Task-Restricted Symmetries in Recurrent Weight Space

Simon Dräger [1]

Abstract

Recurrent networks can contain substantial functional redundancy in weight space: changing a recurrent matrix may leave the input-output rollout nearly unchanged on a task distribution, while similar-scale changes can destroy the same behavior. We study this redundancy in one-layer tanh RNNs using ordered real Schur coordinates. The Schur form separates spectral blocks from directed nonnormal couplings, giving a diagnostic basis for structured ablations that keep the input and readout maps fixed. In a fixed-length copy task, selected nonnormal Schur couplings can be removed with little loss in some trained solutions, whereas other couplings are necessary for accurate autonomous replay. Across flip-flop, sine generation, and context-dependent integration, the loss-preserving ablation profile varies across tasks and trained solutions. These results identify candidate approximate functional invariances, not universal symmetries of recurrent weight space. Schur-coordinate ablations provide a practical diagnostic for which structured perturbations preserve a trained recurrent solution and which ones disrupt its computation.

1. Introduction

Exact weight-space symmetries have become a practical tool for comparing neural networks and for learning directly in parameter space (Entezari et al., 2022; Ainsworth et al., 2023; Navon et al., 2023; 2024). Those symmetries identify transformations that preserve the realized function exactly, and recent work builds such structure directly into models that operate on trained networks as inputs (Zhou et al., 2023; Kofinas et al., 2024). Recurrent networks can also admit large structured changes to the recurrent matrix that preserve task behavior only approximately and only on the task distribution. These directions fall outside exact group-theoretic symmetries, while still shaping the functional geometry of weight space.

Ordered Schur coordinates reveal candidate approximate functional invariances under structured perturbation. Because the resulting ablation profiles vary by task and by trained solution, they should not be read as evidence that nonnormal components can usually be ignored. They identify which Schur-coordinate couplings a particular recurrent solution can lose while preserving its original input-output rollout, and which couplings carry task-specific function.

Because tanh RNNs do not admit arbitrary orthogonal changes of basis as exact symmetries, raw recurrent coordinates make nonnormal structure hard to compare across runs. The real Schur decomposition represents every real recurrent matrix by an orthogonal basis, diagonal or quasi-diagonal spectral blocks, and strictly upper-triangular nonnormal interactions. Such interactions are known to shape transient recurrent computations (Murphy & Miller, 2009; Hennequin et al., 2012; Bondanelli & Ostojic, 2020; Pattadkal et al., 2024), and ordered Schur coordinates make them comparable and ablatable.

Schur-coordinate ablations preserve the rollout function for some blocks and not for others. In the copy task, selected ablations produce nearly identical autonomous replay accuracy, while directed cross-sector ablations move the model to lower-accuracy behavior. The neuroscience-style tasks provide a scope test for the same interventions. The copy task supplies an explicit temporal symmetry; the flip-flop, sine-generation, and context-dependent integration tasks ask whether the same diagnostic basis also localizes fragile directions in other recurrent computations (Sussillo & Barak, 2013; Mante et al., 2013; Maheswaranathan et al., 2019; Schuessler et al., 2024). Task-dependent ablation profiles tie approximate invariance to the rollout distribution rather than to a task-independent property of Schur blocks.

[1]Salk Institute for Biological Studies, La Jolla, CA, USA. Correspondence to: Simon Dräger <sfdraeger@gmail.com>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

2. Ordered Schur Coordinates

A one-layer tanh RNN maps input $x_t \in \mathbb{R}^{N_x}$, hidden state $h_t \in \mathbb{R}^{N_h}$, output $\hat{y}_t \in \mathbb{R}^{N_y}$,

$$h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1}), \qquad h_0 = 0, \tag{1}$$

$$\hat{y}_t = W_{hy}h_t, \tag{2}$$

with $W_{xh} \in \mathbb{R}^{N_h \times N_x}$, $W_{hh} \in \mathbb{R}^{N_h \times N_h}$, and $W_{hy} \in \mathbb{R}^{N_y \times N_h}$. All reported experiments set the recurrent and readout biases to zero, $b_h = b_y = 0$.

