FROM REDUNDANCY TO MINIMALITY: FIXED-POINT-GUIDED HIERARCHICAL REDUCTION OF LEARNED PIECEWISE-LINEAR DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding a nonlinear dynamical system from time series requires not only reproducing its trajectories, but also identifying a simple representation that preserves its essential dynamical structure. Almost-linear recurrent neural networks (AL-RNNs) are piecewise-linear RNNs in which only a subset of units use ReLU nonlinearities, while the remaining units are linear, so that nonlinear capacity is explicitly controlled by the number of ReLU units. Their activation patterns define linear regions, represented as symbols, whose observed transitions form a symbolic transition graph. However, directly training AL-RNNs with few ReLU units to realize minimal dynamical representations can be unreliable. We ask whether an AL-RNN with more ReLU units can instead be trained first and systematically reduced to a minimal dynamical representation. We introduce a fixed-point-guided hierarchical reduction procedure that progressively linearizes selected ReLU units, thereby merging neighboring linear regions and their corresponding graph nodes while preserving distinct symbols containing fixed points (FPs). The resulting reduction tree defines a hierarchy of progressively simpler candidate models. Each reduced candidate is initialized from the parent parameters and retrained under guidance from the parent dynamics, providing a practical route from structural reduction to dynamical realization. We also prove that reproducing Q distinct fixed points requires at least Q FP-containing symbols, providing a certificate of symbol-level minimality when this bound is attained. On the 3-scroll Chua system, direct training with the theoretical minimum of three ReLU units achieves high-fidelity minimal realizations in only 20% of seeds, whereas our learn–reduce–retrain strategy increases the seed-macro success rate to approximately 71% at the same final nonlinear capacity. These results show that redundant nonlinear capacity can serve as a scaffold for discovering and realizing minimal dynamical representations.

1 INTRODUCTION

Learning dynamical systems from time-series observations is a central problem in scientific machine learning. When the governing equations of a system are unknown, recurrent neural networks (RNNs) can serve as flexible generative surrogates that reproduce complex nonlinear dynamics directly from data (Durstewitz et al., 2023). For scientific understanding, however, trajectory reconstruction alone is often insufficient: many different models may reproduce similar observations while providing very different descriptions of the underlying dynamics. A complementary goal is therefore to identify a *parsimonious* representation, one that explains the observed dynamics using as few essential dynamical components as possible (Brunton et al., 2016; Champion et al., 2019). Such representations are attractive not only for interpretability, but also for subsequent dynamical analysis and control, where reduced structural complexity can be advantageous.

Fixed points provide a natural starting point for mechanistic analysis of learned dynamics and have been widely used to reverse-engineer RNNs (Sussillo & Barak, 2013; Golub & Sussillo, 2018; Smith et al., 2021). Stable fixed points characterize attracting states, while unstable fixed points and saddles can organize transitions and shape trajectories far beyond their immediate neighborhoods (Sussillo

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

& Barak, 2013). However, fixed-point analysis alone remains local and does not describe transitions between different local dynamics.

Piecewise-linear recurrent neural networks (PL-RNNs) avoid these ambiguities by using piecewise-linear nonlinearities, such as ReLU units, to explicitly partition the state space into linear regions (Durstewitz, 2017; Monfared & Durstewitz, 2020). Each state is unambiguously assigned to a region whose affine dynamics are exact throughout that region, while candidate fixed points can be obtained analytically by solving a linear equation within each region. Moreover, transitions across region boundaries induce a discrete sequence of visited linear regions, providing a symbolic description of how the learned dynamics evolve globally across the state space. This combination of region-wise analytical tractability and explicit inter-region transitions makes PL-RNNs attractive for turning time-series reconstruction into an interpretable dynamical representation.

A key obstacle, however, is that a conventional PL-RNN with many nonlinear units can partition its state space into an enormous number of linear regions, making the resulting symbolic dynamics difficult to interpret. Almost-linear RNNs (AL-RNNs) address this problem by restricting the ReLU nonlinearity to only P units while keeping the remaining recurrent state variables linear (Brenner et al., 2024). The parameter P therefore provides explicit control over the nonlinear complexity of the model. By training AL-RNNs over different values of P, previous work showed that surprisingly few nonlinear units can suffice to reconstruct complex chaotic systems and recover compact piecewise-linear descriptions of their dynamics (Brenner et al., 2024). More recently, AL-RNNs have also been used as a general probe for systematically identifying when recurrence nonlinearity is computationally necessary and what mechanisms it enables across sequence-modeling tasks (Brenner & Koppe, 2026). Together, these results establish AL-RNNs as a framework not only for reconstructing interpretable piecewise-linear dynamics, but also for controlling and dissecting the functional role of nonlinearity in recurrent computation.

Yet controlling the amount of nonlinearity does not by itself reveal how dynamical representations at different levels of complexity are structurally related, nor does it guarantee that a minimally nonlinear model can be reliably learned. Existing AL-RNN approaches primarily address *how much* nonlinearity is needed by independently training models with different values of P (Brenner et al., 2024; Brenner & Koppe, 2026). Consequently, there is no explicit correspondence between the linear regions or transitions of a successful higher-capacity model and those of a simpler one. More importantly, reducing P also makes optimization increasingly difficult near the representational capacity boundary: a compact architecture may be theoretically sufficient to represent the target dynamics yet rarely be discovered by direct training. Thus, simply searching over progressively smaller models conflates two distinct problems: determining the minimal dynamical structure and successfully optimizing a model that realizes it. Rather than discarding an accurately trained higher-capacity model and repeatedly restarting optimization at smaller P, we ask whether the large model itself can serve as a scaffold from which simpler representations can be identified and subsequently realized. Specifically, *can a higher-capacity AL-RNN reveal a hierarchy of increasingly compact dynamical representations, and can this hierarchy guide the reliable training of a minimal realization?*

We show that fixed points provide a principled constraint for this reduction from redundancy to minimality. Each visited linear region can be classified according to whether its affine dynamics contain a fixed point; we refer to these two classes as *fixed-point (FP) symbols* and *fixed-point-free (FP-free) symbols*, respectively. Under standard nondegeneracy conditions, distinct isolated fixed points of the reconstructed system require distinct FP symbols. Thus, for a system with Q fixed points, any faithful piecewise-linear representation requires at least Q FP symbols. These FP symbols form an irreducible fixed-point skeleton: merging regions associated with distinct fixed points necessarily destroys the fixed-point structure. FP-free symbols, in contrast, are not protected by this lower bound and therefore constitute candidates for coarse-graining, although some may still be required to reproduce global transitions.

Building on this structure, we develop a fixed-point-guided hierarchical reduction procedure for trained AL-RNNs. We represent the learned dynamics as a transition graph whose nodes are visited ReLU activation patterns and whose edges record transitions between the corresponding linear regions. For sufficiently resolved trajectories, transitions between adjacent regions are associated with individual ReLU switches. Linearizing one ReLU removes the boundaries induced by that unit and therefore merges the corresponding regions in the symbolic graph. We systematically enumerate distinct sequences of such linearizations, while terminating any branch that would merge distinct FP

symbols. The result is a *reduction tree*: its root is the original high-resolution transition graph, and its descendants represent progressively coarser dynamical descriptions.

Importantly, a reduced graph specifies a candidate low-complexity dynamical structure, but does not by itself establish that the corresponding reduced AL-RNN can reproduce the target dynamics. The reduction tree records which ReLU units must be linearized to realize each candidate structure, allowing us to construct the corresponding reduced AL-RNN directly from its higher-capacity parent. We then retrain these models from the parent parameters under guidance from the parent dynamics and test whether the graph-derived structures can be realized as faithful autonomous dynamics. This learn–reduce–retrain procedure turns the higher-capacity AL-RNN into a scaffold not only for identifying hierarchical and minimal symbolic representations, but also for realizing compact models that can be difficult to learn directly.

Prior AL-RNN work (Brenner et al., 2024) demonstrated that direct low-capacity training can robustly recover topologically minimal representations of the Lorenz-63 (Lorenz, 1963) and Rössler (Rössler, 1976) systems. Here, we focus on the more challenging 3-scroll Chua system (Tang et al., 2001), whose five target fixed points imply lower bounds of five visited symbols and three ReLU units. At this theoretical minimum of three ReLU units, direct training produces high-fidelity minimal realizations in only 20% of seeds. In contrast, hierarchical reduction of reliably learned higher-capacity models frequently reaches both the symbol-level and nonlinear-capacity lower bounds. After reduction-guided retraining, the complete pipeline achieves approximately 71% seed-macro success for high-fidelity minimal realization at the same final nonlinear capacity, a $3.5\times$ improvement over direct minimal-capacity training. These results show that redundant nonlinear capacity can serve as both a structural and an optimization scaffold for discovering and realizing minimal dynamical representations.

Our main contributions are:

- We derive a fixed-point lower bound on symbolic complexity: reproducing Q distinct target fixed points requires at least Q FP symbols, yielding a certificate of symbol-level minimality when all target fixed points are preserved and exactly this many symbols are visited.
- We introduce a fixed-point-guided hierarchical reduction procedure that reveals progressively coarser representations while preserving distinct FP symbols.
- We show that higher-capacity AL-RNNs can serve as scaffolds for minimal-model learning: the learn–reduce–retrain pipeline substantially improves the reliability of high-fidelity minimal realization over direct minimal-capacity training.

2 AL-RNNs and Fixed-Point Structure

We first review the AL-RNN formulation and its symbolic representation following Brenner et al. (2024), and then derive the fixed-point constraints that underlie our reduction method.

2.1 AL-RNNs and symbolic dynamics

We consider the Almost-Linear Recurrent Neural Network (AL-RNN) (Brenner et al., 2024),

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

where $\mathbf{z}_t \in \mathbb{R}^M$, $\mathrm{ReLU}(u) := \max(0, u)$, and $\boldsymbol{\phi}_P$ applies ReLU nonlinearities only to the last P coordinates,

$$\boldsymbol{\phi}_P(\mathbf{z}) = [z_1, \ldots, z_{M-P}, \mathrm{ReLU}(z_{M-P+1}), \ldots, \mathrm{ReLU}(z_M)]^\top. \tag{2}$$

Thus, P controls the number of nonlinear switching variables while the latent dimension M remains fixed.

Following the identity teacher-forcing formulation of Brenner et al. (2024), the first N linear coordinates serve directly as readout units for the N-dimensional observed system. The model output is therefore

$$\hat{\mathbf{x}}_t = \mathcal{I}\mathbf{z}_t, \qquad \mathcal{I} = [\mathbf{I}_N \quad \mathbf{0}] \in \mathbb{R}^{N\times M}, \tag{3}$$

so that the readout units are part of the recurrent state rather than a separate output layer. The latent state therefore consists of N linear readout units, $M-N-P$ additional linear units, and P nonlinear ReLU units.

The activation state of the nonlinear units defines a binary symbol

$$\boldsymbol{\sigma}(\mathbf{z}) = \big(\mathbb{I}[z_{M-P+1} > 0], \ldots, \mathbb{I}[z_M > 0]\big) \in \{0,1\}^P, \tag{4}$$

which defines a linear region $\Omega_{\boldsymbol{\sigma}}$. Within this region, the dynamics are exactly affine,

$$\mathbf{z}_t = \mathbf{F}_{\boldsymbol{\sigma}} \mathbf{z}_{t-1} + \mathbf{h}, \qquad \mathbf{F}_{\boldsymbol{\sigma}} = \mathbf{A} + \mathbf{W}\mathbf{D}_{\boldsymbol{\sigma}}, \tag{5}$$

where

$$\mathbf{D}_{\boldsymbol{\sigma}} = \mathrm{diag}(\underbrace{1, \ldots, 1}_{M-P}, \sigma_1, \ldots, \sigma_P) \in \mathbb{R}^{M \times M}$$

is the diagonal activation matrix, with the first $M - P$ entries fixed to one for the linear coordinates.

