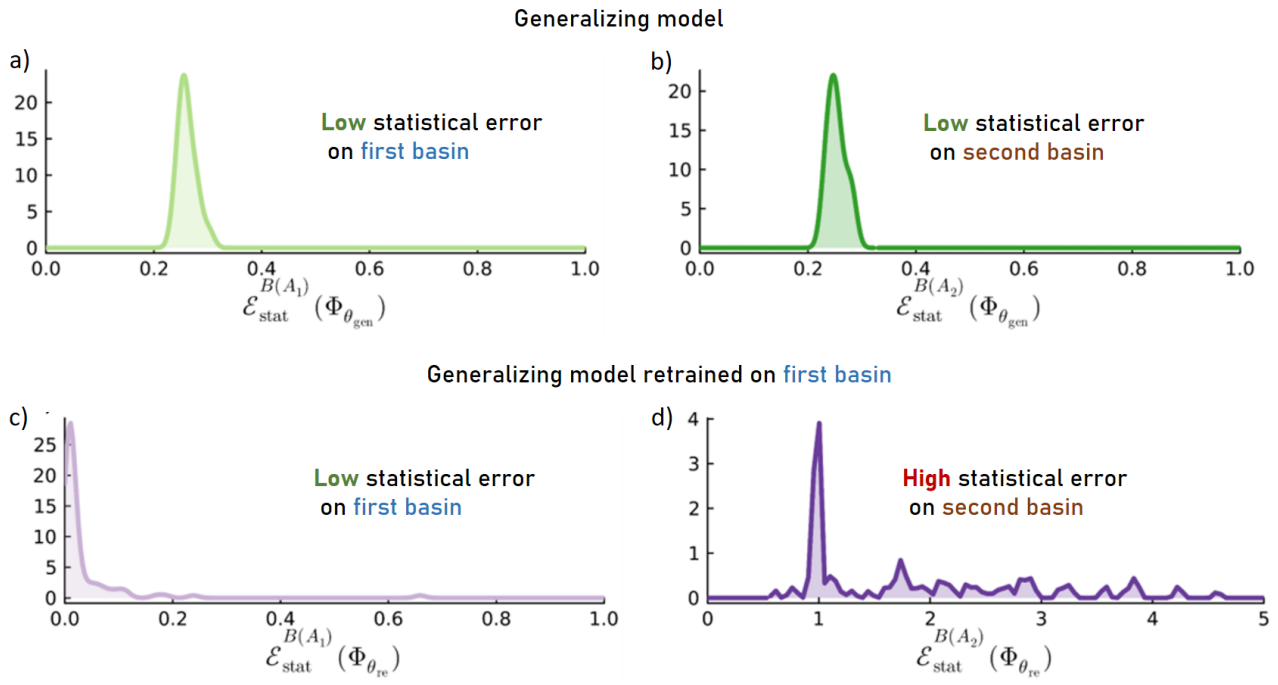

Figure A13. Similar to Fig. 5a, this figure illustrates the statistical error of generalizing and retrained models for the multistable Lorenz-like system described by Eq. (30). The upper two density plots depict the statistical error of generalizing models on $B(A_1)$ and $B(A_2)$. Meanwhile, the lower row illustrates the same for retrained models. A surge in error is observed for $B(A_2)$ in the retrained models, suggesting an unlearning of the second attractor. This finding confirms that the results presented in the main text can be reproduced for a ground-truth system with completely different dynamics.

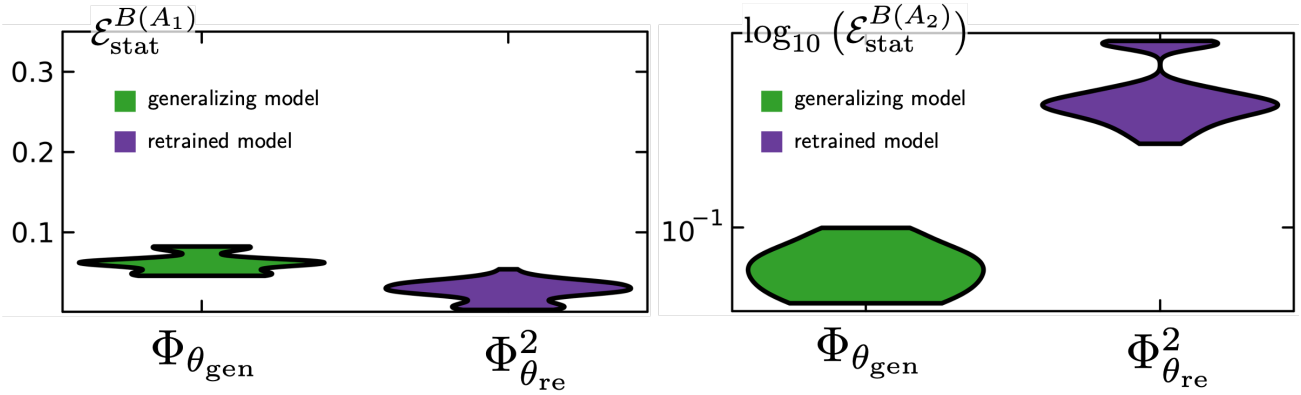

Figure A14. Statistical error distribution on basins $B(A_1)$ and $B(A_2)$ for 10 generalizing models (green) and 10×5 models retrained (purple) using only $B(A_1)$ data, similar to Fig. 5a in the main text, but based on a $6d$ Lorenz-96 system (Pelzer & Sterk, 2020) of the form $x_j = x_{j-1} \cdot (x_j - x_{j-1}) - x_j + F,\ j = 1 \ldots 6$, where periodic boundary conditions are applied. Using $F = 0.654502$, the system has two coexisting chaotic attractors (see sect. 4.3 in (Pelzer & Sterk, 2020) for more details on the system). After retraining on data from $B(A_1)$, the models effectively unlearn the dynamics on $B(A_2)$.

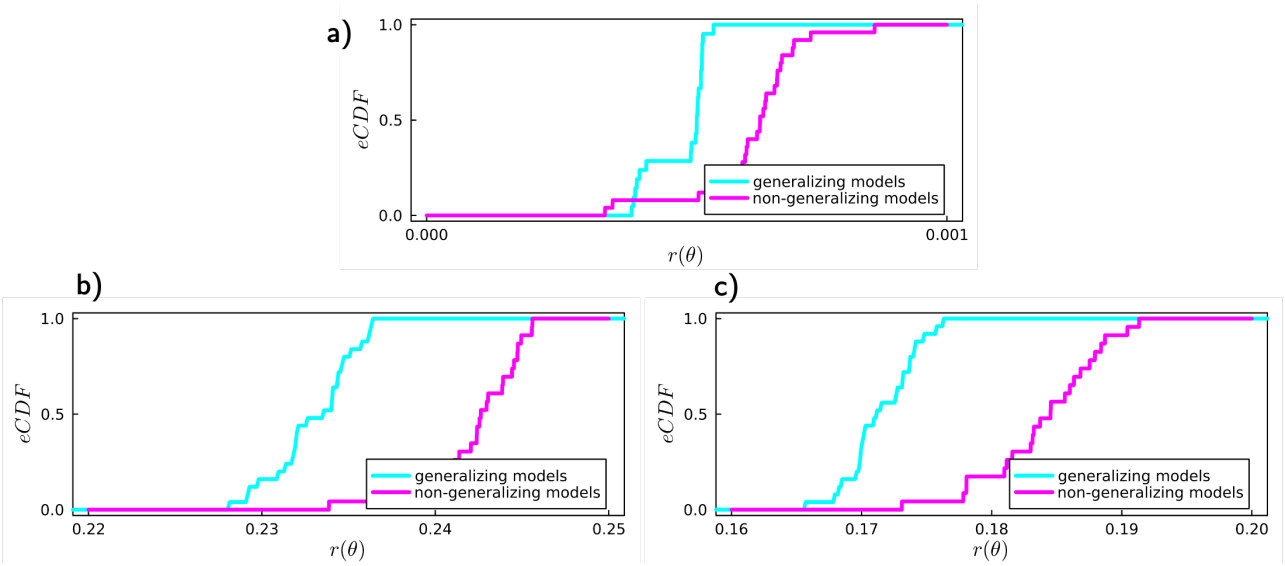

Figure A15. **a)** Similar results as in Fig. 6 for the Duffing system, but for a lower threshold $p_{th} = 1\%$, as described in Appx. D.7. **b)** Similar graphs as in Fig. 6 for the multistable Lorenz system (Eq. (30)) for $p_{th} = 1\%$. **c)** Same as b) but for $p_{th} = 5\%$.

