Multiple temporal credit assignment rules achieve comparable neural data similarity

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the quest to understand how the brain's learning capabilities stem from its ingredients, developing biologically plausible learning rules presents a promising approach. These rules, often relying on gradient approximations, need to be examined for their effectiveness in areas other than task accuracies. This study assesses whether models trained with biologically plausible learning rules can emulate neural data similarity achieved by models trained with Backpropagation Through Time (BPTT). Employing methods such as Procrustes Analysis, we compare well-known neuroscience datasets and discover that models using approximate gradient-based rules show neural data similarities comparable to those trained with BPTT at equal accuracies. Our findings reveal that model architecture and initial conditions have a more pronounced impact on these similarities than the learning rules themselves. Furthermore, our analysis indicates that BPTT-trained models and their biologically plausible counterparts exhibit similar dynamical properties at comparable accuracies. Overall, these results demonstrate the capability of biologically plausible models to not only approximate gradient descent learning in terms of task performance but also emulate its ability to capture neural activity patterns.

1 Introduction

Understanding how animals learn complex behaviors that span multiple temporal scales is a fundamental question in neuroscience. Effectively updating synaptic weights to achieve such learning requires solving the temporal credit assignment problem: determining how to assign the contribution of past neural states to future outcomes. In pursuit of answers, neuroscientists have increasingly adopted the mathematical framework of training recurrent neural networks (RNNs) as a model for brain learning mechanisms, inspired by seminal works that have laid the foundation for this approach [1–4]. This pivot has ushered in a variety of biologically plausible (or bio-plausible for short) learning rules, proposing mechanisms by which learning can be achieved using only known biological processes [5, 6]. However, there has been little work on how these proposals connect to neural data, especially in light of the recent growing availability of neural data [7].

Navigating the vast space of computational models — which vary not only in learning rules but also in architecture and tasks [6, 8, 9] — necessitates a systematic comparison of model representations with empirical brain data. To address this challenge, a variety of methods have been developed, aiming to quantify the similarity between computational models and neural data. Among these, popular methodologies include linear regression [10], Representational Similarity Analysis (RSA) [11], Centered Kernel Alignment (CKA) [12], Singular Vector Canonical Correlation Analysis [13], Procrustes distance [14–16], and Dynamical Similarity Analysis (DSA) [17]. By comparing the

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

geometry of state representations or the dynamics of neural activity, these methods provide a critical framework for evaluating the extent to which models approximate neural systems.

Leveraging existing comparison methodologies, we compute the similarity scores of RNN models trained with bio-plausible learning rules to experimental data. Specifically, we evaluate those similarity scores by comparing them against those achieved by Backpropagation Through Time (BPTT)-trained models. This comparison enables us to assess the efficacy of bio-plausible learning rules as approximations of gradient-descent learning in terms of data similarity. Importantly, the widespread use of task-trained RNNs for modeling brain functions predominantly relies on BPTT [18], despite its bio-plausibility being under scrutiny. It remains an open question whether bio-plausible learning algorithms yield networks with neural similarity comparable to those of BPTT trained networks. Has the pursuit of more biologically plausible learning rules gained biological plausibility at the level of synaptic implementation and parameter updates, but lost biological realism at the level of neural activity?

Main contributions: Our findings reveal that the distance between data and models trained with truncation-based bio-plausible learning rules is comparable to the distance achieved by models trained using Backpropagation Through Time (BPTT). We specifically focus on learning rules that approximate the gradient by truncating bio-implausible terms, as these truncation-based bio-plausible rules have demonstrated efficacy and versatility in learning non-trivial tasks [19, 20]. Other training strategies for RNNs either face bio-plausibility issues, or have limited success and flexibility on non-trivial tasks (see Related Works in Appendix A). Specifically, our contributions include:

- First, we benchmark well-known neuroscience datasets (Mante 2013 [4] and Sussillo 2015 [21]) using state-of-the-art similarity methods (particularly Procrustes distance) to demonstrate that at equal accuracies, RNNs trained with truncation-based bio-plausible rules achieve a level of similarity to data that is comparable to those trained with their deep learning counterpart, BPTT (Figure 1 and Appendix Figure 7).
- Second, we further highlight the indistinguishability of different learning rules by demonstrating that the impact of architectural and initial condition variations — particularly initial weight settings — can surpass the differences in Procrustes distances observed across the learning rules (Figure 2).
- To explain the comparable similarities, we investigate the commonalities between BPTT and its bio-plausible counterparts. Specifically, we demonstrate that BPTT exhibits increased similarity to bio-plausible models at a lower learning rate, as illustrated in Figure 3. Furthermore, we analyze their resemblance in terms of the post-training weight eigenspectrum and dynamical properties (explored via DSA) in Appendix Figure 8.

2 Results

In our study, we analyze the similarity between task-trained RNN models and two neural datasets: Sussillo *et al.*[21] and Mante *et al.*[4]. An overview of our methodology is provided in Figure 1A, with detailed information about our RNN model setup, similarity measure, and datasets in Appendix B. We examine the similarity of RNN models, across different learning rules, to neural data, leveraging Procrustes analysis. Figure 1B shows that multiple learning rules, specifically BPTT and its truncation-based biologically plausible alternative (e-prop), achieve similar Procrustes distances from neural data across two distinct tasks: Sussillo 2015 [21] and Mante 2013 [4]. Although the error bars for BPTT and e-prop do not appear to overlap near perfect accuracy in the Sussillo 2015 task, we demonstrate that such differences are minimal compared to other potential confounding factors in the brain, as shown in Appendix Figure 7.

Also, as a sanity check, we verified whether the observed similarity in data proximity is confined to specific learning rules, we also evaluated older learning methods such as node perturbation and evolutionary strategies. Results show that these methods resulted in greater Procrustes distances compared to the aforementioned rules at equivalent accuracy levels, checking that not all learning rules are equally effective. This also indicate the effectiveness of newer bio-plausible gradient-approximating learning rules over some of the older methods (Appendix Figure 9).

