Almost-Linear RNNs Yield Highly Interpretable Symbolic Codes in Dynamical Systems Reconstruction

Manuel Brenner[1,2]*, **Christoph Jürgen Hemmer**[1,3]*, **Zahra Monfared**[2], **Daniel Durstewitz**[1,2,3]
[1]Dept. of Theoretical Neuroscience, Central Institute of Mental Health, Medical Faculty, Heidelberg University, Germany
[2]Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Germany
[3]Faculty of Physics and Astronomy, Heidelberg University, Heidelberg, Germany

Abstract

Dynamical systems (DS) theory is fundamental for many areas of science and engineering. It can provide deep insights into the behavior of systems evolving in time, as typically described by differential or recursive equations. A common approach to facilitate mathematical tractability and interpretability of DS models involves decomposing nonlinear DS into multiple linear DS separated by switching manifolds, i.e. piecewise linear (PWL) systems. PWL models are popular in engineering and a frequent choice in mathematics for analyzing the topological properties of DS. However, hand-crafting such models is tedious and only possible for very low-dimensional scenarios, while inferring them from data usually gives rise to unnecessarily complex representations with very many linear subregions. Here we introduce Almost-Linear Recurrent Neural Networks (AL-RNNs) which automatically and robustly produce most parsimonious PWL representations of DS from time series data, using as few PWL nonlinearities as possible. AL-RNNs can be efficiently trained with any SOTA algorithm for dynamical systems reconstruction (DSR), and naturally give rise to a symbolic encoding of the underlying DS that provably preserves important topological properties. We show that for the Lorenz and Rössler systems, AL-RNNs discover, in a purely data-driven way, the known topologically minimal PWL representations of the corresponding chaotic attractors. We further illustrate on two challenging empirical datasets that interpretable symbolic encodings of the dynamics can be achieved, tremendously facilitating mathematical and computational analysis of the underlying systems.

1 Introduction

Dynamical systems (DS) underlie many real-world phenomena of scientific and practical relevance. Complex chaotic DS are believed to govern market dynamics [65], the rhythms of the brain [16], climate systems [100], or ecosystems [71]. A by now rapidly growing field in scientific ML is dynamical systems reconstruction (DSR), where the goal is to learn a DS model directly from data that constitutes a generative surrogate model of the data-generating DS. DSR increasingly relies on deep learning, especially in contexts where dynamics are too complex to be captured by simple equations or where the underlying processes are not fully understood.

One way of making DSR models mathematically accessible is piecewise linear (PWL) designs, popular among engineers for decades [8, 17, 43, 81, 91]. In the mathematical theory of DS, PWL models also play a special role and simplify many types of analysis [3, 54], such as the characterization of

*These authors contributed equally to this work.
Corresponding authors: {manuel.brenner, christoph.hemmer, daniel.durstewitz}@zi-mannheim.de

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

bifurcations [27, 10, 42, 40, 77, 70]. This is because linear DS are well-understood and straightforward to analyze, while nonlinear DS lack an equally simple description [78, 94, 13]. RNNs based on PWL activation functions, like rectified-linear units (ReLUs), have been proposed recently for learning mathematically tractable DSR models. Such piecewise-linear RNNs (PLRNNs), combined with effective training techniques for controlling gradient flows [67, 39], achieve state-of-the-art (SOTA) performance across a wide range of DSR tasks, including challenging (high-dimensional, noisy, chaotic, partially observed ...) empirical time series [25, 11, 39, 12]. However, while featuring a PWL design, the resulting constructions are often complex, with a large number of linear subregions required to capture the data properly, hence still impeding effective analysis. On the other hand, a class of switching linear DS has been proposed to decompose nonlinear DS into linear regions combined with switching states that determine the transitions between these regions [1, 31, 28, 58, 56, 57, 2]. However, the underlying assumptions of these models and the complexity of the inference mechanisms these entail, often make their training challenging and impede their efficient application to DSR problems, especially when moving to real-world scenarios and higher-dimensional systems.

Here we propose almost-linear RNNs (AL-RNNs) which combine linear units and ReLUs, but can use as few of the latter as necessary to achieve a most parsimonious representation in terms of linear subregions. AL-RNNs are easy and effective to train by any SOTA algorithm for DSR. Through this, they are able to robustly identify topologically or geometrically minimal representations of well-known chaotic systems. Their structure translates naturally into a symbolic coding that preserves important topological properties. These features make AL-RNNs highly interpretable and mathematically tractable, enabling to harvest tools from symbolic dynamics [74, 55], including the representation of empirically observed DS via minimal topological and computational graphs.

2 Related Work

Dynamical systems reconstruction The field of data-driven DSR has been rapidly expanding in recent years. On the one hand there are approaches based on function libraries for approximating unknown vector fields, which have become particularly popular in some areas like physics [53, 64]. Among these, Sparse Identification of Nonlinear Dynamics (SINDy) and its variants [14, 60, 45, 21, 66, 44] is probably the most popular. Since in these models sets of differential equations are directly formulated in terms of known, predefined function libraries, instead of using NN black-box approximators, they have some level of interpretability in the sense that they are human-readable and can easily be related to established mathematical building blocks in physical or biological theories [37, 83]. This does not necessarily make them mathematically tractable, however, since systems of nonlinear differential equations are in themselves usually hard to analyze (in fact, their behavior is much of the core topic of DS theory [78]). They also have other limitations, including a difficulty in capturing complex and noisy empirical data [11, 39], as they usually require considerable prior knowledge about the system's underlying structure (i.e., which terms to include in the function library). This somewhat limits their applicability for discovering novel phenomena. On the other hand, many recent powerful DSR methods rely on universal approximators, in particular the fact that sufficiently large RNNs can approximate any underlying DS [29, 47, 34]. Such methods may be grouped into several broad classes, including reservoir computers [76, 79, 80], neural ODEs/PDEs [20, 46, 4, 48], neural/ Koopman operators [15, 63, 75, 6, 73, 30, 104], and RNNs [99, 25, 101, 19, 11, 84, 39]. The latter are commonly trained by variants of backpropagation through time (BPTT, [101, 102, 11]), combined with specific control techniques [67] to remedy the exploding/ vanishing gradients problem [9, 67, 11, 39]. While DSR algorithms based on universal approximators achieve SOTA performance on DSR tasks, and often work particularly well on empirical time series [11, 39], they commonly deliver a complex model structure that is difficult to interpret and parse mathematically.

Nonlinear dynamics via linear DS The idea of approaching nonlinear DS through our good grasp of linear DS has been around for quite a long time, reflected in important theoretical results like the Hartman-Grobman theorem [36][1] or Koopman operator theory [49, 15]. While linear DS are easy to analyze and well understood [78, 94], however, they cannot properly capture most real-world systems, as they cannot produce many important DS phenomena such as limit cycles, chaos, or multistability. This has motivated the modeling of complex dynamics in terms of compositions

[1]The Hartman-Grobman theorem states that nonlinear hyperbolic DS are topologically conjugate to a linear DS in some neighborhood of the system's equilibrium points.

of locally linear dynamics, as the next best alternative, i.e. piecewise-linear (PWL) maps or ODE systems [18, 41, 22]. PWL models have been popular in engineering and mathematics for several decades for these reasons, including earlier attempts for learning such models directly from data [93, 24]. Switching linear DS are one particular brand of PWL models with a long tradition in DS and control theory [23, 95, 1, 31, 28]. These systems model nonlinear dynamics through a set of linear (or affine) DS combined with a switching rule which decides which linear DS is currently active. Likewise, in mathematics PWL models served well for investigating generic properties of nonlinear systems, e.g. the tent map which is topologically conjugate to the logistic map [3]. PWL models often lend themselves to particularly convenient symbolic representations [54, 3], based on which important topological properties of the underlying system, e.g. the nature and number of unstable periodic orbits embedded within a chaotic attractor, can be analyzed [69, 74, 98].

More recently, various modern approaches for inferring PWL models from data have been formulated. For instance, switching state space models combine hidden Markov models with linear DS, jointly inferring the state of a switching (random) variable with the linear DS parameters conditioned on these states [31]. Various extensions of this basic setup like recurrent and hierarchical switching linear DS and fully Bayesian inference methods have been advanced in recent years [92, 58, 56]. However, inference in these models is often complex and not necessarily optimized for DSR, limiting their applicability to mainly low-dimensional scenarios with comparatively simple dynamics. Most of these models are also discontinuous in their dynamics across switches, while most commonly we would require the state variables to evolve continuously across the whole of state space. Piecewise-linear RNNs (PLRNNs), and, relatedly, threshold-linear networks [33, 105, 109], on the other hand, are based on familiar ReLUs and hence change continuously across their switching manifolds [25, 51, 88]. They also have some biological justification [33, 25]. Commonly they are trained by variants of BPTT backed up by specific control-theoretic approaches like sparse [67] or generalized [39] teacher forcing which make them SOTA on many DSR tasks. Different PLRNN architectures have been proposed to enhance expressivity or reduce the dimensionality of trained models [11, 39]. Yet, while these advances may yield comparatively low-dimensional state spaces, the number of different linear subregions that need to be allocated usually remains very high, hampering efficient mathematical analysis.

3 Methodological and Theoretical Prerequisites

3.1 AL-RNN Model

Consider a piecewise linear recurrent neural network (PLRNN, [25]):

$$\boldsymbol{z}_t = F_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1}) = \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{W}\phi(\boldsymbol{z}_{t-1}) + \boldsymbol{h}, \tag{1}$$

where diagonal $\boldsymbol{A} \in \mathbb{R}^{M \times M}$ contains linear self-connections, $\boldsymbol{W} \in \mathbb{R}^{M \times M}$ are nonlinear connections between units, $\boldsymbol{h} \in \mathbb{R}^{M}$ is a bias term, and $\phi(\boldsymbol{z}) = \max[0, \boldsymbol{z}]$ is an element-wise ReLU nonlinearity. To expose the piecewise linear structure of this model more clearly, by noting that the slope of the ReLU is either 0 or 1 depending on the sign of $z_{m,t}$, one can reformulate this as

$$\boldsymbol{z}_t = (\boldsymbol{A} + \boldsymbol{W}\boldsymbol{D}_{\Omega(t-1)})\boldsymbol{z}_{t-1} + \boldsymbol{h} =: \boldsymbol{W}_{\Omega(t-1)}\, \boldsymbol{z}_{t-1} + \boldsymbol{h}, \tag{2}$$

where $\boldsymbol{D}_{\Omega(t)} := diag(\boldsymbol{d}_{\Omega(t)})$ is a diagonal matrix and $\boldsymbol{d}_{\Omega(t)} = (d_1, d_2, \cdots, d_M)$ an indicator vector with $d_m(z_{m,t}) = 1$ whenever $z_{m,t} > 0$ and zero otherwise [26]. For the 2^M different configurations of $\boldsymbol{D}_{\Omega(t)}$, $\boldsymbol{D}_{\Omega^k}$, $k \in \{1, 2, \cdots, 2^M\}$, the phase space of system eq. 2 is divided into 2^M subregions with linear dynamics

$$\boldsymbol{z}_{t+1} \;=\; \boldsymbol{W}_{\Omega^k}\, \boldsymbol{z}_t \;+\; \boldsymbol{h}, \qquad\qquad \boldsymbol{W}_{\Omega^k} := \boldsymbol{A} + \boldsymbol{W}\boldsymbol{D}_{\Omega^k}. \tag{3}$$

Empirically, M often needs to be quite large (at least on the order of the number of observations) for achieving good reconstructions of observed DS. Since the number of subregions grows as 2^M, analyzing inferred models in terms of the subregions can thus become very challenging. We therefore introduce a novel variant of the PLRNN in which only a subset of $P << M$ units are equipped with a ReLU nonlinearity, yielding

$$\boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{W}\Phi^{*}(\boldsymbol{z}_{t-1}) + \boldsymbol{h} \tag{4}$$

where

$$\Phi^*(\boldsymbol{z}_t) = [z_{1,t}, \cdots, z_{M-P,t}, \max(0, z_{M-P+1,t}), \cdots, \max(0, z_{M,t})]^T . \quad (5)$$

In this formulation, we thus only have 2^P different linear subregions, while still accommodating a sufficiently large number of latent states for capturing unobserved dimensions in the data and disentangling trajectories sufficiently [96, 86].[2]

The model is trained on the N-dimensional observations $\{\boldsymbol{x}_t\}, t = 1 \dots T, \boldsymbol{x}_t \in \mathbb{R}^N$, by a variant of sparse teacher forcing called identity teacher forcing [67, 11]. In identity teacher forcing, the first N latent states ('readout neurons') are replaced by the N-dimensional observations every τ time steps, where τ is chosen such as to optimally control trajectory and gradient flows, avoiding exploding gradients while providing the model sufficient opportunity to unroll into the future to capture the underlying DS' long-term behavior (see [67, 39] for details); see Appx. A.2 for details on training. We emphasize that sparse teacher forcing is *only used for training* the model, and is turned off at test time where the model generates new trajectories completely independent from the data.

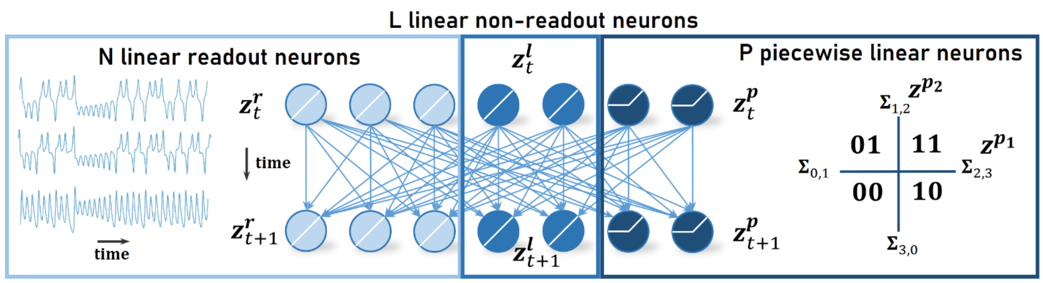

Figure 1: Illustration of the AL-RNN architecture.

3.2 Theoretical Background: Symbolic Dynamics and Symbolic Coding of AL-RNN

The mathematical field of symbolic dynamics formulates conditions under which a DS has a unique symbolic representation and discusses how to harvest this symbolic representation to prove certain properties of the underlying system, which otherwise may be more difficult to address [74, 55]. In fact, symbolic dynamics has led to many powerful insights and formal results in DS theory, e.g. about the properties of chaos or type and number of periodic orbits [32, 106]. An appealing feature of symbolic dynamics for the field of ML/AI is that it links concepts in DS theory to computational concepts like finite state automata or formal languages, as well as graph theory [55, 35]. It can thus facilitate the computational interpretation of natural or trained dynamical systems, like RNNs.

Assume we have an alphabet of n symbols $\mathcal{A} = \{0, \dots, n-1\}$, from which we form infinite sequences (bidirectionally or only in forward-time) $\boldsymbol{a} = \dots a_{-2}a_{-1}.a_0a_1a_2 \dots$ with $a_k \in \mathcal{A}$, and the dot separating past from future (i.e., indices $k < 0$ indicate backward time, and $k \geq 0$ present and forward time). Then the space of all possible sequences, together with the so-called (left) shift operator given by

$$\sigma(\boldsymbol{a}) = \sigma(\dots a_{-2}a_{-1}.a_0a_1a_2\dots) = \dots a_{-1}a_0.a_1a_2a_3\dots \quad (6)$$

defines the full shift space $\mathcal{A}^{\mathbb{Z}}$. We denote by $\sigma^k = \sigma \circ \sigma \circ \dots \circ \sigma$ the k-times iteration of the shift. Now consider a DS (S, ϕ) consisting of a metric (state) space S and a recursive (flow) map ϕ. The flow map $\phi_{\Delta t}(\boldsymbol{x})$ advances the system's current state $\boldsymbol{x}$ by Δt and may be thought of as the solution operator of the underlying DS $\dot{\boldsymbol{x}} = f(\boldsymbol{x})$ [78]. When training RNNs $\boldsymbol{z}_t = F_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1})$ on time series of observations $\{g(\boldsymbol{x}_{k\Delta t})\}, k = 1 \dots T$, from the underlying DS, where g is the observation function, we are trying to approximate this flow map. Assume the whole state space S can be partitioned into a finite set $\mathcal{U} = \{U_0 \dots U_{n-1}\}$ of disjoint open sets U_e, such that $S = \bigcup_{e=0}^{n-1} \overline{U_e}$, i.e. S is covered by the union of the closures of these sets. We call this a *topological partition* of S [55].