For a generated trajectory of length T, let

$$\Sigma = \{\boldsymbol{\sigma}(\mathbf{z}_t) : 1 \le t \le T\}$$

denote the set of visited linear regions, which we refer to as *symbols*. Figure 1(a) schematically illustrates this piecewise-linear partition of the latent state space and a trajectory moving across the resulting regions.

The symbolic trajectory induces a directed transition graph $G = (\Sigma, \mathcal{E})$, with an edge for each observed transition between symbols, as illustrated in figure 1(b).

We measure symbolic complexity by the number of visited symbols, $|\Sigma|$. We refer to a representation attaining the minimum possible $|\Sigma|$ while reproducing the target dynamics as a *minimal piecewise-linear representation*. This notion of minimality concerns the number of visited linear regions rather than the number of nonlinear units P.

2.2 Fixed-Point Theory

For a linear region $\Omega_{\boldsymbol{\sigma}}$, if $\mathbf{I} - \mathbf{F}_{\boldsymbol{\sigma}}$ is nonsingular, its unique candidate fixed point is

$$\mathbf{z}^*_{\boldsymbol{\sigma}} = (\mathbf{I} - \mathbf{F}_{\boldsymbol{\sigma}})^{-1} \mathbf{h}. \tag{6}$$

We therefore define

$$\Sigma_{\mathrm{FP}} := \{\boldsymbol{\sigma} \in \Sigma \mid \mathbf{z}^*_{\boldsymbol{\sigma}} \in \Omega_{\boldsymbol{\sigma}}\}, \tag{7}$$

$$\Sigma_{\mathrm{free}} := \Sigma \setminus \Sigma_{\mathrm{FP}}. \tag{8}$$

Symbols in Σ_{FP} are called *FP symbols*, whereas those in Σ_{free} are called *FP-free symbols*. Figure 1(a) illustrates the distinction between FP and FP-free symbols according to whether the corresponding linear region contains a valid fixed point.

Let Q denote the number of distinct fixed points of the target dynamical system that the learned piecewise-linear representation is required to reproduce. Here, "preserving" these fixed points means that the reconstructed dynamics contain Q distinct fixed points corresponding to those of the target system.

Assuming that these fixed points are isolated and lie in the interiors of their corresponding regions, together with the nonsingularity condition above, each linear region contains at most one valid fixed point. It follows that:

Proposition 1 (Fixed-point lower bound)**.** *Suppose that a piecewise-linear representation reproduces Q distinct target fixed points and that each of these fixed points lies in a visited region, so that its corresponding symbol belongs to Σ_{FP}. Then*

$$|\Sigma| \ge |\Sigma_{\mathrm{FP}}| \ge Q. \tag{9}$$

The proof and further discussion of the consequences and scope of this bound are provided in AppendixA.

Therefore, if a representation reproducing the same Q target fixed points attains

$$|\Sigma| = Q, \tag{10}$$

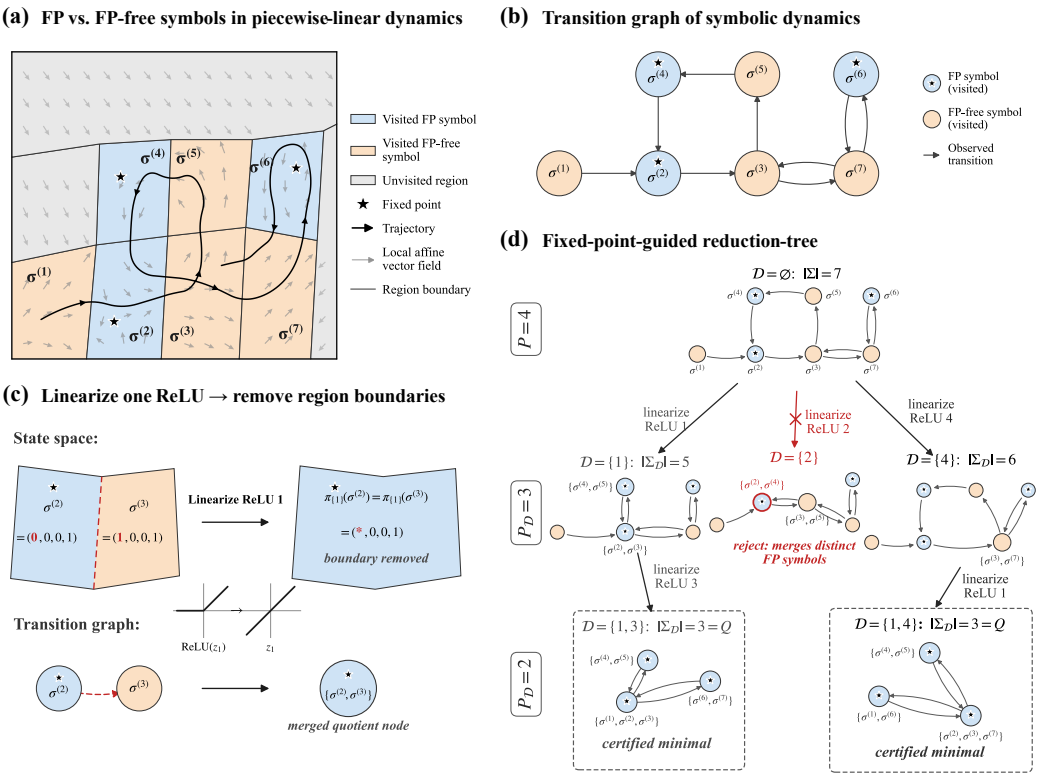


Figure 1: Proposed fixed-point-guided hierarchical reduction. (a) FP and FP-free symbols in the piecewise-linear state-space partition. (b) Symbolic transition graph constructed from visited ReLU activation patterns. (c) Region merging induced by ReLU linearization. (d) Reduction-tree construction under the fixed-point-preservation constraint.

then it is minimal in the number of visited linear regions among all piecewise-linear representations preserving this fixed-point structure. The lower bound need not be tight, however: additional FP-free regions may still be required to organize the global flow. Our reduction procedure therefore preserves the target fixed-point structure while testing constructively which remaining symbolic regions can be removed.

The fixed-point count also yields a necessary lower bound on the number of nonlinear units P. Because P ReLUs can generate at most 2^P activation patterns,

$$P \geq \lceil \log_2 Q \rceil \,. \tag{11}$$

This condition is necessary but not sufficient: having enough activation patterns does not guarantee that training will realize the required fixed points or reproduce the target dynamics.

3 Methods: Fixed-Point-Guided Hierarchical Reduction and Retraining

Our framework follows a simple strategy: *learn large, reduce structurally, and retrain compactly*. We first train a higher-capacity AL-RNN that reliably reconstructs the target dynamics. From its symbolic transition graph, we then construct a hierarchy of progressively reduced representations by linearizing ReLU units while preventing distinct FP symbols from being merged. Finally, each selected reduced graph is mapped back to an AL-RNN and retrained from the parent parameters. Thus, graph reduction identifies structurally admissible compact representations, while retraining determines whether they can be realized as actual dynamical systems.

3.1 FIXED-POINT-GUIDED GRAPH REDUCTION

The reduction does not assume that fixed points are provided with the observed time series. Once a parent AL-RNN has been trained, candidate fixed points can be computed analytically from its piecewise-affine dynamics using equation 6. We therefore construct each reduction hierarchy directly from the fixed-point structure recovered by the parent model, rather than from ground-truth fixed-point annotations or a prescribed fixed-point count.

Let $G = (\Sigma, \mathcal{E})$ be the symbolic transition graph of a trained parent AL-RNN, as defined in Sec. 2.1, and let P_{parent} denote the number of nonlinear ReLU units in the parent model. Each node is a visited activation pattern $\boldsymbol{\sigma} \in \{0,1\}^{P_{\text{parent}}}$, and is classified as an FP or FP-free symbol according to $\boldsymbol{\sigma} \in \Sigma_{\text{FP}}$ or $\boldsymbol{\sigma} \in \Sigma_{\text{free}}$, respectively (figure 1(a,b)).

For interpreting graph edges in terms of individual ReLU switches, we assume that the trajectory is sufficiently resolved such that transitions typically cross a single switching boundary. Empirically, 97.1% of observed symbol transitions ($7{,}024/7{,}233$) across the 30 $P_{\text{parent}} = 10$ Chua parents changed exactly one ReLU activation bit, consistent with this assumption.

We reduce the representation by replacing selected ReLU activations by the identity,

$$\text{ReLU}(z_j) \to z_j. \tag{12}$$

This operation leaves the latent dimension unchanged but removes the switching boundary associated with ReLU unit j, thereby allowing the corresponding symbolic regions to merge (figure 1(c)).

Let $\mathcal{D}$ denote the index set of ReLU units that have been linearized. Removing the activation bits associated with $\mathcal{D}$ defines a projection

$$\pi_{\mathcal{D}} : \{0,1\}^{P_{\text{parent}}} \to \{0,1\}^{P_{\text{parent}}-|\mathcal{D}|}.$$

The projected symbols $\Sigma_{\mathcal{D}} := \pi_{\mathcal{D}}(\Sigma)$ form the node set of the quotient graph $G_{\mathcal{D}} = (\Sigma_{\mathcal{D}}, \mathcal{E}_{\mathcal{D}})$, whose transitions are obtained by projecting both endpoints of each parent transition in $\mathcal{E}$ through $\pi_{\mathcal{D}}$.

For two original symbols $\boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \Sigma$, if

$$\pi_{\mathcal{D}}(\boldsymbol{\sigma}) = \pi_{\mathcal{D}}(\boldsymbol{\sigma}') = \tilde{\boldsymbol{\sigma}},$$

they are merged into the same node $\tilde{\boldsymbol{\sigma}} \in \Sigma_{\mathcal{D}}$ of $G_{\mathcal{D}}$.

The fixed-point theory in Sec. 2.2 provides the reduction constraint: for all distinct FP symbols $\boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \Sigma_{\text{FP}}$, we require

$$\pi_{\mathcal{D}}(\boldsymbol{\sigma}) \neq \pi_{\mathcal{D}}(\boldsymbol{\sigma}'). \tag{13}$$

Starting from $\mathcal{D} = \varnothing$, we enumerate admissible single-ReLU linearizations and terminate branches violating equation 13. Duplicate reductions are removed. The resulting *reduction tree* connects the original high-resolution graph to progressively coarser symbolic representations while recording the exact ReLU units linearized along each path (figure 1(d)).

We characterize each reduction by the remaining nonlinear capacity $P_{\mathcal{D}}$ and symbolic complexity $|\Sigma_{\mathcal{D}}|$. Candidate-selection criteria are fixed before retraining and depend only on structural information. For the 3-scroll Chua experiments, we focus on candidates attaining both theoretical lower bounds in section 2.2, $|\Sigma_{\mathcal{D}}| = 5$ and $P_{\mathcal{D}} = 3$. Further details on reduction-tree enumeration, computational cost, and candidate selection are provided in Appendix B.

3.2 REDUCTION-GUIDED RETRAINING

A quotient graph specifies which nonlinear switching boundaries are removed, but does not guarantee that the corresponding reduced AL-RNN will realize the target dynamics. For each selected reduction with linearization set $\mathcal{D}$, we initialize the corresponding reduced AL-RNN from the trained parent model and permanently replace the ReLU units in $\mathcal{D}$ by linear activations before retraining.

During retraining, the autonomous parent trajectory serves as the teacher. Rather than supervising only the observed readout coordinates, we use the full latent state of the parent as the teacher signal,

with equal weight per coordinate. This transfers the dynamical realization learned by the higher-capacity parent while the linearization mask fixes the reduced nonlinear structure.