For a trained recurrent matrix, write $W = W_{hh}$. Its real Schur decomposition is

$$W = QTQ^\top, \tag{3}$$

where Q is orthogonal and T is real quasi-upper-triangular (Trefethen & Embree, 2005). We decompose

$$T = B + N, \tag{4}$$

where B contains the block-diagonal 1×1 and 2×2 real Schur eigenvalue blocks, and N contains the strictly block-upper-triangular nonnormal couplings between those blocks.

The Schur blocks are ordered by nonincreasing eigenvalue modulus. A relative threshold α separates leading spectral blocks from their complement:

$$R = \{i : |\lambda_i| \geq \alpha\rho(W)\}, \qquad C = \{1, \ldots, N_h\} \setminus R.$$

Here λ_i is the eigenvalue associated with the ith Schur block and $\rho(W) = \max_j |\lambda_j|$ is the spectral radius of W. R indexes the leading rotation-like subspace used as the reference sector, while C indexes the remaining Schur blocks whose couplings to R and to each other are tested by ablation. In this ordered partition,

$$B = \begin{pmatrix} B_R & 0 \\ 0 & B_C \end{pmatrix}, \qquad N = \begin{pmatrix} T_{RR} & T_{C\to R} \\ 0 & T_{CC} \end{pmatrix}. \tag{5}$$

T_{RR}, $T_{C\to R}$, and T_{CC} are blocks of the nonnormal coupling matrix N, not separate eigenvalue blocks. The cross block $T_{C\to R}$ is the upper-right coupling from the complement sector into the leading sector in the ordered Schur coordinates.

For a set S of Schur-coupling blocks, the intervention zeros the corresponding entries of N, reconstructs

$$\widetilde{W}_{hh}(S) = Q\widetilde{T}(S)Q^\top, \tag{6}$$

and reevaluates the original network without changing input or readout weights. Let f_W denote the rollout function of the trained network on a task distribution $\mathcal{D}$. This fixed-encoder/fixed-decoder intervention tests whether the original input-output map is preserved in the original readout coordinates. Refitting a linear or ridge decoder after the ablation would answer a different question: whether the perturbed latent dynamics still contain task information up to a new readout.

Table 1. Sensitivity to the Schur split threshold. Values are mean autonomous replay accuracy over 128 lags. The main experiments use $\alpha = 0.9$.

α	model	n_R	full	$-T_{CC}$	$-T_{C\to R}$
0.85	dense orth.	64	1.000	1.000	0.634
0.90	dense orth.	64	1.000	1.000	0.634
0.95	dense orth.	64	1.000	1.000	0.634
0.85	Cayley	68	1.000	1.000	1.000
0.90	Cayley	68	1.000	1.000	1.000
0.95	Cayley	68	1.000	1.000	1.000

For a rollout discrepancy $d_{\mathcal{D}}$ and tolerance ϵ, an intervention S is an ϵ-stabilizer on $\mathcal{D}$ when $d_{\mathcal{D}}(f_W, f_{\widetilde{W}_{hh}(S)}) \leq \epsilon$. A Schur-coupling block is a candidate approximate functional invariance when zeroing it gives small discrepancy while removing non-negligible Schur mass. If performance changes sharply, the block lies in a fragile functional direction for that trained solution.

For neuroscience-style tasks, held-out error is measured by

$$\mathrm{FVU} = \frac{\mathbb{E}\|\hat{y} - y\|^2}{\mathbb{E}\|y - \bar{y}\|^2}. \tag{7}$$

The expectation is over held-out rollouts, y is the target trajectory, $\hat{y}$ is the model output, and $\bar{y}$ is the empirical mean target over the evaluation set. For those tasks two summaries are reported:

$$\Delta\mathrm{FVU} = \mathrm{FVU}(\widetilde{W}_{hh}) - \mathrm{FVU}(W_{hh}), \tag{8}$$

$$S_{\Delta T} = \frac{\Delta\mathrm{FVU}}{\|\Delta T\|_F / \|T\|_F}. \tag{9}$$

$\Delta T = T - \widetilde{T}(S)$, and $\|\cdot\|_F$ denotes the Frobenius norm. ΔFVU captures the effect at the trained scale, whereas $S_{\Delta T}$ measures effect per unit removed Schur mass. The perturbations are evaluated after training; no input or readout weights are refit.