The central idea now is to assign a unique symbol $a_e \in \mathcal{A}$ to each set $U_e \in \mathcal{U}$, with $n = |\mathcal{U}| = |\mathcal{A}|$. As a trajectory $\boldsymbol{x}(t)$ of the underlying DS travels through the system's state space S, observed at times

[2]Strictly, in this formulation, for the linear units the diagonal entries in $\boldsymbol{W}$ are redundant to those in $\boldsymbol{A}$ and could be omitted, but we found in practice this hardly makes a difference.

$k\Delta t$ as it passes through different subregions U_e, it gives rise to a specific symbolic sequence $\boldsymbol{a}_{\boldsymbol{x},\phi}$ (with a unique symbol assigned at each time step via $h : S \to \mathcal{A}, \boldsymbol{x}_{k\Delta t} \mapsto a_k$). We may thus think of the shift operator σ as moving along a trajectory in correspondence with the flow map $\phi_{\Delta t}(\boldsymbol{x})$. If the symbolic coding of each trajectory is unique, $\mathcal{U}$ may constitute a *Markov partition* (see Appx. B for a formal definition). We denote by $(A_{S,\phi}, \sigma)$ the *shift of finite type* induced by the flow ϕ which picks out from the full shift space $\mathcal{A}^{\mathbb{Z}}$ only those *admissible* symbolic sequences that correspond to valid trajectories of (S, ϕ) (we will use the term 'induced by' to refer to this property). The set of admissible blocks constitutes the *language* of $(A_{S,\phi}, \sigma)$. Every shift of finite type has a graph representation $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ (Fig. 2), with either the edges $e_{ij} \in \mathcal{E}$ or vertices $v_i \in \mathcal{V}$ of the graph encoding the permitted transitions among symbols $a_k \in \mathcal{A}$ of admissible sequences $\boldsymbol{a} \in A$ [55].

The finite collection $\mathcal{U} = \{U_0 \dots U_{n-1}\},\ n = 2^P$, of *linear subregions* of an AL-RNN defined in eq. 4, separated by the switching manifolds $\Sigma_{i,j} = \overline{U_i} \cap \overline{U_j}$ between every pair of neighboring subregions U_i and U_j, forms a topological partition. Here we use this partition as the basis of our symbolic coding and the respective symbolic dynamics $(A_{\mathcal{U},F_{\boldsymbol{\theta}}}, \sigma)$ induced by the AL-RNN, where we assign to each state $\boldsymbol{z}_t \in S$ (i.e., at each time point) the unique symbol $a_t \in \mathcal{A}$ such that $a_t = a_i$ iff $\boldsymbol{z}_t \in U_i$ (Fig. 2).[3] In the corresponding symbolic graphs, we identify vertices v_i with symbols $a_i \in \mathcal{A}$ and draw a directed edge e_{ij} from v_i to v_j whenever $F_{\boldsymbol{\theta}}(U_i \cap B) \cap U_j \neq \varnothing$, where B is the attracting set of interest (see Appx. A.1 for details). As we will show further below, this particular partition has useful theoretical properties that makes the symbolic coding topologically interpretable w.r.t. the AL-RNN map $F_{\boldsymbol{\theta}}$. In fact, a large literature in symbolic dynamics has dealt with the relation between the dynamics in a finite shift space and that of a PWL map, like, e.g., the tent map, with a partition of S into the map's different linear subregions as we use here for the AL-RNN [55, 3, 68].

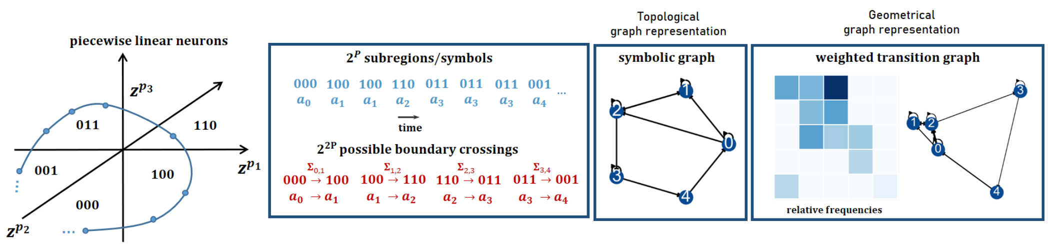

Figure 2: Illustration of symbolic approach (3 panels on the left) and geometrical graphs (right).

4 Theoretical Results

Recall that within each subregion U_e the map $F_{\boldsymbol{\theta}}$ is monotonic and the dynamics are linear (ruling out certain possibilities, like chaos or isolated cycles occurring within just one subregion). We furthermore assume that the dynamics are globally non-diverging (this could be strictly enforced through 'state-clipping' and constraints on matrix $\boldsymbol{A}$ in eq. 4, see Hess et al. [39], but will also be the case for a well-trained AL-RNN). Here we claim that for hyperbolic AL-RNNs $F_{\boldsymbol{\theta}}$[4], we have 1:1 relations between important topological objects in the AL-RNN's state space and those of the symbolic coding formed from the linear subregions U_e of the AL-RNN, as expressed in the following theoretical results.

Consider a hyperbolic, non-globally-diverging AL-RNN $F_{\boldsymbol{\theta}}$, eq. 4, and a topological partition $\mathcal{U}$ of the state space into its linear subregions $U_e \subseteq S, e = 0 \dots 2^P - 1$. Denote by $(A_{\mathcal{U},F_{\boldsymbol{\theta}}}, \sigma)$ the finite shift induced by $(S, F_{\boldsymbol{\theta}})$, with each $a_e \in \mathcal{A}$ of its alphabet $\mathcal{A}$ associated with exactly one linear subregion $U_e \in \mathcal{U}$, and let us consider the system's evolution only in forward time. Then the following holds:

Theorem 1. *An orbit $\Omega_S = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ of the AL-RNN $F_{\boldsymbol{\theta}}$ is asymptotically fixed (i.e., converges to a fixed point) if and only if the corresponding symbolic sequence* $\boldsymbol{a}$ =

[3]For simplicity we will ignore here and in the following the borders between subregions, on which the coding is ambiguous.

[4]By *hyperbolic AL-RNN* we mean the AL-RNN is hyperbolic in each of its linear subregions, i.e. none of its Jacobians $\boldsymbol{A} + \boldsymbol{W}\boldsymbol{D}_{\Omega^k}$ has eigenvalues of absolute magnitude 1. The chances that this condition is *not* met in practice in trained AL-RNNs, i.e. the non-hyperbolic case, are close to zero numerically.

$(a_1a_2a_3\dots a_{N-1})(a^*)^\infty \in A_{\mathcal{U},F_{\boldsymbol{\theta}}}$ *is an eventually fixed point of the shift map* σ *(where by 'eventually' we mean it exactly lands on the point in the limit, see Appx. B for a precise definition).*

Proof. See Appx. B. □

Theorem 2. *An orbit* $\Omega_S = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ *of the AL-RNN* F_θ *is asymptotically* p*-periodic if and only if the corresponding symbolic sequence*

$$\boldsymbol{a} \;=\; (a_1a_2\dots a_{N-1})(a_1^*a_2^*\dots a_p^*)^\infty \in A_{\mathcal{U},F_{\boldsymbol{\theta}}}$$

is an eventually p*-periodic orbit of the shift map* σ.

Proof. See Appx. B. □

Theorem 3. *An orbit* $\Omega_S = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ *is an asymptotically* ***aperiodic*** *(irregular) orbit of the AL-RNN* $F_{\boldsymbol{\theta}}$ *if and only if the corresponding symbolic sequence* $(a_1, \dots, a_n, \dots)$ *is aperiodic.*

Proof. See Appx. B. □

Loosely speaking, these results confirm that fixed points of our symbolic coding correspond to fixed points of the AL-RNN, cycles to cycles, and chaos to chaos, thus preserving important topological properties in the symbolic representation.

5 Experimental Results

To assess the quality of DSR, we employed established performance criteria based on long-term, invariant topological, geometrical, and temporal features of DS [51, 11, 39]. Due to exponential trajectory divergence in chaotic systems, mean-squared prediction errors rapidly grow even for well-trained systems, and hence are only of limited suitability for evaluating DSR quality [108, 67]. Thus, we prioritize the geometric agreement between true and reconstructed attractors, quantified by a Kullback-Leibler divergence (D_{stsp}, Appx. A.2) [51]. Additionally, we examine the long-term temporal agreement between true and reconstructed time series by evaluating the average dimension-wise Hellinger distance (D_{H}) between their power spectra (Appx. A.2). We first confirmed that the AL-RNN is at least on par with other SOTA methods for DSR. We then tested AL-RNNs on two commonly employed benchmark DS for which minimal PWL representations are known, the famous Lorenz-63 model of atmospheric convection [61] and the chaotic Rössler system [85]. We finally explored the suitability of our approach on two real-world examples, human electrocardiogram (ECG) and human functional magnetic resonance imaging (fMRI) data.

5.1 SOTA performance

While our goal here is a technique that constructs topologically minimal, interpretable DS representations, at the same time we do not want to compromise on DSR performance which should still be within the same ballpark as existing SOTA methods. We checked this on the Lorenz-63, Rössler and ECG data noted above, and in addition on the higher-dimensional chaotic Lorenz-96 system [62] and on human electroencephalogram (EEG) data. Table 1 in the Appx. confirms that the AL-RNN is not only on par with, but indeed outperforms most other techniques when trained with sparse teacher forcing (which may be rooted in its simple and parsimonious design).

5.2 Reconstructed Systems Occupy a Small Number of Subregions

Fig. 3 illustrates reconstruction performance for varying numbers of ReLU nonlinearities at constant network size M. We found that a small number of PWL units already significantly improves performance, especially for the Lorenz-63 and Rössler systems, and that beyond that number performance starts to plateau (or even briefly decrease again). Additionally, some linear units are necessary to sufficiently expand the space, but they cannot compensate for an insufficient number of PWL units (Fig. 10).[5] Moreover, as shown in Fig. 4 (left), the number of linear subregions explored by the

[5]Note that the number of linear units does not affect the interpretability or symbolic coding of the model.

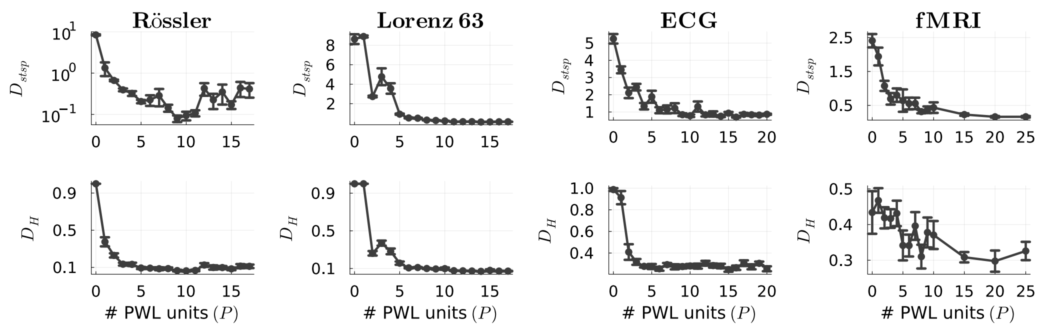

Figure 3: Quantification of DSR quality in terms of attractor geometry disagreement (D_{stsp}, top row) and disagreement in temporal structure (D_{H}, bottom row) as a function of the number of ReLUs (P) in the AL-RNN (Rössler: $M = 20$, Lorenz-63: $M = 20$, ECG: $M = 100$, fMRI: $M = 50$). The little humps at $P = 3$ for the Lorenz-63 indicate that performance may sometimes first degrade again when passing the number of minimally necessary PWL units (see also Fig. 9). Error bars = SEM.

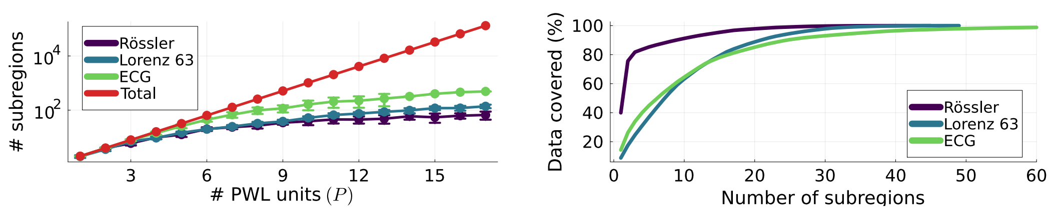

Figure 4: Left: Number of linear subregions traversed by trained AL-RNNs as a function of the number P of ReLUs. Theoretical limit (2^P) in red. Right: Cumulative number of data (trajectory) points covered by linear subregions in trained AL-RNNs (Rössler: $M = 20, P = 10$, Lorenz-63: $M = 20, P = 10$, ECG: $M = 100, P = 10$), illustrating that trajectories on an attractor live in a relatively small subset of subregions.

trained dynamics saturates well below the theoretical limit of 2^P once this performance threshold is reached. Within this already small subset of explored subregions, generated network activity is furthermore concentrated within an even smaller number of dominant subregions: For instance, for the Rössler system 4 out of 45 subregions used cover 80% of the data (Fig. 4, right). This substantial reduction of necessary linear subregions strongly facilitates the analysis of trained models with respect to fixed points and k-cycles, which naively would require examining 2^P and 2^{kP} combinations of subregions, respectively. To select the optimal number of PWL units, the point where performance starts plateauing (as in Fig. 3) may be chosen. Alternatively, one may restrict the number of linear subregions employed through regularization, adding a penalty for ReLU nonlinearities. This approach, as Fig. 9 illustrates, results in the same number of selected PWL units.

5.3 Minimal PWL Reconstructions of Chaotic attractors

Topologically minimal reconstructions Investigating reconstructions with the minimal number of PWL units needed for close-to-optimal performance (Fig. 3), we found that the AL-RNN would deliver reconstructions capturing the overall structure of the attractor using only three (Lorenz-63 system) or two (Rössler system) linear subregions (Fig. 5**a**), explaining the strong performance gains in Fig. 3 for 2 and 1 PWL units, respectively. These representations, and their symbolic coding (Fig. 5**c**), expose the mechanisms of chaotic dynamics (Fig. 5**b**). Notably, these closely agree with the minimal topologically equivalent PWL representations of the two chaotic DS as described in Amaral et al. [5]: The Lorenz-63 system has at its core two unstable spiral points in the two lobes, separated by the saddle node in the center (Fig. 5**b**). For the Rössler system, the topologically minimal PWL representation indeed consists of just two subregions [5], one containing an unstable spiral in the x-y plane and the other a 'half-spiral' almost orthogonal to that plane (Fig. 5**b**). The AL-RNN automatically and robustly discovers these representations from data: across multiple training runs, performance values are very similar (Figs. 23, 25), the assignment of subregions to different parts of the attractor remains almost the same (Figs. 24 & 25), and the regions with linear dynamics

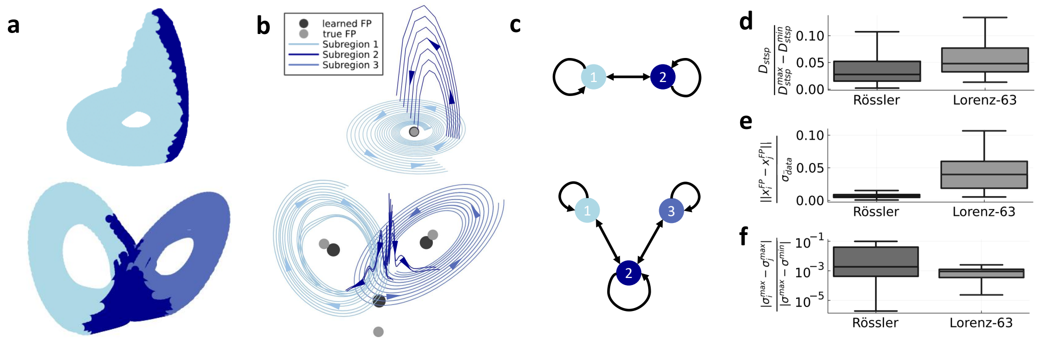

Figure 5: **a**: Color-coded linear subregions of minimal AL-RNNs representing the Rössler (top) and Lorenz-63 (bottom) chaotic attractor. **b**: Illustration of how the AL-RNN creates the chaotic dynamics. For the Rössler, trajectories diverge from an unstable spiral point (true position in gray, learned position in black) into the second subregion, where after about half a cycle they are propelled back into the first. For the Lorenz-63, two unstable spiral points (true: gray; learned: black) create the diverging spiraling dynamics in the two lobes, separated by the saddle node in the center. **c**: Topological graphs of the symbolic coding. While for the Rössler it is fully connected, for the Lorenz-63 the crucial role of the center saddle region in distributing trajectories onto the two lobes is apparent. **d**: Geometrical divergence (D_{stsp}) among repeated trainings of AL-RNNs ($n = 20$), separately evaluated within each subregion, shows close agreement among different training runs. Likewise, low **e**: normalized distances between fixed point locations and **f**: relative differences in maximum absolute eigenvalues $\sigma^{\max}$ across 20 trained models indicate that these topologically minimal representations are robustly identified.

closely agree both in terms of their topology and geometry (in fact, the topological graphs remained identical). This is in contrast to the standard PLRNN, where assignments strongly varied among multiple training runs (Figs. 23, 25). We quantified this further by computing across training runs separately for each subregion D_{stsp} (Fig. 5**d**), the normalized distances between fixed points (Fig. 5**e**), and the normalized differences between the maximum absolute eigenvalues $\sigma_{\max}$ of the AL-RNN's Jacobians (Fig. 5**f**), obtaining values close to zero in all three cases (see Fig. 22 for absolute values). While Amaral et al. [5] explicitly handcrafted such minimal PWL representations, the AL-RNN extracts them automatically without the provision of any prior knowledge about the system.