Analogous to the sparse teacher-forcing scheme used in standard AL-RNN training (Brenner et al., 2024), we also periodically synchronize the reduced-model state during training. Here, the reference state is taken from the autonomous parent trajectory rather than from the observed data.

Implementation details on reduced-model initialization, teacher-trajectory construction, phase alignment, synchronization, and retraining objectives are provided in Appendices C.3–C.7.

3.3 Evaluation

Each model is evaluated using an autonomous rollout of $10{,}000$ steps, with the first $1{,}000$ steps discarded as transient. We define the post-transient visited-symbol set as

$$\widehat{\Sigma} = \{\boldsymbol{\sigma}(\mathbf{z}_t) : 1001 \leq t \leq 10000\},$$

so that $|\widehat{\Sigma}|$ measures the symbolic complexity actually realized by the long-run autonomous dynamics.

Let Q_{vis} denote the number of analytically admissible fixed points of the learned model whose corresponding symbolic regions are visited after the transient. We define *minimal-realization success* as

$$\mathcal{S}_{\min} = \mathbb{I}\left[Q_{\mathrm{vis}} = Q \ \wedge \ |\widehat{\Sigma}| = Q\right]. \tag{14}$$

Thus, success requires the learned dynamics to realize Q visited admissible fixed points, together with a post-transient symbolic realization that attains the fixed-point lower bound. For the 3-scroll Chua system considered here, $Q = 5$.

We additionally assess continuous-state fidelity using the state-space distance E_{stsp}, which measures discrepancy between the generated and target attractor distributions. As a stringent high-fidelity criterion, we define

$$\mathcal{S}_{\mathrm{fid}} = \mathbb{I}\left[\mathcal{S}_{\min} = 1 \ \wedge \ E_{\mathrm{stsp}} < 2.0\right]. \tag{15}$$

The threshold $E_{\mathrm{stsp}} < 2.0$ is used as a conservative representative criterion; sensitivity to this choice is examined in Appendix D. As a complementary diagnostic of temporal fidelity, we also evaluate the spectral Hellinger distance E_{H}, but do not include it in the definition of $\mathcal{S}_{\mathrm{fid}}$.

We use 30 independent random seeds for each training configuration. Because different parent models can yield different numbers of reduced graphs, the source parent seed is treated as the primary statistical unit for the reduction pipeline. Full metric definitions, threshold-sensitivity analyses, and candidate-level statistics are provided in Appendix D.

4 Experiments and Results

We evaluate the proposed framework on the 3-scroll Chua system. Throughout the experiments, we use a latent dimension of $M = 20$ and an observed/readout dimension of $N = 3$; data-generation and training configurations are detailed in Appendices C.1-C.2.

4.1 Minimal nonlinear capacity is difficult to optimize directly

The 3-scroll Chua system has five target fixed points. The lower bound in equation 11 therefore gives $P \geq \lceil \log_2 5 \rceil = 3$, making $P = 3$ the smallest nonlinear capacity permitted by the fixed-point count.

Direct training becomes increasingly reliable as P increases (figure 2(a)). At the capacity boundary $P = 3$, only $7/30$ independently trained models achieve minimal realization, $\mathcal{S}_{\min} = 1$. Of these, $6/30 = 20\%$ also satisfy the high-fidelity criterion, $\mathcal{S}_{\mathrm{fid}} = 1$. Thus, although the target dynamics can be realized with only three nonlinear units, optimization at this minimal capacity is unreliable.

In contrast, higher-capacity $P = 10$ models recover the five-fixed-point structure in all $30/30$ runs. These results reveal a clear gap between the existence of a minimal-capacity representation and the reliability with which it can be learned directly.

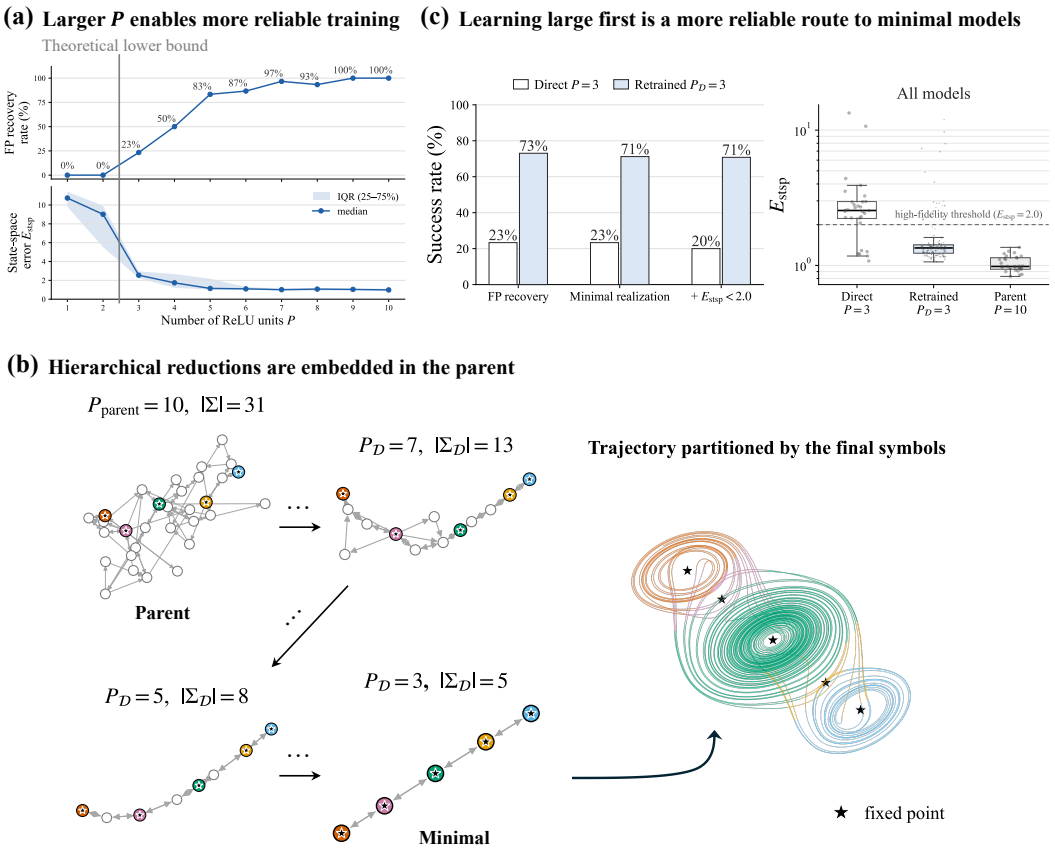


Figure 2: Main results for the 3-scroll Chua system. (a) Direct training becomes more reliable as the number of ReLU units P increases. (b) Hierarchical reduction reaches $|\Sigma_{\mathcal{D}}| = 5$ and $P_{\mathcal{D}} = 3$, attaining both the symbol-level and nonlinear-capacity lower bounds. (c) Learning large first provides a more reliable route to minimal models than direct training at the minimal capacity: seed-macro success rates are substantially higher after reduction and retraining (left), while the resulting $P_{\mathcal{D}} = 3$ models also exhibit substantially lower E_{stsp} than directly trained $P = 3$ models (right). The $P = 10$ parents serve as a reference for the geometric fidelity attainable before reduction.

4.2 A HIERARCHY OF REDUCED REPRESENTATIONS IS EMBEDDED IN THE PARENT

We use the successfully trained $P = 10$ models as parents for the hierarchical reduction procedure. Reduction yields not only a final compact graph but a hierarchy of intermediate quotient graphs. Different linearization paths progressively merge neighboring symbolic regions while preserving all distinct FP symbols, thereby providing explicit structural paths from the higher-capacity parent to progressively coarser representations (figure 2(b)).

This hierarchy frequently reaches the theoretical lower bounds. Among the 30 parent models with $P_{\text{parent}} = 10$, $26/30$ contain at least one reduced graph with $|\Sigma_{\mathcal{D}}| = 5$, thereby attaining the fixed-point lower bound and certifying symbol-level minimality of the quotient representation. Moreover, $25/30$ contain at least one reduced graph with both $|\Sigma_{\mathcal{D}}| = 5$ and $P_{\mathcal{D}} = 3$, thereby attaining both the symbolic-complexity and nonlinear-capacity lower bounds. The remaining parents cannot always be reduced to both bounds, indicating that the minimal structure accessible through reduction depends on the representation learned by the parent.

4.3 LEARNING LARGE FIRST PROVIDES A MORE RELIABLE ROUTE TO MINIMAL MODELS

We next test whether the structurally minimal quotient candidates identified by hierarchical reduction can be realized as actual dynamical models, and whether this learn–reduce–retrain strategy

provides a more reliable route to minimal dynamics than direct minimal-capacity training. Each selected candidate is retrained using the reduction-guided protocol described in Section 3.2.

To account for the different numbers of reduced candidates generated by each parent, we first average each success indicator across the selected candidates within a parent seed and then average across the 30 parent seeds. Selected candidates that fail during training are counted as failures, and parents producing no reduced graph with $|\Sigma_{\mathcal{D}}| = 5$ and $P_{\mathcal{D}} = 3$ are assigned zero success. Unless stated otherwise, the main comparison is restricted to the capacity-matched subset with $P_{\mathcal{D}} = 3$ (126 of the 141 selected five-symbol candidates), so that the comparison against direct $P = 3$ training is made at equal nonlinear capacity.

As shown in figure 2(c, left), the seed-macro evaluation yields FP-recovery, minimal-realization, and high-fidelity minimal-realization success rates of approximately 73%, 71%, and 71% for reduction-guided retraining, compared with 23%, 23%, and 20% for direct $P = 3$ training, respectively. Thus, the learn–reduce–retrain pipeline increases the high-fidelity minimal-realization success rate by approximately $3.5\times$ at the same final nonlinear capacity.

The advantage is also evident in the geometric fidelity of the reconstructed attractor (figure 2(c, right)). Across the evaluated models, the retrained $P_{\mathcal{D}} = 3$ models exhibit substantially lower E_{stsp} than directly trained $P = 3$ models, approaching the geometric fidelity of the higher-capacity $P = 10$ parents.

Notably, high-fidelity minimal realizations need not reproduce the quotient transition graph exactly. Additional structural results, retraining ablations, fidelity-threshold analyses, and transition-topology reorganization are provided in Appendix E.

5 Discussion and Conclusion

Our results show that higher-capacity AL-RNNs can serve as scaffolds for minimal dynamical modeling. Rather than optimizing a minimal-capacity model directly, we first learn a higher-capacity AL-RNN, reduce its symbolic structure, and retrain the resulting compact model under guidance from the parent dynamics. This learn–reduce–retrain strategy separates structural simplification from optimization at the capacity boundary.

Beyond identifying minimal representations, the reduction tree records how successive ReLU linearizations remove switching boundaries and merge symbolic regions, while rejecting reductions that would merge distinct FP symbols. This provides an explicit hierarchy from the parent representation to progressively coarser dynamical descriptions. Interestingly, successful retraining need not reproduce the quotient transition graph exactly. Even when a quotient candidate is non-path-like, retraining can reorganize the realized transitions into the five-node path-like symbolic organization observed in high-fidelity Chua realizations. This suggests that reduction provides a structural scaffold rather than fixing the final transition topology.

The fixed-point criterion provides a lower bound rather than a complete characterization of minimal dynamics. FP-free regions may still be required to organize the global flow. Because reduction preserves the fixed-point structure recovered by the parent, its effectiveness depends on the parent having correctly captured the relevant fixed points of the target system. Moreover, not every successful parent can be reduced to both the symbolic-complexity and nonlinear-capacity lower bounds, indicating that reducibility depends on the representation learned by the parent.