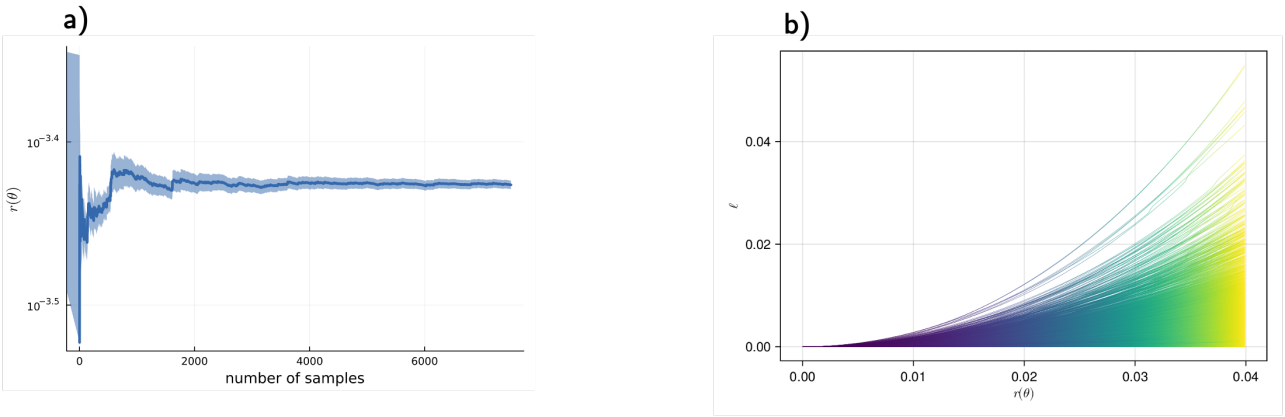

Figure A16. **a)** Average radius as a function of the number of samples drawn from the loss landscape, indicating convergence to a constant radius at around 3000 samples. **b)** Loss as a function of the radius for a shPLRNN $(N = 2, M = 100)$ trained on the Duffing system (Eq. (29)) and 5000 sampled points. Curves are with kernel density smoothing.

E. Proofs

E.1. Proof of Theorem 3.3

We first clarify the mathematical interpretation of 'not reconstructed'. We consider two different scenarios, assuming the ground truth model Φ to be given (fixed):

(i) All trajectories starting in $B(A_k)$ converge to another reconstructed attractor $A_j,\ j \neq k$. This is often the case for trained models, as illustrated in Fig. 3 and Fig. A6, and corresponds to a missing basin. In this case $\mathcal{E}_{\text{gen}}^{M_{\text{test}}}(\Phi_R) \propto \text{vol}(B(A_k))$.

(ii) There is an open set $B(A_{n+1}) \subset M$ corresponding to a new basin of a new attractor A_{n+1} of Φ_R, where we further assume $A_i \cap B(A_{n+1}) = \emptyset, \quad \forall i \leq n$. This is the case when some additional dynamics is learned which is not present in the ground-truth system. In this case $\mathcal{E}_{\text{gen}}^{M_{\text{test}}}(\Phi_R) \propto \text{vol}(B(A_{n+1}))$.

Proof of (i) We assume the decomposition as in Eq. (7) for the ground-truth system Φ:

$$M = \sqcup_{e=1}^{N} B_e \sqcup \tilde{M} \quad \text{such that} \quad \mu(\tilde{M}) = 0 \tag{51}$$

We will first prove the theorem for the **statistical error**. First, we want to show that the error between the occupation measure of the ground-truth system and the reconstructed system on $B(A_k)$ is non-zero:

$$\text{SW}_1(\mu_{\boldsymbol{x},T}^{\Phi}, \mu_{\boldsymbol{x},T}^{\Phi_R}) \neq 0 \quad \forall \boldsymbol{x} \in B(A_k). \tag{52}$$

To do so, we take some $\boldsymbol{x} \in B(A_k)$. By assumption $\exists j \neq k : \omega(\boldsymbol{x}, \Phi_R) \subseteq A_j$. We take some open subset $U \supset A_j$ of $B(A_j)$, which has to exist as $B(A_j)$ is open. But for Φ it still holds that $\omega(\boldsymbol{x}, \Phi) \subseteq A_k$. It follows that

$$\mu_{\boldsymbol{x},T}^{\Phi_R}(U) \neq \mu_{\boldsymbol{x},T}^{\Phi}(U), \tag{53}$$

as, by definition of an attractor, there is some time T' such that $\Phi_R(T', \boldsymbol{x})$ enters U, while $\Phi(t, \boldsymbol{x})$ never does for any t. Also note that $\mu_{\boldsymbol{x},T}^{\Phi_R}(U) \neq 0$ as Φ_R enters U, making the occupation measure of U non-zero. Consequently, we found a Borel set on which the two occupation measures disagree, hence $\mu_{\boldsymbol{x},T}^{\Phi_R} \neq \mu_{\boldsymbol{x},T}^{\Phi}$. This construction can be repeated for any $\boldsymbol{x} \in B(A_k)$.
Since SW_1 is a metric on the space of measures (Kolouri et al., 2016), this implies

$$\forall \boldsymbol{x} \in B(A_k), \quad \text{SW}_1(\mu_{\boldsymbol{x},T}^{\Phi}, \mu_{\boldsymbol{x},T}^{\Phi_R}) \neq 0. \tag{54}$$

Since $B(A_k)$ is connected, also $\overline{B(A_k)}$ is connected. As $\overline{B(A_k)}$ is closed and M compact, also $\overline{B(A_k)}$ is compact. By the mean-value theorem for integrals, there exists some $\boldsymbol{x}' \in \overline{B(A_k)}$ such that

$$\mathcal{E}_{stat}^{B(A_k)}(\Phi_R) = \text{SW}_1(\mu_{\boldsymbol{x}',T}^{\Phi}, \mu_{\boldsymbol{x}',T}^{\Phi_R}) \cdot \text{vol}(\overline{B(A_k)}). \tag{55}$$

$\text{SW}_1(\mu_{\boldsymbol{x}',T}^{\Phi}, \mu_{\boldsymbol{x}',T}^{\Phi_R})$ is a constant, consequently $\mathcal{E}_{\text{stat}}^{B(A_k)} \propto \text{vol}(\overline{B(A_k)})$ which proves the theorem.
We go on to prove the statement for the **topological error**. Denote by $D_r^n(\boldsymbol{y})$ a unit ball of radius r in $\mathbb{R}^n$ centered on $\boldsymbol{y}$. Again, we take some $\boldsymbol{x} \in B(A_k)$ and by assumption $\exists j \neq k : \omega(\boldsymbol{x}, \Phi_R) \subseteq A_j$, while for Φ it holds that $\omega(\boldsymbol{x}, \Phi) \subseteq A_k$. As M is equipped with the Euclidean distance, the Hausdorff distance between two sets measures the farthest possible Euclidean distance between a point in one set to the closest point in the other set. Denote by S the set of points in A_j being ε_{d_H}-close to A_k,

$$S = \{\boldsymbol{y} \in A_j | d_H(\{\boldsymbol{y}\}, A_k) \leq \varepsilon_{d_H}\} \tag{56}$$

For reasonable choices of our hyperparameter ε_{d_H} this set will be empty. Only if both attractors A_j, A_k, lie very close to the boundary separating their basins of attraction, it could be non-zero. In this case we have

$$d_H(\omega(\boldsymbol{x}, \Phi_R), \omega(\boldsymbol{x}, \Phi)) \leq \varepsilon_{d_H}, \quad \forall \boldsymbol{x} \in \cup_{\boldsymbol{y} \in S} D_{\varepsilon_{d_H}}^n(\boldsymbol{y}) \tag{57}$$

and

$$d_H(\omega(\boldsymbol{x}, \Phi_R), \omega(\boldsymbol{x}, \Phi)) > \varepsilon_{d_H}, \quad \forall \boldsymbol{x} \in B(A_k) \setminus \cup_{\boldsymbol{y} \in S} D_{\varepsilon_{d_H}}^n(\boldsymbol{y}). \tag{58}$$

Hence, the indicator function in Eq. (10) $\mathbb{1}_{\Phi_R(\boldsymbol{x})}$ is 0 on the set $B(A_k)\setminus \cup_{\boldsymbol{y}\in S} D^n_{\varepsilon_{d_H}}(\boldsymbol{y})$, as the condition for the closeness of limit sets in the Hausdorff metric is violated.
Consequently,

$$\mathcal{E}^M_{\text{top}}(\Phi_R) = 1 - \frac{1}{\text{vol}(M)} \int_M \mathbb{1}_{\Phi_R}(\boldsymbol{x})\, d\boldsymbol{x} = \frac{\text{vol}(B(A_k)) - \text{vol}(\cup_{\boldsymbol{y}\in S} D^n_{\varepsilon_{d_H}}(\boldsymbol{y}))}{\text{vol}(M)} \propto \frac{\text{vol}(B(A_k))}{\text{vol}(M)} \tag{59}$$

This result proves the theorem when choosing $M_{\text{test}} = M$, given a choice of ε_{d_H} that ensures that $\text{vol}(B(A_k)) \gg \text{vol}(\cup_{\boldsymbol{y}\in S} D^n_{\varepsilon_{d_H}}(\boldsymbol{y}))$. This should be generally fulfilled by a reasonably small choice of ε_{d_H}.