Geometrically minimal reconstructions While these reconstructions capture the topology of the underlying DS, they do not yet capture the full geometry and temporal structure of the attractor (Fig. 3). Fig. 6 illustrates for the Rössler system that as the number of PWL units is further increased to $P = 10$, the geometrical agreement becomes almost perfect. Although the mapping from latent to observation space is not 1:1 (since $M > N$), points close in observation space still tended to fall into the same latent subregion, such that the observed system's attractor still decomposed into distinct subregions, as confirmed by proximity matching (see Appx. A.1). For the Rössler system there is just one nonlinearity, the $x \cdot z$ term in the temporal derivative of z (eq. 14). Accordingly, the AL-RNN devotes most of its subregions to the lobe along the z coordinate, while dynamics in the (x, y) plane is geometrically faithfully represented by only 4 subregions. Hence, the AL-RNN utilizes additional subregions to express finer geometric details where dynamics are more nonlinear. This is apparent from a more geometrical graph representation (see Fig. 2, right), where – in addition to topological information – transition probabilities among subregions are being used to construct node distances via the graph Laplacian (see Appx. A.1), see Fig. 6 for the Rössler and Fig. 11 for the Lorenz-63.

5.4 PWL Reconstructions of Real-World Systems

Topologically minimal reconstructions We next considered two experimental datasets, human ECG data (with $1d$ membrane potential recordings delay-embedded into $N = 5$, see Appx. A.3) and fMRI recordings (with $N = 20$ time series extracted, cf. Appx. A.3) from human subjects performing three different types of cognitive task [50, 52].

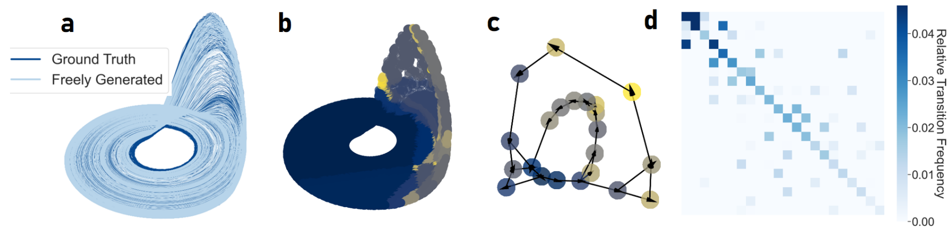

Figure 6: Geometrically minimal reconstruction and graph representation of the Rössler attractor ($M = 30, P = 10, D_{stsp} = 0.08, D_H = 0.06$). **a**: Provided a sufficient number of linear subregions, the geometry of the attractor is almost perfectly captured. **b**: Reconstruction with linear subregions color-coded by frequency of visits (dark: most frequently visited regions, yellow: least frequent regions). **c**: Corresponding geometrical graph, which contains information about transition frequencies via node distances, visualized using the spectral layout in `networkx`. Note that self-connections were omitted in this representation. **d**: Connectome of relative transition frequencies between subregions.

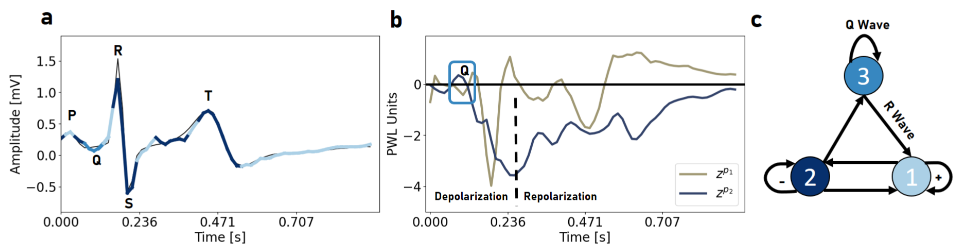

Figure 7: **a**: Freely generated ECG activity using an AL-RNN with 3 linear subregions (color-coded according to subregion) and ground truth time series in black. **b**: After activation of the Q wave in the third subregion, the second PWL unit is driven far below 0, whose activity, consistent with the known physiology [89], mimics the latent de- and re-polarization process of the interventricular septum. **c**: Symbolic graph representation of the trained AL-RNN.

As for the Lorenz-63, for the ECG data we observed a strong performance gain for just $P = 2$ PWL units, see Fig. 3. Indeed, Fig. 7**a** confirms that the complex activity pattern and positive max. Lyapunov exponent ($\lambda_{\text{max}} = 1.96\ s^{-1}$, ground truth: $\lambda_{\text{max}}^{\text{true}} = 2.19\ s^{-1}$ [39]) of the ECG time series could be achieved with $P = 2$ in only 3 linear subregions. These subregions corresponded to distinct parts of the ECG activity: ramping-up phases (light blue, node $\#1$), declining activity (dark blue, node $\#2$), represented by two unstable spirals with shifted phases (Fig. 12), and the Q wave (medium blue, node $\#3$). The activity in the third region ($\sigma_{max} \approx 1.34$, other two: $\sigma_{max} \approx 1.02$, Fig. 12) caused a strong inhibition in the second PWL unit (Fig. 7**b**) that captures the critical transition initiated by the Q wave. The Q wave triggers the depolarization of the interventricular septum during the QRS complex [89], indicating that this latent depolarization process is captured by the model. These results suggest that the AL-RNN cannot only learn dynamically but also biologically interpretable latent representations. The core aspects of this representation were furthermore consistent across successful reconstructions (Fig. 13). The symbolic sequences corresponding to the graph in Fig. 7 reveal the nearly periodic yet chaotic dynamics of the ECG (Fig. 14).[6]

For the short ($T = 360$) fMRI time series, $P = 3$ often resulted in reconstructions matching the complex activity patterns reasonably well (cf. Fig. 3, right). Fig. 15 illustrates results for an example subject using 8 linear subregions. The second most visited subregion implemented an unstable spiral, while the most visited region had a stable *virtual* fixed point also located in the second subregion. These two regions covered over 50% of the data and were strongly connected (Fig. 15**b**). This balance between stable and unstable activity suggests a mechanism through which the network implements chaotic dynamics, with the stable virtual fixed point pulling activity into the second subregion from which it then diverges again (Fig. 15**d**).

[6]In general, we also found that quantities computed from symbolic sequences like the topological entropy correlated highly with the maximum Lyapunov exponent, Fig. 16.

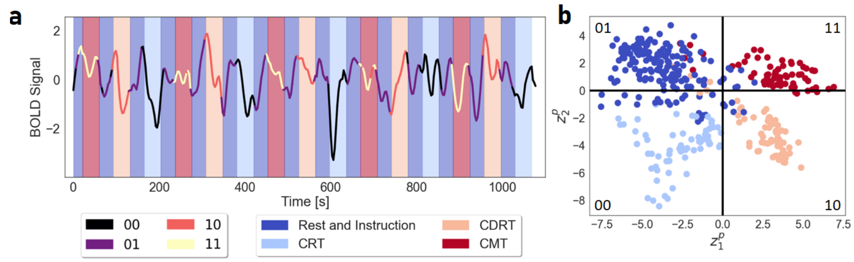

Figure 8: Reconstructions from human fMRI data using an AL-RNN with $M = 100$ total units and $P = 2$ PWL units. **a**: Mean generated BOLD activity color-coded according to the linear subregion. Background color shadings indicate the task stage. **b**: Generated activity (trajectory points) in the latent space of PWL units with color-coding indicating task stage as in **a**.

Task stages align with linear subregions To integrate the fMRI signal with cognitive (task-stage) information, in addition to the linear decoder for the BOLD signal, we coupled the PWL units $\boldsymbol{z}_t^P$ to a categorical decoder model (eq. 12) which predicts the three cognitive task stages and the 'Rest and Instruction' period, as in [52, 12]. Using only $P = 2$ PWL units on the fMRI data made it challenging to capture the complex, chaotic long-term activity patterns in the freely generated activity of the AL-RNN. However, the dynamics were still well-approximated locally (Fig. 17). To maintain temporal alignment with the task stages when sampling from the AL-RNN, the AL-RNN's readout units were reset to the observations every 7 time steps (Fig. 8**a**). Fig. 8**a**-**b** show that in this setup, the four linear subregions of the AL-RNN often closely aligned with the different task stages of the experimental time series. Similar results were obtained across different subjects, with an average classification accuracy of $p = 0.78 \pm 0.05$ (mean $\pm$ SEM) (see Appx. A.4 for details). While the categorical decoder aids in separating latent states according to task stage, there is nothing that would bias this separation to align with the linear subregions. The observed alignment therefore suggests that the AL-RNN learns to leverage distinct linear dynamics in each subregion to represent differences across cognitive tasks. Furthermore, as shown in Fig. 18, the network weights of the PWL units were significantly larger than those of other units, indicating their critical role in modulating the dynamics and representing task-related variations in brain activity. This approach also demonstrates how local context-aligned linear approximations can be achieved using the AL-RNN, which is useful in areas such as model-predictive control [72, 23, 58].

6 Conclusion

Here we introduced a novel variant of a PLRNN, the AL-RNN, which learns to represent nonlinear DS with as few PWL nonlinearities as possible. Despite its simple design and the minimal hyperparameter tuning required, the AL-RNN robustly and automatically identifies highly interpretable, topologically minimal representations of complex nonlinear DS, reproducing known minimal PWL forms of chaotic attractors [5]. Such minimal PWL forms that allow for an interpretable symbolic and graph-theoretical representation were discovered even from challenging physiological and neuroscientific data. They also profoundly ease subsequent model analysis. For instance, with only a few linear subregions to consider, the search for fixed points or cycles becomes very fast and efficient [26].

Limitations While this seems promising, how to determine whether a topologically minimal and valid reconstruction from empirical data has truly been achieved remains an open topic. Performance curves as in Fig. 3 or Fig. 9 give an indication of how many PWL units may be required to yield an optimal minimal representation, but whether there is a more principled way of automatically inferring the optimal number P of PWL nonlinearities from data may be an interesting future direction. Finally, the current finding that even for empirical ECG and fMRI data a few linear subregions (≤ 8) suffice for faithful reconstructions is encouraging. Whether this more generally will be the case in empirical scenarios is another interesting and open question. Not all types of (empirically observed) dynamical systems may easily allow for such topologically minimal representations.

All code created is available at `https://github.com/DurstewitzLab/ALRNN-DSR`.

Acknowledgements

This work was funded by the Federal Ministry of Science, Education, and Culture (MWK) of the state of Baden-Württemberg within the AI Health Innovation Cluster, by the German Research Foundation (DFG) within Germany's Excellence Strategy EXC 2181/1 – 390900948 (STRUCTURES), and through DFG individual grant Du 354/15-1 to DD. ZM was funded by the Federal Ministry of Education and Research (BMBF) through project OIDLITDSM, 01IS24061.

References

[1] G. Ackerson and K. Fu. On state estimation in switching environments. *IEEE Transactions on Automatic Control*, 15(1):10–17, February 1970. ISSN 1558-2523. doi: 10.1109/TAC.1970.1099359. URL https://ieeexplore.ieee.org/document/1099359. Conference Name: IEEE Transactions on Automatic Control.

[2] Xavier Alameda-Pineda, Vincent Drouard, and Radu Patrice Horaud. Variational Inference and Learning of Piecewise Linear Dynamical Systems. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):3753–3764, August 2022. ISSN 2162-2388. doi: 10.1109/TNNLS.2021.3054407. URL https://ieeexplore.ieee.org/document/9353398. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[3] Kathleen T. Alligood, Tim D. Sauer, and James A. Yorke. *Chaos: An Introduction to Dynamical Systems*. Textbooks in Mathematical Sciences. Springer, 1996. ISBN 978-0-387-94677-1 978-0-387-22492-3. doi: 10.1007/b97589.

[4] Victor M. Martinez Alvarez, Rareş Roşca, and Cristian G. Fălcuţescu. DyNODE: Neural Ordinary Differential Equations for Dynamics Modeling in Continuous Control, September 2020.

[5] Gleison F. V. Amaral, Christophe Letellier, and Luis Antonio Aguirre. Piecewise affine models of chaotic attractors: the Rossler and Lorenz systems. *Chaos (Woodbury, N.Y.)*, 16(1):013115, March 2006. ISSN 1054-1500. doi: 10.1063/1.2149527.

[6] Omri Azencot, N. Benjamin Erichson, Vanessa Lin, and Michael W. Mahoney. Forecasting Sequential Data using Consistent Koopman Autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL http://arxiv.org/abs/2003.02236.

[7] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/hash/f106b7f99d2cb30c3db1c3cc0fde9ccb-Abstract.html.

[8] Alberto Bemporad, Francesco Borrelli, and Manfred Morari. Piecewise linear optimal controllers for hybrid systems. In *Proceedings of the 2000 American Control Conference. ACC (IEEE Cat. No. 00CH36334)*, volume 2, pages 1190–1194. IEEE, 2000.

[9] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994. ISSN 1045-9227. doi: 10.1109/72.279181.

[10] Mario Di Bernardo, Mark I. Feigin, Stephen J. Hogan, and Martin E. Homer. Local Analysis of C-Bifurcations in n-Dimensional Piecewise-Smooth Dynamical Systems. *Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena*, 11(10):1881–1908, 1999. ISSN 0960-0779. URL https://www.infona.pl//resource/bwmeta1.element.elsevier-b61cfd87-9650-310f-bd8f-4bd7e7174946.

[11] Manuel Brenner, Florian Hess, Jonas M. Mikhaeil, Leonard F. Bereska, Zahra Monfared, Po-Chen Kuo, and Daniel Durstewitz. Tractable Dendritic RNNs for Reconstructing Nonlinear Dynamical Systems. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2292–2320. PMLR, June 2022. URL https://proceedings.mlr.press/v162/brenner22a.html. ISSN: 2640-3498.

[12] Manuel Brenner, Florian Hess, Georgia Koppe, and Daniel Durstewitz. Integrating Multimodal Data for Joint Generative Modeling of Complex Dynamics. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4482–4516. PMLR, July 2024. URL https://proceedings.mlr.press/v235/brenner24a.html. ISSN: 2640-3498.

[13] Steven L Brunton and J Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2019.

[14] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences USA*, 113(15):3932–3937, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1517384113.

[15] Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman Theory for Dynamical Systems, October 2021. arXiv:2102.12086 [cs, eess, math].

[16] Gyorgy Buzsaki. *Rhythms of the Brain*. Oxford University Press, August 2006. ISBN 978-0-19-804125-2. Google-Books-ID: ldz58irprjYC.

[17] Victoriano Carmona, Emilio Freire, Enrique Ponce, and Francisco Torres. On simplifying and classifying piecewise-linear systems. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 49(5):609–620, 2002.

[18] Martin Casdagli. Nonlinear prediction of chaotic time series. *Physica D: Nonlinear Phenomena*, 35(3):335–356, May 1989. ISSN 0167-2789. doi: 10.1016/0167-2789(89)90074-2.

[19] Rok Cestnik and Markus Abel. Inferring the dynamics of oscillatory systems using recurrent neural networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(6):063128, June 2019. ISSN 1054-1500. doi: 10.1063/1.5096918. URL https://doi.org/10.1063/1.5096918.

[20] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems 31*, 2018. URL http://arxiv.org/abs/1806.07366.

[21] Alexandre Cortiella, Kwang-Chun Park, and Alireza Doostan. 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.

[22] Antonio C. Costa, Tosif Ahamed, and Greg J. Stephens. Adaptive, locally linear models of complex dynamics. *Proceedings of the National Academy of Sciences*, 116(5):1501–1510, January 2019. doi: 10.1073/pnas.1813476116. Publisher: Proceedings of the National Academy of Sciences.