Additionally, Figure 2 delves into the impact of initial weight settings on model-data distances, revealing that such initial condition nuances exert a more pronounced influence than the choice of

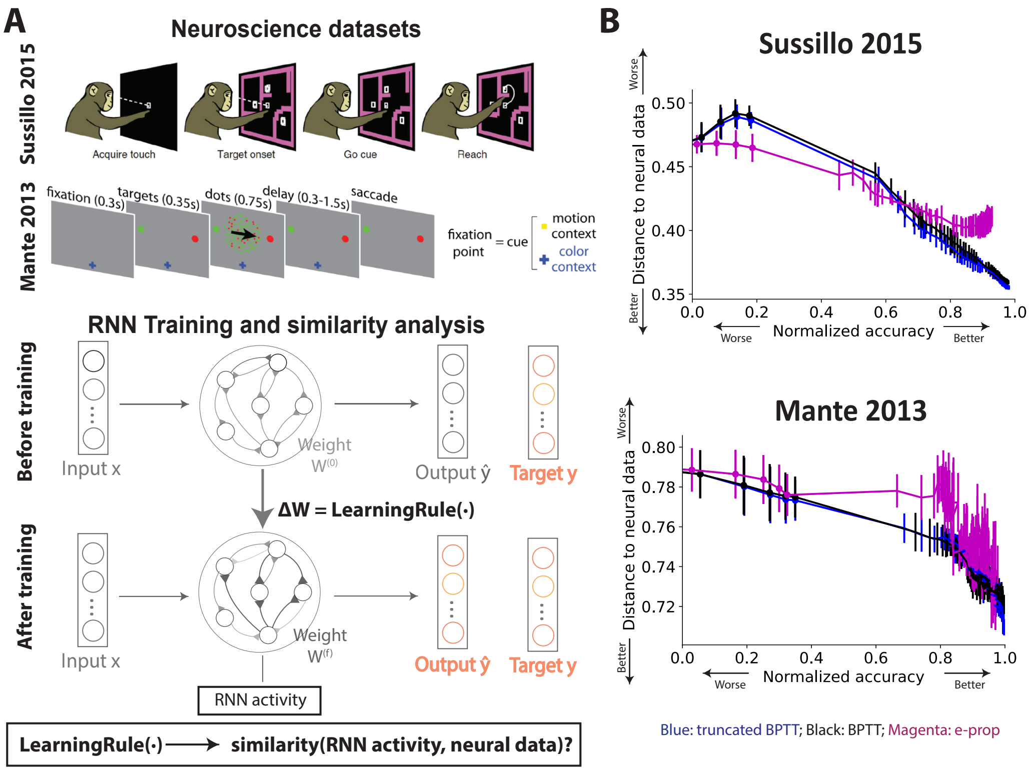

Figure 1: (A) Setup overview: analysis of two neural datasets. We computed similarity scores between RNN activity and electrode recordings from (1) Mante *et al.* (2013) [4] and (2) Sussillo *et al.* (2015) [21]. Schematics have been modified from those in the original papers. RNNs are trained on these respective tasks using various learning rules, including BPTT and bio-plausible alternatives. Subsequently, we evaluate the similarity between RNN activity post-training and the neural recordings to compare model-data similarity across different learning rules. (B) The Procrustes distance vs. accuracy plots for the Sussillo 2015 (top) and Mante 2013 (bottom) tasks illustrate multiple learning rules achieve comparable data similarity. Here, magenta is for e-prop, blue is for truncated BPTT, and black is for BPTT. The mean is plotted with error bars denoting the standard deviations across four different seeds. The x-axis, normalized accuracy, is defined in Appendix B.5. Although there is a slight difference in the distances between e-prop and BPTT at higher accuracies for Sussillo 2015, we demonstrate that such differences are minimal compared to other potential confounding factors in the brain (Figure 2 and Appendix Figure 7).

learning rule itself. Initial weight gain is a crucial attribute, as it significantly affects the dynamical properties of RNNs, particularly the Lyapunov exponents that govern the rates of expansion and contraction. It can also interpolate between rich and lazy learning regimes, imparting distinct inductive biases [23–30]. This finding further underscores the significant role of model initialization in shaping learning outcomes, with particular initial conditions facilitating a closer approximation to neural data than others.

Figure 3 explores the impact of learning rates on model-data distances across learning rules. In Figure 3A, Procrustes distances remain consistent across learning rates for BPTT. Given that e-prop can be decomposed into a lower learning rate BPTT and an approximation error [22], which is further illustrated here by the similarity between a lower learning rate BPTT and e-prop (Figure 3B), this shared component of a lower learning rate BPTT could partly explain their similar distances. Additionally, post-training weight eigenspectrums and distances, analyzed via Dynamical Similarity Analysis (DSA), further reinforce the similarity between BPTT and e-prop (Appendix Figure 8). This similarity is further explored in Appendix Figure 5, where top demixed principle components show a qualitative match between the neural data and the models. We also display the similarity among

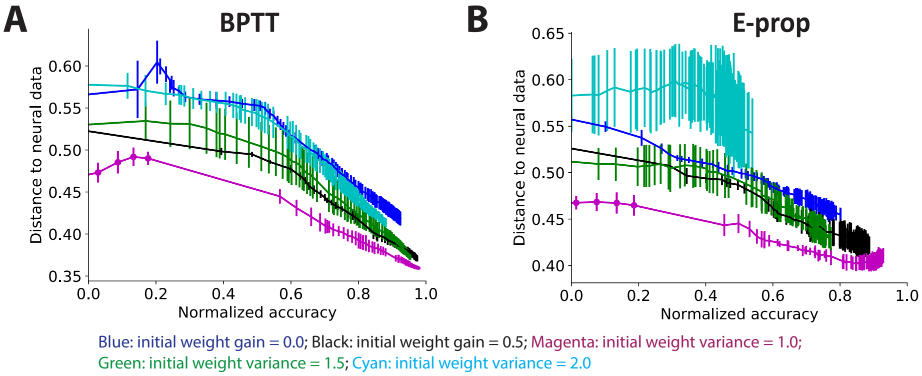

Figure 2: **Impact of Initial Weight Magnitude on Model-Data Distances Exceeds Variation from Learning Rules**. Model-data distances versus normalized accuracy for various initial gain values (depicted by different colors) for (A) BPTT and (B) e-prop. Initial weight gain refers to the multiplier applied to the default initializations for recurrent and readout weights. The results shown are for the Sussillo 2015 task, with similar trends observed for the Mante 2013 task. The mean is plotted with error bars representing the standard deviation.