We use $\alpha = 0.9$ throughout the main experiments. This value was chosen a priori as a simple relative spectral-radius cutoff for grouping high-modulus Schur blocks into R, rather than tuned for an ablation outcome. The threshold controls only the R/C partition used to assign nonnormal couplings to T_{RR}, $T_{C\to R}$, and T_{CC}. A nearby-threshold check on the copy-task controllers preserves the same qualitative profile (Table 1).

Coordinate choice. The Schur basis remains orthogonal even for strongly nonnormal matrices (Trefethen & Embree,

2005). Direct eigencoordinates are often ill-conditioned when transient amplification is large, making cross-run comparison unstable and turning component ablations into basis-sensitive operations. By separating spectral blocks from non-normal couplings and ordering them by eigenvalue modulus, the real Schur form turns those couplings into structured perturbation directions. Compared with eigencoordinates, Schur coordinates provide a reproducible diagnostic basis for perturbing and interpreting recurrent dynamics.

3. Approximate Stabilizers in the Copy Task

The copy task is a fixed-delay variant of the copying-memory benchmark for long-range recurrent memory (Hochreiter & Schmidhuber, 1997; Arjovsky et al., 2016), and related fixed-length copy tasks have been used to study traveling-wave recurrent models (Keller et al., 2024). It presents a sequence of $s = 8$ symbols in $\{-1, +1\}^d$, with $d = 8$, then sets inputs to zero while the network autonomously reproduces the stored sequence. Replay accuracy is measured over the first 128 generated symbols after the input sequence. The copy task experiments train one-layer tanh RNNs at $N_h \in \{56, 64, 72\}$ under four recurrent constructions. Let $m = N_h^{-1/2}$. The three dense constructions optimize an unconstrained matrix $W_{hh} \in \mathbb{R}^{N_h \times N_h}$ and differ only in $W_{hh}^{(0)}$:

$$\begin{aligned} \text{dense default:} \quad & W_{hh,ij}^{(0)} \sim \mathrm{Unif}[-m, m], \\ \text{dense orthogonal:} \quad & W_{hh}^{(0)} = Q, \qquad Q^\top Q = I, \\ \text{dense normal:} \quad & W_{hh}^{(0)} = Q D_{\mathrm{norm}} Q^\top, \end{aligned}$$

where

$$\begin{aligned} D_{\mathrm{norm}} &= \mathrm{blockdiag}(B_1, \ldots, B_{N_h/2}), \\ B_i &= \begin{pmatrix} a_i & -b_i \\ b_i & a_i \end{pmatrix}, \\ a_i, b_i &\sim \mathcal{N}(0, 1/6). \end{aligned}$$

For the Cayley construction, every optimization iterate satisfies $W_{hh}^{(k)} = O(A^{(k)}) D^{(k)} O(A^{(k)})^\top$, where $(A^{(k)})^\top = -A^{(k)}$ and

$$O(A) = (I - A)(I + A)^{-1}.$$

At initialization,

$$\begin{aligned} U_{ij} &\sim \mathrm{Unif}[-m, m], \\ A^{(0)} &= (U - U^\top)/2, \\ \widetilde{W}_{ij} &\sim \mathrm{Unif}[-m, m], \\ D^{(0)} &= \mathrm{realblock}\Big(\mathrm{eig}(\widetilde{W})\Big), \end{aligned}$$

with $\mathrm{realblock}(\cdot)$ converting conjugate eigenvalue pairs into 2×2 real blocks of the form above. For $\mathcal{Z} = \{T_{RR}, T_{C\to R}, T_{CC}\}$ and $S \subseteq \mathcal{Z}$, the intervention is

$$\begin{aligned} \widetilde{W}_{hh}(S) &= Q Z_S(T) Q^\top, \\ \big(Z_S(T)\big)_B &= \begin{cases} 0, & B \in S, \\ T_B, & B \notin S, \end{cases} \qquad B \in \mathcal{Z}. \end{aligned}$$