[23] J. Daafouz, P. Riedinger, and C. Iung. Stability analysis and control synthesis for switched systems: a switched Lyapunov function approach. *IEEE Transactions on Automatic Control*, 47(11):1883–1887, November 2002. ISSN 1558-2523. doi: 10.1109/TAC.2002.804474. URL https://ieeexplore.ieee.org/document/1047016. Conference Name: IEEE Transactions on Automatic Control.

[24] Oscar De Feo and Marco Storace. Piecewise-Linear Identification of Nonlinear Dynamical Systems in View of Their Circuit Implementations. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 54(7):1542–1554, July 2007. ISSN 1558-0806. doi: 10.1109/TCSI.2007.899613. URL https://ieeexplore.ieee.org/document/4268404. Conference Name: IEEE Transactions on Circuits and Systems I: Regular Papers.

[25] Daniel Durstewitz. A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements. *PLoS Comput. Biol.*, 13(6): e1005542, 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005542.

[26] Lukas Eisenmann, Zahra Monfared, Niclas Göring, and Daniel Durstewitz. Bifurcations and loss jumps in RNN training. *Advances in Neural Information Processing Systems*, 36, 2024.

[27] Mark I Feigin. The increasingly complex structure of the bifurcation tree of a piecewise-smooth system. *Journal of Applied Mathematics and Mechanics*, 59(6):853–863, 1995. ISSN 0021-8928. doi: 10.1016/0021-8928(95)00118-2. URL https://www.sciencedirect.com/science/article/pii/0021892895001182.

[28] Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. Nonparametric Bayesian Learning of Switching Linear Dynamical Systems. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://papers.nips.cc/paper_files/paper/2008/hash/950a4152c2b4aa3ad78bdd6b366cc179-Abstract.html.

[29] Ken-ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 1989. doi: 10.1016/0893-6080(89)90003-8.

[30] Nicholas Geneva and Nicholas Zabaras. Transformers for modeling physical systems. *Neural Networks*, 146:272–289, February 2022. ISSN 0893-6080. doi: 10.1016/j.neunet.2021.11.022.

[31] Z. Ghahramani and G. E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):831–864, April 2000. ISSN 0899-7667. doi: 10.1162/089976600300015619.

[32] John Guckenheimer and Philip Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, volume 42 of *Applied Mathematical Sciences*. Springer, New York, NY, 1983. ISBN 978-1-4612-7020-1 978-1-4612-1140-2. doi: 10.1007/978-1-4612-1140-2. URL http://link.springer.com/10.1007/978-1-4612-1140-2.

[33] Richard Hahnloser and H. Sebastian Seung. Permitted and Forbidden Sets in Symmetric Threshold-Linear Networks. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[34] Joshua Hanson and Maxim Raginsky. Universal simulation of stable dynamical systems by recurrent neural nets. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 384–392. PMLR, 10–11 Jun 2020. URL https://proceedings.mlr.press/v120/hanson20a.html.

[35] Bailin Hao and Weimou Zheng. *Applied symbolic dynamics and chaos*. World Scientific, 1998.

[36] Philip Hartman. A lemma in the theory of structural stability of differential equations. *Proceedings of the American Mathematical Society*, 11(4):610–620, 1960. ISSN 0002-9939, 1088-6826. doi: 10.1090/S0002-9939-1960-0121542-7. URL https://www.ams.org/proc/1960-011-04/S0002-9939-1960-0121542-7/.

[37] Niklas Heim, Václav Šmídl, and Tomáš Pevný. Rodent: Relevance determination in differential equations. *arXiv preprint arXiv:1912.00656*, 2019. URL http://arxiv.org/abs/1912.00656.

[38] Christoph Jürgen Hemmer, Manuel Brenner, Florian Hess, and Daniel Durstewitz. Optimal Recurrent Network Topologies for Dynamical Systems Reconstruction. In *Proceedings of the 41st International Conference on Machine Learning*, pages 18174–18204. PMLR, July 2024. URL https://proceedings.mlr.press/v235/hemmer24a.html. ISSN: 2640-3498.

[39] Florian Hess, Zahra Monfared, Manuel Brenner, and Daniel Durstewitz. Generalized Teacher Forcing for Learning Chaotic Dynamics. In *Proceedings of the 40th International Conference on Machine Learning*, pages 13017–13049. PMLR, July 2023. URL https://proceedings.mlr.press/v202/hess23a.html. ISSN: 2640-3498.

[40] Stephen J. Hogan, L. Higham, and T. C. L. Griffin. Dynamics of a piecewise linear map with a gap. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2077):49–65, 2007. doi: 10.1098/rspa.2006.1735. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2006.1735.

[41] Anthony R. Ives and Vasilis Dakos. Detecting dynamical changes in nonlinear time series using locally linear state-space models. *Ecosphere*, 3(6):art58, 2012. ISSN 2150-8925. doi: 10.1890/ES11-00347.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/ES11-00347.1.

[42] Parag Jain and Soumitro Banerjee. Border-collision bifurcations in one-dimensional discontinuous maps. *Int. J. Bifurcation Chaos*, 13(11):3341–3351, 2003. ISSN 0218-1274. doi: 10.1142/S0218127403008533. URL https://www.worldscientific.com/doi/abs/10.1142/S0218127403008533.

[43] A.L. Juloski, S. Weiland, and W.P.M.H. Heemels. A Bayesian approach to identification of hybrid systems. *IEEE Transactions on Automatic Control*, 50(10):1520–1533, October 2005. ISSN 1558-2523. doi: 10.1109/TAC.2005.856649. URL https://ieeexplore.ieee.org/document/1516255. Conference Name: IEEE Transactions on Automatic Control.

[44] Kadierdan Kaheman, Steven L. Brunton, and J. Nathan Kutz. 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. Publisher: IOP Publishing.

[45] E. Kaiser, J. N. Kutz, and S. L. Brunton. 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. Publisher: Royal Society.

[46] Daniel Karlsson and Olle Svanström. Modelling Dynamical Systems Using Neural Ordinary Differential Equations, 2019. URL https://hdl.handle.net/20.500.12380/256887.

[47] Masahiro Kimura and Ryohei Nakano. Learning dynamical systems by recurrent neural networks from orbits. *Neural Networks*, 11(9):1589–1599, 1998.

[48] Joon-Hyuk Ko, Hankyul Koh, Nojun Park, and Wonho Jhe. Homotopy-based training of NeuralODEs for accurate dynamics discovery, May 2023. URL http://arxiv.org/abs/2210.01407. arXiv:2210.01407 [physics].

[49] B. O. Koopman and J. v. Neumann. Dynamical Systems of Continuous Spectra. *Proceedings of the National Academy of Sciences*, 18(3):255–263, March 1932. doi: 10.1073/pnas.18.3.255. Publisher: Proceedings of the National Academy of Sciences.

[50] Georgia Koppe, Harald Gruppe, Gebhard Sammer, Bernd Gallhofer, Peter Kirsch, and Stefanie Lis. Temporal unpredictability of a stimulus sequence affects brain activation differently depending on cognitive task demands. *NeuroImage*, 101:236–244, 2014. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2014.07.008.

[51] Georgia Koppe, Hazem Toutounji, Peter Kirsch, Stefanie Lis, and Daniel Durstewitz. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. *PLOS Computational Biology*, 15(8):e1007263, 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007263.

[52] Daniel Kramer, Philine L Bommer, Carlo Tombolini, Georgia Koppe, and Daniel Durstewitz. Reconstructing nonlinear dynamical systems from multi-modal time series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11613–11633. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kramer22a.html.

[53] William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H. Moore. Contemporary Symbolic Regression Methods and their Relative Performance, July 2021. URL http://arxiv.org/abs/2107.14351. arXiv:2107.14351 [cs].

[54] Douglas Lind and Brian Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, Cambridge, 1995. doi: 10.1017/CBO9780511626302. URL https://www.cambridge.org/

core/books/an-introduction-to-symbolic-dynamics-and-coding/331DF5E9DC464B340DED80431BD6D186.

[55] Douglas Lind and Brian Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge Mathematical Library. Cambridge University Press, 2 edition, 2021.

[56] Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 914–922. PMLR, April 2017. URL https://proceedings.mlr.press/v54/linderman17a.html. ISSN: 2640-3498.

[57] Scott W. Linderman and Matthew J. Johnson. Structure-Exploiting variational inference for recurrent switching linear dynamical systems. In *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5, December 2017. doi: 10.1109/CAMSAP.2017.8313132. URL https://ieeexplore.ieee.org/document/8313132.

[58] Scott W. Linderman, Andrew C. Miller, Ryan P. Adams, David M. Blei, Liam Paninski, and Matthew J. Johnson. Recurrent switching linear dynamical systems. *arXiv:1610.08466 [stat]*, October 2016. URL http://arxiv.org/abs/1610.08466. arXiv: 1610.08466.

[59] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgz2aEKDr.

[60] Jean-Christophe Loiseau and Steven L. Brunton. 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.

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

[62] Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

[63] Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nat Commun*, 9(1):4950, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07210-0. URL http://arxiv.org/abs/1712.09707. arXiv: 1712.09707.

[64] Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review*, 57(1):2, January 2024. ISSN 1573-7462. doi: 10.1007/s10462-023-10622-0. URL https://doi.org/10.1007/s10462-023-10622-0.

[65] Benoit Mandelbrot and Richard L. Hudson. *The Misbehavior of Markets: A Fractal View of Financial Turbulence*. Basic Books, March 2007. ISBN 978-0-465-00468-3. Google-Books-ID: GMKeUqufPQ0C.

[66] Daniel A. Messenger and David M. Bortz. 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.

[67] Jonas Mikhaeil, Zahra Monfared, and Daniel Durstewitz. On the difficulty of learning chaotic dynamics with RNNs. *Advances in Neural Information Processing Systems*, 35:11297–11312, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/495e55f361708bedbab5d81f92048dcd-Abstract-Conference.html.

[68] John Milnor. 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.

[69] K. Mischaikow, M. Mrozek, J. Reiss, and A. Szymczak. Construction of Symbolic Dynamics from Experimental Time Series. *Physical Review Letters*, 82(6):1144–1147, February 1999. doi: 10.1103/PhysRevLett.82.1144. URL https://link.aps.org/doi/10.1103/PhysRevLett.82.1144. Publisher: American Physical Society.

[70] Zahra Monfared and Daniel Durstewitz. Existence of n-cycles and border-collision bifurcations in piecewise-linear continuous maps with applications to recurrent neural networks. *Nonlinear Dyn*, 101(2):1037–1052, 2020. ISSN 1573-269X. doi: 10.1007/s11071-020-05841-x. URL https://doi.org/10.1007/s11071-020-05841-x.

[71] Peter J. Mumby, Alan Hastings, and Helen J. Edwards. 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.

[72] Kenneth R. Muske and James B. Rawlings. Model predictive control with linear models. *AIChE Journal*, 39(2):262–287, 1993. ISSN 1547-5905. doi: 10.1002/aic.690390208. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.690390208. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690390208.

[73] Ilan Naiman and Omri Azencot. A Koopman Approach to Understanding Sequence Neural Models. *arXiv:2102.07824 [cs, math]*, October 2021. URL http://arxiv.org/abs/2102.07824. arXiv: 2102.07824.

[74] George Osipenko and Stephen Campbell. Applied symbolic dynamics: attractors and filtrations. *Discrete and Continuous Dynamical Systems*, 5(1):43–60, September 1998. ISSN 1078-0947. doi: 10.3934/dcds.1999.5.43. URL https://www.aimsciences.org/en/article/doi/10.3934/dcds.1999.5.43. Publisher: Discrete and Continuous Dynamical Systems.

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

[76] Jaideep Pathak, Zhixin Lu, Brian R. Hunt, Michelle Girvan, and Edward Ott. 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.

[77] Mahashweta Patra. Multiple Attractor Bifurcation in Three-Dimensional Piecewise Linear Maps. *Int. J. Bifurcation Chaos*, 28(10):1830032, 2018. ISSN 0218-1274. doi: 10.1142/S021812741830032X. URL https://www.worldscientific.com/doi/abs/10.1142/S021812741830032X.

[78] Lawrence Perko. *Differential equations and dynamical systems*. Number 7 in Texts in applied mathematics. Springer, New York, 3rd ed edition, 2001. ISBN 978-0-387-95116-4.

[79] Jason A. Platt, Stephen G. Penny, Timothy A. Smith, Tse-Chun Chen, and Henry D. I. Abarbanel. A Systematic Exploration of Reservoir Computing for Forecasting Complex Spatiotemporal Dynamics, January 2022. URL http://arxiv.org/abs/2201.08910. arXiv:2201.08910 [cs].

[80] Jason A Platt, Stephen G Penny, Timothy A Smith, Tse-Chun Chen, and Henry DI Abarbanel. Constraining chaos: Enforcing dynamical invariants in the training of reservoir computers. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(10), 2023.

[81] Anders Rantzer and Mikael Johansson. Piecewise linear quadratic optimal control. *IEEE transactions on automatic control*, 45(4):629–637, 2000.

[82] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

[83] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://www.nature.com/articles/s42256-019-0048-x. Publisher: Nature Publishing Group.

[84] T Konstantin Rusch, Siddhartha Mishra, N Benjamin Erichson, and Michael W Mahoney. Long expressive memory for sequence modeling. In *International Conference on Learning Representations*, 2022.

[85] O. E. Rössler. An equation for continuous chaos. *Physics Letters A*, 57(5):397–398, 1976. ISSN 0375-9601. doi: 10.1016/0375-9601(76)90101-8. URL https://www.sciencedirect.com/science/article/pii/0375960176901018.

[86] Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. *Journal of statistical Physics*, 65(3):579–616, 1991.

[87] Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE transactions on bio-medical engineering*, 51(6):1034–1043, 2000. ISSN 0018-9294. doi: 10.1109/TBME.2004.827072.

[88] Dominik Schmidt, Georgia Koppe, Zahra Monfared, Max Beutelspacher, and Daniel Durstewitz. Identifying nonlinear dynamical systems with multiple time scales and long-range dependencies. In *Proceedings of the 9th International Conference on Learning Representations*, 2021. URL http://arxiv.org/abs/1910.03471.

[89] Demetrio Sodi-Pallares, María Isabel Rodriguez, Leonardo O. Chait, and Rudolf Zuckermann. The activation of the interventricular septum. *American Heart Journal*, 41(4):569–608, April 1951. ISSN 0002-8703. doi: 10.1016/0002-8703(51)90024-5. URL https://www.sciencedirect.com/science/article/pii/0002870351900245.

[90] Justin Solomon. PDE Approaches to Graph Analysis, April 2015. URL http://arxiv.org/abs/1505.00185. arXiv:1505.00185 [cs, math].

[91] E. Sontag. Nonlinear regulation: The piecewise linear approach. *IEEE Transactions on Automatic Control*, 26(2):346–358, April 1981. ISSN 1558-2523. doi: 10.1109/TAC.1981.1102596. URL https://ieeexplore.ieee.org/document/1102596. Conference Name: IEEE Transactions on Automatic Control.

[92] Ioan Stanculescu, Christopher K. I. Williams, and Yvonne Freer. A Hierarchical Switching Linear Dynamical System Applied to the Detection of Sepsis in Neonatal Condition Monitoring. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, 2014. URL https://www.research.ed.ac.uk/en/publications/a-hierarchical-switching-linear-dynamical-system-applied-to-the-d.

[93] M. Storace and O. De Feo. Piecewise-linear approximation of nonlinear dynamical systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 51(4):830–842, April 2004. ISSN 1558-0806. doi: 10.1109/TCSI.2004.823664. Conference Name: IEEE Transactions on Circuits and Systems I: Regular Papers.

[94] Steven H. Strogatz. *Nonlinear Dynamics and Chaos*. CRC Press, 1 edition, 2015. ISBN 978-0-429-96111-3. doi: 10.1201/9780429492563. URL https://www.taylorfrancis.com/books/9780429961113.

[95] Zhendong Sun. *Switched Linear Systems: Control and Design*. Springer Science & Business Media, March 2006. ISBN 978-1-84628-131-0. Google-Books-ID: u4GArZN1bmsC.

[96] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, volume 898, pages 366–381. Springer, 1981. ISBN 978-3-540-11171-9 978-3-540-38945-3. URL http://link.springer.com/10.1007/BFb0091924.

[97] Sachin S. Talathi and Aniket Vartak. Improving performance of recurrent neural network with relu nonlinearity. In *Proceedings of the 4th International Conference on Learning Representations*, 2016. URL http://arxiv.org/abs/1511.03771.