Proof of (ii) The proof is analogous to the one of (i) if we replace A_k with A_{n+1}, but the main steps will be stated again for completeness. For the **statistical error** we want to prove a result similar to Eq. (52):

$$\text{SW}_1(\mu^{\Phi}_{\boldsymbol{x},T}, \mu^{\Phi_R}_{\boldsymbol{x},T}) \neq 0 \quad \forall \boldsymbol{x} \in B(A_{n+1}). \tag{60}$$

Assume some $\boldsymbol{x} \in B(A_{n+1})$. We know that $\nexists j \leq n : \omega(\boldsymbol{x}, \Phi_R) \subseteq A_j$, since by assumption $A_i \cap B(A_{n+1}) = \emptyset \quad \forall i \leq n$. That is, while $\omega(\boldsymbol{x}, \Phi_R) \subset A_{n+1}$, for Φ we have $\exists j \leq n : \omega(\boldsymbol{x}, \Phi) \subseteq A_j$. In accordance with the proof of (i), we can construct a set $U \supset A_{n+1}$ and conclude with a similar line of arguments as above that

$$\text{SW}_1(\mu^{\Phi}_{\boldsymbol{x},T}, \mu^{\Phi_R}_{\boldsymbol{x},T}) \neq 0 \quad \forall \boldsymbol{x} \in B(A_{n+1}). \tag{61}$$

Using the mean-value theorem for integrals we can conclude

$$\mathcal{E}^{B(A_{n+1})}_{\text{stat}}(\Phi_R) \propto \text{vol}(B(A_{n+1})) \tag{62}$$

proving the statement if we set $M_{\text{test}} = B(A_{n+1})$.
For the **topological error** we take an $\boldsymbol{x} \in B(A_{n+1})$. We know that $\omega(\boldsymbol{x}, \Phi_R) \in A_{n+1}$ and $\exists j \leq n : \omega(\boldsymbol{x}, \Phi) \subseteq A_j$. In the construction of the set S, we need to be cautious as $B(A_{n+1})$ could have many neighbouring basins. Accordingly, we have to construct a seperate set S_i for each attractor. We denote by S_i the set of points in A_i, ε_{d_H}-close to A_{n+1},

$$S_i = \{\boldsymbol{y} \in A_i | d_H(\{\boldsymbol{y}\}, A_{n+1}) \leq \varepsilon_{d_H}\}, \quad i \leq n \tag{63}$$

Again, each S_i will be empty for reasonable choices of ε_{d_H} and only non-empty for systems where A_i lies very close to A_{n+1}. Now, by an argument analogous to (i), we can conclude that

$$\mathcal{E}^M_{\text{top}}(\Phi_R) = \frac{\text{vol}(B(A_{n+1})) - \text{vol}(\cup_{i=1}^n \cup_{\boldsymbol{y}\in S_i} D^n_{\varepsilon_{d_H}}(\boldsymbol{y}))}{\text{vol}(M)} \propto \frac{\text{vol}(B(A_{n+1}))}{\text{vol}(M)} \tag{64}$$

proving the result when choosing $M_{\text{test}} = M$, given a reasonable choice of ε_{d_H} (as discussed in (i)).

E.2. Proof of Theorem 4.1

For the proof of Theorem 4.1, we need the following definitions:

Definition E.1. Given a trajectory $\Gamma_{\boldsymbol{x}_0}$, we define the graph of this trajectory by

$$\Omega_{\boldsymbol{x}_0} = I^f_{\boldsymbol{x}_0} \times \Gamma_{\boldsymbol{x}_0}. \tag{65}$$

Definition E.2. We define the following map, mapping the hypothesis class $\mathcal{H}$, a set of initial conditions and a set of times to the graph of the solution:

$$\sigma : \mathcal{H} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R} \times M, \quad (f, \boldsymbol{x}_0, I^f_{\boldsymbol{x}_0}) \mapsto \Omega_{\boldsymbol{x}_0} \tag{66}$$

The proof of theorem 4.1 is based on the following lemmata:

(i) Using Lemma E.3, we will first show that the set of solutions to the parameter estimation in SINDy (or any library-based algorithm) which is based on a minimization problem, can be rewritten as the pre-image $\pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0}))$. π_1 denotes the projection on the first argument of the pre-image.

(ii) Lemma E.4 then establishes how the pre-image and the associated minimization problem can be rewritten as a linear (matrix) equation.

(iii) Lemma E.5 shows under which conditions this equation will have a unique solution.

(iv) Finally, Lemma E.6 is used to characterize the set $\mathcal{H}_0$ for $\mathcal{H} = \mathcal{B}_L$, which will be used to prove the theorem.

Lemma E.3. *Let $\Gamma_{\boldsymbol{x}_0} = \{\boldsymbol{x}(t) | t \in I^f_{\boldsymbol{x}_0}\}$ be the solution (with graph $\Omega_{\boldsymbol{x}_0}$) of some ground-truth VF $f_{GT} \in \mathcal{B}_L$, with parameters $\boldsymbol{\theta}_{GT}$, $f_{GT} = f(\boldsymbol{x}; \boldsymbol{\theta}_{GT})$. For linearly parameterized function spaces (see Eq. (14)) like $\mathcal{B}_L$ we can write the (first projection of the) pre-image of a single trajectory as*

$$\pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0})) = \left\{ f \in \mathcal{B}_L | \Delta_j = \int_{t_0}^{t_1} \left(\boldsymbol{x}_j(t) - \boldsymbol{x}_j(0) - \int_0^t f_j(\boldsymbol{x}(s); \boldsymbol{\theta}) ds \right)^2 dt = 0, \quad j \in \{1, ..., n\} \right\} \tag{67}$$

$$= \min_{\boldsymbol{\theta}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}) | \int_{t_0}^{t_1} \| \dot{\boldsymbol{x}}(t) - f(\boldsymbol{x}; \boldsymbol{\theta}) \|^2_{L^2} dt \} \tag{68}$$

Note that the second equation exactly corresponds to the minimization problem stated by SINDy (without regularization).

Proof. By the definition of Eq. (66), it holds that

$$\pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0})) = \left\{ f(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathcal{B}_L \,|\, \dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t); \boldsymbol{\theta}), \, \forall t \in I^f_{\boldsymbol{x}_0} \quad \& \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \right\}. \tag{69}$$

The following equivalence holds for all $j \in \{1, ..., n\}$:

$$\dot{x}_j(t) = f_j(\boldsymbol{x}(t); \boldsymbol{\theta}), \quad x_j(0) = \boldsymbol{x}_{0,j} \tag{70}$$

$$\Leftrightarrow \quad x_j(t) - x_j(0) = \int_0^t f_j(\boldsymbol{x}(s); \boldsymbol{\theta}) ds \tag{71}$$

$$\Leftrightarrow \quad \int_{t_0}^{t_1} \left(x_j(t) - x_j(0) - \int_0^t f_j(\boldsymbol{x}(s); \boldsymbol{\theta}) ds \right)^2 dt = 0 \tag{72}$$

where we used in the last step that for a continuous function f with $f(x) \geq 0, \forall x \in I \subseteq \mathbb{R}$ we have $\int f(x) dx = 0 \quad \Rightarrow f(x) = 0$. This establishes Eq. (67).
For Eq. (68), let us assume $f \in \pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0}))$. By the definition of σ, we note that f satisfies for any $j \leq n$

$$\dot{x}_j(t) = f_j(\boldsymbol{x}(t); \boldsymbol{\theta}), \quad x_j(0) = \boldsymbol{x}_{0,j} \tag{73}$$

$$\Leftrightarrow \quad \dot{x}_j(t) - f_j(\boldsymbol{x}(t); \boldsymbol{\theta}) = 0, \quad x_j(0) = \boldsymbol{x}_{0,j}. \tag{74}$$