Computationally, exhaustive reduction scales combinatorially with P_{parent}, but remains inexpensive in the regime considered here: full tree construction took a median of 0.2 s at $P_{\mathrm{parent}} = 10$ and 10.2 s at $P_{\mathrm{parent}} = 14$ (Appendix B.4). Moreover, exhaustive enumeration is not required to obtain a compact realization, and the combinatorial dependence is on the nonlinear switching variables P rather than directly on the ambient dimension.

Our empirical study focuses on the 3-scroll Chua system, whose five fixed points impose comparatively strong lower bounds on both symbolic complexity and nonlinear capacity, while direct training at the capacity boundary remains unreliable. Extending the framework to systems with different attractor organizations, as well as to noisy and partially observed real-world data, remains an important direction for future work.

AI USE STATEMENT

We used generative AI tools, including OpenAI ChatGPT and Anthropic Claude, to assist with manuscript drafting and editing, literature retrieval and discovery, discussion and refinement of theoretical and methodological formulations, experimental design, and software implementation and debugging. We also used these tools to assist with the formulation and refinement of mathematical claims and derivations, and with the preparation and refinement of scientific figures.

The authors critically reviewed all AI-assisted content. Mathematical claims and derivations were independently checked by the authors, and AI-assisted code was executed, tested, and validated against the intended methodology. Experimental results reported in the paper were obtained from the authors' implementations and experiments rather than generated by AI, and cited literature was checked against the original sources. The authors take full responsibility for the final content of this work, including all text, claims, figures, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed definitions of the AL-RNN architecture, the fixed-point-guided reduction procedure, reduction-guided retraining, and the evaluation criteria in the main text and appendix. The appendix documents the experimental settings, hyperparameters, random-seed protocol, fixed-point computation and evaluation procedures, and other implementation details needed to reproduce the reported experiments, together with additional analyses and derivations supporting the theoretical claims. Anonymized source code and experiment scripts for reproducing all reported results, covering training, hierarchical reduction, retraining, evaluation, and figure generation, are available at `https://anonymous.4open.science/r/alrnn-fp-guided-reduction-DB32/`. The repository additionally contains the dataset and all trained model checkpoints, so that every reported figure and statistic can be regenerated directly from the released models without retraining.

ACKNOWLEDGMENTS

Omitted for anonymous review.

REFERENCES

Manuel Brenner and Georgia Koppe. Uncovering the computational roles of nonlinearity in sequence modeling using almost-linear RNNs. *Transactions on Machine Learning Research*, 2026.

Manuel Brenner, Christoph Jürgen Hemmer, Zahra Monfared, and Daniel Durstewitz. Almost-linear RNNs yield highly interpretable symbolic codes in dynamical systems reconstruction. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 36829–36868. Curran Associates, Inc., 2024. doi: 10.52202/079017-1161.

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*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116 (45):22445–22451, 2019. doi: 10.1073/pnas.1906995116.

Daniel Durstewitz. A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements. *PLOS Computational Biology*, 13(6): e1005542, 2017. doi: 10.1371/journal.pcbi.1005542.

Daniel Durstewitz, Georgia Koppe, and Max Ingo Thurm. Reconstructing computational system dynamics from neural data with recurrent neural networks. *Nature Reviews Neuroscience*, 24 (11):693–710, 2023. doi: 10.1038/s41583-023-00740-7.

Matthew D. Golub and David Sussillo. Fixedpointfinder: A tensorflow toolbox for identifying and characterizing fixed points in recurrent neural networks. *Journal of Open Source Software*, 3(31): 1003, 2018. doi: 10.21105/joss.01003.

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.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2): 130–141, 1963. doi: 10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2.

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 Dynamics*, 101:1037–1052, 2020. doi: 10.1007/s11071-020-05841-x.

Otto E. Rössler. An equation for continuous chaos. *Physics Letters A*, 57(5):397–398, 1976. doi: 10.1016/0375-9601(76)90101-8.

Jimmy T. H. Smith, Scott W. Linderman, and David Sussillo. Reverse engineering recurrent neural networks with jacobian switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16700–16713. Curran Associates, Inc., 2021.

David Sussillo and Omri Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649, 2013. doi: 10.1162/NECO_a_00409.

Wallace K. S. Tang, G. Q. Zhong, Guanrong Chen, and K. F. Man. Generation of N-scroll attractors via sine function. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 48(11):1369–1372, 2001. doi: 10.1109/81.964432.

A Additional Theory and Proofs

A.1 Assumptions and proof of the fixed-point lower bound

We assume that the target fixed points considered in the analysis are isolated and lie in the interiors of their corresponding linear regions. We further assume that $\mathbf{I} - \mathbf{F}_{\boldsymbol{\sigma}}$ is nonsingular for every relevant region. These conditions exclude fixed points lying exactly on switching boundaries and affine subsystems containing continua of fixed points.

Lemma 1 (At most one fixed point per region)**.** *Under these assumptions, each linear region $\Omega_{\boldsymbol{\sigma}}$ contains at most one fixed point.*

Proof. Within a fixed region, a fixed point $\mathbf{z}^*$ satisfies

$$(\mathbf{I} - \mathbf{F}_{\boldsymbol{\sigma}})\mathbf{z}^* = \mathbf{h}.$$

Because $\mathbf{I} - \mathbf{F}_{\boldsymbol{\sigma}}$ is nonsingular, this linear system has a unique solution. Hence the region contains at most one valid fixed point. □

Proof of Proposition 1. By assumption, each of the Q target fixed points lies in a visited region, and hence its corresponding symbol belongs to Σ_{FP}. By Lemma 1, no two distinct target fixed points can occupy the same linear region. They therefore require at least Q distinct FP symbols:

$$|\Sigma_{\mathrm{FP}}| \geq Q.$$

Since $\Sigma_{\mathrm{FP}} \subseteq \Sigma$,

$$|\Sigma| \geq |\Sigma_{\mathrm{FP}}| \geq Q.$$

□

A.2 Consequences for symbolic and nonlinear complexity

If a representation reproducing the same Q target fixed points satisfies

$$|\Sigma| = Q,$$

then no representation preserving these fixed points can contain fewer visited linear regions. It is therefore minimal in symbolic complexity under the stated fixed-point-preservation requirement.

This certificate concerns the number of visited regions rather than the number of nonlinear units. Because P ReLU units admit at most 2^P activation patterns, preserving Q distinct FP symbols also requires

$$P \geq \lceil \log_2 Q \rceil .$$

This is a necessary but not sufficient condition: meeting this bound does not guarantee that the required fixed points or global dynamics can be realized or learned.

A.3 Scope and possible non-tightness of the certificate

The fixed-point lower bound specifies a minimum number of symbolic regions required by the fixed-point structure, but it does not imply that this number is sufficient to reproduce the full attractor dynamics. Additional FP-free regions may still be required to organize the global flow.

Consequently, the absence of a fixed point does not imply that a symbol can be removed. Our reduction procedure therefore preserves the fixed-point skeleton while exploring which additional FP-free regions can be merged. The corresponding reduced models are then retrained and evaluated to determine whether they can realize the target dynamics faithfully.

Importantly, a quotient graph accepted by the fixed-point-preservation constraint is only a structurally admissible candidate. Retraining may introduce or eliminate visited symbolic regions, so the resulting autonomous dynamics must be evaluated separately. Whether the bound is tight must therefore be determined through retraining and autonomous dynamical evaluation.

B HIERARCHICAL REDUCTION DETAILS

B.1 RELU LINEARIZATION AND SYMBOL PROJECTION

Let $\boldsymbol{\sigma} \in \{0,1\}^{P_{\text{parent}}}$ denote an original activation pattern and let $\mathcal{D}$ be the index set of linearized ReLU units. For each $j \in \mathcal{D}$, we replace

$$\text{ReLU}(z_j) \longrightarrow z_j. \tag{A1}$$

This operation does not remove a latent-state dimension. Instead, it removes the switching boundary induced by the jth ReLU while preserving the full latent dimension M.

The set $\mathcal{D}$ induces the projection

$$\pi_{\mathcal{D}} : \{0,1\}^{P_{\text{parent}}} \to \{0,1\}^{P_{\text{parent}}-|\mathcal{D}|}, \tag{A2}$$

which removes all activation-pattern coordinates indexed by $\mathcal{D}$. Two original symbols are therefore equivalent after the linearization if

$$\boldsymbol{\sigma} \sim_{\mathcal{D}} \boldsymbol{\sigma}' \iff \pi_{\mathcal{D}}(\boldsymbol{\sigma}) = \pi_{\mathcal{D}}(\boldsymbol{\sigma}'). \tag{A3}$$

The nodes of the reduced quotient graph are the resulting equivalence classes.

The reduced transition set is

$$\mathcal{E}_{\mathcal{D}} := \{(\pi_{\mathcal{D}}(\boldsymbol{\sigma}), \pi_{\mathcal{D}}(\boldsymbol{\sigma}')) : (\boldsymbol{\sigma}, \boldsymbol{\sigma}') \in \mathcal{E}\}. \tag{A4}$$

Thus, the reduced transition graph is obtained directly by projecting both endpoints of each parent transition onto the retained activation bits. Self-transitions created by quotienting may be omitted from graph visualizations.

B.2 FP-SYMBOL SEPARATION CONSTRAINT

A reduction is structurally admissible only if no two distinct FP symbols are projected onto the same quotient node:

$$\forall\, \boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \Sigma_{\text{FP}}, \quad \boldsymbol{\sigma} \neq \boldsymbol{\sigma}', \qquad \pi_{\mathcal{D}}(\boldsymbol{\sigma}) \neq \pi_{\mathcal{D}}(\boldsymbol{\sigma}'). \tag{A5}$$

This condition allows an FP-free symbol to merge with another FP-free symbol or with a single FP symbol, but rejects any reduction that would merge two distinct FP symbols.

The condition protects only the fixed-point skeleton. It does not imply that every accepted quotient graph can be realized as a faithful reduced dynamical model. Dynamical realization is assessed separately through reduction-guided retraining and autonomous evaluation.

B.3 REDUCTION-TREE ENUMERATION

Starting from $\mathcal{D} = \varnothing$, we systematically consider each ReLU unit that has not yet been linearized. For a candidate ReLU unit j, we form $\mathcal{D}' = \mathcal{D} \cup \{j\}$ and compute the induced symbol projection $\pi_{\mathcal{D}'}$. If equation A5 is violated, the branch is terminated. Because further linearization can only remove additional activation bits, once two distinct FP symbols collide they cannot be separated again by any descendant reduction. Otherwise, the corresponding quotient graph is added to the reduction tree and further linearizations are considered.

Because different linearization orders can lead to the same set $\mathcal{D}$, each linearization set is stored only once. A concise pseudocode description is given below.

Input: $G = (\Sigma, \mathcal{E}),\ \Sigma_{\mathrm{FP}},\ P_{\mathrm{parent}}$

Initialize: $\mathcal{D} \leftarrow \varnothing, \quad \mathrm{Queue} \leftarrow \{\mathcal{D}\}$

while $\mathrm{Queue} \neq \varnothing$:

    pop $\mathcal{D}$

    for each $j \in \{1, \ldots, P_{\mathrm{parent}}\} \setminus \mathcal{D}$:

        $\mathcal{D}' \leftarrow \mathcal{D} \cup \{j\}$

        project all symbols with $\pi_{\mathcal{D}'}$

        if distinct FP symbols collide:

            reject $\mathcal{D}'$

        else:

            construct $G_{\mathcal{D}'}$

            if $\mathcal{D}'$ has not been visited:

                add $(\mathcal{D}', G_{\mathcal{D}'})$ as a child of $(\mathcal{D}, G_{\mathcal{D}})$

                push $\mathcal{D}'$

Output: tree of admissible $(\mathcal{D}, G_{\mathcal{D}})$ pairs,
with edges corresponding to single-ReLU linearizations.