[98] M. Tomas-Rodriguez and S. P. Banks. Linear approximations to nonlinear dynamical systems with applications to stability and spectral theory. *IMA Journal of Mathematical Control and Information*, 20(1):89–103, March 2003. ISSN 0265-0754. doi: 10.1093/imamci/20.1.89. URL https://doi.org/10.1093/imamci/20.1.89.

[99] Adam P. Trischler and Gabriele M.T. D'Eleuterio. Synthesis of recurrent neural networks for dynamical system simulation. *Neural Networks*, 80:67–78, 2016. ISSN 08936080. doi: 10.1016/j.neunet.2016.04.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608016300314.

[100] Eli Tziperman, Harvey Scher, Stephen E. Zebiak, and Mark A. Cane. Controlling spatiotemporal chaos in a realistic el niño prediction model. *Phys. Rev. Lett.*, 79:1034–1037, Aug 1997. doi: 10.1103/PhysRevLett.79.1034. URL https://link.aps.org/doi/10.1103/PhysRevLett.79.1034.

[101] Pantelis R. Vlachas, Wonmin Byeon, Zhong Y. Wan, Themistoklis P. Sapsis, and Petros Koumoutsakos. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proc. R. Soc. A.*, 474(2213):20170844, 2018. ISSN 1364-5021, 1471-2946. doi: 10.1098/rspa.2017.0844. URL https://royalsocietypublishing.org/doi/10.1098/rspa.2017.0844.

[102] Pantelis R. Vlachas, Jaideep Pathak, Brian R. Hunt, Themistoklis P. Sapsis, Michelle Girvan, Edward Ott, and Petros Koumoutsakos. Backpropagation Algorithms and Reservoir Computing in Recurrent Neural Networks for the Forecasting of Complex Spatiotemporal Dynamics. *arXiv:1910.05266 [physics]*, February 2020. URL http://arxiv.org/abs/1910.05266. arXiv: 1910.05266.

[103] Ryan Vogt, Maximilian Puelma Touzel, Eli Shlizerman, and Guillaume Lajoie. 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.

[104] Rui Wang, Yihe Dong, Sercan Ö Arik, and Rose Yu. Koopman Neural Forecaster for Time Series with Temporal Distribution Shifts, October 2022. URL http://arxiv.org/abs/2210.03675. arXiv:2210.03675 [cs, stat].

[105] H. Wersing, W. J. Beyn, and H. Ritter. Dynamical stability conditions for recurrent neural networks with unsaturating piecewise linear transfer functions. *Neural Computation*, 13(8): 1811–1825, August 2001. ISSN 0899-7667. doi: 10.1162/08997660152469350.

[106] Stephen Wiggins. *Global Bifurcations and Chaos*, volume 73 of *Applied Mathematical Sciences*. Springer, New York, NY, 1988. ISBN 978-1-4612-1041-2 978-1-4612-1042-9. doi: 10.1007/978-1-4612-1042-9. URL http://link.springer.com/10.1007/978-1-4612-1042-9.

[107] Alan Wolf, Jack B. Swift, Harry L. Swinney, and John A. Vastano. Determining lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317, 1985. ISSN 0167-2789. doi: 10.1016/0167-2789(85)90011-9. URL https://www.sciencedirect.com/science/article/pii/0167278985900119.

[108] Simon N. Wood. 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.

[109] Zhang Yi, K. K. Tan, and T. H. Lee. Multistability Analysis for Recurrent Neural Networks with Unsaturating Piecewise Linear Transfer Functions. *Neural Computation*, 15(3):639–662, March 2003. ISSN 0899-7667. doi: 10.1162/089976603321192112. URL https://doi.org/10.1162/089976603321192112.

A Appendix

A.1 Graph Representation

The symbolic graph construction followed the rules given in Sect. 3.2: Most generally, each linear subregion U_i is assigned a node (symbol), and a directed edge is drawn between nodes $i, j, i \to j$, whenever $F_{\boldsymbol{\theta}}(U_i) \cap U_j \neq \varnothing$. Here, however, we are mostly interested in the topological graphs representing particular chaotic attractors (like the Lorenz-63 or Rössler attractor), and hence restrict the graph representation to the nodes corresponding to subregions visited by trajectories on the attractor B, and edges drawn when $F_{\boldsymbol{\theta}}(U_i \cap B) \cap U_j \neq \varnothing$. More specifically, we first sampled a long trajectory $\boldsymbol{Z} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_T\}$ with $T = 100,000$ time steps, removed the first 1000 time steps as transients, and counted all transitions between any two subregions U_i, U_j. To obtain a more nuanced geometrical/ statistical picture, we also evaluated the relative number of time steps the trajectory spent in each subregion U_i (i.e., an estimate of the occupation measure), $|\{\boldsymbol{z}_t \in U_i, t > 1000\}|/(100,000 - 1000)$, as well as the relative frequency of transitions between any two subregions U_i, U_j, $|\{\boldsymbol{z}_t, \boldsymbol{z}_{t+1}| t > 1000, \boldsymbol{z}_t \in U_i, F_{\boldsymbol{\theta}}(\boldsymbol{z}_t) \in U_j\}|/(100,000 - 1001)$, yielding a weight for each edge, or a distance between nodes through the graph's Laplacian (see below). To enhance the readability of the larger graphs in Figs. 6, 11, 15 and 21, we removed self-connections (time steps where the trajectory remained within a single subregion).

Laplacian matrix The Laplacian matrix of a graph is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ where $\boldsymbol{A}$ is the adjacency matrix of the graph containing the weights (transition probabilities), and $\boldsymbol{D}$ is the out-degree matrix, which is a diagonal matrix where each element is the sum of the outgoing edge weights of node i, $D_{ii} = \sum_j A_{ij}$. The spectral layout in `networkx` uses the eigenvectors of the Laplacian matrix corresponding to the smallest non-zero eigenvalues as positions for the nodes. This tends to group more tightly connected nodes closer together. The Laplacian is more widely used as a dimensionality reduction technique in ML, for example in Laplacian eigenmaps [7], and has also been used to represent discretizations of PDEs as graphs [90].

Proximity matching The mapping from latent space to observation space is not unique because $M > N$, and hence the nullspace of the linear observation model is non-empty (does not contain just the $\mathbf{0}$ vector). For the AL-RNN this is evident from the fact that the non-readout neurons, particularly the PWL neurons, do not contribute to the observations. However, in practice, for the freely generated activity of trained AL-RNNs, points that fall into the same linear subregion in latent space also were close in observation space. This implies that attractors are segmented into different subregions in latent space in accordance with the observable dynamics. To numerically confirm this, we found that proximal points in observation space were typically related to the same linear subregion when generating activity from trained AL-RNNs: We conducted proximity matching by defining a threshold distance (e.g. $d = 0.05$, corresponding to 5% of the data variance) and assessing whether generated latent trajectory points proximal in observation space fall into the same or different subregions. We found that for the geometrically minimal reconstruction of the Rössler system (Fig. 6), only 6% of proximal data points (within d) belonged to different subregions, while for the Lorenz-63 attractor (Fig. 11) 4% of proximal data points (within d) belonged to different subregions, confirming that the attractors were segmented into relatively distinct patches.

A.2 Methodological Details

Training method Our training method rests on a variant of sparse teacher forcing. This approach has recently been proven to effectively tackle gradient divergence when training on observations from chaotic DS [67] and has shown SOTA performance on benchmark and real-world systems in DSR [11, 39]. In sparse teacher forcing, the idea is to directly replace latent states (or a subset of them) with states inferred from observations at intervals τ while leaving the network to evolve freely otherwise. To obtain forced states, the observation model needs to be 'pseudo-inverted'. Here we employ a specific variant of sparse teacher forcing called identity teacher forcing [67, 11], where this pseudo-inversion becomes trivial by adopting an identity mapping as the observation model:

$$\hat{\boldsymbol{x}}_t = \mathcal{I}\boldsymbol{z}_t, \tag{7}$$

with $\mathcal{I} \in \mathbb{R}^{N \times M}$, and $\mathcal{I}_{rr} = 1$ for the N read-out neurons, $r \leq N$, and zeroes elsewhere. During training, the read-out states are replaced with observations every τ time steps:

$$\boldsymbol{z}_{t+1} = \begin{cases} F_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_t) & \text{if } t \in \mathcal{T} \\ F_{\boldsymbol{\theta}}(\boldsymbol{z}_t) & \text{else} \end{cases} \tag{8}$$

with $\mathcal{T} = \{l\tau + 1\}_{l \in \mathbb{N}_0}$, and $\tilde{\boldsymbol{z}}_t = (\boldsymbol{x}_t, \boldsymbol{z}_{N+1:M,t})^T$. Employing identity teacher forcing splits the AL-RNN into essentially three types of units, the N linear readout-neurons $\boldsymbol{z}_t^r$ which are equivalent to the predicted observations and teacher-forced during training, the remaining $M - P - N$ linear neurons $\boldsymbol{z}_t^l$, and the P nonlinear neurons $\boldsymbol{z}_t^p$. The overall model and architecture are illustrated in Fig. 1. The AL-RNN can be trained using Mean Squared Error (MSE) loss over model predictions and observations:

$$\ell_{MSE}(\hat{\boldsymbol{X}}, \boldsymbol{X}) = \frac{1}{N \cdot T} \sum_{t=1}^{T} \|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t\|_2^2, \tag{9}$$

where $\hat{\boldsymbol{X}}$ are the model predictions and $\boldsymbol{X}$ denotes the training sequence of length T. We found that performance was better when the read-out neurons were linear rather than ReLU-based. Note that the $M - N$ non-readout neurons, including the PWL neurons, do not explicitly contribute to the loss function. We used rectified adaptive moment estimation (RADAM) [59] as the optimizer, with $L = 50$ batches with $S = 16$ sequences per epoch. Further, we chose $M = \{20, 20, 100, 100, 100, 130\}$, $\tau = \{16, 8, 10, 7, 20, 10\}$, $T = \{200, 300, 50, 72, 50, 100\}$, initial learning rates $\eta_{\text{start}} = \{10^{-3}, 5 \cdot 10^{-3}, 2 \cdot 10^{-3}, 5 \cdot 10^{-3}, 10^{-3}, 10^{-3}\}$, $\eta_{\text{end}} = 10^{-5}$ and $epochs = \{2000, 3000, 4000, 2000, 3000, 2000\}$ for the {Lorenz-63, Rössler, ECG, fMRI,Lorenz-96,EEG} dataset, respectively. Parameters in $\boldsymbol{W}$ were initialized using a Gaussian initialization with $\sigma = 0.01$, $\boldsymbol{h}$ as a vector of zeros, and $\boldsymbol{A}$ as the diagonal of a normalized positive-definite random matrix [11, 97]. The initial latent state $\boldsymbol{z}_1 = [\boldsymbol{x}_1, \boldsymbol{L}\boldsymbol{x}_1]^T$ is estimated from $\boldsymbol{x}_1$ using a matrix $\boldsymbol{L} \in \mathbb{R}^{(M-N) \times N}$ which is jointly learned with the other model parameters. Additionally, for the Rössler and Lorenz systems, we added 5% observation noise during training. Across all training epochs of a given run, we consistently selected the model with the lowest D_{stsp}. Each individual training run of the AL-RNN was performed on a single CPU. Depending on the training sequence length, a single epoch took between 0.5 to 3 seconds.

Geometric agreement For evaluating attractor geometries, we use a state space measure D_{stsp} based on the Kullback-Leibler (KL) divergence, which assesses the (mis)match between the ground truth spatial distribution of data points, $p_{\text{true}}(\boldsymbol{x})$, and the distribution $p_{\text{gen}}(\boldsymbol{x}|\boldsymbol{z})$ of trajectory points freely generated by the inferred DSR model. These probability distributions can be approximated in different ways from the observed/ generated trajectories. Here, we usually sampled long trajectories corresponding to the test set length (usually $T = 100{,}000$ time steps, but sometimes shorter for the empirical time series) from trained systems, removing transients to ensure that the system has reached a limit set. For low-dimensional systems, the KL divergence can be straightforwardly calculated through a discrete binning approximation [11]:

$$D_{\text{stsp}}\left(p_{\text{true}}(\boldsymbol{x}), p_{\text{gen}}(\boldsymbol{x} \mid \boldsymbol{z})\right) \approx \sum_{k=1}^{K} \hat{p}_{\text{true}}^{(k)}(\boldsymbol{x}) \log\left(\frac{\hat{p}_{\text{true}}^{(k)}(\boldsymbol{x})}{\hat{p}_{\text{gen}}^{(k)}(\boldsymbol{x} \mid \boldsymbol{z})}\right), \tag{10}$$

where $K = m^N$ is the total number of bins, with m bins per dimension and N being the dimension of the ground truth system. A bin number of $m = 30$ per dimension was chosen as a good compromise for distinguishing between successful and bad reconstructions for 3d systems. Since the number of data required to fill the bins scales exponentially with N, for the ECG time series ($N = 5$) we reduced the number of bins to $m = 8$, as suggested in [38].

Temporal agreement To assess temporal agreement, we computed Hellinger distances D_H between power spectra [67, 39]. We first simulated long time series $T = 100{,}000$ (as with D_{stsp} above). After standardization, we computed dimension-wise Fast Fourier Transforms (FFT). The power spectra were smoothened using a Gaussian kernel and normalized, and the extended, high-frequency tails, which predominantly contained noise, were truncated. The Hellinger distance between smoothed ground-truth spectra $F(\omega)$ and generated spectra $G(\omega)$ is given by:

$$H(F(\omega), G(\omega)) = \sqrt{1 - \int_{-\infty}^{\infty} \sqrt{F(\omega)G(\omega)}d\omega} \in [0, 1] \tag{11}$$

Maximum Lyapunov exponent The maximum Lyapunov exponent of a system quantifies how fast nearby trajectories diverge, and for a flow map can be computed in the limit $T \to \infty$ from the system's product of Jacobians. To approximate the maximum exponent numerically, we first iterated a trained model forward from some randomly drawn initial condition and discarded transients. Given that for chaotic systems the product of Jacobians itself explodes [67], we employed a numerical algorithm from [107, 103] that re-orthogonalizes the series of Jacobian products at regular intervals using a QR decomposition.

Categorical decoder We coupled categorical observations to the P PWL neurons via a link function given by

$$\begin{aligned} \pi_i &= \frac{\exp\left(\boldsymbol{\beta}_i^T \mathbf{z}_t^p\right)}{1+\sum_{j=1}^{K-1} \exp\left(\boldsymbol{\beta}_j^T \mathbf{z}_t^p\right)} \quad \forall i \in \{1 \dots K-1\} \\ \pi_K &= \frac{1}{1+\sum_{j=1}^{K-1} \exp\left(\boldsymbol{\beta}_j^T \mathbf{z}_t^p\right)}. \end{aligned} \tag{12}$$

The regression weights $\boldsymbol{\beta}_i \in \mathbb{R}^{P\times 1}$ are learned for each category $i = 1 \dots K-1$, while the normalization ensures that the total probability over all categories sums to one, $\sum_{i=1}^{K} \pi_i = 1$.

A.3 Benchmark datasets

Lorenz-63 The Lorenz-63 system, formulated by Edward Lorenz in 1963 [61] to model atmospheric convection, is defined by

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \sigma(y-x) \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= x(\rho - z) - y \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= xy - \beta z, \end{aligned} \tag{13}$$

where σ, ρ, β, are parameters that control the dynamics of the system (here set to $\sigma = 10$, $\beta = \frac{8}{3}$, and $\rho = 28$, in the chaotic regime). The system was solved numerically with integration time step $\Delta t = 0.01$ using `scipy.integrate` with the default `RK45` solver.

Rössler The Rössler system, intended by Otto Rössler in 1976 [85] as a further simplification of the Lorenz model, produces chaotic dynamics using a single nonlinearity in one state variable:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= -y - z \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= x + ay \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= b + z(x-c), \end{aligned} \tag{14}$$

where a, b, c, are parameters controlling the dynamics of the system (here set to $a = 0.2$, $b = 0.2$, and $c = 5.7$, in the chaotic regime). The system was solved with integration time step $\Delta t = 0.08$ using `scipy.integrate` with the default `RK45` solver.

Human electrocardiogram The electrocardiogram (ECG) time series was taken from the PPG-DaLiA dataset [82]. With a sampling frequency of $700Hz$, the recording duration spanned 600 seconds, yielding a time series of $T = 419{,}973$ time points. Initially, a Gaussian smoothing filter with $\sigma = 6$ was applied, resulting in a filter length of $l = 8\sigma + 1 = 49$. We standardized the time series and applied temporal delay embedding using the `DynamicalSystems.jl` Julia library, resulting in an embedding dimension of $m = 5$. For model training, the first $T = 100{,}000$ time steps (approximately 143 seconds) were used, downsampled to every 10th datapoint.