By the same argument we used to get from Eq. (71) to Eq. (72), it holds that f fulfills

$$\int_{t_0}^{t_1} \| \dot{\boldsymbol{x}}(t) - f(\boldsymbol{x}; \boldsymbol{\theta}) \|^2_{L^2} dt = 0. \tag{75}$$

As the L^2-norm (its square) $\| \cdot \|^2_{L^2}$ is a positive function, the minimum value it can attain is 0. Consequently, $f \in \min_{\boldsymbol{\theta}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}) | \int_{t_0}^{t_1} \| \dot{\boldsymbol{x}}(t) - f(\boldsymbol{x}; \boldsymbol{\theta}) \|^2_{L^2} dt \}$. For the other direction of the set-inclusion, assume $f \in \min_{\boldsymbol{\theta}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}) | \int_{t_0}^{t_1} \| \dot{\boldsymbol{x}}(t) - f(\boldsymbol{x}; \boldsymbol{\theta}) \|^2_{L^2} dt \}$. In general, f could be different from f_{GT} as there could exist many different f solving the minimization problem. However, for f_{GT} we know that the following holds:

$$\int_{t_0}^{t_1} \| \dot{\boldsymbol{x}}(t) - f_{GT}(\boldsymbol{x}; \boldsymbol{\theta}_{GT}) \|^2_{L^2} dt = 0. \tag{76}$$

It follows that $\forall f \in \min_{\boldsymbol{\theta}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}) | \int_{t_0}^{t_1} \| \dot{\boldsymbol{x}}(t) - f(\boldsymbol{x}; \boldsymbol{\theta}) \|^2_2 dt \}$ we must have

$$\int_{t_0}^{t_1} \| \dot{\boldsymbol{x}}(t) - f(\boldsymbol{x}; \boldsymbol{\theta}) \|^2_{L^2} dt = 0, \tag{77}$$

as the minimum value for $\int_{t_0}^{t_1} \| \dot{\boldsymbol{x}}(t) - f(\boldsymbol{x}; \boldsymbol{\theta}) \|^2_{L^2} dt$ is zero when choosing $\boldsymbol{\theta} = \boldsymbol{\theta}_{GT}$. This establishes that any $f \in \min_{\boldsymbol{\theta}} \{ f(\boldsymbol{x}; \boldsymbol{\theta}) | \int_{t_0}^{t_1} \| \dot{\boldsymbol{x}}(t) - f(\boldsymbol{x}; \boldsymbol{\theta}) \|^2_{L^2} dt \}$ has $\boldsymbol{x}(t)$ as a solution. By definition of σ, we conclude that $f \in \pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0}))$. □

Based on Lemma E.3, studying the solution of the (unregularized) SINDy minimization problem, Eq. (68), comes down to studying the pre-image $\pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0}))$. Further, note that there is a bijection between parameters $\boldsymbol{\theta} \in \mathbb{R}^{m\times n}$ and functions $f(\boldsymbol{x};\boldsymbol{\theta})$ for a linearly parameterized function space such as $\mathcal{B}_L$. In the following, we will identify $\boldsymbol{\theta}$ with $f(\boldsymbol{x};\boldsymbol{\theta})$, allowing us to write proofs for f in terms of proofs for $\boldsymbol{\theta}$.
In Kunze (1999), the solutions to Eq. (68) were studied for a one-dimensional setting. For higher dimensional settings, this takes the following form:

Lemma E.4. *$f(\boldsymbol{x};\boldsymbol{\theta})$ is an element of $\pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0}))$ if and only if $\boldsymbol{\theta}$ solves the following matrix equation*

$$\underbrace{\begin{bmatrix} \langle\Psi_0\cdot\Psi_0\rangle & \langle\Psi_0\cdot\Psi_1\rangle & \cdots & \langle\Psi_0\cdot\Psi_N\rangle \\ \langle\Psi_1\cdot\Psi_0\rangle & \langle\Psi_1\cdot\Psi_1\rangle & & \langle\Psi_1\cdot\Psi_N\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle\Psi_N\cdot\Psi_0\rangle & \langle\Psi_N\cdot\Psi_1\rangle & \cdots & \langle\Psi_N\cdot\Psi_N\rangle \end{bmatrix}}_{\bar{\Psi}} \underbrace{\begin{bmatrix} \theta_{1,j} \\ \theta_{2,j} \\ \vdots \\ \theta_{N,j} \end{bmatrix}}_{\boldsymbol{\theta}_j} = \underbrace{\begin{bmatrix} \langle\bar{\boldsymbol{x}}_j\Psi_0\rangle \\ \langle\bar{\boldsymbol{x}}_j\Psi_1\rangle \\ \vdots \\ \langle\bar{\boldsymbol{x}}_j\Psi_N\rangle \end{bmatrix}}_{X_j} \quad \forall j=1,\dots,n \tag{78}$$

with $\bar{\boldsymbol{x}}_j = \boldsymbol{x}_j(t) - \boldsymbol{x}(0)$ and

$$\Psi_k = \int_0^t \psi_k(\boldsymbol{x}(s))ds \tag{79}$$

$$\langle\Psi_k\cdot\Psi_i\rangle = \int_{t_0}^{t_1} \Psi_k(t)\cdot\Psi_i(t)dt = \int_{t_0}^{t_1}\int_0^t \psi_k(\boldsymbol{x}(s))ds\cdot\int_0^t \psi_i(\boldsymbol{x}(s))dsdt. \tag{80}$$

Proof. Consider the stationarity conditions

$$\frac{\partial\Delta_j}{\partial(\boldsymbol{x}_0)_j} = 0 \quad \& \quad \frac{\partial\Delta_j}{\partial\theta_{k,j}} = 0 \quad \forall k=1,\dots,m \quad \forall j=1,...,n.$$

The second condition leads to

$$\frac{\partial\Delta_j}{\partial\theta_{k,j}} = \int_{t_0}^{t_1}\left(2\left(\boldsymbol{x}_j(t)-\boldsymbol{x}_j(0)-\int_0^t f_j(\boldsymbol{x}(s);\boldsymbol{\theta})ds\right)\cdot\left(\frac{\partial}{\partial\theta_{k,j}}\int_0^t f_j(\boldsymbol{x}(s);\boldsymbol{\theta})ds\right)\right)dt \tag{81}$$

$$= \int_{t_0}^{t_1} 2\left((\boldsymbol{x}_j(t)-\boldsymbol{x}_j(0))\cdot\int_0^t \psi_k(\boldsymbol{x}(s))ds - \left(\sum_{i=1}^m \theta_{i,j}\int_0^t \psi_i(\boldsymbol{x}(s))ds\right)\cdot\int_0^t \psi_k(\boldsymbol{x}(s))ds\right)dt = 0. \tag{82}$$

Using the linearity of the integral this can be rewritten as

$$\sum_{i=1}^m \theta_{i,j}\int_{t_0}^{t_1}\left(\int_0^t \psi_i(\boldsymbol{x}(s))ds\cdot\int_0^t \psi_k(\boldsymbol{x}(s))ds\right)dt = \int_{t_0}^{t_1}\left((\boldsymbol{x}_j(t)-\boldsymbol{x}_j(0))\cdot\int_0^t \psi_k(\boldsymbol{x}(s))ds\right)dt \tag{83}$$

leading to Eq. (78). □

Lemma E.5. *Let $\Omega_{\boldsymbol{x}_0}$ be the graph of a trajectory from a VF $f \in \mathcal{B}_L$ with $\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t))$. It holds*

$$\pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0})) \text{ is unique } \iff \nexists\boldsymbol{\theta}\in\mathbb{R}^m:\ \Lambda(\boldsymbol{x}(t)) = \sum_{i=1}^m \theta_i\psi_i(\boldsymbol{x}(t)) = 0 \quad \forall t\in I^f_{\boldsymbol{x}_0}. \tag{84}$$

That is, the given trajectory does not solve an algebraic equation in the basis functions. By 'unique' we mean that $\pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0}))$ contains only one element.