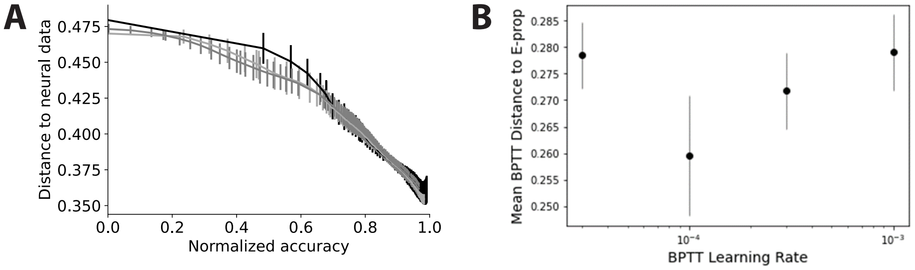

Figure 3: (A) Procrustes distances remain consistent across various learning rates when employing the same rule (BPTT). Different color shades represent different learning rates: $1e-3$, $3e-4$, $1e-4$, and $3e-5$. These rates result in nearly indistinguishable Procrustes distances. The analysis in this figure is done using the Sussillo 2015 task. (B) E-prop — has been viewed as BPTT with a reduced learning rate plus some degree of gradient approximation error [22] — aligns more closely with BPTT at a lower learning rate ($1e-4$) compared to the default setting ($1e-3$). Here, the mean distance from BPTT to e-prop is plotted, with error bars denoting the standard deviation.

models in terms of their pairwise distances and their embeddings across different sampled training snapshots in Appendix Figure 6.

It is noteworthy that if all models were equally far from the data, it might also suggest random noise. However, that is not the case, as our models are significantly closer to the neural data after training (Figure 4). Additionally, what does it mean for a model to be close to the data? To interpret model-data closeness, we need a baseline based on data-to-data similarity, which reflects how close the models are to the data relative to other data points (subsamples within the dataset). Due to limited subjects, we generated this baseline by splitting the data by neurons, though this may create an overly stringent baseline due to potential neuron dependence (details in Appendix B.5). For the Hatsopoulos2007 dataset [31], the final trained models match the neural data as closely as other neurons (Figure 4). For the Sussillo2015 dataset, trained models approach the noise floor compared to untrained models; the remaining differences from the baseline offer insights for improving learning algorithms and architectures in future work.

3 Discussion

To decipher how the brain's intricate learning capabilities emerge from its biological processes, various biologically plausible learning rules have been proposed [6, 5], leaving their connection to

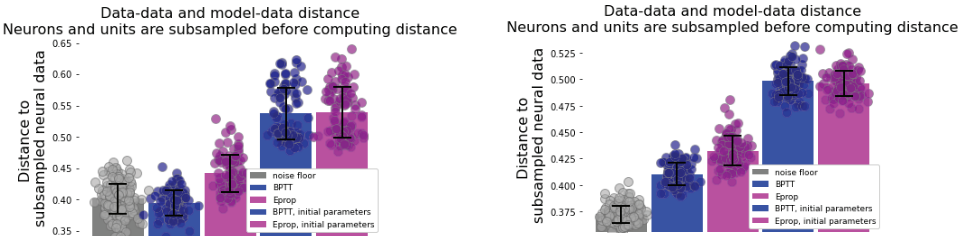

Figure 4: Data-to-data distance (noise floor) vs model-to-model distance (BPTT and e-prop before and after training). Left: Hatsopolous 2007; right: Sussillo 2015. The data-splitting procedure for obtaining the baseline (i.e. noise floor) is detailed in Appendix B.5. We note that these distances are computed with fewer neurons (about half) and units than the previous plots, so the exact distance values here may differ.

neural activity as an open question. This study investigates RNN models trained with approximate gradient-based biologically plausible learning rules, comparing their neural data similarity to models trained using the standard BPTT algorithm. Grounded in state-of-the-art comparison methods like Procrustes Analysis, our analysis reveals that at equal accuracies, RNNs employing truncation-based bio-plausible learning rules exhibit levels of similarity to empirical neural data strikingly comparable to those achieved by BPTT-trained models (Figures 1 and 2). Further probing into this similarity, we find that BPTT shows an increased resemblance to bio-plausible models at lower learning rates (Figure 3), with further examination of their congruence in post-training weight eigenspectrum and dynamical properties through Dynamical Similarity Analysis (DSA) (Appendix Figure 8). Moreover, our research reveals that architectural nuances and initial condition variations can significantly influence model-data similarity, overshadowing the impact of the learning rule choice itself (Figures 2 and 7). Such insights affirm the efficacy of bio-plausible learning rules and encourage a reevaluation of the factors most critical for aligning model activity with real neural systems.

Extending our approach to encompass a broader spectrum of learning rules, architectures, datasets, and comparison methods is a crucial direction for future research. A comprehensive evaluation across these dimensions exceeds the scope of a single paper, especially in a rapidly evolving research landscape. Our study demonstrates the existence of scenarios where biologically plausible rules and their deep learning counterparts achieve comparable data similarities. Furthermore, our pipeline is flexible, allowing for expansion across these various facets in future investigations. On the learning rule front, we primarily examined rules involving gradient truncations, chosen for their biological plausibility, proven efficacy in task learning, and versatility in settings that eschew the equilibrium assumption [32, 33], as detailed in the Related Works section in Appendix A. These rules have been the subject of several recent studies within the computational neuroscience community [34, 22]. Additionally, our analysis is predicated on the concept of learning through synaptic credit assignment, yet other approaches — e.g. in-context learning [35] if it can be implemented biologically — warrant future examination. In addition to learning rules, other model attributes — particularly architecture and initialization, as illustrated in Figure 2 — are crucial areas for future research. Although our results demonstrate comparable similarities at equal accuracies, this does not imply that e-prop is indistinguishable from BPTT. In fact, e-prop accuracies falls short on some of the more challenging tasks [36]. Future experimental neuroscience research could focus on obtaining data from these challenging tasks where e-prop training fails to perform well and conduct further comparisons using these tasks. Furthermore, we chose to focus on Procrustes distance for its ability to provide a proper metric for comparing the geometry of state representations, and its stringency in only allowing for rotations and a global stretching to align neural trajectories. We were also motivated to emphasize Procrustes distance because several weaknesses have been identified in other similarity measures that are, for example, biased due to high dimensionality, or may rely on low variance noise components of the data [12, 37–39]. That said, like all scalar measures, it focuses on specific structures, and it remains uncertain whether these structures accurately capture the computational properties of interest. Therefore, developing new measures remains a crucial and intriguing endeavor [40–45]. Altogether, this vibrant area — which focuses on comparing neurally plausible learning rules with neural data — is ripe for exploration across various knobs including learning rules, architecture, tasks/datasets, and comparison methodologies.