B.4 Reduction complexity and candidate selection

For each linearization set $\mathcal{D}$, we characterize the reduced representation by two complementary quantities:

$$P_{\mathcal{D}} = P_{\mathrm{parent}} - |\mathcal{D}|, \qquad |\Sigma_{\mathcal{D}}|.$$

The first is the number of ReLU units remaining after linearization, whereas the second is the symbolic complexity of the quotient graph. These quantities need not attain their minima simultaneously.

Candidate selection is fixed before retraining and never uses post-retraining reconstruction performance. For the 3-scroll Chua system, the fixed-point theory gives a five-symbol lower bound and a three-ReLU lower bound. The principal capacity-boundary analysis therefore uses candidates satisfying

$$|\Sigma_{\mathcal{D}}| = 5, \qquad P_{\mathcal{D}} = 3.$$

Five-symbol candidates with larger $P_{\mathcal{D}}$ still attain the certified symbol-count lower bound, but do not attain the nonlinear-capacity lower bound.

Computational cost of reduction. Each reduction is uniquely specified by a subset $\mathcal{D}$ of the P_{parent} nonlinear units, so the number of distinct reduction masks is at most $2^{P_{\mathrm{parent}}}$. The exhaustive search therefore has a combinatorial worst-case dependence on P_{parent}, although branches are pruned once the fixed-point-separation constraint is violated. In practice, each reduction step consists only of symbolic projection, fixed-point-collision checks, and quotient-graph construction, and requires no gradient-based optimization.

To characterize the empirical scaling of the structural search, we measured exhaustive reduction-tree construction for $P_{\mathrm{parent}} = 10, \ldots, 14$, using 30 independently trained parent models for each value of P_{parent}. All timings were measured single-threaded on one thread of an Intel Core i9-10980XE (3.0 GHz) desktop CPU. The resulting tree sizes and wall-clock times are summarized in Table A1.

For the $P_{\mathrm{parent}} = 10$ setting used in the main experiments, exhaustive tree construction required only a median of 0.2 s per parent. As expected from the combinatorial growth of the mask space, the runtime increased rapidly with P_{parent}. Nevertheless, the absolute cost remained modest over the range examined: even for $P_{\mathrm{parent}} = 14$, corresponding to $2^{14} = 16{,}384$ possible masks, exhaustive construction required a median of only 10.2 s per parent. Thus, for the nonlinear capacities considered here, the structural reduction stage is computationally inexpensive despite its combinatorial worst-case scaling.

The exhaustive traversal used in this study is primarily an analysis choice rather than a requirement of the learn–reduce–retrain framework. We enumerate the full admissible hierarchy to characterize

Table A1: Empirical scaling of exhaustive reduction-tree construction. Values are medians over 30 parent models, with interquartile ranges in brackets.

P_{parent}	$2^{P_{\mathrm{parent}}}$	Admissible nodes	Tree time [s]
10	1,024	476 [382–586]	0.2 [0.1–0.3]
11	2,048	1,133 [861–1,270]	0.5 [0.4–0.7]
12	4,096	2,324 [1,936–2,685]	1.4 [1.2–1.7]
13	8,192	5,145 [3,554–5,498]	3.3 [2.7–4.3]
14	16,384	10,616 [8,543–11,779]	10.2 [8.1–11.5]

variability across parent representations and to support the consensus analysis in Appendix B.5. If the objective is simply to obtain a compact realization, the search can instead terminate after one or several candidates satisfying a prescribed structural criterion have been identified. Such targeted search avoids exploring the full reduction tree, although it does not provide the exhaustive candidate distribution required by the consensus estimator.

B.5 Consensus estimation of symbolic granularity

For an unknown system, the fixed-point structure of a trained parent model still provides a data-driven lower bound on symbolic complexity. For parent r, let $Q_{\mathrm{vis}}^{(r)}$ denote the number of analytically admissible fixed points whose symbolic regions are visited during autonomous dynamics. Any reduction preserving these fixed-point symbols must satisfy

$$|\Sigma_{\mathcal{D}}| \geq Q_{\mathrm{vis}}^{(r)}. \tag{A6}$$

This bound is parent-dependent and need not be tight, since additional FP-free regions may be required to organize the global flow. We therefore consider a simple consensus-based strategy for estimating the relevant symbolic granularity from reduction hierarchies across independently trained parents.

Our consensus procedure is conditioned on a prescribed reduced nonlinear capacity $P_{\mathcal{D}} = P_{\min}$. For each independently trained parent r, we collect all distinct admissible reduction nodes at this capacity. Rejected nodes, including reductions that violate the fixed-point preservation constraint, are excluded, and duplicate linearization masks are counted only once within each parent. Let $n_r(k)$ denote the number of such admissible nodes from parent r with symbolic complexity $|\Sigma_{\mathcal{D}}| = k$.

This procedure estimates symbolic granularity conditional on a chosen reduced nonlinear capacity $P_{\min}$; selecting $P_{\min}$ itself is a separate model-selection problem. The parent-inferred fixed-point count provides the necessary capacity lower bound

$$P_{\mathcal{D}} \geq \left\lceil \log_2 Q_{\mathrm{vis}}^{(r)} \right\rceil, \tag{A7}$$

but this bound need not identify the reduced capacity required for faithful realization of the full dynamics.

Directly pooling all reduction nodes would give disproportionate weight to parents whose reduction trees happen to contain more admissible branches. We therefore first normalize the vote distribution within each parent:

$$h_r(k) = \frac{n_r(k)}{\sum_i n_r(i)}. \tag{A8}$$

The parent-balanced consensus support for symbolic granularity k is then

$$H(k) = \frac{1}{|\mathcal{P}_{\mathrm{cons}}|} \sum_{r \in \mathcal{P}_{\mathrm{cons}}} h_r(k), \tag{A9}$$

where $\mathcal{P}_{\mathrm{cons}}$ is the set of parent models that contribute at least one admissible node to the $P_{\mathcal{D}} = P_{\min}$ slice. The consensus estimate is

$$\widehat{k} = \arg\max_k H(k), \tag{A10}$$

and we define the corresponding consensus strength as

$$C = H(\widehat{k}). \tag{A11}$$

This construction gives each independently trained parent equal total weight, irrespective of the number of admissible descendants it generates. Hence, $\widehat{k}$ identifies the symbolic granularity that is most consistently supported across independently learned parent representations, rather than the granularity that merely occurs most often in the pooled reduction tree.

When the consensus distribution $H(k)$ is sharply concentrated, we retain the dominant layer $|\Sigma_{\mathcal{D}}| = \widehat{k}$ for subsequent candidate selection. When the distribution is diffuse, we additionally inspect neighboring well-supported symbolic layers as a robustness analysis rather than interpreting the mode as a sharp estimate of the true minimum.

For the 3-scroll Chua reduction hierarchy at $P_{\mathcal{D}} = 3$, the consensus mode is

$$\widehat{k} = 5, \qquad H(5) = 0.370, \qquad H(6) = 0.221, \qquad H(7) = 0.320. \tag{A12}$$

The support is relatively diffuse, with substantial mass also assigned to $k = 6$ and $k = 7$. Together, the $k = 5, 6, 7$ layers account for 91.1% of the parent-balanced consensus support. For the 3-scroll Chua system, the modal estimate $\widehat{k} = 5$ can be independently validated by the known five-symbol lower bound. For an unknown system, however, such an external reference would not generally be available. When the consensus distribution is similarly diffuse, it is therefore useful to retain other highly supported symbolic granularities for subsequent retraining and dynamical evaluation.

The consensus estimate $\widehat{k}$ should be interpreted as an empirical estimate of compact symbolic granularity rather than a certificate of global minimality. The fixed-point lower bound may be non-tight, and the same symbolic granularity can occur at different nonlinear capacities. Accordingly, candidate models selected from the consensus layer must still be retrained and evaluated for autonomous dynamical fidelity.

C Training and Reduction-Guided Retraining Details

C.1 3-scroll Chua data generation

3-scroll Chua system. The 3-scroll Chua trajectory used in our experiments was generated by the continuous-time system

$$\dot{x} = 10\,(y - 0.05x - 0.6\sin x)\,, \tag{A13}$$

$$\dot{y} = x - y + z, \tag{A14}$$

$$\dot{z} = -16y. \tag{A15}$$

We numerically integrated these equations using a time step of $\Delta t = 0.01$. The sinusoidal nonlinearity follows the multi-scroll Chua construction of Tang et al. (2001).

The fixed points satisfy

$$y^* = 0, \qquad z^* = -x^*, \qquad 0.05x^* + 0.6\sin x^* = 0.$$

Within the region occupied by the target 3-scroll attractor, this equation yields five fixed points with

$$x^* \approx 0,\ \pm 3.4316,\ \pm 5.7806.$$

These five fixed points constitute the target fixed-point set used in our analysis, and hence $Q = 5$.

C.2 Parent and direct-training configurations

For the 3-scroll Chua experiments, the parent AL-RNN uses

$$M = 20, \qquad N = 3, \qquad P_{\text{parent}} = 10, \tag{A16}$$

where M is the latent-state dimension, N is the number of observed/readout dimensions, and P_{parent} is the number of ReLU units, which occupy the final P_{parent} coordinates of the latent state.

We train 30 independent parent seeds for 2000 epochs, with 50 optimization steps per epoch. Each optimization step uses a batch of 16 trajectory segments of length 200. Models are optimized using RAdam (Liu et al., 2020) with an exponentially decaying learning rate from 10^{-3} to 10^{-5}. Gradients are clipped to a maximum norm of 10 at every optimization step. No noise is added during training.

Parent training uses the standard sparse teacher-forcing protocol of Brenner et al. (2024), with a forcing interval of 128 time steps. The observed coordinates are overwritten by the corresponding target values at the beginning of each training segment and every 128 steps thereafter. The forcing strength is 1.0.

All directly trained baselines use the same latent dimension, optimization schedule, sparse teacher-forcing protocol, and checkpoint selection rule as the parent models, varying only the number of ReLU units P.

Checkpoint selection. Every 10 epochs, we evaluate an autonomous rollout of 10,000 steps and retain the checkpoint minimizing

$$S_{\text{ckpt}} = E_{\text{stsp}} + 5.0\,|Q_{\text{vis}} - Q|\,, \tag{A17}$$

where Q is the known number of target fixed points ($Q = 5$ for the 3-scroll Chua system) and Q_{vis} is the number of analytically admissible fixed points of the learned model whose corresponding symbolic regions are visited by the autonomous trajectory. Importantly, no fixed-point term is included in the training loss; the known target fixed-point count is used for checkpoint selection.

C.3 Reduced-model initialization

Let θ_{parent} denote the complete trainable parameter set of the parent checkpoint. For every reduction candidate induced by $\mathcal{D}$, the reduced student is initialized as

$$\theta_{\mathcal{D}}^{(0)} = \theta_{\text{parent}}. \tag{A18}$$

Thus, all trainable parameters are copied exactly from the higher-capacity parent.

The architectural change is imposed only through the candidate's linearization mask: ReLU activations selected by the reduction procedure are permanently replaced by identity activations. The mask is fixed throughout retraining and is not trainable.

Consequently, candidates derived from the same parent begin with the same trainable parameter values and differ only in which nonlinear units have been linearized.

Unless otherwise stated, reduced candidates are retrained using the same optimization schedule as the parent models described in Appendix C.2. Each selected candidate is retrained once using a fixed retraining seed. Because the trainable parameters are initialized deterministically from the parent checkpoint, the only stochasticity within a retraining run comes from sampling trajectory-segment start times. The optimizer state is initialized from scratch for each run, and checkpoints are selected using the same criterion as in equation A17.