Human fMRI data The functional magnetic resonance imaging (fMRI) data from human subjects performing three cognitive tasks is publicly available on GitHub [52]. We followed Kramer et al. [52] and selected the first principal component of BOLD activity in each of the 20 brain regions. Subjects underwent five rounds of three cognitive tasks ('Choice Reaction Task [CRT]', 'Continuous Delayed Response Task [CDRT]' and 'Continuous Matching Task [CMT]'), together with a 'Rest' and 'Instruction' period. The time series per subject were short ($T = 360$) and reconstructions in [52] exhibited a positive maximum Lyapunov exponent, indicating the chaotic nature of the underlying system.

Lorenz-96 The Lorenz-96 system, formulated by Edward Lorenz in 1996 [62], is defined by

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F, \tag{15}$$

with system variables x_i, $i = 1, ..., N$, and forcing term F (here, $F = 8$, in the chaotic regime). Furthermore, cyclic boundary conditions are assumed with $x_{-1} = x_{N-1}$, $x_0 = x_N$, $x_{N+1} = x_1$, and the system was solved with integration step $\Delta t = 0.04$ using `scipy.integrate` with the default `RK45` solver.

Human EEG data Electroencephalogram (EEG) data were taken from a study by Schalk et al. [87]. These are 64-channel EEG recordings obtained from human subjects during different motor and imagery tasks. The signal was standardized and smoothed using a Hann function and a window length of 15, as in [11].

A.4 Further Results

Table 1: Comparison of AL-RNN to different SOTA models for dynamical systems reconstruction. Comparison values, datasets and evaluation measures as in [39], based on code provided on GitHub by the authors. id-TF: identity teacher forcing, GTF: generalized teacher forcing, D_{stsp}: state space divergence, D_H: Hellinger distance across power spectra. Reported values are median $\pm$ median absolute deviation.

Dataset	Method	$D_{\text{stsp}} \downarrow$	$D_H \downarrow$	$\lvert\boldsymbol{\theta}\rvert$
Lorenz-63 (3d)	AL-RNN + id-TF	0.34 ± 0.05	0.081 ± 0.012	360
	shPLRNN + GTF	0.26 ± 0.03	0.090 ± 0.007	365
	dend-RNN + id-TF	0.9 ± 0.2	0.15 ± 0.03	361
	Reservoir Computer	0.52 ± 0.12	0.19 ± 0.04	603
	LSTM-TBPTT	0.46 ± 0.22	0.11 ± 0.03	1188
	SINDy	0.24 ± 0.00	0.091 ± 0.000	30
	Neural-ODE	0.63 ± 0.2	0.15 ± 0.05	353
	Long Expressive Memory	0.39 ± 0.24	0.12 ± 0.05	360
ECG	AL-RNN + id-TF	3.0 ± 0.7	0.29 ± 0.04	2808
	shPLRNN + GTF	4.3 ± 0.6	0.34 ± 0.02	2785
	dendPLRNN + id-TF	5.8 ± 0.6	0.37 ± 0.06	3245
	Reservoir Computer	5.3 ± 1.7	0.39 ± 0.05	5000
	LSTM-TBPTT	15.2 ± 0.5	0.73 ± 0.02	5920
	SINDy	diverging	diverging	3960
	Neural-ODE	12.2 ± 0.7	0.7 ± 0.03	4955
	Long Expressive Memory	16.3 ± 0.2	0.56 ± 0.04	4872
Lorenz-96 (20d)	AL-RNN + id-TF	1.64 ± 0.03	0.089 ± 0.001	4623
	shPLRNN + GTF	1.68 ± 0.06	0.072 ± 0.001	4540
	dendPLRNN + id-TF	1.65 ± 0.05	0.083 ± 0.005	5740
	Reservoir Computer	2.40 ± 0.15	0.14 ± 0.02	12000
	LSTM-TBPTT	5 ± 1	0.31 ± 0.04	10580
	SINDy	1.59 ± 0.00	0.06 ± 0.00	4620
	Neural-ODE	1.77 ± 0.07	0.076 ± 0.01	4530
	Long Expressive Memory	7.2 ± 2.3	0.54 ± 0.13	4620
EEG (64d)	AL-RNN + id-TF	2.6 ± 0.3	0.14 ± 0.03	17955
	shPLRNN + GTF	2.1 ± 0.2	0.11 ± 0.01	17952
	dendPLRNN + id-TF	3 ± 1	0.13 ± 0.04	18099
	Reservoir Computer	14 ± 7	0.54 ± 0.15	28672
	LSTM-TBPTT	30 ± 21	0.2 ± 0.1	51584
	SINDy	diverging	diverging	133120
	Neural-ODE	20 ± 0.5	0.47 ± 0.01	17995
	Long Expressive Memory	10.2 ± 1.5	0.38 ± 0.06	18304

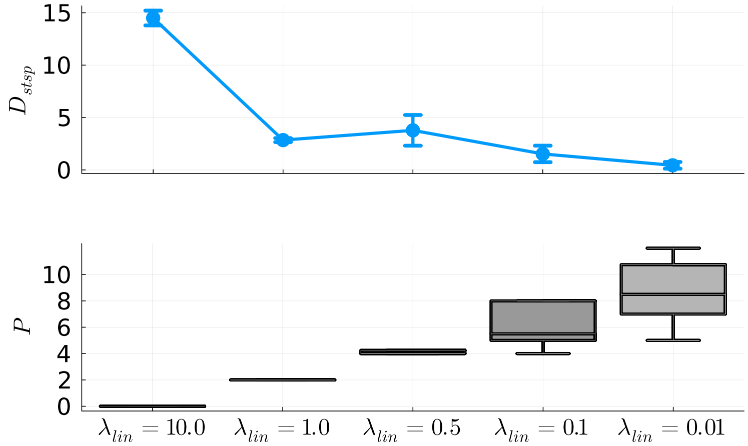

Figure 9: Top: DSR quality (assessed by D_{stsp}) as a function of strength of regularization on the number of nonlinearities for the AL-RNN trained on Lorenz-63. Bottom: Number of selected piecewise-linear units P as a function of regularization strength. As in Fig. 3, a first optimum consistently occurs for $P = 2$. To select the number of nonlinear units through regularization, we replaced the standard ReLU by a leaky ReLU $\max(\alpha_i z_i, z_i), \alpha_i \in (0, 1)$, for each of the $i = 1 \dots M$ units. The slope $\alpha_i = \sigma(\gamma_i)$ is determined through a steep sigmoid, $\sigma(\gamma_i) = 1/(1 + \exp(-500(\gamma_i - 0.5)))$, via trainable parameter γ_i, ensuring that it is either close to 0 or close to 1. To encourage linearity, we include a loss term $\mathcal{L}_{\text{lin}} = \lambda_{\text{lin}} \sum_{i=1}^{M} |\alpha_i - 1|$, pushing slopes towards 1. After training, units with $\alpha_i \approx 1$ are classified as linear, while all remaining units were considered nonlinear to provide an estimate for P.

Task stages align with subregions To test the alignment of the reconstructed AL-RNN activity in the subspace of the $P = 2$ PWL units with the 4 task stages, we trained 10 models on each of the 10 subjects without visible movement artifacts (yielding 100 trained models). We then checked which assignment of the four subregions (00, 01, 10, and 11) to the four task stages (Rest and Instruction, CRT, CDRT, and CMT) gave the highest alignment (Fig. 8), and used this to determine the average classification accuracy as the percentage of time points correctly assigned to their respective task stage based on the four subregions.

Geometrically minimal reconstructions of empirical time series Increasing the number of PWL units also improved geometric agreement for the empirical time series (Fig. 3, $P = 10$ for the ECG data), while dynamics remained confined to a relatively small subset of linear subregions (Fig. 4). This confinement within only a few subregions allows for the efficient computation of dynamical objects like fixed points. For the ECG data, real and virtual fixed points in the linear subregions were primarily located within a 3d hyperplane of the 5d data. Principal component analysis showed that this hyperplane harboring the fixed points aligned with the first principal component of the data, explaining approximately 40% of the data's variance (Fig. 20**a/b**). A similar pattern was observed in the fMRI data, where fixed point locations often aligned with PC1 of the data (Fig. 21).

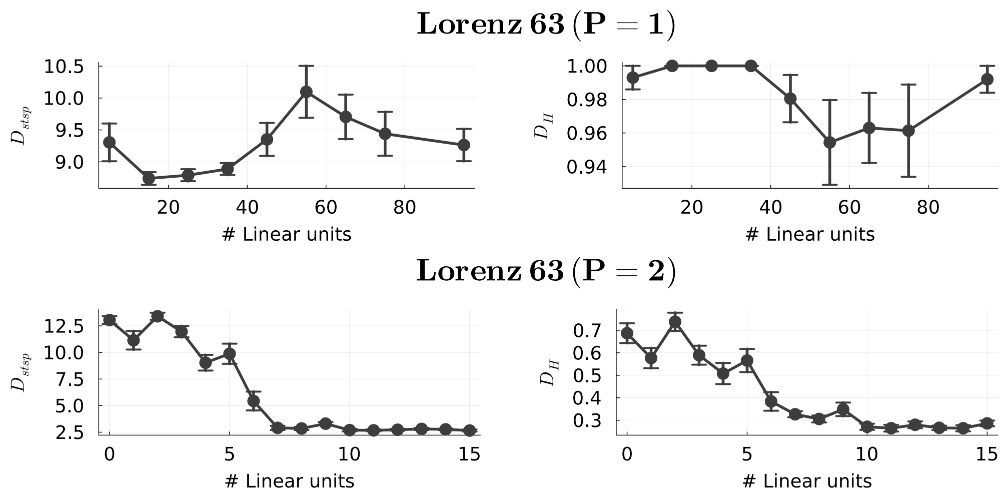

Figure 10: Reconstruction performance on Lorenz-63 system for an AL-RNN ($M = 20$) as a function of the number of linear units, once for the case where the number of PWL units was insufficient for a topologically accurate reconstruction ($P = 1$, top), and once for the case where it was sufficient ($P = 2$, bottom). Results indicate performance cannot be improved by adding more linear units if P is too small, but can be – up to some saturation level – when P is sufficiently large. Error bars = SEM.

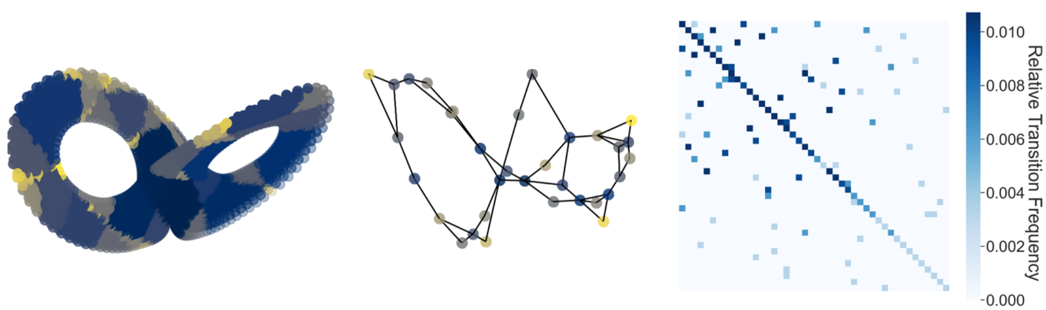

Figure 11: Optimal geometric reconstruction of a Lorenz-63 using the AL-RNN with $P = 8$ PWL units. Left: reconstruction with subregions color-coded by frequency of trajectory visits (dark: most frequently visited regions, yellow: least frequent regions). Center: Resulting geometrical graph structure (using transition probabilities for placing the nodes) visualized using the spectral layout in `networkx`. Note that self-connections and directedness of edges were omitted in this representation. The resulting graph shadows the layout of the reconstructed system. Right: Connectome of transitions between subregions.

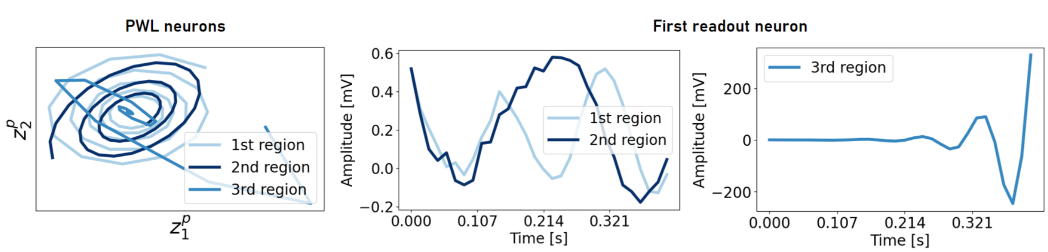

Figure 12: 'Linearized' dynamics (i.e., considering the linear map from each subregion) within the three linear subregions of the AL-RNN trained on the ECG data from Fig. 7. The first two subregions host weakly unstable spirals with shifted phase, corresponding to the excitatory/ inhibitory phases of the ECG. The strongly divergent activity in the third subregion induces the Q wave.

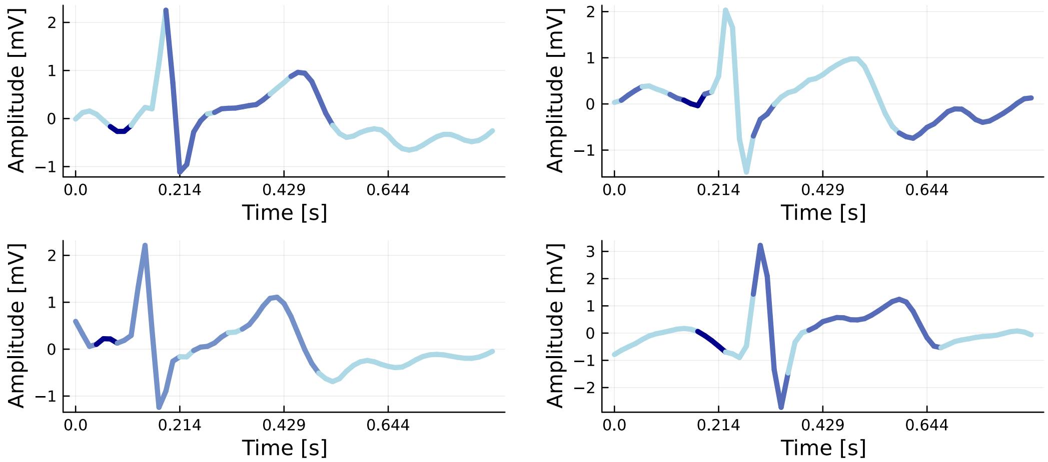

Figure 13: Freely generated ECG activity using an AL-RNN with 3 linear subregions (color-coded) shows consistent assignment of the Q wave to a distinct subregion across multiple successful reconstructions.

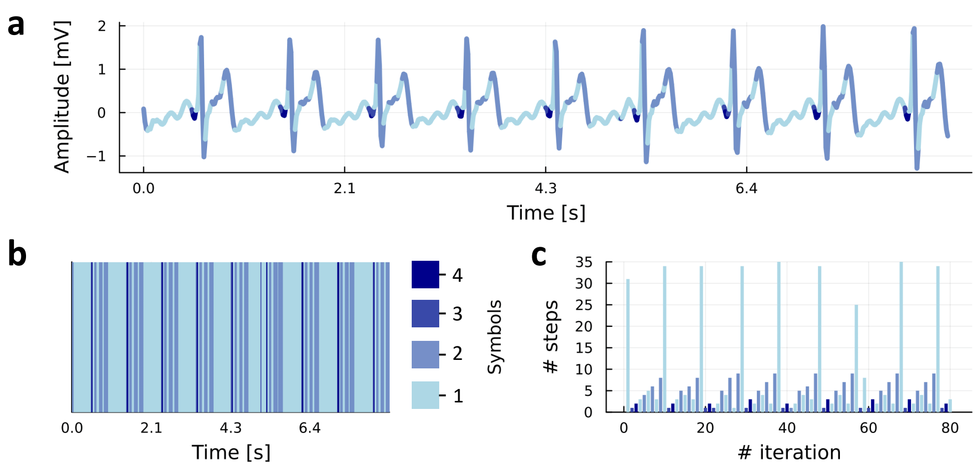

Figure 14: **a**: Freely generated ECG activity by the AL-RNN with $M = 100$ total units and $P = 2$ PWL units. **b**: Symbolic coding of the dynamics (with each shade of blue a different symbol/ linear subregion), reflecting the QRS phase with alternating excitation/inhibition (lighter shades of blue) following the short Q wave burst (dark blue). **c**: Time histogram of distinct symbols along the symbolic trajectory, exposing the mildly chaotic nature of the reconstructed ECG signal [39].