References

[1] David Zipser. Recurrent network model of the neural mechanism of short-term active memory. *Neural Computation*, 3(2):179–193, 1991.

[2] Eberhard Fetz. Are movement parameters recognizably coded in the activity of single neurons?, 1992.

[3] Sohie Lee Moody, Steven P. Wise, Giuseppe di Pellegrino, and David Zipser. A model that accounts for activity in primate frontal cortex during a delayed matching-to-sample task. *Journal of Neuroscience*, 18(1):399–410, 1998.

[4] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[5] Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020.

[6] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[7] Christof Koch, Karel Svoboda, Amy Bernard, Michele A Basso, Anne K Churchland, Adrienne L Fairhall, Peter A Groblewski, Jérôme A Lecoq, Zachary F Mainen, Mackenzie W Mathis, et al. Next-generation brain observatories. *Neuron*, 110(22):3661–3666, 2022.

[8] Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):1–7, 2019.

[9] Guangyu Robert Yang and Manuel Molano Mazon. Next-generation of recurrent neural network models for cognition. 2021.

[10] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[11] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

[12] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

[13] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

[14] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.

[15] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021.

[16] Lyndon R Duong, Jingyang Zhou, Josue Nassar, Jules Berman, Jeroen Olieslagers, and Alex H Williams. Representational dissimilarity metric spaces for stochastic neural networks. *arXiv preprint arXiv:2211.11665*, 2022.

[17] Mitchell Ostrow, Adam Eisen, Leo Kozachkov, and Ila Fiete. Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Guangyu Robert Yang and Xiao-Jing Wang. Artificial neural networks for neuroscientists: a primer. *Neuron*, 107(6):1048–1070, 2020.

[19] Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 11(1):3625, 2020.

[20] Roy Henha Eyono, Ellen Boven, Arna Ghosh, Joseph Pemberton, Franz Scherr, Claudia Clopath, Rui Ponte Costa, Wolfgang Maass, Blake A Richards, Cristina Savin, et al. Current state and future directions for learning in biological recurrent neural networks: A perspective piece. *Neurons, Behavior, Data analysis, and Theory*, 1, 2022.

[21] David Sussillo, Mark M Churchland, Matthew T Kaufman, and Krishna V Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, 18(7):1025–1033, 2015.

[22] Yuhan Helena Liu, Arna Ghosh, Blake A Richards, Eric Shea-Brown, and Guillaume Lajoie. Beyond accuracy: generalization properties of bio-plausible temporal credit assignment rules. *arXiv preprint arXiv:2206.00823*, 2022.

[23] Lukas Braun, Clémentine Dominé, James Fitzgerald, and Andrew Saxe. Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, 35:6615–6629, 2022.

[24] Timo Flesch, Andrew Saxe, and Christopher Summerfield. Continual task learning in natural and artificial agents. *Trends in Neurosciences*, 46(3):199–210, 2023.

[25] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

[26] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[27] Friedrich Schuessler, Francesca Mastrogiuseppe, Srdjan Ostojic, and Omri Barak. Aligned and oblique dynamics in recurrent neural networks. *arXiv preprint arXiv:2307.07654*, 2023.

[28] Blake Bordelon and Cengiz Pehlevan. The influence of learning rule on representation dynamics in wide neural networks. *arXiv preprint arXiv:2210.02157*, 2022.

[29] Yuhan Helena Liu, Aristide Baratin, Jonathan Cornford, Stefan Mihalas, Eric Shea-Brown, and Guillaume Lajoie. How connectivity structure shapes rich and lazy learning in neural circuits. *arXiv preprint arXiv:2310.08513*, 2023.

[30] Jonas Paccolat, Leonardo Petrini, Mario Geiger, Kevin Tyloo, and Matthieu Wyart. Geometric compression of invariant manifolds in neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(4):044001, 2021.

[31] Nicholas G Hatsopoulos, Qingqing Xu, and Yali Amit. Encoding of movement fragments in the motor cortex. *Journal of Neuroscience*, 27(19):5105–5114, 2007.

[32] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24, 2017.

[33] Alexander Meulemans, Nicolas Zucchet, Seijin Kobayashi, Johannes Von Oswald, and João Sacramento. The least-control principle for local learning at equilibrium. *Advances in Neural Information Processing Systems*, 35:33603–33617, 2022.

[34] Jacob Portes, Christian Schmid, and James M Murray. Distinguishing learning rules with brain machine interfaces. *Advances in neural information processing systems*, 35:25937–25950, 2022.

[35] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[36] Yuhan Helena Liu, Stephen Smith, Stefan Mihalas, Eric Shea-Brown, and Uygar Sümbül. Cell-type–specific neuromodulation guides synaptic credit assignment in a spiking neural network. *Proceedings of the National Academy of Sciences*, 118(51):e2111821118, 2021.

[37] MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of cka as a similarity measure in deep learning, 2022.

[38] Marin Dujmović, Jeffrey S Bowers, Federico Adolfi, and Gaurav Malhotra. Obstacles to inferring mechanistic similarity using representational similarity analysis. *bioRxiv*, 2023.

[39] Eric Elmoznino and Michael F. Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20(1):1–23, 01 2024.

[40] Nicholas J. Sexton and Bradley C. Love. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, 8(28):eabm2219, 2022.

[41] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

[42] Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations, 2023.

[43] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures, 2023.

[44] Jeffrey S. Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E. Hummel, Rachel F. Heaton, and et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385, 2023.

[45] Andrew Kyle Lampinen, Stephanie C. Y. Chan, and Katherine Hermann. Learned feature representations are biased by complexity, learning order, position, and more, 2024.

[46] Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

[47] Axel Laborieux and Friedemann Zenke. Holomorphic equilibrium propagation computes exact gradients through finite size oscillations. *arXiv preprint arXiv:2209.00530*, 2022.

[48] Will Greedy, Heng Wei Zhu, Joseph Pemberton, Jack Mellor, and Rui Ponte Costa. Single-phase deep learning in cortico-cortical networks. *Advances in Neural Information Processing Systems*, 35:24213–24225, 2022.

[49] João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *arXiv preprint arXiv:1810.11393*, 2018.

[50] Alexandre Payeur, Jordan Guerguiev, Friedemann Zenke, Blake A Richards, and Richard Naud. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature neuroscience*, pages 1–10, 2021.