C.4 Autonomous teacher trajectory

For all teacher-guided retraining conditions, teacher targets are extracted from a single continuous autonomous free-run trajectory of the frozen parent model. No teacher forcing, raw-data overwrite, or segment reset is applied during this trajectory.

The teacher rollout is initialized from the first sample of the training trajectory using the same latent-state initialization procedure as in standard autonomous evaluation. The resulting teacher trajectory contains $400{,}000$ time steps.

For every time step, we store the parent latent state, readout, and activation pattern of the original nonlinear units. A single cache is constructed for each source parent and shared across all reduction candidates, retraining conditions, and retraining seeds derived from that parent.

The same cache also supplies the initial student state for each training segment and the reference states used for sparse synchronization. Specifically, for a segment beginning at teacher time s, the student is initialized as

$$\mathbf{z}_s^{\mathcal{D}} = \mathbf{z}_s^{\text{parent}}, \tag{A19}$$

and subsequent synchronization events use the corresponding parent states from the same cached trajectory. Thus, all trajectory-based teacher signals used during retraining are obtained from this single autonomous parent rollout.

C.5 Phase-aligned student rollouts

During teacher-guided retraining, a random segment start s is sampled from the autonomous teacher cache and the student is initialized from the corresponding teacher state:

$$\mathbf{z}_s^{\text{student}} = \mathbf{z}_s^{\text{teacher}}. \tag{A20}$$

The student is then rolled out autonomously for $L = 200$ steps. The teacher targets are taken from the same free-running teacher trajectory at

$$s + 1, \ldots, s + L. \tag{A21}$$

The phase-alignment procedure is important for chaotic dynamics. Because nearby chaotic trajectories quickly diverge in phase, teacher targets must be taken from the same parent trajectory used to initialize the student segment. Otherwise, the student would be trained to match parent states corresponding to a different phase of the attractor.

C.6 Periodic parent-state synchronization

In addition to initializing each training segment from the corresponding parent state, the main reduction-guided protocol applies sparse parent-state synchronization during the student rollout. This follows the same general idea as the sparse teacher-forcing procedure used in standard AL-RNN training (Brenner et al., 2024), but the reference state is taken from the autonomous parent trajectory rather than from the observed data.

Synchronization is applied every 128 time steps to the entire latent state:

$$\mathbf{z}_t^{\mathcal{D}} \leftarrow \mathbf{z}_t^{\text{parent}}. \tag{A22}$$

Because the training-segment length is 200, each segment is initialized from the corresponding parent state and contains one additional synchronization event at $t = 128$. Between synchronization events, the reduced model evolves autonomously.

For the retained-guidance comparison condition, synchronization is instead restricted to the readout coordinates and the retained ReLU coordinates,

$$\mathcal{R}_{\mathrm{sync}} = \{1, \ldots, N\} \cup \mathcal{R}, \tag{A23}$$

with

$$\mathbf{z}^{\mathcal{D}}_{t,\mathcal{R}_{\mathrm{sync}}} \leftarrow \mathbf{z}^{\mathrm{parent}}_{t,\mathcal{R}_{\mathrm{sync}}}, \tag{A24}$$

while the remaining coordinates continue to evolve freely.

Thus, sparse teacher forcing in direct training and parent-state synchronization in reduction-guided retraining both periodically realign the model trajectory, but use different reference signals. Direct training uses the observed trajectory, whereas reduction-guided retraining uses the autonomous parent trajectory. No raw-data overwrite is applied during teacher-guided retraining.

C.7 Retraining objectives

We consider four retraining conditions:

1. no teacher guidance;
2. full-preactivation guidance;
3. retained-preactivation guidance;
4. symbol guidance.

The full-preactivation condition with fixed coordinate-equal weighting and full-state synchronization is used as the main reduction-guided retraining protocol. The remaining conditions are evaluated as ablations.

No teacher guidance. The no-guidance condition retains the parent-derived parameter initialization and the fixed linearization mask, but does not use the autonomous parent trajectory during retraining. Instead, it is trained on the original observed trajectory using the same output objective and sparse teacher-forcing procedure as standard AL-RNN training. This condition therefore isolates the effect of parent-trajectory guidance from that of parent initialization and structural reduction.

Output matching. All teacher-guided conditions use the readout-matching loss

$$\mathcal{L}_{\mathrm{out}} = \frac{1}{TN} \sum_{\ell=1}^{T} \left\| \mathbf{x}^{\mathcal{D}}_{s+\ell} - \mathbf{x}^{\mathrm{parent}}_{s+\ell} \right\|_2^2, \tag{A25}$$

where s denotes the starting time of the sampled segment in the cached autonomous parent trajectory.

Full-preactivation guidance. The full-preactivation objective additionally matches all non-readout latent coordinates:

$$\mathcal{L}_{\mathrm{full}} = \frac{1}{T(M-N)} \sum_{\ell=1}^{T} \left\| \mathbf{z}^{\mathcal{D}}_{s+\ell,N:M} - \mathbf{z}^{\mathrm{parent}}_{s+\ell,N:M} \right\|_2^2. \tag{A26}$$

The total loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{out}} + \lambda_{\mathrm{full}} \mathcal{L}_{\mathrm{full}}. \tag{A27}$$

For the main protocol, we use the fixed coordinate-equal weighting

$$\lambda_{\mathrm{full}} = \frac{M-N}{N}. \tag{A28}$$

Because both losses are normalized by their respective numbers of coordinates, this choice gives each readout and non-readout coordinate the same weight in the total squared-error objective. For $M = 20$ and $N = 3$, $\lambda_{\mathrm{full}} = 17/3$.

Retained-preactivation guidance. Let $\mathcal{R}$ denote the retained ReLU coordinates after reduction. The retained-preactivation loss is

$$\mathcal{L}_{\mathrm{ret}} = \frac{1}{T|\mathcal{R}|} \sum_{\ell=1}^{T} \left\| \mathbf{z}^{\mathcal{D}}_{s+\ell,\mathcal{R}} - \mathbf{z}^{\mathrm{parent}}_{s+\ell,\mathcal{R}} \right\|_2^2, \tag{A29}$$

with total loss

$$\mathcal{L} = \mathcal{L}_{\mathrm{out}} + \lambda_{\mathrm{ret}} \mathcal{L}_{\mathrm{ret}}. \tag{A30}$$

Coordinates corresponding to linearized ReLU units and originally linear latent variables receive no auxiliary state-matching loss. Thus, this objective constrains only the nonlinear coordinates retained after reduction while allowing the remaining linear latent coordinates to reorganize during retraining.

Symbol guidance. For each retained ReLU coordinate $j \in \mathcal{R}$, define the parent activation label

$$b^{\mathrm{parent}}_{t,j} = \mathbb{I}\left[z^{\mathrm{parent}}_{t,j} > 0 \right], \qquad y_{t,j} = 2b^{\mathrm{parent}}_{t,j} - 1. \tag{A31}$$

We use the smooth margin loss

$$\mathcal{L}_{\mathrm{sym}} = \frac{1}{T|\mathcal{R}|} \sum_{\ell=1}^{T} \sum_{j \in \mathcal{R}} \mathrm{softplus}\left(m - y_{s+\ell,j} z^{\mathcal{D}}_{s+\ell,j} \right), \tag{A32}$$

with margin

$$m = 1. \tag{A33}$$

The total loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{out}} + \lambda_{\mathrm{sym}} \mathcal{L}_{\mathrm{sym}}. \tag{A34}$$

A systematic comparison of these retraining variants is reported in Section E.2.

D Evaluation Protocol and Statistics

D.1 Autonomous free-run evaluation

Each selected model is evaluated using an autonomous rollout of $10{,}000$ time steps without teacher forcing or external state correction. The first $1{,}000$ steps are discarded as transient for all final evaluation quantities, including Q_{vis}, $|\widehat{\Sigma}|$, E_{stsp}, and E_{H}.

We define the post-transient visited-symbol set as

$$\widehat{\Sigma} = \{\boldsymbol{\sigma}(\mathbf{z}_t) : 1001 \leq t \leq 10000\}\,, \tag{A35}$$

whose cardinality $|\widehat{\Sigma}|$ measures the symbolic complexity realized by the long-run autonomous dynamics.

The visited fixed-point count Q_{vis} is computed over the same post-transient interval and counts admissible fixed points of the learned model whose symbolic regions are visited during that interval.

D.2 State-space and spectral fidelity

We use two complementary measures of autonomous dynamical fidelity, following the prior AL-RNN work (Brenner et al., 2024). Both metrics are computed from the post-transient portion of the autonomous rollout, after discarding the first $1{,}000$ steps.

The state-space error is the Kullback–Leibler divergence

$$E_{\mathrm{stsp}} = D_{\mathrm{KL}}\left(\widehat{p}_{\mathrm{target}} \,\|\, \widehat{p}_{\mathrm{gen}}\right), \tag{A36}$$

where $\widehat{p}_{\mathrm{target}}$ and $\widehat{p}_{\mathrm{gen}}$ are histogram-based empirical state-space distributions in the observed coordinates. We use 30 bins per dimension over the coordinate ranges of the target trajectory, with additive smoothing of 10^{-5} applied to each histogram cell before normalization. Generated samples outside the target histogram range are excluded. The target distribution is estimated from the full training trajectory, whereas the generated distribution is estimated from the post-transient $9{,}000$-step autonomous trajectory.

As a complementary measure of temporal structure, we compute the spectral Hellinger distance E_{H}. For each observed coordinate, we compare the generated post-transient segment of $9{,}000$ steps with a target segment of equal length, taken as the first $9{,}000$ samples of the training trajectory. Each segment is standardized to zero mean and unit variance, and its power spectrum is computed using a real FFT. Because the generated and target segments have identical lengths, their spectra share the same frequency grid and require no interpolation or bin matching. Each spectrum is smoothed with a Gaussian kernel with standard deviation $\sigma = 20$ frequency bins and normalized to unit mass.

The Hellinger distance between the normalized target and generated spectra is

$$D_{\mathrm{H}}(p, q) = \sqrt{1 - \sum_f \sqrt{p_f q_f}}, \tag{A37}$$

where p_f and q_f denote the normalized target and generated power spectra, respectively, at frequency bin f. The spectral error E_{H} is obtained by averaging $D_{\mathrm{H}}(p, q)$ over the N observed coordinates.

Because E_{stsp} and E_{H} measure complementary aspects of the autonomous dynamics, only E_{stsp} enters the high-fidelity success criterion. The spectral error E_{H} is reported as an independent diagnostic.

D.3 Chua success criteria

For the 3-scroll Chua benchmark, the target fixed-point count is $Q = 5$. We first define fixed-point-count success as

$$S_{\mathrm{FP}} = \mathbb{I}\left[Q_{\mathrm{vis}} = 5\right]. \tag{A38}$$

Thus, $S_{\mathrm{FP}} = 1$ indicates that the autonomous trajectory visits symbolic regions containing exactly five admissible fixed points of the learned model.

Our primary structural criterion is minimal-realization success,

$$S_{\min} = \mathbb{I}\left[Q_{\text{vis}} = 5 \ \wedge \ |\widehat{\Sigma}| = 5\right]. \tag{A39}$$

For Chua, this criterion requires both the target number of visited admissible fixed points and the certified minimum of five post-transient symbolic regions.

To additionally require faithful state-space dynamics, we define

$$S_{\text{fid}} = \mathbb{I}\left[S_{\min} = 1 \ \wedge \ E_{\text{stsp}} < 2.0\right]. \tag{A40}$$

The threshold $E_{\text{stsp}} = 2.0$ is used as a conservative representative fidelity criterion for the Chua benchmark. Its sensitivity is examined separately in Appendix E.3.