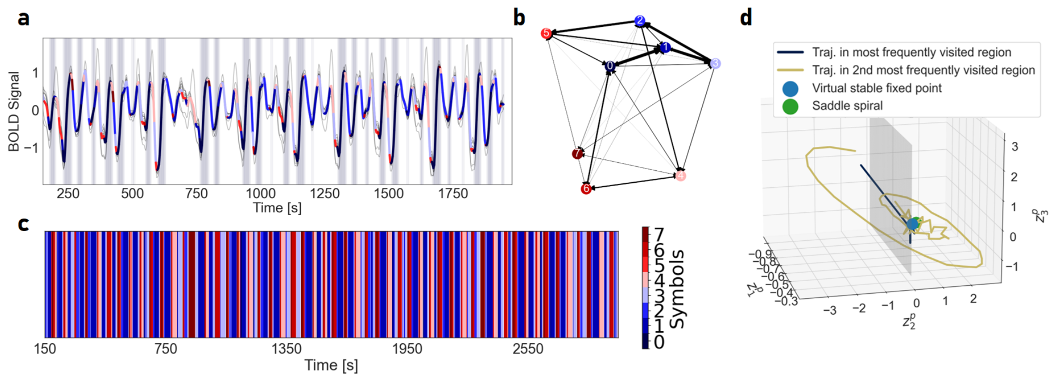

Figure 15: Freely generated fMRI activity using an AL-RNN with $M = 100$ total units and $P = 3$ PWL units. **a**: Mean generated activity color-coded according to linear subregions, with background shading highlighting the most frequently visited subregion. **b**: Geometrical graph representation of connections between linear subregions, with edge weights representing relative transition frequencies (self-connections omitted). **c**: Time series of the symbolic coding of dynamics according to linear subregions. **d**: Dynamics in the two most frequently visited linear subregions in the subspace of the three PWL units, with the boundary between subregions in gray. The dark blue trajectory bit in the first subregion moves towards a virtual stable fixed point located near the center of the saddle spiral in the second subregion. The yellow trajectory illustrates an orbit cycling away from this spiral point and eventually crossing into the first subregion. From there, trajectories are pulled back into the second subregion through the virtual stable fixed point located close to the saddle spiral (see also activity with background shading in **a**). This dynamical behavior is similar to the one observed in the chaotic benchmark systems, where locally divergent activity of the AL-RNN is propelled back into the center of an unstable manifold within another subregion.

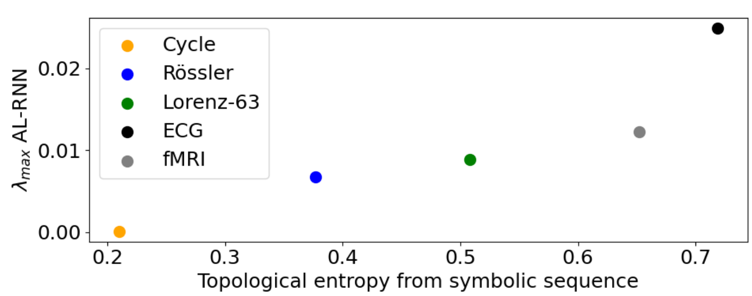

Figure 16: Topological entropy computed from symbolic sequences (Figs. 14 & 15) versus λ_{max}, calculated from corresponding topologically minimal AL-RNNs (Figs. 5 & 7).

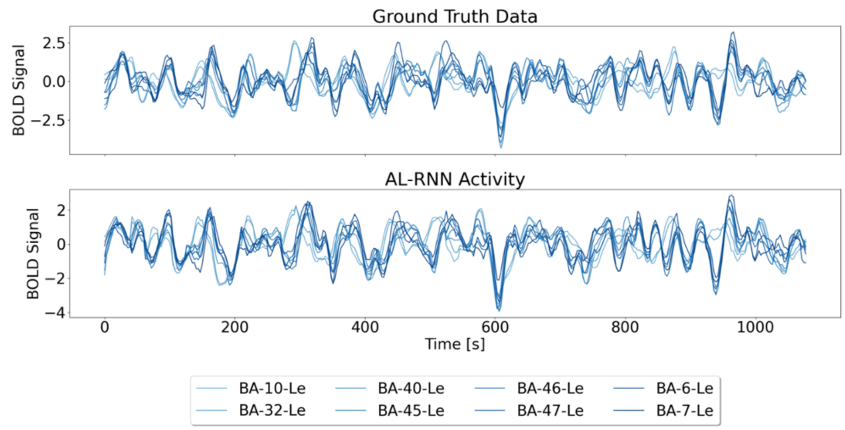

Figure 17: Generated fMRI activity using an AL-RNN with $M = 100$ total units and $P = 2$ PWL units, with the readout unit states replaced by observations every 7 time steps.

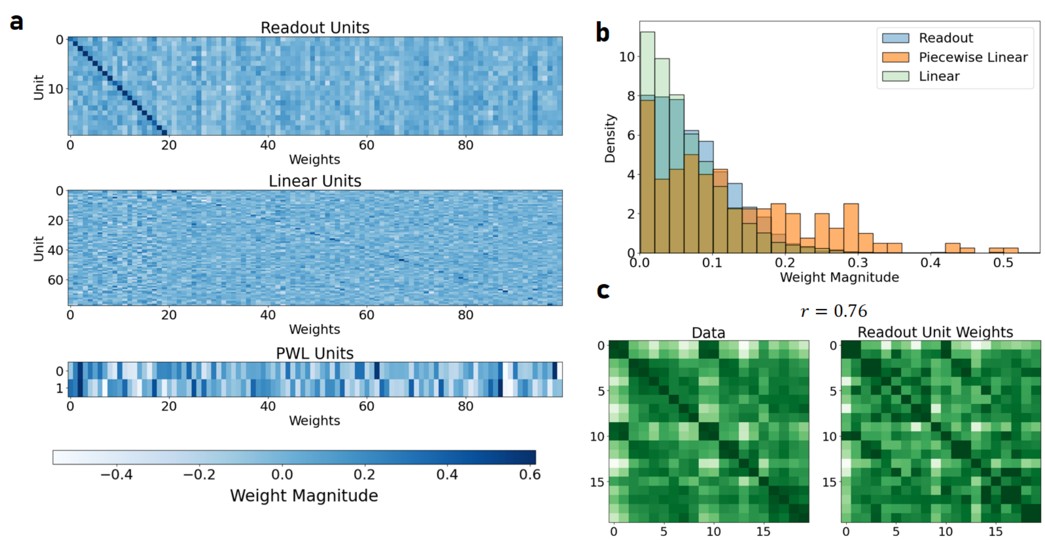

Figure 18: **a**: Weights of the reconstructed AL-RNN. **b**: Histogram of the absolute weight distributions for the different types of AL-RNN units. On average, weight magnitudes of the PWL units are much higher than those of the other unit types. **c**: The correlation structure among the weights of the $N = 20$ readout units (rows in **a**, top) reflects the correlation structure within the observed time series variables (correlation between both matrices $r \approx 0.76$).

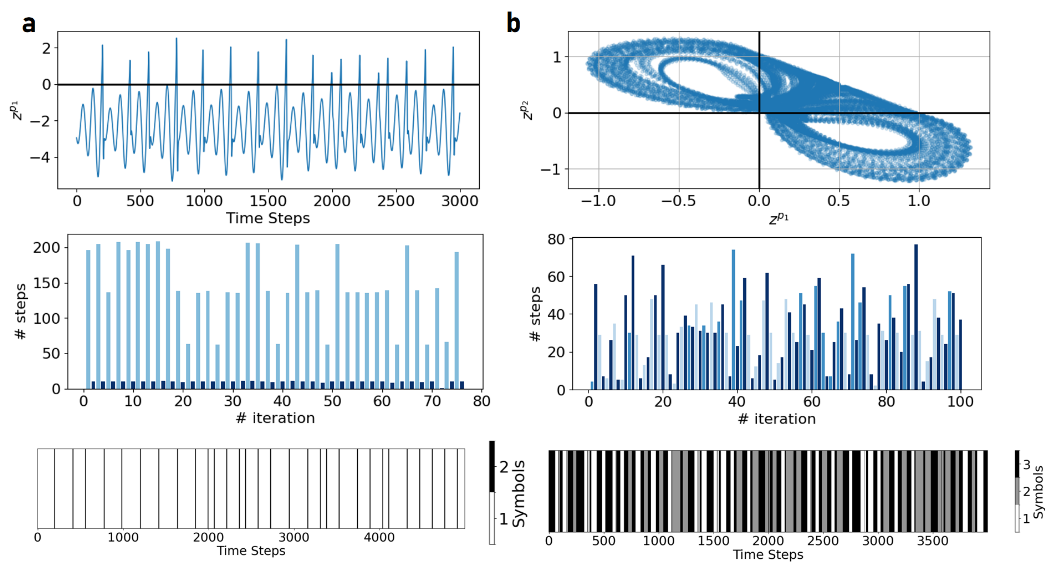

Figure 19: Top row: Activity of the PWL units for the topologically minimal representations of the Rössler (**a**) and Lorenz-63 attractor (**b**) from Fig. 5. Center row: Time histogram of discrete symbols of the symbolic trajectory. Bottom row: Time series of the symbolic trajectory.

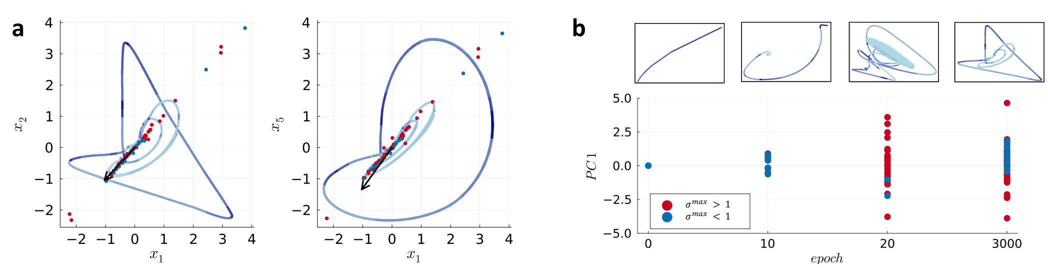

Figure 20: **a**: Variation in the location of the analytically computed real and virtual fixed points of the linear subregions aligns with the first PC of the data (generated trajectories in bluish, color-coded according to linear subregion). **b**: Fixed point location along the first principal component (with corresponding dynamics within (x_1,x_2)-plane of observation space on top) at different characteristic stages of training. At the early stages of training, fixed points of the linear subregions are distributed along the data manifold within the subspace of readout units. Around epoch 20, the maximum absolute eigenvalues $\sigma^{\max}$ of the Jacobians in many subregions become larger than one, inducing local divergence necessary for producing the observed chaotic dynamics.

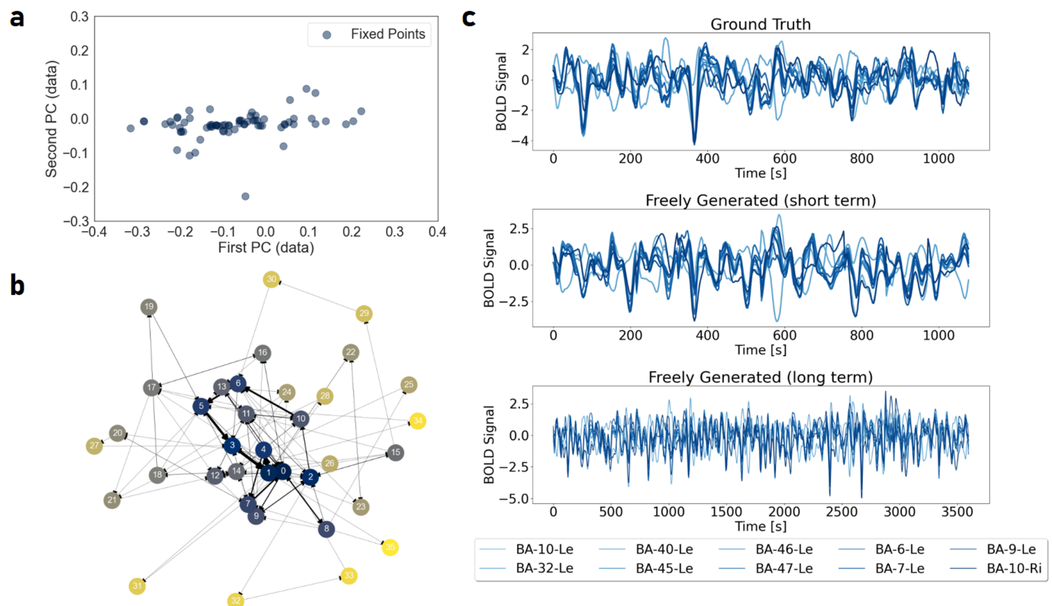

Figure 21: **a**: Analytically computed real and virtual fixed points of the linear subregions of a geometrically minimal AL-RNN ($M = 100, P = 10$) align with the first PC of the data within the subspace of readout units. The BOLD time series for different brain regions were highly correlated, so PC1 accounted for approximately 80% of the data variance. **b**: Geometrical graph representation with relative frequency of transitions between linear subregions indicated by line thickness of edges, showing a central highly connected subgraph of frequently visited (bluish) dominant subregions, as in Fig. 4. **c**: Example freely generated activity from ten simulated brain regions.

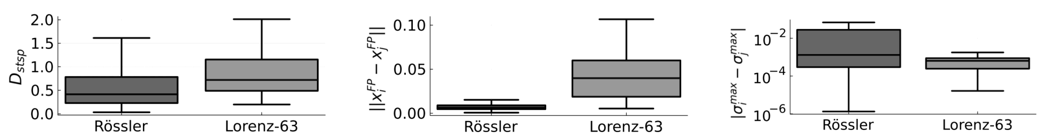

Figure 22: Same as Fig. 5**d-f**, but showing absolute instead of relative deviations (for D_{stsp}, values < 1.0 indicate tight agreement).

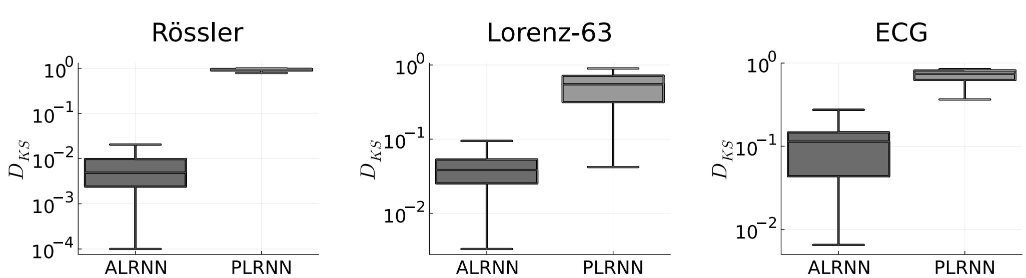

Figure 23: Pairwise differences in sorted (from lowest to highest probability) cumulative trajectory point distributions across linear subregions across all valid pairs from 20 training runs (quantified by the Kolmogorov–Smirnov distance, D_{KS}) for the AL-RNN vs. PLRNN, revealing much higher consistency for AL-RNN. Note the log-scale on the y-axis.

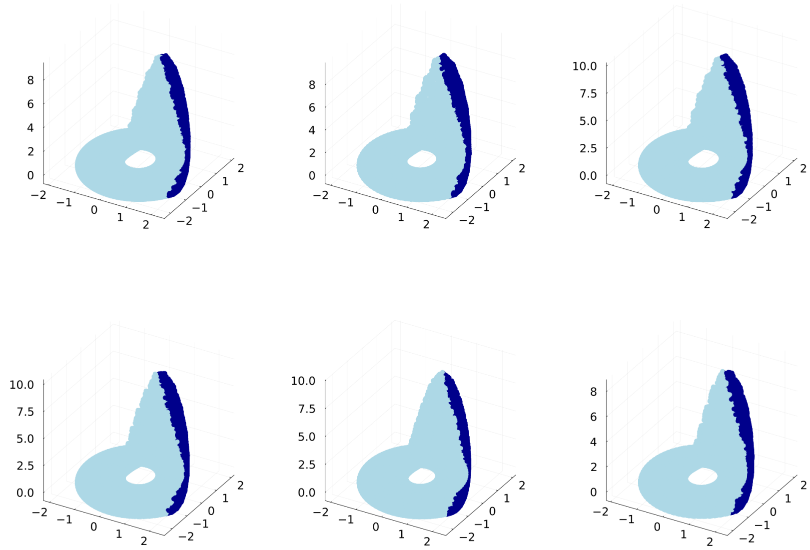

Figure 24: Linear subregions (color-coded) mapped to observation space of the Rössler system, showing a robust representation of the individual subregions across multiple training runs/ models.