[51] Pieter R Roelfsema and Anthony Holtmaat. Control of synaptic plasticity in deep cortical networks. *Nature Reviews Neuroscience*, 19(3):166–180, 2018.

[52] Johnatan Aljadeff, James D'amour, Rachel E Field, Robert C Froemke, and Claudia Clopath. Cortical credit assignment by hebbian, neuromodulatory and inhibitory plasticity. *arXiv preprint arXiv:1911.00307*, 2019.

[53] James M Murray. Local online learning in recurrent networks with random feedback. *Elife*, 8:e43299, 2019.

[54] Yuhan Helena Liu, Stephen Smith, Stefan Mihalas, Eric Shea-Brown, and Uygar Sümbül. Biologically-plausible backpropagation through arbitrary timespans via local neuromodulators. *arXiv preprint arXiv:2206.01338*, 2022.

[55] Owen Marschall, Kyunghyun Cho, and Cristina Savin. A unified framework of online learning algorithms for training recurrent neural networks. *The Journal of Machine Learning Research*, 21(1):5320–5353, 2020.

[56] Arna Ghosh, Yuhan Helena Liu, Guillaume Lajoie, Konrad Kording, and Blake Aaron Richards. How gradient estimator variance and bias impact learning in neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

[57] Basile Confavreux, Poornima Ramesh, Pedro J Goncalves, Jakob H Macke, and Tim Vogels. Meta-learning families of plasticity rules in recurrent spiking networks using simulation-based inference. *Advances in Neural Information Processing Systems*, 36, 2024.

[58] Aran Nayebi, Sanjana Srivastava, Surya Ganguli, and Daniel L Yamins. Identifying learning rules from neural network observables. *Advances in Neural Information Processing Systems*, 33:2639–2650, 2020.

[59] Zoe Ashwood, Nicholas A Roy, Ji Hyun Bak, and Jonathan W Pillow. Inferring learning rules from animal decision-making. *Advances in Neural Information Processing Systems*, 33:3442–3453, 2020.

[60] Sukbin Lim, Jillian L McKee, Luke Woloszyn, Yali Amit, David J Freedman, David L Sheinberg, and Nicolas Brunel. Inferring learning rules from distributions of firing rates in cortical neurons. *Nature neuroscience*, 18(12):1804–1810, 2015.

[61] Daniel R Kepple, Rainer Engelken, and Kanaka Rajan. Curriculum learning as a tool to uncover learning principles in the brain. In *International Conference on Learning Representations*, 2021.

[62] Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020.

[63] Matthew G Perich, Charlotte Arlt, Sofia Soares, Megan E Young, Clayton P Mosher, Juri Minxha, Eugene Carter, Ueli Rutishauser, Peter H Rudebeck, Christopher D Harvey, et al. Inferring brain-wide interactions using data-constrained recurrent neural network models. *bioRxiv*, pages 2020–12, 2021.

[64] Friedrich Schuessler, Francesca Mastrogiuseppe, Alexis Dubreuil, Srdjan Ostojic, and Omri Barak. The interplay between randomness and structure during learning in rnns. *Advances in neural information processing systems*, 33:13352–13362, 2020.

[65] Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, 22(2):297–306, 2019.

[66] Elia Turner, Kabir V Dabholkar, and Omri Barak. Charting and navigating the space of solutions for recurrent neural networks. *Advances in Neural Information Processing Systems*, 34:25320–25333, 2021.

[67] Adrian Valente, Srdjan Ostojic, and Jonathan Pillow. Probing the relationship between linear dynamical systems and low-rank recurrent neural network models. *arXiv preprint arXiv:2110.09804*, 2021.

[68] Christopher Langdon and Tatiana A Engel. Latent circuit inference from heterogeneous neural responses during cognitive tasks. *bioRxiv*, 2022.

[69] Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current opinion in neurobiology*, 46:1–6, 2017.

[70] H Francis Song, Guangyu R Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS computational biology*, 12(2):e1004792, 2016.

[71] Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *Advances in neural information processing systems*, 32, 2019.

[72] Stephen J. Smith, Michael Hawrylycz, Jean Rossier, and Uygar Sümbül. New light on cortical neuropeptides and synaptic network plasticity. *Current Opinion in Neurobiology*, 63:176–188, aug 2020.

[73] Jacob Menick, Erich Elsen, Utku Evci, Simon Osindero, Karen Simonyan, and Alex Graves. A practical sparse approximation for real time recurrent learning. *arXiv preprint arXiv:2006.07232*, 2020.

[74] David Sussillo and Larry F Abbott. Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544–557, 2009.

[75] Thomas Miconi. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *Elife*, 6:e20899, 2017.

[76] SueYeon Chung and Larry F Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70:137–144, 2021.

[77] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[78] Rishidev Chaudhuri, Berk Gerçek, Biraj Pandey, Adrien Peyrache, and Ila Fiete. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature neuroscience*, 22(9):1512–1520, 2019.

[79] Marino Pagan, Vincent D Tang, Mikio C Aoi, Jonathan W Pillow, Valerio Mante, David Sussillo, and Carlos D Brody. A new theoretical framework jointly explains behavioral and neural variability across subjects performing flexible decision-making. *bioRxiv*, pages 2022–11, 2022.

[80] Alan D Degenhart, William E Bishop, Emily R Oby, Elizabeth C Tyler-Kabara, Steven M Chase, Aaron P Batista, and Byron M Yu. Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nature biomedical engineering*, 4(7):672–685, 2020.

[81] Mahdiyar Shahbazi, Ali Shirali, Hamid Aghajan, and Hamed Nili. Using distance on the riemannian manifold to compare representations in brain and in models. *NeuroImage*, 239:118271, 2021.

[82] Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. *arXiv preprint arXiv:2311.09466*, 2023.

[83] Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations. *arXiv preprint arXiv:2309.11028*, 2023.

[84] Dean A Pospisil, Brett W Larsen, Sarah E Harvey, and Alex H Williams. Estimating shape distances on neural representations with limited samples. *arXiv preprint arXiv:2310.05742*, 2023.

[85] Jeffrey C. Magee and Christine Grienberger. Synaptic Plasticity Forms and Functions. *Annual Review of Neuroscience*, 43(1):95–117, jul 2020.

[86] Wulfram Gerstner, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea. Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. *Frontiers in Neural Circuits*, 12:53, jul 2018.