D.4 Parent-balanced statistical aggregation

Different source parents can generate different numbers of selected reduction candidates. Pooling all candidates would therefore give greater weight to parent models that happen to generate more descendants. We instead treat the independently trained source parent as the primary statistical unit.

Let $\mathcal{C}_r$ denote the set of candidates selected *before retraining* from parent seed r. For a binary success indicator S, we define the within-parent success fraction as

$$q_r(S) = \frac{1}{|\mathcal{C}_r|} \sum_{c \in \mathcal{C}_r} S(c), \qquad |\mathcal{C}_r| > 0. \tag{A41}$$

Candidates that fail during retraining or evaluation remain in $\mathcal{C}_r$ and are assigned $S(c) = 0$. If a parent produces no candidate satisfying the selection criterion, we set

$$q_r(S) = 0. \tag{A42}$$

Let $\mathcal{P}$ denote the set of the 30 independently trained parent seeds. The parent-balanced seed-macro success rate is

$$\bar{q}(S) = \frac{1}{|\mathcal{P}|} \sum_{r \in \mathcal{P}} q_r(S). \tag{A43}$$

Thus, each independently trained parent contributes equal total weight, irrespective of the number of reduction candidates it generates.

For continuous candidate-level quantities such as E_{stsp}, each successfully evaluated candidate retains the weight determined by the number of candidates originally selected from its parent. Let $\mathcal{C}_r^{\text{eval}} \subseteq \mathcal{C}_r$ denote the candidates from parent r for which a valid continuous measurement is available. The corresponding parent-balanced empirical cumulative distribution is

$$F_{\text{bal}}(e) = \frac{\displaystyle\sum_{r \in \mathcal{P}} \sum_{c \in \mathcal{C}_r^{\text{eval}}} \frac{1}{|\mathcal{C}_r|} \mathbb{I}[E_{r,c} \le e]}{\displaystyle\sum_{r \in \mathcal{P}} \frac{|\mathcal{C}_r^{\text{eval}}|}{|\mathcal{C}_r|}}. \tag{A44}$$

Candidates with missing or non-finite continuous measurements are excluded from both the numerator and denominator, without renormalizing the weights of their successfully evaluated siblings.

Directly trained models and higher-capacity parent models contribute one observation per independent seed and therefore require no additional candidate-level weighting.

E ADDITIONAL 3-SCROLL CHUA RESULTS

This section provides additional results for the 3-scroll Chua benchmark that complement the main-text analysis. We first summarize the structural outcomes of hierarchical reduction, then compare alternative retraining protocols, examine the robustness of the state-space fidelity criterion, and finally analyze how the symbolic transition structure changes after retraining.

E.1 STRUCTURAL OUTCOMES OF HIERARCHICAL REDUCTION

We apply hierarchical reduction to 30 independently trained parent models with $P_{\mathrm{parent}} = 10$, producing multiple fixed-point-preserving quotient candidates with different combinations of nonlinear capacity $P_{\mathcal{D}}$ and symbolic complexity $|\Sigma_{\mathcal{D}}|$. For the 3-scroll Chua system, the theoretical lower bounds are

$$|\Sigma_{\mathcal{D}}| \geq 5, \qquad P_{\mathcal{D}} \geq 3. \tag{A45}$$

Among the selected five-symbol quotient candidates, 126 attain both bounds simultaneously,

$$|\Sigma_{\mathcal{D}}| = 5, \qquad P_{\mathcal{D}} = 3, \tag{A46}$$

whereas an additional 15 candidates attain the certified five-symbol bound with $P_{\mathcal{D}} = 4$. The latter are therefore minimal in symbolic complexity but do not attain the nonlinear-capacity lower bound.

These results also illustrate that the two notions of complexity are distinct. A quotient graph can already attain the minimum possible number of symbols while retaining more nonlinear units than required by the capacity lower bound. For the main comparison with direct minimal-capacity training, we therefore restrict the reduction-guided models to the $P_{\mathcal{D}} = 3$, $|\Sigma_{\mathcal{D}}| = 5$ candidates, so that both methods are evaluated at the same final nonlinear capacity $P = 3$.

E.2 REDUCTION-GUIDED RETRAINING ABLATIONS

We next examine how the form of parent-derived guidance affects the realization of compact Chua models. For each reduction candidate, all methods start from the same parent-derived parameter initialization and the same reduced architecture, differing only in the form of guidance and synchronization used during retraining. The corresponding retraining objectives and synchronization schemes are detailed in Appendix C.7. We compare training without parent-trajectory guidance against three guided variants: symbol guidance with $\lambda_{\mathrm{sym}} = 1$, retained-preactivation guidance with $\lambda_{\mathrm{ret}} = 1$, and full-preactivation guidance with the coordinate-equal weighting $\lambda_{\mathrm{full}} = (M - N)/N$. We use the latter protocol for the main results.

Unlike the capacity-matched main comparison ($P_{\mathcal{D}} = 3$; 126 candidates), this ablation uses all 141 selected five-symbol candidates, including 15 candidates with $P_{\mathcal{D}} = 4$ contributed by the two parents whose reduction trees reach $|\Sigma_{\mathcal{D}}| = 5$ only at $P_{\mathcal{D}} = 4$. The success rates reported here therefore differ slightly from those in the capacity-matched main comparison.

As shown in figure A1, parent-derived trajectory guidance provides a strong optimization scaffold after structural reduction. Without trajectory guidance, fixed-point recovery reaches only 19%, and both minimal and high-fidelity minimal realizations are obtained in approximately 14% of parent-balanced trials. All three guided variants substantially improve these rates. Symbol guidance yields success rates of 74%, 72%, and 71%, while retained-preactivation guidance yields 74%, 68%, and 68%, respectively. Full-preactivation guidance gives the highest rates across the three criteria, at 77%, 74%, and 73%.

The state-space fidelity measures show the same qualitative effect. Without guidance, E_{stsp} is broadly distributed and frequently exceeds the high-fidelity threshold. Guided retraining shifts the distribution toward substantially lower state-space error, with the retained- and full-preactivation variants showing particularly concentrated distributions near the low-error regime. The improvement is therefore not limited to recovering the desired fixed-point count or symbolic complexity; parent guidance also makes the resulting autonomous dynamics more faithful to the target attractor.

In contrast, the E_{H} distributions overlap more strongly across the guided variants. This is consistent with E_{H} capturing a different aspect of dynamical fidelity from E_{stsp}, rather than simply tracking

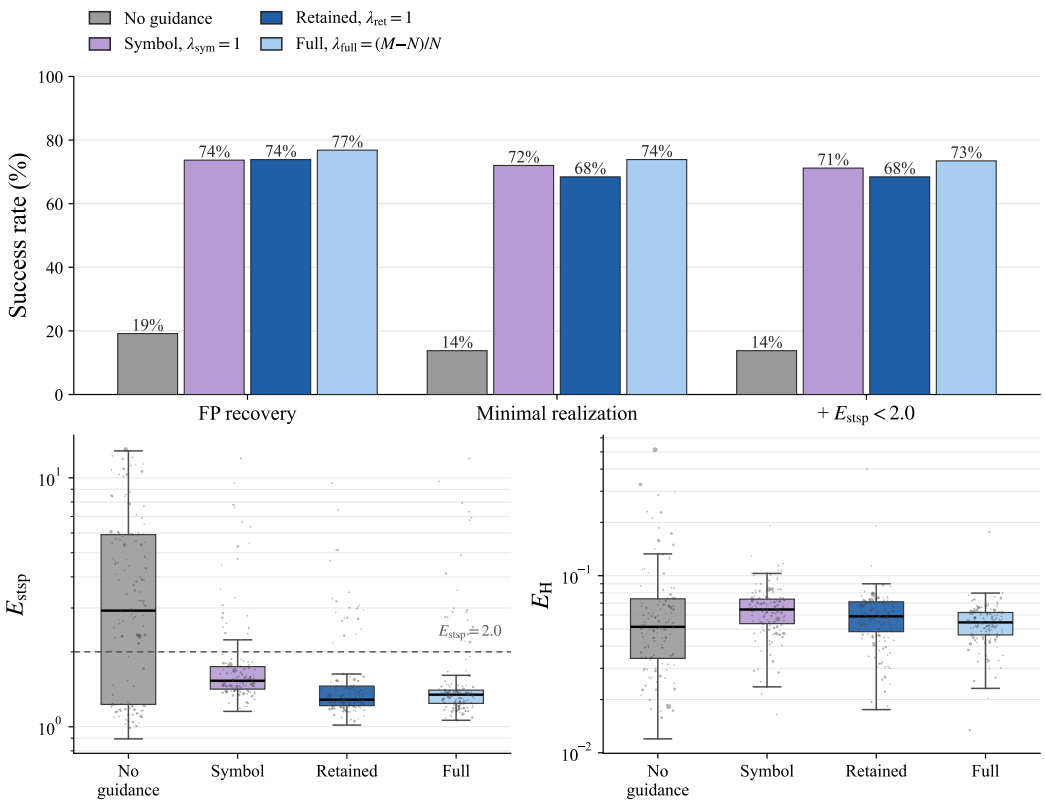


Figure A1: **Ablation of reduction-guided retraining on the 3-scroll Chua system.** All methods use the same selected five-symbol reduction candidates and parent-derived initialization. **Top:** parent-balanced seed-macro success rates for fixed-point recovery, minimal realization, and high-fidelity minimal realization ($S_{\text{fid}} = 1$, requiring $E_{\text{stsp}} < 2.0$). Full-preactivation guidance achieves 77%, 74%, and 73%, respectively, compared with 19%, 14%, and 14% without parent-trajectory guidance. Symbol and retained-preactivation guidance also substantially improve success rates. **Bottom left:** post-transient state-space error E_{stsp}; the dashed line marks the $E_{\text{stsp}} = 2.0$ high-fidelity threshold. **Bottom right:** post-transient spectral Hellinger error E_{H}, shown as a complementary dynamical diagnostic. Points denote individual candidates with valid finite measurements, and boxplots summarize the corresponding parent-balanced distributions.

the state-space reconstruction criterion. We therefore retain E_{H} as a complementary diagnostic and use E_{stsp} alone in the binary high-fidelity criterion.

Overall, several forms of parent-derived guidance are effective, and the differences among the guided variants are considerably smaller than the gap between guided and unguided retraining. We use full-preactivation guidance with $\lambda_{\text{full}} = (M - N)/N$ as the main protocol because it has a simple, directly interpretable coordinate-equal weighting, requires no additional calibration, and performs consistently well across structural and fidelity criteria.

E.3 State-space fidelity threshold

The main analysis uses $E_{\text{stsp}} < 2.0$ as a conservative criterion for high-fidelity realization. This threshold is used as a practical criterion for identifying clearly reconstructed 3-scroll attractors, rather than as a theoretically defined dynamical boundary. As illustrated in Fig. A2, models with $E_{\text{stsp}} < 2.0$ generally retain well-formed 3-scroll geometries, whereas larger values increasingly correspond to distorted or incomplete attractors.