Entries outside $\{T_{RR}, T_{C\to R}, T_{CC}\}$ are unchanged. For $\mathcal{D}_{\mathrm{rc}}$ and $\mathcal{L} = \{1, \ldots, 128\}$,

$$\hat{y}_{\ell j}^S(x) := \hat{y}_{\ell j}(x; \widetilde{W}_{hh}(S)),$$

$$\mathrm{Acc}_{\mathrm{rc}} = \frac{1}{|\mathcal{D}_{\mathrm{rc}}|\,|\mathcal{L}|\,d} \sum_{\substack{(x,y)\in\mathcal{D}_{\mathrm{rc}} \\ \ell\in\mathcal{L},\ j\in[d]}} \mathbf{1}\{\mathrm{sgn}(\hat{y}_{\ell j}^S(x)) = y_{\ell j}\}.$$

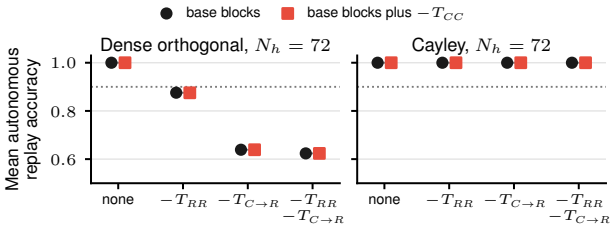

Figure 1. Candidate approximate functional invariances in the copy task. Points connected by gray line segments differ only by additionally zeroing T_{CC}. In the dense orthogonal model, T_{CC} removal leaves the autonomous replay function nearly unchanged conditional on the other removed blocks, while T_{RR} and $T_{C\to R}$ move the network between lower-accuracy functional classes. The Cayley-transform representative has negligible complement blocks and changes little under the shown ablations.

In the dense orthogonal $N_h = 72$ model, removing T_{CC} alone leaves mean replay accuracy at 1.00, matching the full model (Figure 1). The same near-equivalence holds after other Schur blocks have already been removed: $-T_{RR}$ and $-T_{RR}, -T_{CC}$ give 0.876 and 0.875; $-T_{C\to R}$ and $-T_{C\to R}, -T_{CC}$ both give 0.639; $-T_{RR}, -T_{C\to R}$ and zeroing all three blocks both give 0.624. Selected structured changes to nonnormal Schur couplings can therefore preserve the task behavior once the other ablated blocks are fixed.

T_{CC} is close to a stabilizer for this solved copy task controller conditional on the other removed blocks. Removing $T_{C\to R}$ moves the dense model to a different functional class, and removing T_{RR} produces a distinct intermediate class. The Cayley representative has negligible complement blocks at this width, so the same ablations leave replay accuracy unchanged.

These pairs define task-restricted approximate equivalence classes in which multiple recurrent matrices with different nonnormal coordinates realize nearly identical rollout functions on the copy task distribution. The copy task panels

evaluate two representative solved $N_h = 72$ controllers, using Schur-coordinate ablations as mechanistic interventions on trained controllers.

Takeaway: in the dense orthogonal copy solution, T_{CC} is nearly loss-preserving while $T_{C\to R}$ is not; in the Cayley-transform solution, the tested nonnormal couplings are nearly absent and the same ablations have little effect.

4. Task Dependence Beyond the Copy Task

The cross-task suite tests whether the Schur-coordinate interventions remain informative beyond the explicit temporal symmetry of the copy task. The three tasks require discrete memory, oscillatory generation, and context-dependent accumulation, so they probe distinct recurrent computations in the same one-layer architecture. The experiments use one-layer tanh RNNs with $N_h = 64$, $W_{hh}^{(0)} = Q$, $Q^\top Q = I$, Adam with learning rate 10^{-3}, batch size 64, 30 epochs, and 128 batches per epoch. Three seeds are trained for each of 3-bit flip-flop with length 25, frequency-cued sine generation with length 50, and context-dependent integration with four inputs and length 48. Full models have held-out mean $\mathrm{FVU} = 0.0048$ for flip-flop, 0.0036 for sine generation, and 0.0104 for context-dependent integration.