[87] Sarah E Harvey, Brett W. Larsen, and Alex H Williams. Duality of bures and shape distances with implications for comparing neural representations. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023.

[88] Manuel Molano-Mazon, Joao Barbosa, Jordi Pastor-Ciurana, Marta Fradera, Ru-Yuan Zhang, Jeremy Forest, Jorge del Pozo Lerida, Li Ji-An, Christopher J Cueva, Jaime de la Rocha, et al. Neurogym: An open resource for developing and sharing neuroscience tasks. 2022.

[89] Justin Werfel, Xiaohui Xie, and H Seung. Learning curves for stochastic gradient descent in linear feedforward networks. *Advances in neural information processing systems*, 16, 2003.

[90] Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1):1–10, 2016.

[91] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[92] Nathan Cloos, Moufan Li, Guangyu Robert Yang, and Christopher J Cueva. Scaling up the evaluation of recurrent neural network models for cognitive neuroscience. 2022.

[93] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[94] Brian DePasquale, Christopher J Cueva, Kanaka Rajan, G Sean Escola, and LF Abbott. full-force: A target-based method for training recurrent networks. *PloS one*, 13(2):e0191527, 2018.

A Extended discussions on related works

Understanding the mechanisms through which the brain learns, utilizing its myriad elements, remains a perennial quest in neuroscience. Recent years have seen a resurgence of interest in proposing biologically plausible learning rules [5, 32, 46–52, 33, 53, 19, 36, 54–57, 6], suggesting potential neural algorithms that leverage known neural components. Despite these advances, relatively little research has focused on how such proposals might connect back to neural circuits. A prevailing line of work concentrates on inferring learning rules directly from neural data [58–61, 34]. In contrast, our approach evaluates different learning rules based on their post-learning activity similarity to neural data, offering a flexible methodology that prioritizes the outcome of learning without necessitating data from before or during the training process.

Our research focuses on learning rules for recurrent neural networks (RNNs), which are extensively used in brain modeling [62–65, 4, 66–71, 9]. This study specifically investigates local learning rules that truncate gradients, as these have shown promising results in task learning and offer versatility across various network architectures. A systematic review [55] recognized random feedback local online (RFLO) as the only fully local (hence bio-plausible) rule. Post-review developments include e-prop, an adaptation of RFLO for non-vanilla (particularly spike-based) RNNs [19], and MDGL [36] with its extension ModProp [54], which further refine the gradient approximation by considering local modulatory signals [72]. These rules are notable for their effectiveness in bio-plausible temporal credit assignment, matching the performance of the more traditional backpropagation through time (BPTT) in many settings [20]. Our study will, therefore, concentrate on these specific learning rules due to their demonstrated efficacy and bio-plausibility. Further details of these rules are explained in Appendix B.4.

Alternative training strategies for RNNs exist, but they either face bio-plausibility issues, lack versatility across settings, or struggle to scale to complex tasks. For instance, equilibrium propagation and related rules depend on the equilibrium assumption [32, 33]. Within truncation-based methods, the SnAP-n algorithm introduced in [73] allows customization by selecting the truncation level n. While SnAp-1 aligns closely with e-prop/RFLO, SnAp-2 and higher n require storing a triple tensor, which poses $O(N^3)$ storage demands not yet proven feasible for neural circuits. Therefore, SnAp-n ($n \geq 2$) remains biologically implausible, while SnAp-1 effectively reduces to e-prop/RFLO under certain conditions. Beyond truncation, the KeRNL algorithm approximates long-term dependencies using first-order low-pass filters and updates parameters via node perturbation, yet this also challenges biological plausibility by requiring frequent meta-parameter updates. Other strategies like FORCE learning [74] offer alternatives, but our scope assumes recurrent weight adjustment and the non-reservoir version faces issues with locality. This study focuses on supervised learning, setting aside the broader field of reinforcement learning for future work, thus not covering certain learning rules like the one in [75].

Comparing high-dimensional neural responses across different systems and contexts is crucial in neuroscience [76] for assessing model quality, determining invariant neural states, and aligning brain-machine interface recordings, among other tasks [77–80]. Among the myriad of methods developed to quantify representational dissimilarity [10, 77, 11, 13, 81, 14–16, 82–84] — such as linear regression, Canonical Correlation Analysis (CCA), Centered Kernel Alignment (CKA), Representational Similarity Analysis (RSA), shape metrics, and Riemannian distance — we focus on Procrustes distance for its ability to provide a proper metric for comparing the geometry of state representations, and because several weaknesses have been identified in other similarity measures that are, for example, biased due to high dimensionality, or may rely on low variance noise components of the data [12, 37–39]. Additionally, we extend our investigation to include Dynamical Similarity Analysis (DSA [17]) in the Appendix, assessing system dynamics to complement our geometric analyses. Overall, the value of these existing measures stems from their ability to compare complex systems without fully understanding them by capturing key structures. However, this strength also poses a limitation: they focus on specific structures, and it remains uncertain whether these structures accurately capture the computational properties of interest. Therefore, developing new measures remains a crucial and intriguing endeavor [40–45].

B Methods

B.1 RNN training setup

Our RNN architecture consists of N_{in} input units, N hidden units, and N_{out} readout units. The update mechanism for the hidden state at time t, $h_t \in \mathbb{R}^N$, follows the equation:

$$h_{t+1} = \beta h_t + (1-\beta)(W_h f(h_t) + W_x x_t), \tag{1}$$

where $\beta = 1 - \frac{dt}{\tau_m} \in \mathbb{R}$ is the leak factor determined by the simulation time step dt and membrane time constant τ_m; $f(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$ represents the $retanh$ activation function; $W_h \in \mathbb{R}^{N \times N}$ and $W_x \in \mathbb{R}^{N \times N_{in}}$ are the recurrent and input weight matrices, respectively; and $x_t \in \mathbb{R}^{N_{in}}$ is the input at time t. The readout, $\hat{y}_t \in \mathbb{R}^{Nout}$, is calculated as a linear combination of the hidden state's activation, $f(h_t)$, with the readout weights $w \in \mathbb{R}^{N_{out} \times N}$.