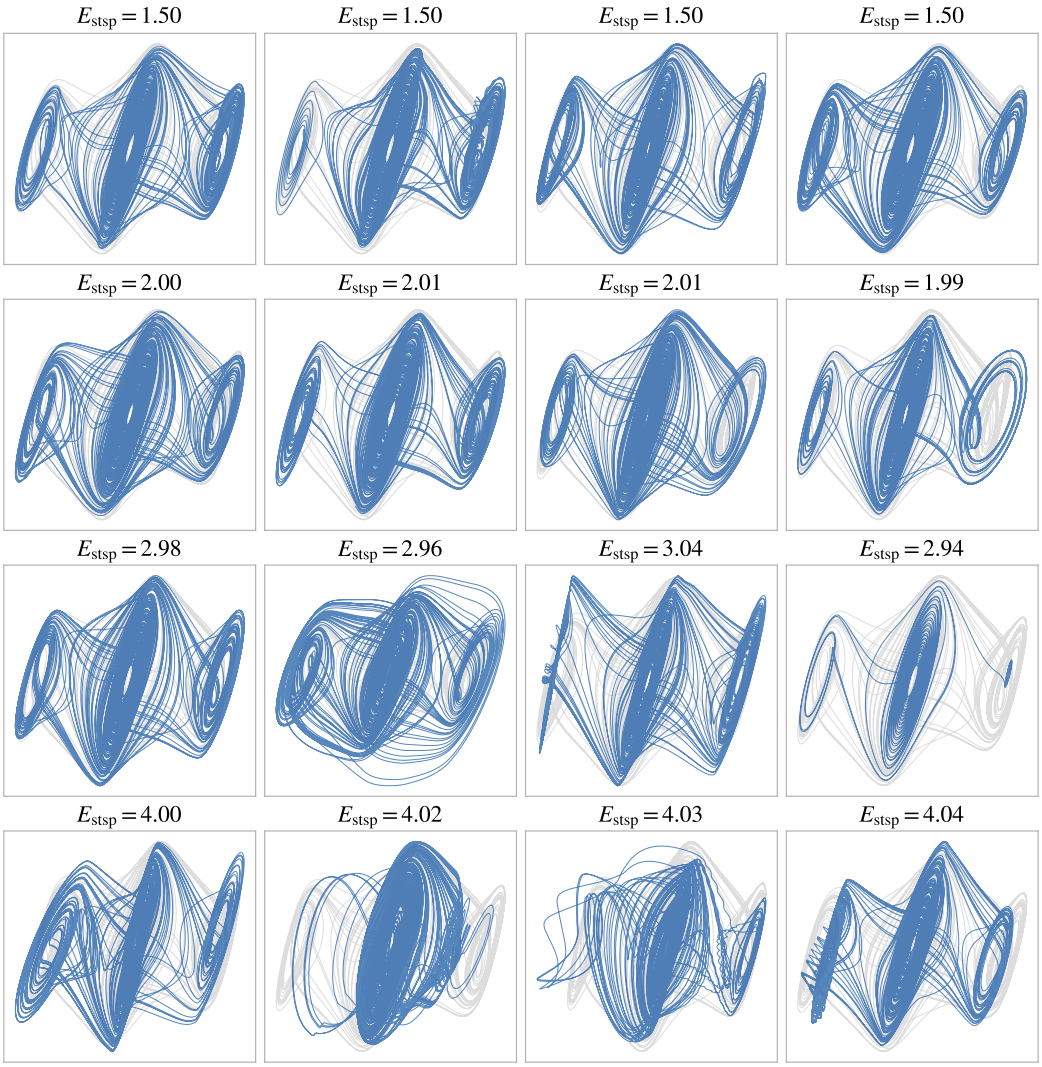


Figure A2: **Representative autonomous 3-scroll Chua trajectories across state-space reconstruction errors.** Gray trajectories show the ground-truth attractor, and blue trajectories show the corresponding autonomous model outputs. From top to bottom, the rows contain four representative models with $E_{\mathrm{stsp}} \approx 1.5$, 2.0, 3.0, and 4.0, respectively; the exact E_{stsp} value of each model is shown above its panel. Models with $E_{\mathrm{stsp}} < 2.0$ generally preserve the overall 3-scroll geometry, whereas larger errors increasingly include distorted, imbalanced, or incomplete attractors. These examples illustrate the practical motivation for using $E_{\mathrm{stsp}} < 2.0$ as the high-fidelity criterion.

To assess whether the quantitative conclusions depend on this particular cutoff, we evaluate

$$H(\tau) = \Pr\left(S_{\min} = 1 \wedge E_{\mathrm{stsp}} < \tau\right) \tag{A47}$$

over a range of fidelity thresholds τ.

As shown in Fig. A3, the qualitative comparison is insensitive to the precise choice of fidelity threshold. Direct training at the minimal nonlinear capacity $P = 3$ remains substantially below the guided retraining variants over a broad range of τ. Likewise, retraining without parent-trajectory guidance remains at low $H(\tau)$ values, whereas all three guided variants rise sharply as the fidelity threshold is relaxed.

At the threshold used in the main analysis, $\tau = 2.0$, full-preactivation guidance reaches approximately 73% high-fidelity minimal realizations, compared with approximately 20% for direct $P = 3$

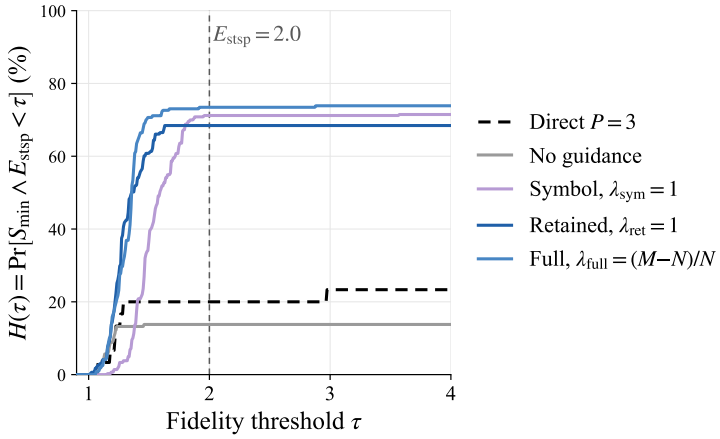


Figure A3: **Sensitivity of high-fidelity minimal-realization rates to the state-space-error threshold on the Chua system.** The curves show $H(\tau) = \Pr[S_{\min} = 1 \wedge E_{\mathrm{stsp}} < \tau]$ for direct minimal-capacity training, retraining without parent-trajectory guidance, and the three guided retraining variants. The vertical dashed line marks the threshold $E_{\mathrm{stsp}} = 2.0$ used in the main analysis. Reduction-guided retraining maintains substantially higher success rates than direct $P = 3$ training over a broad range of thresholds, and the guided curves are already close to their plateaus by $\tau = 2.0$.

training. Symbol and retained-preactivation guidance yield similarly high rates of approximately 71% and 68%, respectively. Thus, the advantage of reduction-guided retraining over direct minimal-capacity training is not specific to the particular cutoff $E_{\mathrm{stsp}} = 2.0$.

Importantly, the separation between direct training and guided retraining is already evident below $\tau = 2.0$ and persists for more permissive thresholds. Moreover, the guided curves are close to their asymptotic plateaus around the selected threshold. The reported improvement is therefore not caused by a narrow cluster of models lying immediately on opposite sides of the chosen cutoff.

E.4 Emergence of a common chain-like symbolic organization

An additional structural regularity emerges after reduction-guided retraining. Although the selected quotient candidates can differ substantially in their pre-retraining transition graphs, every high-fidelity minimal realization ($S_{\mathrm{fid}} = 1$) that we observed exhibits the same five-node chain-like symbolic organization in its autonomous dynamics.

Figure A4 illustrates representative examples. The selected quotient candidates on the left have transition graphs that differ structurally from the final realized organization, including cases with two cycles and no leaf and cases with one cycle plus a tail. After retraining, however, the corresponding autonomous models all preserve the five-symbol minimal realization ($Q_{\mathrm{vis}} = 5$, $|\widehat{\Sigma}| = 5$) and reorganize into the same chain-like transition structure.

Importantly, this chain-like organization is not explicitly imposed by either stage of the pipeline. The reduction procedure constrains only which activation-pattern mergers are admissible under fixed-point preservation, and the retraining objective matches continuous trajectories rather than directly optimizing graph topology. The common chain-like structure therefore emerges from the autonomous dynamics learned after retraining.

This observation clarifies the role of the quotient graph. A fixed-point-preserving quotient should be interpreted as a structurally admissible starting point for retraining, rather than as a literal prediction of the final realized transition topology. Different non-chain quotient candidates can reorganize during retraining into the same chain-like symbolic structure while preserving the five target fixed-point symbols and reproducing the 3-scroll attractor faithfully.

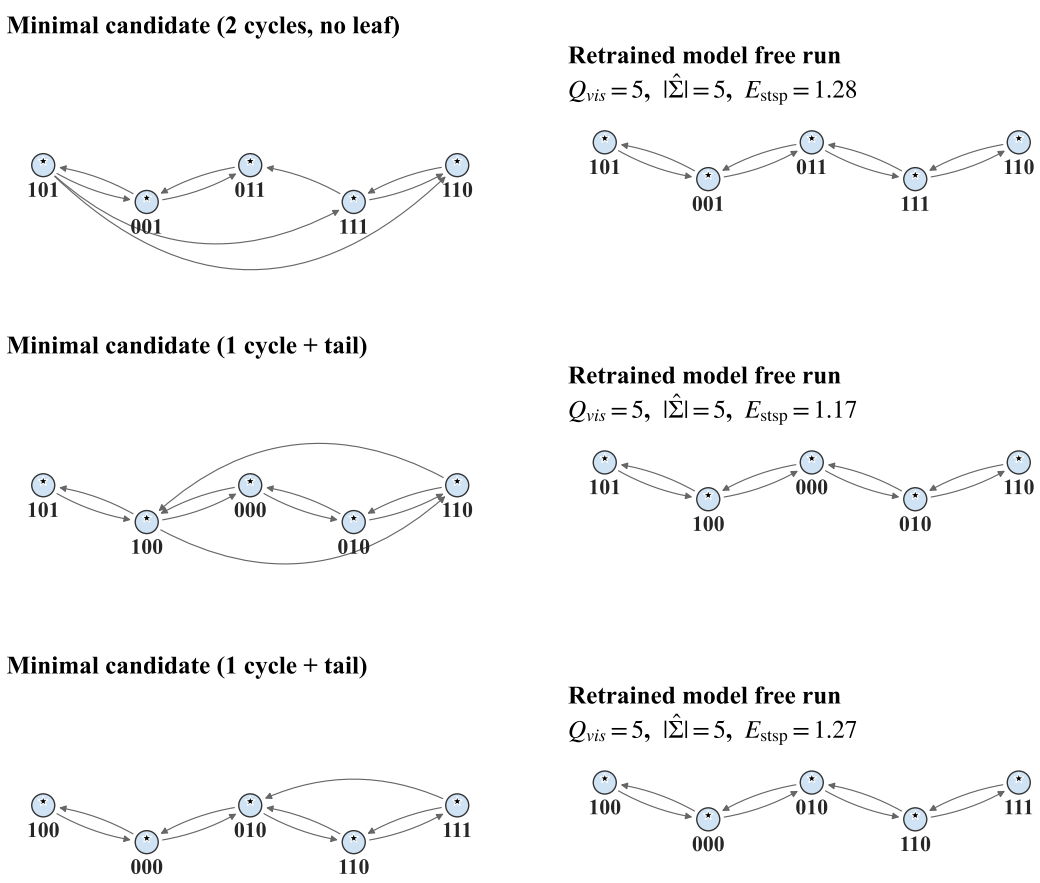


Figure A4: **Reorganization of non-chain quotient candidates into a common five-node chain after retraining.** The left column shows representative selected quotient graphs before retraining, including a two-cycle topology without a leaf and one-cycle-plus-tail topologies. The right column shows the corresponding autonomous transition graphs after reduction-guided retraining. Despite the different predicted quotient structures, all three retrained models realize $Q_{\text{vis}} = 5$ and $|\widehat{\Sigma}| = 5$, with low post-transient state-space error ($E_{\text{stsp}} = 1.17$–1.28), and converge to the same path-like five-node organization. Stars indicate symbols associated with admissible fixed points.

For the 3-scroll Chua system, all $n = 114$ observed high-fidelity minimal candidate runs converge to the same five-node chain-like organization, revealing a common symbolic skeleton of the target dynamics.