Values are mean $\pm$ SEM over seeds (Figure 2). For flip-flop, zeroing $T_{C\to R}$ increases held-out error by $9.45 \times 10^{-2} \pm 9.35 \times 10^{-3}$, while zeroing T_{CC} increases error by $4.96 \times 10^{-2} \pm 5.39 \times 10^{-3}$. The ring-internal block T_{RR} has almost no raw effect.

For sine generation, zeroing T_{CC} raises held-out error by 2.08 ± 0.23, and zeroing $T_{C\to R}$ raises it by 1.73 ± 0.34. The normalized sensitivity is largest for $T_{C\to R}$, 21.1 ± 5.1, with T_{CC} still substantial at 12.3 ± 1.5. Removing T_{RR} has little effect at this width. In context-dependent integration, zeroing T_{CC} raises held-out error by 0.94 ± 0.03, while zeroing $T_{C\to R}$ raises it by 0.37 ± 0.16. The raw effect is dominated by T_{CC}, consistent with a slow accumulated variable supported by within-complement recurrence.

Across tasks, selected Schur couplings can be removed with little loss when they avoid task-relevant directions, as in the copy task T_{CC} pairs. The same coordinates localize fragile directions when a block is required, as in sine generation and context-dependent integration.

Takeaway: the loss-preserving ablation profile varies across tasks and trained solutions; no single Schur coupling is uniformly safe to remove.

Metric interpretation. Raw degradation measures loss at the trained operating point, whereas $S_{\Delta T}$ measures loss per unit removed Schur mass. We treat $\Delta\mathrm{FVU}$ as the primary behavioral effect and use $S_{\Delta T}$ to identify small sectors with disproportionate impact.

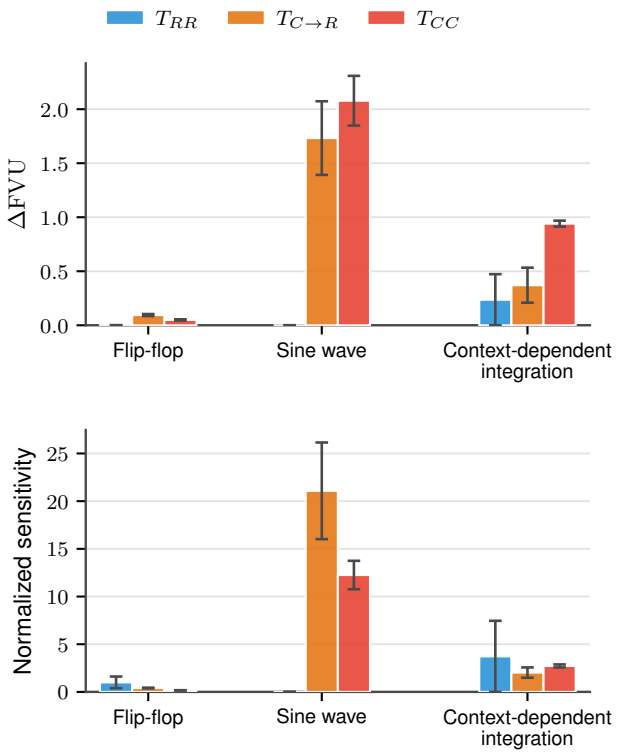

Figure 2. Single-block Schur ablations across neuroscience-style tasks. Top: raw degradation $\Delta\mathrm{FVU}$. Bottom: normalized sensitivity $S_{\Delta T}$. The loss-preserving ablation profile depends on the computation: raw degradation is largest for $T_{C\to R}$ in flip-flop and for complement-linked blocks in sine generation and context-dependent integration.