To train this RNN for the specific tasks in the datasets, we used synthetic input and target output detailed in Appendix B.4. Our objective is to minimize the scalar loss $L \in \mathbb{R}$. For loss minimization, we examine various learning rules, including BPTT (our benchmark) that computes the exact gradient, $\nabla_W L(W_h) \in \mathbb{R}^{N \times (N_{in}+N+N_{out})}$, as well as bio-plausible learning rules that apply approximate gradients, $\tilde{\nabla}_W L(W) \in \mathbb{R}^{N \times (N_{in}+N+N_{out})}$:

$$\Delta W = -\eta \nabla_W L(W), \tag{2}$$

$$\widehat{\Delta W} = -\eta \tilde{\nabla}_W L(W), \tag{3}$$

where $W = [W_h \quad W_x \quad w^T] \in \mathbb{R}^{N \times (N_{in}+N+N_{out})}$ encompasses all trainable parameters and $\eta \in \mathbb{R}$ is the learning rate.

The learning rules investigated in this study are elaborated upon in Appendix B.4. Our analysis centers on how training RNNs with different algorithms influences their similarity to neural data. Predominantly, we concentrate on the truncation-based, bio-plausible rule known as e-prop [19], which simplifies the gradient by retaining only those terms that align with a three-factor learning rule. This includes a Hebbian eligibility trace modulated by a top-down instructive factor, potentially attributable to neuromodulators [85, 86]. It is noteworthy that e-prop is equivalent to the RFLO learning rule introduced in [53] under most conditions. Additionally, we explore ModProp [54], which incorporates cell-type-specific local modulatory signals [72] to recover terms omitted by e-prop. However, due to ModProp's limitations (it is constrained to settings that adhere to Dale's law and employ the $ReLU$ activation function), our examination of this rule is restricted to such specific contexts in Appendix Figure 7.

B.2 Similarity measures

As mentioned in the Introduction, we utilize the metric Procrustes distance [14] to quantify the similarity between the hidden states of RNN models, denoted by $H \in \mathbb{R}^{B*T \times N}$, and the experimentally recorded neural responses, represented as $\tilde{H} \in \mathbb{R}^{B*T \times N'}$. Here, B represents the number of trials or experimental conditions, T denotes the number of time steps in each trial, and N and N' correspond to the number of RNN hidden units and recorded neurons, respectively. The metric Procrustes distance can be viewed as the residual distance after the two neural representations are aligned with an optimal rotation, and is quantified as

$$\theta(H, \tilde{H}) = \min_{Q \in \mathcal{O}} \arccos \left(\frac{< H^{\phi}, \tilde{H}^{\phi} Q >}{\|H^{\phi}\| \|\tilde{H}^{\phi}\|} \right) \tag{4}$$

where $\mathcal{O}$ is the group of orthogonal linear transformations [15, 87].

B.3 Further details on the neural datasets and synthetic data for RNN training

The **Mante 2013** dataset was downloaded from `https://www.ini.uzh.ch/en/research/groups/mante/data.html`. We trained RNNs using a synthetic task setup from Neurogym [88], which included a 350 ms fixation period, a 750 ms stimulus presentation period, a 300 ms delay period, and a 300 ms decision period. The activity of the trained RNNs during the stimulus period was then compared to the downloaded neural dataset using the aforementioned similarity measures. A grid search on the fixation and decision interval durations revealed only minor differences in distances and a consistent trend across learning rules.

The **Sussillo 2015** dataset consisted of electrode recordings from primary motor (M1) and dorsal premotor cortex (PMd) taken while a monkey performed a maze-reaching task consisting of 27 differerent reaching conditions [21]. To assess the similarity between the neural activity and RNNs we compared activity from -1450 ms to 400 ms relative to movement onset. The inputs and outputs to train the models were described in Sussillo et al. 2015, but in brief, for each reach condition there were 16 inputs and 7 target outputs. The 7 outputs were the electromyographic (EMG) signals recorded from 7 muscles as the monkey performed a reaching movement. 15 inputs specified the upcoming reach condition, and were derived from preparatory period neural activity. The remaining input was a hold-cue that took a value of +1 before movemement onset and then a value of 0 to initiate the movement, whereupon the model generated the 7 EMG signals.

B.4 Further details on the learning rule

This subsection aims to clarify the approximation mechanisms employed by each bio-plausible learning rule. For comprehensive descriptions, we recommend consulting the detailed references provided. We begin by expressing the gradient via real-time recurrent learning (RTRL) factorization (an equivalent but causal alternative to the BPTT factorization of the gradient):

$$\frac{\partial L}{\partial W_{h,ij}} \quad = \quad \sum_{l,t} \frac{\partial L}{\partial h_{l,t}} \frac{\partial h_{l,t}}{\partial W_{h,ij}}, \tag{5}$$

The primary challenge with RTRL in terms of biological plausibility and computational efficiency lies in the term $\frac{\partial h_{l,t}}{\partial W_{h,ij}}$ from the gradient decomposition (Eq. 5). This term tracks all recursive dependencies of $h_{l,t}$ on the weight $W_{h,ij}$ due to recurrent connections, calculated recursively as:

$$\begin{aligned} \frac{\partial h_{l,t}}{\partial W_{h,ij}} &= \frac{\partial h_{j,t}}{\partial W_{h,ij}} + \sum_{m} \frac{\partial h_{l,t}}{\partial h_{m,t-1}} \frac{\partial h_{m,t-1}}{\partial W_{h,ij}} \\ &= \frac{\partial h_{l,t}}{\partial W_{h,ij}} + \frac{\partial h_{l,t}}{\partial h_{l,t-1}} \frac{\partial h_{l,t-1}}{\partial W_{h,ij}} + \underbrace{\textstyle\sum_{m \neq l} W_{h,lm} f'(h_{m,t-1}) \frac{\partial h_{m,t-1}}{\partial W_{h,ij}}}_{\text{involving all weights } W_{h,lm}}. \end{aligned} \tag{6}$$

Consequently, $\frac{\partial h_{l,t}}{\partial W_{h,ij}}$ presents a significant challenge for biological plausibility as it includes nonlocal terms, necessitating knowledge of all other network weights for updating each $W_{h,ij}$. **For a learning rule to be biologically plausible, all information required to update a synaptic weight must be physically accessible to that synapse. However, it remains unclear how neural circuits could make such extensive information readily available to every synapse.**

Approaches like **e-prop** [19] and equivalently, **RFLO** [53], address this by truncating the problematic nonlocal terms in Eq. 6, ensuring that updates to $W_{h,ij}$ follow a three-factor framework — the updates rely solely on local pre- and post-synaptic activity and a third top-down instructive signal (e.g. from neuromodulators):

$$\widehat{\frac{\partial h_{l,t}}{\partial W_{h,ij}}} = \begin{cases} \frac{\partial h_{i,t}}{\partial W_{h,ij}} + \frac{\partial h_{i,t}}{\partial h_{i,t-1}} \widehat{\frac{\partial h_{i,t-1}}{\partial W_{h,ij}}}, & l = i \\ 0, & l \neq i \end{cases} \tag{7}$$

which yields a much simpler factor than the comprehensive tensor depicted in Eq. 6. This truncation can be achieved in PyTorch using $h.detach()$, preventing gradient propagation through the recurrent weights.