Proof. We directly prove the equivalence stated in Eq. (84). Assuming $\pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0}))$ is not unique, by lemma E.4 this is equivalent to matrix $\bar{\Psi}$ being singular. This implies $\ker(\bar{\Psi}) \neq \{0\}$, which in turn is equivalent to

$$\exists\boldsymbol{\theta}\in\mathbb{R}^m\backslash\{0\}:\ \bar{\Psi}\cdot\boldsymbol{\theta} = 0$$

Hence, each row of the vector $\bar{\Psi} \cdot \boldsymbol{\theta}$ needs to be zero, implying that [6]

$$\forall k \leq m \quad : \int_{t_0}^{t_1} \left[\int_0^t \psi_k(\boldsymbol{x}(s))ds \cdot \int_0^t \sum_{i=1}^m \theta_i \psi_i(\boldsymbol{x}(s))ds \right] dt = 0 \tag{85}$$

We need to show that the integral in Eq. (85) can only become zero iff $\sum_{i=1}^m \theta_i \psi_i(\boldsymbol{x}(t)) = 0 \quad \forall t \in [t_0, t_1]$. By linearity of the integral over t, we can choose an arbitrary vector $\boldsymbol{\eta} \in \mathbb{R}^m$ and sum its components such that Eq. (85) is equal to the following:

$$\int_{t_0}^{t_1} \left[\sum_{k=1}^m \eta_k \int_0^t \psi_k(\boldsymbol{x}(s))ds \cdot \int_0^t \sum_{i=1}^m \theta_i \psi_i(\boldsymbol{x}(s))ds \right] dt = 0 \tag{86}$$

We are free to choose $\boldsymbol{\eta} = \boldsymbol{\theta}$, leading to

$$\int_{t_0}^{t_1} \left[\int_0^t \sum_{i=1}^m \boldsymbol{\theta}_i \psi_i(\boldsymbol{x}(s))ds \right]^2 dt = 0 \tag{87}$$

By the same argument as above, as the integral of a continuous, non-negative function can only be zero if the function itself is zero, we have

$$\int_0^t \sum_{i=1}^m \boldsymbol{\theta}_i \psi_i(\boldsymbol{x}(s))ds = 0 \quad \forall s \in [t_0, t_1]. \tag{88}$$

Since this has to be true $\forall t \in [t_0, t_1]$, we finally conclude

$$\sum_{i=1}^m \boldsymbol{\theta}_i \psi_i(\boldsymbol{x}(t)) = 0 \quad \forall t \in [t_0, t_1] \tag{89}$$

□

Lemma E.6. *Consider two trajectories,* $\Gamma_{\boldsymbol{y}_0} = \{\boldsymbol{y}(t) | t \in T, \boldsymbol{y}(0) = \boldsymbol{y}_0\}$ *and* $\Gamma_{\boldsymbol{x}_0} = \{\boldsymbol{x}(t) | t \in T, \boldsymbol{x}(0) = \boldsymbol{x}_0\}$*, with initial conditions* $\boldsymbol{x}_0$ *and* $\boldsymbol{y}_0$ *such that* $\boldsymbol{x}_0 = \boldsymbol{y}_0$*. We further assume these trajectories arise from two evolution operators associated to the VFs* f, g*,* $\Phi^f, \Phi^g : \mathbb{R} \times M \to M$ *with* $\boldsymbol{x}(t) = \Phi^f(t, \boldsymbol{x}_0)$ *and* $\boldsymbol{y}(t) = \Phi^g(t, \boldsymbol{y}_0)$*. If, for all times* $T \in \mathbb{R}$ *and Borel sets* $B \subseteq M$*, the occupation measures associated with these trajectories satisfy*

$$\mu_{\boldsymbol{x}_0,T}(B) = \mu_{\boldsymbol{y}_0,T}(B) \quad \forall B, \forall T \in \mathbb{R}, \tag{90}$$

then the underlying trajectories are identical:

$$\Gamma_{\boldsymbol{x}_0} = \Gamma_{\boldsymbol{y}_0}. \tag{91}$$

Proof. The proof of this Lemma E.6 was adapted from Dawkins (2024). As M is compact and Hausdorff and the Lebesgue measure is regular, Eq. (90) is equal to a different condition. Let φ be a bounded continuous function $\varphi : M \to \mathbb{R}$, where we denote the set of all bounded continuous functions by $\mathcal{C}_0(M)$. Then, $\mu_{\boldsymbol{x}_0,T} = \mu_{\boldsymbol{y}_0,T}$ implies

$$\int_0^T \varphi(\boldsymbol{x}(s))\, ds = \int_0^T \varphi(\boldsymbol{y}(s))\, ds, \qquad \forall T \geq 0,\ \forall \varphi \in \mathcal{C}_0(M) \tag{92}$$

where $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t)$ are the trajectories corresponding to Φ^f and Φ^g. Differentiating with respect to T yields:

$$\varphi(\boldsymbol{x}(T)) = \varphi(\boldsymbol{y}(T)), \qquad \forall T \geq 0. \tag{93}$$

Bounded continuous functions separate points, which means that $\forall \boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in M$ and $\exists \varphi \in \mathcal{C}_0(M) :\ \varphi(\boldsymbol{x}_1) \neq \varphi(\boldsymbol{x}_2)$. It follows that

$$\boldsymbol{x}(T) = \boldsymbol{y}(T), \qquad \forall T \geq 0. \tag{94}$$

□

[6] Note that in this proof we deviate from our standard notation in that $\boldsymbol{\theta}$ does not refer to the full matrix $\boldsymbol{\theta} \in \mathbb{R}^{m \times n}$ as in Eq. (14), but just to a vector in $\mathbb{R}^m$.

Proof of Theorem 4.1 Based on these four lemmas, we can now prove Theorem 4.1.

Proof. For the sake of simplicity, we focus on a bistable system where $M_{\text{train}} = B(A_1)$ and $M_{\text{test}} = B(A_2)$, but the generalization to n basins immediately follows. Let us denote the ground truth VF by $f_{GT} \in \mathcal{B}_L$ and write $f_{GT}(\boldsymbol{x}; \boldsymbol{\theta}_{GT})$ to make the parameter dependence explicit.
Now, consider an arbitrary element g in $\mathcal{B}_{L,0} = \{g \in \mathcal{B}_L | \mathcal{E}_{\text{gen}}^{M_{\text{train}}}(\Phi_g) = 0\}$, where Φ_g denotes the evolution operator associated with g. $g \in \mathcal{B}_{L,0}$ implies $\mathcal{E}_{\text{stat}}^{B(A_1)}(\Phi_g) = 0$. It holds that

$$\mathcal{E}_{\text{stat}}^{B(A_1)}(\Phi_g) = 0 \iff \mu_{\boldsymbol{x}_0,T}^{\Phi_f} = \mu_{\boldsymbol{x}_0,T}^{\Phi_g}, \tag{95}$$

based on SW_1 being a metric on the space of measures (Kolouri et al., 2016). This equality holds for all $T \in I_{\boldsymbol{x}_0}^f$ and for all $\boldsymbol{x}_0 \in B(A_1)$. According to Lemma E.6, this implies that all trajectories in $B(A_1)$ of f_{GT} and g coincide. Further, applying Lemma E.5 and considering the assumption that at least one trajectory does not satisfy an algebraic equation, which we will denote by $\Gamma'_{\boldsymbol{x}'_0}$ (with graph $\Omega'_{\boldsymbol{x}'_0}$), we can uniquely determine the parameters $\boldsymbol{\theta}$ of $g(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathcal{B}_L$ via $\pi_1(\sigma^{-1}(\Omega'_{\boldsymbol{x}'_0}))$. Since g and f_{GT} share the same non-algebraic trajectory, f and g must be identical as $\pi_1(\sigma^{-1}(\Omega_{\boldsymbol{x}_0}))$ only contains one element (Lemma E.5). Consequently,

$$\mathcal{B}_{L,0} = \{f_{GT}\}. \tag{96}$$

Thus, $\mathcal{B}_{L,0}$ solely consists of f_{GT}. Since f_{GT} has zero generalization error on M, it follows that any model in $\mathcal{B}_{L,0}$ exhibits zero generalization error on M. Therefore, $(\mathcal{B}_L, \mathcal{D})$ is strictly learnable. □

Corollary E.7. *When the observed trajectory $\Gamma_{\boldsymbol{x}_0} \subset \mathcal{D}$ solves an algebraic equation, we can nevertheless restrict $\mathcal{B}_L$ to $\mathcal{B}'_L \subset \mathcal{B}_L$ in a way that $(\mathcal{B}'_L, \mathcal{D})$ is strictly learnable.*