5. Discussion and Limitations

Interpretation. Exact symmetries characterize functional equivalence classes in weight space (Entezari et al., 2022; Ainsworth et al., 2023; Navon et al., 2023). Ordered Schur coordinates play a complementary role by fixing an orthogonal coordinate system in which recurrent matrices can be compared and nonnormal sectors can be causally ablated. The resulting equivalences are task-restricted and approximate, because they are defined by rollout behavior on $\mathcal{D}$ rather than by a global parameter-space group action. For recurrent networks, raw parameter distance can miss both large structured changes that preserve the task function and small directed changes that alter it.

Because the tasks studied here are low-dimensional, the trained networks may use only a low-dimensional hidden-state subspace. A Schur ablation can then preserve performance because it avoids the activity directions aligned with the readout or the dominant hidden-state principal components, rather than because the removed coupling has no

computational role. The experiments do not separate this subspace explanation from the Schur-coordinate account. Separating the two would require measuring how the ablated Schur directions project onto hidden-state PCs, readout-aligned subspaces, and task-conditioned activity manifolds.

Scope. The experiments use vanilla one-layer tanh RNNs, simple low-dimensional tasks, a narrow width range, and a small number of trained solutions. They do not test LSTMs, GRUs, gated architectures, large sequence models, or high-dimensional real-world sequence tasks, so the evidence supports Schur-coordinate ablation as a diagnostic for trained recurrent controllers rather than a universal statement about nonnormal structure.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Arjovsky, M., Shah, A., and Bengio, Y. Unitary evolution recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1120–1128. PMLR, 2016. URL https://proceedings.mlr.press/v48/arjovsky16.html.

Bondanelli, G. and Ostojic, S. Coding with transient trajectories in recurrent neural networks. *PLOS Computational Biology*, 16(2):e1007655, 2020. doi: 10.1371/journal.pcbi.1007655.

Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dNigytemkL.

Hennequin, G., Vogels, T. P., and Gerstner, W. Non-normal amplification in random balanced neuronal networks. *Physical Review E*, 86(1):011909, 2012. doi: 10.1103/PhysRevE.86.011909.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Keller, T. A., Muller, L., Sejnowski, T., and Welling, M. Traveling waves encode the recent past and enhance sequence learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=p4S5Z6Sah4.

Kofinas, M., Knyazev, B., Zhang, Y., Chen, Y., Burghouts, G. J., Gavves, E., Snoek, C. G. M., and Zhang, D. W. Graph neural networks for learning equivariant representations of neural networks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oO6FsMyDBt.

Maheswaranathan, N., Williams, A. H., Golub, M. D., Ganguli, S., and Sussillo, D. Universality and individuality in neural dynamics across large populations of recurrent networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503:78–84, 2013. doi: 10.1038/nature12742.

Murphy, B. K. and Miller, K. D. Balanced amplification: A new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009. doi: 10.1016/j.neuron.2009.02.005.

Navon, A., Shamsian, A., Achituve, I., Fetaya, E., Chechik, G., and Maron, H. Equivariant architectures for learning in deep weight spaces. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25790–25816. PMLR, 2023. URL https://proceedings.mlr.press/v202/navon23a.html.

Navon, A., Shamsian, A., Fetaya, E., Chechik, G., Dym, N., and Maron, H. Equivariant deep weight space alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37376–37395. PMLR, 2024. URL https://proceedings.mlr.press/v235/navon24a.html.

Pattadkal, J. J., Zemelman, B. V., Fiete, I., and Priebe, N. J. Primate neocortex performs balanced sensory amplification. *Neuron*, 112(4):661–675.e7, 2024. doi: 10.1016/j.neuron.2023.11.005.

Schuessler, F., Mastrogiuseppe, F., Ostojic, S., and Barak, O. Aligned and oblique dynamics in recurrent neural networks. *eLife*, 13:RP93060, 2024. doi: 10.7554/eLife.93060.3.

Sussillo, D. and Barak, O. 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.

Trefethen, L. N. and Embree, M. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005. ISBN 9780691119465. doi: 10.1515/9780691213101.

Zhou, A., Yang, K., Burns, K., Cardace, A., Jiang, Y., Sokota, S., Kolter, J. Z., and Finn, C. Permutation equivariant neural functionals. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24966–24992. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=fmYmXNPmhv.