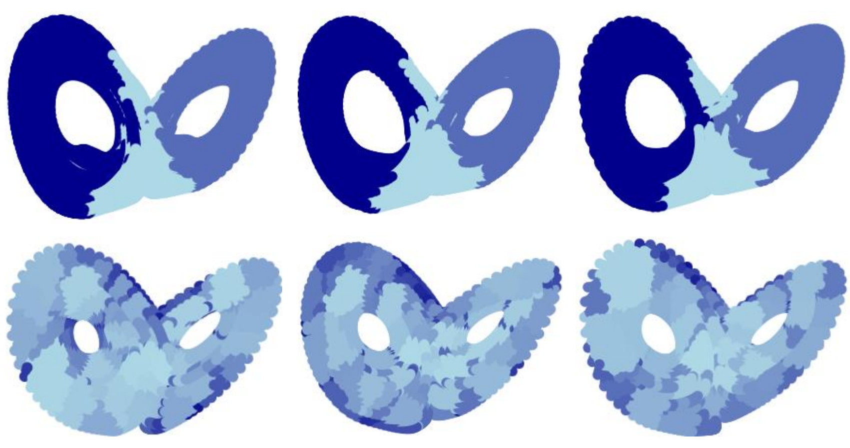

Figure 25: Top row: Robust placing of linear subregions (color coded) mapped to observation space across training runs using the AL-RNN. Model recovery experiments further confirmed the robustness of the model solutions, with very similar overall performance measures across different experiments (original: $D_{stsp} = 3.14$, $D_H = 0.28$; recovered: $D_{stsp} = 3.38 \pm 0.18$, $D_H = 0.28 \pm 0.03$; 3 linear subregions in all cases). Bottom row: In contrast, for the PLRNN linear subregions are differently assigned (with varying boundaries) on each run.

B Proofs of Theorems

Define a shift space of finite type $(A_{\mathcal{F}}, \sigma)$ such that each symbol of the alphabet of $\mathcal{A}$ is associated with exactly one set U_e of the topological partition of S, i.e. we have a total of $|\mathcal{U}|$ symbols in $\mathcal{A}$. Further define for $\boldsymbol{a} \in A_{\mathcal{F}}$ the sets [54]

$$D_l(\boldsymbol{a}) = \bigcap_{k=-l}^{l} \phi^{-k}(U_{a_k}) \subseteq S, \tag{16}$$

where a_k is the kth symbol in the sequence $\boldsymbol{a}$: Think of this as building the intersection of U_{a_0} with k-times forward iterates of subsets associated with a_{-k} and k-times backward iterates of subsets associated with a_k, such that in the limit $l \to \infty$ hopefully we end up with a single point corresponding to a unique trajectory in S, i.e. considering $D(\boldsymbol{a}) = \bigcap_{l=0}^{\infty} \overline{D_l}(\boldsymbol{a})$ (see Lind and Marcus [54] for details). We now define [54]

Definition 1. *A symbolic representation of an invertible DS (S, ϕ) with topological partition $\mathcal{U} = \{U_0 \dots U_{n-1}\}$ is a shift space $(A_{\mathcal{F}}, \sigma)$ with alphabet $\mathcal{A} = \{0 \dots n-1\}$, such that each symbol $a_k \in \mathcal{A}$ is associated with exactly one subset $U_k \in \mathcal{U}$, and $D(\boldsymbol{a}) = \bigcap_{l=0}^{\infty} \overline{D_l}(\boldsymbol{a})$ contains exactly one point $\boldsymbol{x} \in S$ for each $\boldsymbol{a} \in A_{\mathcal{F}}$. If $A_{\mathcal{F}}$ is a finite shift, we call this a Markov partition.*

Ideally, we would like our symbolic coding of the DS to be a symbolic representation or Markov partition according to this definition, but in practice this may entail other unfavorable properties (e.g., too fine-grained) and we contend here with the properties given by Theorems 1 - 3.

In the statement of our theorems, we further used the terms 'eventually periodic' and 'asymptotically periodic'. Let us now strictly define them (according to Alligood et al. [3]):

Definition 2. *An orbit $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ of the map ϕ is said to be **asymptotically periodic** if it converges to a periodic orbit as $n \to \infty$. This implies that there is a periodic orbit $\Gamma_k = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_k, \boldsymbol{y}_1, \boldsymbol{y}_2, \dots\}$ such that $\lim_{n\to\infty} d(\boldsymbol{z}_n, \Gamma_k) = 0$.*

For instance, any orbit that is attracted to a stable fixed point or to a saddle fixed point (evolving on its stable manifold) is asymptotically periodic (fixed).

Definition 3. *A point $\boldsymbol{z}$ is called **eventually periodic** with period p for the map ϕ, if for some positive integer N, $\phi^{n+p}(\boldsymbol{z}) = \phi^n(\boldsymbol{z})$ for all $n \geq N$, and p is the smallest positive integer with this exact property. This means that the orbit of $\boldsymbol{z}$ eventually maps **exactly onto a periodic orbit**.*

Note: The term 'eventually periodic' describes the extreme case where an orbit coincides *precisely* with a periodic orbit. Thus, any eventually periodic orbit is also asymptotically periodic, but the reverse is not always true: An asymptotically periodic orbit comes arbitrarily close to a periodic orbit, but may not land precisely on it.

With these definitions we are now ready to prove the theorems.

Proof of Theorem 1

Proof. "$\Rightarrow$": Suppose that the orbit $\Omega_S = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ is asymptotically fixed. Hence, there exists a fixed point $\boldsymbol{z}^* \in U_{a^*}$ of the AL-RNN $F_{\boldsymbol{\theta}}$ (i.e., $\boldsymbol{z}^* = F_\theta(\boldsymbol{z}^*)$) such that $\lim_{n\to\infty} \boldsymbol{z}_n = \boldsymbol{z}^*$. Let $\boldsymbol{a} = (a_1 a_2 a_3 \dots)$ be the corresponding symbolic sequence of the orbit Ω_S with $\boldsymbol{z}_n \in U_{a_n}, \forall n$. Since $\lim_{n\to\infty} \boldsymbol{z}_n = \boldsymbol{z}^*$, so there exists some $N \in \mathbb{N}$ such that for every $n \geq N$ the points $\boldsymbol{z}_n$ will remain arbitrarily close to $\boldsymbol{z}^*$. This means the points $\boldsymbol{z}_n$ eventually enter the same linear subregion that contains $\boldsymbol{z}^*$ and will remain there for all future iterations. Thus, $\boldsymbol{a} = (a_1 a_2 a_3 \dots a_{N-1})(a^*)^\infty$ is eventually fixed.

"$\Leftarrow$": Assume $\boldsymbol{a} = (a_1 a_2 a_3 \dots a_{N-1})(a^*)^\infty$ is an eventually fixed sequence of the symbolic encoding. This means for all the corresponding orbits $\Omega_S = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ of $F_{\boldsymbol{\theta}}$, there exists an index N such that for every $n \geq N$ the orbit points $\boldsymbol{z}_n$ must remain in the same subregion, say U_{a^*}. Since by assumption the map F_θ is non-globally-diverging and hyperbolic, the system cannot be expanding in all directions in any of the subregions. Thus, there must be at least one contracting or stable direction in U_{a^*}. Consequently, there must be at least one corresponding orbit

$\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ that converges toward some stable structure in U_{a^*} as $n \to \infty$. Since F_θ is a linear hyperbolic map in each subregion, there cannot be any k-cycle, $k \geq 2$, with all periodic points contained within a single subregion. Therefore, the corresponding orbit converges to a (saddle) fixed point $\boldsymbol{z}^* \in U_{a^*}$ in the stable manifold along stable directions. If **all** directions in U_{a^*} are contracting or stable, then all corresponding orbits will converge to a stable fixed point within that subregion. $\square$

Proof of Theorem 2

Proof. "$\Rightarrow$": Let $\Omega_S = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ be an asymptotically p-periodic orbit of F_θ. Then there is a periodic orbit $\Omega_{S,p} = \{\boldsymbol{z}_1^*, \dots, \boldsymbol{z}_p^*\}$ (i.e., such that all points $\boldsymbol{z}_k^* \in \Omega_{S,p}$ are distinct and $\boldsymbol{z}_{k+p}^* = F_{\boldsymbol{\theta}}^p(\boldsymbol{z}_k^*) = \boldsymbol{z}_k^*, k = 1 \dots p$) and

$$\lim_{n\to\infty} d(\boldsymbol{z}_n, \Omega_{S,p}) \;=\; 0. \tag{17}$$

Since $F_{\boldsymbol{\theta}}$ is a linear and hyperbolic map in each subregion, there cannot be any p-periodic orbit with $p \geq 2$ where all periodic points are contained within a single subregion. On the other hand, due to the definition of the symbolic coding for each trajectory, each point $\boldsymbol{z}_t$ at time step t is assigned its own symbol a_t, depending on its associated linear subregion. Hence, the corresponding symbolic sequence of the periodic orbit $\Omega_{S,p}$ is $(a_1^* a_2^* \dots a_p^*)^\infty$ with $\boldsymbol{z}_k \in U_{a_k^*}, k = 1 \dots p$. Moreover, due to eq. 17, for some large enough index $N \in \mathbb{N}$, the points $\boldsymbol{z}_n$ of the orbit Ω_S become arbitrarily close to the orbit $\Omega_{S,p}$. Thus, for $n \geq N$, the itinerary of Ω_S will follow the same repeating pattern as the periodic orbit's itinerary and will revisit the same subregions as the periodic orbit points $\boldsymbol{z}_k^*$, $k = 1 \dots p$. Therefore, the corresponding symbolic sequence of the orbit Ω_S is $\boldsymbol{a} = (a_1 a_2 \dots a_{N-1})(a_1^* a_2^* \dots a_p^*)^\infty$, which is an eventually p-periodic orbit of the shift map σ.

"$\Leftarrow$": Assume that $\boldsymbol{a} = (a_1 a_2 \dots a_{N-1})(a_1^* a_2^* \dots a_p^*)^\infty \in A_{\mathcal{U}, F_{\boldsymbol{\theta}}}$ is an eventually p-periodic orbit of the shift map σ. Consequently,

$$\exists N \in \mathbb{N} \;\; \forall \;\; n \geq N \; : \qquad a_{n+p} = a_n, \tag{18}$$

which says that the symbolic sequence is repeating every p steps. Therefore, for all the corresponding orbits $\Omega_S = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ of $F_{\boldsymbol{\theta}}$, there exists an index N such that for every $n \geq N$ the orbit points $\boldsymbol{z}_n$ will stay in the same p linear subregions, say $U_{a_1^*}, U_{a_2^*} \dots, U_{a_p^*}$, and revisit each $U_{a_i^*} (i = 1, 2, \dots, p)$ after exactly p time steps:

$$\exists N \in \mathbb{N} \;\; \forall \;\; n \geq N \; : \qquad U_{a_{n+p}^*} = U_{a_n^*}. \tag{19}$$

Since by assumption the map F_θ is non-globally-diverging and hyperbolic, the system cannot be expanding in all directions in any of the subregions. Thus, there exists at least one contracting direction in each of the subregions $U_{a_1^*}, U_{a_2^*} \dots, U_{a_p^*}$, and therefore at least one corresponding orbit $\Omega_S = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ that converges toward some stable structure within $U_{a_1^*}, U_{a_2^*} \dots, U_{a_p^*}$ as $n \to \infty$. Furthermore, as F_θ is a linear and hyperbolic map in each subregion there cannot be any k-cycle, $k \geq 2$, with all periodic points contained within only one of the subregions. Similarly, it cannot be chaotic or quasi-periodic within just one subregion. Now, due to eq. 19, for the corresponding orbit Ω_S

$$\forall \, n \geq N \;\; \forall \, \boldsymbol{z}_n \in U_{a_i^*} (i = 1 \dots p) \;\; : \; \boldsymbol{z}_{n+p} = F_{\boldsymbol{\theta}}^p(\boldsymbol{z}_n) = \boldsymbol{W}_{\boldsymbol{a}} \boldsymbol{z}_n + \boldsymbol{h}_{\boldsymbol{a}} \in U_{a_i^*}, \tag{20}$$

with fixed parameters $\boldsymbol{W}_{\boldsymbol{a}} := \boldsymbol{W}_{\Omega^{a_p^*}} \cdots \boldsymbol{W}_{\Omega^{a_2^*}} \boldsymbol{W}_{\Omega^{a_1^*}}$ and $\boldsymbol{h}_{\boldsymbol{a}} := \sum_{i=2}^{p} \prod_{j=2}^{i} \boldsymbol{W}_{\Omega^{a_{p-j+2}}} \boldsymbol{h} + \boldsymbol{h}$. Since $F_{\boldsymbol{\theta}}^p$ is strictly affine, for $n \geq N$, any sub-sequence $\{\boldsymbol{z}_{n+mp}\}_{m=1}^{\infty}$ of Ω_S cannot be convergent to an aperiodic (i.e., chaotic or quasi-periodic) orbit. Therefore, the corresponding orbit Ω_S converges to a (saddle) p-periodic orbit, with all periodic points within the sub-regions $U_{a_1^*}, U_{a_2^*} \dots, U_{a_p^*}$, in the stable manifold along stable directions. If all directions in U_{a^*} are contracting or stable, then all corresponding orbits will converge to a stable p-periodic orbit with all periodic points within the subregions $U_{a_1^*}, U_{a_2^*} \dots, U_{a_p^*}$. $\square$

Proof of Theorem 3

Proof. "$\Rightarrow$": Let $\Omega_S = \{\boldsymbol{z}_1, \boldsymbol{z}_1, \dots, \boldsymbol{z}_n, \dots\}$ of $F_{\boldsymbol{\theta}}$ be an asymptotically **aperiodic** orbit of F_θ. Then, there is an aperiodic orbit $\Omega = \{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2 \dots\}$ (i.e., with $\bar{\boldsymbol{z}}_k \neq F_{\boldsymbol{\theta}}^p(\bar{\boldsymbol{z}}_k) \forall k, p > 0$) and

$$\lim_{n\to\infty} d(\boldsymbol{z}_n, \Omega) \;=\; 0. \tag{21}$$

According to the proof of Theorem 2 (second part), this orbit cannot have an eventually periodic symbolic representation $(a_1 a_2 \dots a_{N-1})(a_1 a_2 \dots a_p)^\infty$, because if it had, it would need to be asymptotically periodic as well.

"$\Leftarrow$": Assume an aperiodic symbolic sequence $\boldsymbol{a} = (a_1, \dots, a_n, \dots)$, where there is no $p > 0$ such that $a_k = \sigma^p(a_k) \forall k$. This will correspond to an aperiodic succession $U_{a_1} \dots U_{a_n} \dots$ of linear subregions U_{a_k} visited, since each subregion has its unique symbol. However, from the proof of Theorem 2 (first part) we know that any asymptotically periodic orbit of $F_{\boldsymbol{\theta}}$ must have an eventually periodic symbolic encoding, so the orbit Ω_S corresponding to $\boldsymbol{a}$ cannot be asymptotically periodic. Consequently, it cannot be (eventually) periodic either, which implies that it must be aperiodic. $\square$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The results illustrated in the text and figures properly address the presented claims, including formal proof for the theorems and hyperparameter settings and code for reproducing the results of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are addressed in the conclusion or, e.g. in the case of theoretical results, directly in the respective section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper includes formal proofs of all theorems Appx. B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detailed hyperparameter settings for the results in the paper are given in Sect. A.2, which combined with the codebase allow for replication of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: A link to a github repository is provided that contains the implementation of the network architecture that was used for the main results in the paper (https://github.com/DurstewitzLab/ALRNN-DSR).

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All hyperparameter settings are provided in Sect. A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars are reported in figures and with numerical results, i.e. Figs. 3, 4 and 5. The settings for the random initialization of the network architecture are specified in Sect. A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details on computational resources are provided in Sect. A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All datasets considered here were publically available. The research is aimed at facilitating interpretability of machine learning models. We believe our approach has primarily positive ethical implications. Although we have not identified specific ethical concerns, the wide range of potential applications means that possible misuses cannot be entirely ruled out.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work constitutes foundational research. As such, we do not foresee any direct negative societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not foresee a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All code and results were created by us. Datasets were publically sourced and the respective authors cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: A fully documented version of the code is published on github.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.