Proof. If $\Gamma_{\boldsymbol{x}_0}$ solves an algebraic equation, this means that $\ker(\bar{\Psi}) \neq \{0\}$ (see Lemma E.5). Hence, Eq. (78)

$$\bar{\Psi} \cdot \boldsymbol{\theta}_j = X_j \tag{97}$$

does not have a unique solution for $\boldsymbol{\theta}_j \neq 0$. However, for linear function spaces, we can use the fundamental theorem of homomorphisms to make $\bar{\Psi} \setminus \ker(\bar{\Psi}) \to \operatorname{im} \bar{\Psi}$ an isomorphism. Then, the system

$$\bar{\Psi}' \cdot \boldsymbol{\theta}'_j = X'_j \tag{98}$$

has a unique solution, where $\bar{\Psi}'$ corresponds to a matrix where some basis functions $\psi_{i,j}$ are removed. In the corresponding coefficient vector $\boldsymbol{\theta}'_j \in \mathbb{R}^{m'}$ the coefficients for the basis $\psi_{i,j}$ are removed as well, hence $m' < m$. By this process of removing basis functions $\psi_{i,j}$ from $\bar{\Psi}$, thus restricting $\mathcal{B}_L$ to B'_L, we can make the solution unique again. □

E.3. Proof of Theorem 4.2

Proof. Without loss of generality, we assume we have a bistable system, with f the ground truth VF, $M_{\text{train}} = B(A_1)$ and $M_{\text{test}} = B(A_2)$. We aim to construct an infinite family of functions $G = \{g_\alpha | \alpha \in I\}$ with zero reconstruction error on $B(A_1)$ but non-zero error on $B(A_2)$. We denote by Φ_{g_α} the evolution operator associated with g_α. Let $V' = B(A_2) \setminus A_2$. Due to the assumption that f is not topologically transitive on $B(A_2)$, we can choose some non-empty, open subset $V \subseteq V'$. As M is locally Hausdorff, we can define a compact subset $K \subset V$ such that K itself contains an open set. As V is open, we define a bump function $\Lambda \in \mathcal{C}^\infty(M)$ (Lee, 2012) that is zero outside of V and equals 1 on K.
The i-th component of a VF in G is then defined as

$$g_{\alpha,i}(\boldsymbol{x}) = f_i(\boldsymbol{x}) + \Lambda(\boldsymbol{x}) \cdot (-f_i(\boldsymbol{x}) + s_{\alpha,i}(\boldsymbol{x})), \quad i = 1, ..., n, \tag{99}$$

where $\boldsymbol{x}_\alpha \in K$. As K contains an open set, there are infinitely many different $\boldsymbol{x}_\alpha$. By construction, $\mathcal{E}_{\text{gen}}^{B(A_1)}(\Phi_{g_\alpha}) = 0$, since Λ is zero outside of V, thus zero on $B(A_1)$. On $B(A_1)$, g_α has the form $g_{\alpha,i}(\boldsymbol{x}) = f_i(\boldsymbol{x})$ (f has zero generalization error as it is the ground-truth VF). In contrast, on K, the VFs in G reduce to

$$g_\alpha(\boldsymbol{x}) = s_\alpha(\boldsymbol{x}), \tag{100}$$

where we assume s_α to be a differentiable VF with an attracting equilibrium point at $\boldsymbol{x}_\alpha$. This construction is feasible on any compact set K in any dimension. Given that Λ is smooth and s_α is differentiable, g_α remains a member of $\mathcal{X}^1$.

Thus, G forms an infinite family of VFs, each possessing an attracting equilibrium point at distinct positions in K. We will denote by $B(A_{\boldsymbol{x}_\alpha}) \subseteq V$ the basin of attraction of the equilibrium point $\boldsymbol{x}_\alpha$. All functions g_α have an attractor on $B(A_2)$ different from the one of the ground truth VF f. Additionally, $B(A_{\boldsymbol{x}_\alpha}) \cap A_2 = \emptyset$ by construction of V for any α. Thus, we constructed the same setting encountered in the proof of Theorem 3.3, case (ii). Using the proof and setting $B(A_{n+1})$ (in the notation of Theorem 3.3, case (ii)) to $B(A_{\boldsymbol{x}_\alpha})$, we can conclude that

$$\mathcal{E}_{\text{gen}}^{B(A_{\boldsymbol{x}_\alpha})}(\Phi_{g_\alpha}) \geq \varepsilon \tag{101}$$

with $\varepsilon > 0$. Since $V \subset B(A_2)$, it follows that $\mathcal{E}_{\text{gen}}^{B(A_2)}(\Phi_{g_\alpha}) \neq 0$ for all $\alpha \in I$. Consequently, $(\mathcal{X}^1, \mathcal{D})$ is not strictly learnable, since we can construct infinitely many functions in $\mathcal{X}^1$ with non-zero generalization error [7]. □

[7] For the proof that $(\mathcal{X}^1, \mathcal{D})$ is not strictly learnable, a single VF as constructed above would have sufficed.

	DSR	standard ML
data generating process	$\Phi : T \times M \to M$ Φ is the flow on some manifold. Note that no stochasticity is assumed as the OODG problem in DSR persists without noise in the data.	$p(x, y)$ $p(x, y)$ is the true distribution on an input $\mathcal{X}$ and output space $\mathcal{Y}$.
training data	$\cup_{\boldsymbol{x}_0 \in \boldsymbol{X}_0} \Phi\left([0, T], \boldsymbol{X}_0\right), \quad \boldsymbol{X}_0 \subset M_{\text{train}}$ The training data takes the form of trajectories where the initial conditions come from the training domain.	$\{(x_i, y_i)\}_{i=1}^N \sim p^{\text{train}}(x, y)$ N data points are drawn from a training distribution.
test distribution	$\Phi : T \times M_{\text{test}} \to M_{\text{test}}$ The equivalent object to the test distribution is the flow on a subset of the state space which was not seen in training $M_{\text{test}} \subset M$. If Φ is multistable, M_{test} harbors basins different from M_{train}. Hence, the data (trajectories) in M_{test} can follow a different dynamical law than on M_{train}.	$p^{\text{test}} \neq p^{\text{train}}$ In OODG, the test distributions are different from the training distribution.
distribution shift	A distribution shift occurs when the dynamics of Φ on M_{test} differ from the dynamics on M_{train}. This is particularly the case if Φ is multistable. Then, M_{train} and M_{test} contain different basins. We thoroughly discuss this in Sect. 3.1 of the paper.	Can take the form of: covariate shift $p^{\text{test}}(y\|x) = p^{\text{train}}(y\|x), \quad p^{\text{test}}(x) \neq p^{\text{train}}(x)$ label shift $p^{\text{test}}(x\|y) = p^{\text{train}}(x\|y), \quad p^{\text{test}}(y) \neq p^{\text{train}}(y)$ There are also other forms like concept drift or even more general versions of a distribution shift.
bounds on the generalization gap	$p(\varepsilon_{\text{gen}}\|\mathcal{D}) = \frac{1}{\text{vol}(\Theta_0)} \int_{\Theta_0} \mathbb{I}[\mathcal{E}_{\text{gen}}^{M_{\text{test}}}(\Phi_{\boldsymbol{\theta}}) = \varepsilon_{\text{gen}}] d\boldsymbol{\theta}$ In DSR, the role of the generalization gap is taken on by the learnability distribution. In Sect. 3.3 we discuss why such a reformulation is necessary. Essentially, bounds in terms of the distance of two distributions are meaningless as we do not (necessarily) have any noise in our data. Instead, these bounds are based on topological and ergodic properties of the flows.	$\|R^{\text{test}}(f) - R^{\text{train}}(f)\| < d(p^{\text{test}}, p^{\text{train}})$ Taking f as a trained neural network, one generally tries to bound the test vs. the training error by the distance between the test and training distributions given some formal distance measure d .

Figure A17. Comparison of OODG in DSR and standard ML.

Benchmark Survey

Ayed, I., de Bézenac, E., Pajot, A., Brajard, J., and Gallinari, P. Learning Dynamical Systems from Partial Observations. *arXiv:1902.11136 [physics]*, February 2019. URL http://arxiv.org/abs/1902.11136. arXiv: 1902.11136.

Azencot, O., Erichson, N. B., Lin, V., and Mahoney, M. W. Forecasting Sequential Data using Consistent Koopman Autoencoders. *arXiv:2003.02236 [physics]*, June 2020. URL http://arxiv.org/abs/2003.02236. arXiv: 2003.02236.

Brenner, M., Hess, F., Mikhaeil, J. M., Bereska, L. F., Monfared, Z., Kuo, P.-C., and Durstewitz, D. Tractable Dendritic RNNs for Reconstructing Nonlinear Dynamical Systems. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 2292–2320. PMLR, June 2022. URL https://proceedings.mlr.press/v162/brenner22a.html. ISSN: 2640-3498.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data: Sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, April 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1517384113. URL http://arxiv.org/abs/1509.03580. arXiv: 1509.03580.

Champion, K., Lusch, B., Kutz, J. N., and Brunton, S. L. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45):22445–22451, November 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1906995116. URL http://arxiv.org/abs/1904.02107. arXiv:1904.02107 [stat].

Chen, Y., Qian, Y., and Cui, X. Time series reconstructing using calibrated reservoir computing. *Scientific Reports*, 12(1): 16318, September 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-20331-3. URL https://www.nature.com/articles/s41598-022-20331-3. Number: 1 Publisher: Nature Publishing Group.

Duncker, L., Bohner, G., Boussard, J., and Sahani, M. Learning interpretable continuous-time models of latent stochastic dynamical systems. *arXiv:1902.04420 [cs, math, stat]*, February 2019. URL http://arxiv.org/abs/1902.04420. arXiv: 1902.04420.

Farmer, J. D. and Sidorowich, J. J. Predicting chaotic time series. *Physical Review Letters*, 59(8):845–848, August 1987. doi: 10.1103/PhysRevLett.59.845. URL https://link.aps.org/doi/10.1103/PhysRevLett.59.845. Publisher: American Physical Society.

Fu, Y., Saab Jr, S., Ray, A., and Hauser, M. A Dynamically Controlled Recurrent Neural Network for Modeling Dynamical Systems. *arXiv:1911.00089 [cs, stat]*, October 2019. URL http://arxiv.org/abs/1911.00089. arXiv: 1911.00089.

Gauthier, D. J., Bollt, E., Griffith, A., and Barbosa, W. A. S. Next generation reservoir computing. *Nature Communications*, 12(1):5564, September 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25801-2. URL https://www.nature.com/articles/s41467-021-25801-2. Number: 1 Publisher: Nature Publishing Group.

Geneva, N. and Zabaras, N. Transformers for modeling physical systems. *Neural Networks*, 146:272–289, February 2022. ISSN 0893-6080. doi: 10.1016/j.neunet.2021.11.022. URL https://www.sciencedirect.com/science/article/pii/S0893608021004500.

Gilpin, W. Deep reconstruction of strange attractors from time series. *arXiv:2002.05909 [nlin, physics:physics, q-bio, stat]*, October 2020. URL http://arxiv.org/abs/2002.05909. arXiv: 2002.05909.

Gilpin, W. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, January 2022. URL https://openreview.net/forum?id=enYjtbjYJrf.

Goyal, P. and Benner, P. Learning Dynamics from Noisy Measurements using Deep Learning with a Runge-Kutta Constraint, September 2021. URL http://arxiv.org/abs/2109.11446. arXiv:2109.11446 [cs, math].

Hernandez, D., Moretti, A. K., Saxena, S., Wei, Z., Cunningham, J., and Paninski, L. Nonlinear Evolution via Spatially-Dependent Linear Dynamics for Electrophysiology and Calcium Data. *Neurons, Behavior, Data analysis, and Theory*, 3 (3), June 2020. Publisher: The neurons, behavior, data analysis and theory collective.

Hess, F., Monfared, Z., Brenner, M., and Durstewitz, D. Generalized Teacher Forcing for Learning Chaotic Dynamics. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 13017–13049. PMLR, July 2023. URL https://proceedings.mlr.press/v202/hess23a.html. ISSN: 2640-3498.

Jordana, A., Carpentier, J., and Righetti, L. Learning Dynamical Systems from Noisy Sensor Measurements using Multiple Shooting. *arXiv:2106.11712 [cs, stat]*, June 2021. URL http://arxiv.org/abs/2106.11712. arXiv: 2106.11712.

Karlsson, D. and Svanström, O. Modelling Dynamical Systems Using Neural Ordinary Differential Equations, 2019.

Kim, T. D., Luo, T. Z., Pillow, J. W., and Brody, C. Inferring Latent Dynamics Underlying Neural Population Activity via Neural Differential Equations. In *International Conference on Machine Learning*, pp. 5551–5561. PMLR, July 2021. URL http://proceedings.mlr.press/v139/kim21h.html. ISSN: 2640-3498.

Kraemer, K. H., Datseris, G., Kurths, J., Kiss, I. Z., Ocampo-Espindola, J. L., and Marwan, N. A unified and automated approach to attractor reconstruction. *New Journal of Physics*, 23(3):033017, March 2021. ISSN 1367-2630. doi: 10.1088/1367-2630/abe336. URL https://doi.org/10.1088/1367-2630/abe336. Publisher: IOP Publishing.

Lai, Z., Mylonas, C., Nagarajaiah, S., and Chatzi, E. Structural identification with physics-informed neural ordinary differential equations. *Journal of Sound and Vibration*, 508:116196, September 2021. ISSN 0022-460X. doi: 10.1016/j.jsv.2021.116196. URL https://www.sciencedirect.com/science/article/pii/S0022460X21002686.

Lee, K. and Carlberg, K. T. Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders. *Journal of Computational Physics*, 404:108973, 2020. ISSN 0021-9991. doi: 10.1016/j.jcp.2019.108973. URL https://www.sciencedirect.com/science/article/pii/S0021999119306783.

Lejarza, F. and Baldea, M. Data-driven discovery of the governing equations of dynamical systems via moving horizon optimization. *Scientific Reports*, 12(1):11836, July 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-13644-w. URL https://www.nature.com/articles/s41598-022-13644-w. Number: 1 Publisher: Nature Publishing Group.

Li, Y. and Duan, J. A data-driven approach for discovering stochastic dynamical systems with non-Gaussian Lévy noise. *Physica D: Nonlinear Phenomena*, 417:132830, March 2021. ISSN 01672789. doi: 10.1016/j.physd.2020.132830. URL https://linkinghub.elsevier.com/retrieve/pii/S0167278920308319.

Linderman, S. W., Miller, A. C., Adams, R. P., Blei, D. M., Paninski, L., and Johnson, M. J. Recurrent switching linear dynamical systems, October 2016. URL http://arxiv.org/abs/1610.08466. arXiv:1610.08466 [stat].

Linot, A. J., Burby, J. W., Tang, Q., Balaprakash, P., Graham, M. D., and Maulik, R. Stabilized neural ordinary differential equations for long-time forecasting of dynamical systems. *Journal of Computational Physics*, 474:111838, February 2023. ISSN 0021-9991. doi: 10.1016/j.jcp.2022.111838. URL https://www.sciencedirect.com/science/article/pii/S0021999122009019.

Liu, Z. and Jin, L. Model-Free Prediction of Chaotic Systems Using High Efficient Next-generation Reservoir Computing, October 2021. URL http://arxiv.org/abs/2110.13614. arXiv:2110.13614 [nlin].

Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G. E. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, March 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00302-5. URL https://www.nature.com/articles/s42256-021-00302-5. Number: 3 Publisher: Nature Publishing Group.

Lu, Z., Hunt, B. R., and Ott, E. Attractor Reconstruction by Machine Learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(6):061104, June 2018. ISSN 1054-1500, 1089-7682. doi: 10.1063/1.5039508. URL http://arxiv.org/abs/1805.03362. arXiv:1805.03362 [nlin].

Lusch, B., Kutz, J. N., and Brunton, S. L. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07210-0. URL https://www.nature.com/articles/s41467-018-07210-0. Number: 1 Publisher: Nature Publishing Group.

Mehta, V., Char, I., Neiswanger, W., Chung, Y., Nelson, A., Boyer, M., Kolemen, E., and Schneider, J. Neural Dynamical Systems: Balancing Structure and Flexibility in Physical Prediction. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 3735–3742, December 2021. doi: 10.1109/CDC45484.2021.9682807. ISSN: 2576-2370.

Mikhaeil, J. M., Monfared, Z., and Durstewitz, D. On the difficulty of learning chaotic dynamics with RNNs. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, October 2022. URL https://openreview.net/forum?id=-_AMpmyV0Ll.

Mohajerin, N. and Waslander, S. L. Multi-Step Prediction of Dynamic Systems with Recurrent Neural Networks. *arXiv:1806.00526 [cs]*, May 2018. URL http://arxiv.org/abs/1806.00526. arXiv: 1806.00526.

Nguyen, D., Ouala, S., Drumetz, L., and Fablet, R. EM-like Learning Chaotic Dynamics from Noisy and Partial Observations. *arXiv:1903.10335 [cs, stat]*, March 2019. URL http://arxiv.org/abs/1903.10335. arXiv: 1903.10335.

Nguyen, D., Ouala, S., Drumetz, L., and Fablet, R. Variational Deep Learning for the Identification and Reconstruction of Chaotic and Stochastic Dynamical Systems from Noisy and Partial Observations. *arXiv:2009.02296 [cs, stat]*, February 2021. URL http://arxiv.org/abs/2009.02296. arXiv: 2009.02296.

Otto, S. E. and Rowley, C. W. Linearly-Recurrent Autoencoder Networks for Learning Dynamics, January 2019. URL http://arxiv.org/abs/1712.01378. arXiv:1712.01378 [cs, math, stat].

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., and Ott, E. Using Machine Learning to Replicate Chaotic Attractors and Calculate Lyapunov Exponents from Data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(12):121102, December 2017. ISSN 1054-1500, 1089-7682. doi: 10.1063/1.5010300. URL http://arxiv.org/abs/1710.07313. arXiv: 1710.07313.

Qin, T., Wu, K., and Xiu, D. Data driven governing equations approximation using deep neural networks. *Journal of Computational Physics*, 395:620–635, October 2019. ISSN 0021-9991. doi: 10.1016/j.jcp.2019.06.042. URL https://www.sciencedirect.com/science/article/pii/S0021999119304504.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Multistep Neural Networks for Data-driven Discovery of Nonlinear Dynamical Systems. *arXiv:1801.01236 [nlin, physics:physics, stat]*, January 2018. URL http://arxiv.org/abs/1801.01236. arXiv: 1801.01236.

Raissi, M., Perdikaris, P., and Karniadakis, G. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019. ISSN 00219991. doi: 10.1016/j.jcp.2018.10.045. URL https://linkinghub.elsevier.com/retrieve/pii/S0021999118307125.

Rusch, T. K., Mishra, S., Erichson, N. B., and Mahoney, M. W. Long Expressive Memory for Sequence Modeling. *arXiv:2110.04744 [cs, math, stat]*, February 2022. URL http://arxiv.org/abs/2110.04744. arXiv: 2110.04744.

Schlaginhaufen, A., Wenk, P., Krause, A., and Dörfler, F. Learning Stable Deep Dynamics Models for Partially Observed or Delayed Dynamical Systems. *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, October 2021. URL https://arxiv.org/abs/2110.14296v2.

Schmidt, D., Koppe, G., Monfared, Z., Beutelspacher, M., and Durstewitz, D. Identifying nonlinear dynamical systems with multiple time scales and long-range dependencies. *arXiv:1910.03471 [cs, q-bio, stat]*, March 2021. URL http://arxiv.org/abs/1910.03471. arXiv: 1910.03471.

Shalova, A. and Oseledets, I. Deep Representation Learning for Dynamical Systems Modeling. *arXiv:2002.05111 [nlin, physics:physics]*, February 2020. URL http://arxiv.org/abs/2002.05111. arXiv: 2002.05111.

Singh, S. K., Yang, R., Behjat, A., Rai, R., Chowdhury, S., and Matei, I. PI-LSTM: Physics-Infused Long Short-Term Memory Network. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 34–41, December 2019. doi: 10.1109/ICMLA.2019.00015.

Strauss, R. Augmenting Neural Differential Equations to Model Unknown Dynamical Systems with Incomplete State Information. *arXiv:2008.08226 [physics, q-bio]*, August 2020. URL http://arxiv.org/abs/2008.08226. arXiv: 2008.08226.

Sussillo, D., Jozefowicz, R., Abbott, L. F., and Pandarinath, C. LFADS - Latent Factor Analysis via Dynamical Systems. *arXiv:1608.06315 [cs, q-bio, stat]*, August 2016. URL http://arxiv.org/abs/1608.06315. arXiv: 1608.06315.

Tran, G. and Ward, R. Exact Recovery of Chaotic Systems from Highly Corrupted Data. *Multiscale Modeling & Simulation*, 15(3):1108–1129, January 2017. ISSN 1540-3459. doi: 10.1137/16M1086637. URL https://epubs.siam.org/doi/10.1137/16M1086637. Publisher: Society for Industrial and Applied Mathematics.

Tripura, T. and Chakraborty, S. A sparse Bayesian framework for discovering interpretable nonlinear stochastic dynamical systems with Gaussian white noise. *Mechanical Systems and Signal Processing*, 187:109939, March 2023. ISSN 0888-3270. doi: 10.1016/j.ymssp.2022.109939. URL https://www.sciencedirect.com/science/article/pii/S088832702201007X.

Trischler, A. and D'Eleuterio, G. M. Synthesis of recurrent neural networks for dynamical system simulation. *arXiv:1512.05702 [cs]*, June 2016. URL http://arxiv.org/abs/1512.05702. arXiv: 1512.05702.

Uribarri, G. and Mindlin, G. B. Dynamical time series embeddings in recurrent neural networks. *Chaos, Solitons & Fractals*, 154:111612, January 2022. ISSN 0960-0779. doi: 10.1016/j.chaos.2021.111612. URL https://www.sciencedirect.com/science/article/pii/S0960077921009668.

Vlachas, P. R., Byeon, W., Wan, Z. Y., Sapsis, T. P., and Koumoutsakos, P. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2213):20170844, May 2018. doi: 10.1098/rspa.2017.0844. URL https://royalsocietypublishing.org/doi/10.1098/rspa.2017.0844. Publisher: Royal Society.

Vlachas, P. R., Pathak, J., Hunt, B. R., Sapsis, T. P., Girvan, M., Ott, E., and Koumoutsakos, P. Backpropagation algorithms and Reservoir Computing in Recurrent Neural Networks for the forecasting of complex spatiotemporal dynamics. *Neural Networks*, 126:191–217, June 2020. ISSN 0893-6080. doi: 10.1016/j.neunet.2020.02.016. URL https://www.sciencedirect.com/science/article/pii/S0893608020300708.

Voss, H. U., Timmer, J., and Kurths, J. Nonlinear dynamical system identification from uncertain and indirect measurements. *International Journal of Bifurcation and Chaos*, 14(06):1905–1933, June 2004. ISSN 0218-1274. doi: 10.1142/S0218127404010345. URL https://www.worldscientific.com/doi/abs/10.1142/S0218127404010345. Publisher: World Scientific Publishing Co.

Wang, Y.-J. and Lin, C.-T. Runge-Kutta neural network for identification of dynamical systems in high accuracy. *IEEE Transactions on Neural Networks*, 9(2):294–307, March 1998. ISSN 1941-0093. doi: 10.1109/72.661124. Conference Name: IEEE Transactions on Neural Networks.

Yang, L., Sun, X., Hamzi, B., Owhadi, H., and Xie, N. Learning Dynamical Systems from Data: A Simple Cross-Validation Perspective, Part V: Sparse Kernel Flows for 132 Chaotic Dynamical Systems, January 2023. URL https://ui.adsabs.harvard.edu/abs/2023arXiv230110321Y. arXiv e-prints ADS Bibcode: 2023arXiv230110321Y Type: article.

Yin, Y., Ayed, I., de Bézenac, E., Baskiotis, N., and Gallinari, P. LEADS: Learning Dynamical Systems that Generalize Across Environments, February 2022. URL http://arxiv.org/abs/2106.04546. arXiv:2106.04546 [cs, stat].

Zhang, L., Tang, S., and He, G. Learning chaotic systems from noisy data via multi-step optimization and adaptive training. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(12):123134, 2022. doi: 10.1063/5.0114542. URL https://doi.org/10.1063/5.0114542. _eprint: https://doi.org/10.1063/5.0114542.

Zhao, H., Rai, P., Du, L., Buntine, W., Phung, D., and Zhou, M. Variational Autoencoders for Sparse and Overdispersed Discrete Data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1684–1694. PMLR, June 2020. URL https://proceedings.mlr.press/v108/zhao20c.html. ISSN: 2640-3498.