Putting this together, e-prop can be written in terms of known biological processes including — eligibility trace e and top-down instructive signals I — as [19]:

$$\Delta W_{h,ij}|_{e-prop} = \sum_{t} I_{i,t} e_{ij,t}, \tag{8}$$

where $I_{i,t} = \frac{\partial L}{\partial h_{i,t}}$ is the top-down instructive signal (e.g. from neuromodulator dopamine, neuronal firing, etc. [86, 19]) sent to neuron i at time t, and $e_{ij,t} = \widehat{\frac{\partial h_{i,t}}{\partial W_{h,ij}}} = \frac{\partial h_{i,t}}{\partial W_{h,ij}} + \frac{\partial h_{i,t}}{\partial h_{i,t-1}} \widehat{\frac{\partial h_{i,t-1}}{\partial W_{h,ij}}}$ is the eligibility trace for synapse (ij) at time t. This is a three-factor rule, with the pre-and postsynaptic neuron factors in the eligibility trace as well as a third factor from the instructive signal.

Besides eligibility traces and top-down instructive signals, recent transcriptomics data [72] suggest the presence of widespread cell-type-specific local modulatory signals that could convey additional information for guiding synaptic weight updates. **ModProp** is developed to incorporate these processes and restore the gradient terms truncated by e-prop, thereby improving the approximation of the gradient. Specifically, the ModProp update rule is described as follows [54]:

$$\begin{aligned}\Delta W_{h,ij}|_{ModProp} &\propto I_i \times e_{ij} + \left(\sum_{\alpha\in C}\left(\sum_{l\in\alpha} I_l h'_l\right)\times F_{\alpha\beta}\right) * e_{ij},\\ F_{\alpha\beta,s} &= \mu^{s-1}(W^s)_{\alpha\beta},\end{aligned}\tag{9}$$

where I and e again denote the top-down learning signal and the eligibility trace, respectively. Here, neuron j belongs to type α, neuron p to type β, and C denotes the set of cell types. $F_{\alpha\beta}$ is hypothesized to represent type-specific filter taps of GPCRs expressed by cells of type β in response to precursors secreted by cells of type α. The operator $*$ denotes convolution, and s indexes the filter taps. The hyperparameter μ, set to 0.25 in this study, and the genetically predetermined $(W^s)_{\alpha\beta}$ values for different filter taps $F_{\alpha\beta,s}$ could be optimized over evolutionary timescales [54].

We also explored an older learning rule, **node perturbation** [89, 90], which is known to have trouble scaling beyond small-scale networks and simple tasks. Specifically, it is implemented by

$$\Delta W_{h,ij}|_{NP} \propto \sum_t \widehat{I_{i,t}} e_{ij,t},\tag{10}$$

where $\widehat{I_t} = (L_t(h_t + \xi) - L_t(h_t))\xi/\sigma^2$ provides an estimate to $\frac{\partial L}{\partial h_t}$; elements of ξ are chosen independently from a zero-mean Gaussian distribution with variance σ^2.

In addition, we explored **evolutionary strategies** [91] for parameter updates in our model. This method, for a Gaussian distribution, is implemented as follows:

$$\Delta W_{h,ij}|_{ES} \propto \frac{1}{\sigma S}\sum_{s=1}^{S} L^{(s)}\epsilon^{(s)},\tag{11}$$

where $\epsilon^{(s)}$ is sampled from a standard normal distribution $\mathcal{N}(0, I)$ for $s = 1, ..., S$. Here, $L^{(s)}$ represents the loss function evaluated after perturbing the parameter by $\sigma\epsilon^{(s)}$, σ is the standard deviation of the perturbations, and S is the number of samples. Due to computational constraints, we set S to 50 for our experiments.

B.5 Additional details on training and analysis

Our model-data comparison method utilizes Procrustes distances, as implemented in `https://github.com/ahwillia/netrep`, with the configuration set to $metric = LinearMetric(alpha = 1.0, center\_columns = True)$. Additionally, in Appendix Figure 8, we employed Dynamical Systems Analysis (DSA), available at `https://github.com/mitchellostrow/DSA/tree/main`. For this analysis, we tested with hyperparameters $n\_delays \in \{5, 10, 15, 20\}$ and $rank \in \{10, 20, 30, 40\}$, observing consistent trends across settings. We did not test values beyond these ranges due to computational resource limitations. For the loss used in training RNNs, we used cross-entropy loss for the Mante 2013 task and mean-squared error for the Sussillo 2015 task (with EMG outputs as the targets [92]). As mentioned, the Mante 2013 dataset was downloaded from `https://www.ini.uzh.ch/en/research/groups/mante/data.html`. However, we obtained the Sussillo 2015 dataset from the original authors and do not have permission to redistribute it.

Our code is available at `https://anonymous.4open.science/r/XYZ2442-860A/`. We utilized PyTorch Version 1.10.2 [93]. Simulations were executed on a server equipped with two 20-core Intel(R) Xeon(R) CPU E5-2698 v4 at 2.20GHz. The average training duration for tasks was about 10 minutes, and the analysis pipeline required approximately 2 minutes per model. Training employed the Adam optimizer. Unless otherwise noted, the learning rate was set at $1e-3$, optimized through a grid search of $\{3e-3, 1e-3, 3e-4, 1e-4\}$. We used a batch size of around 100; changes in this parameter led to negligible differences in the results. The number of time steps, T, for the

Sussillo task was set to 186, matching the original data. The number of time steps T, for the Mante task was 34, based on $dt = 50$ ms from the original Mante paper and the total task duration in the Neurogym setting. Similar trends were observed when we varied dt and the durations of the fixation and delay periods. We employed 200 hidden units for the Sussillo 2015 task and 400 hidden units for the Mante 2013 task; doubling these numbers resulted in similar trends. Each simulation was repeated with four different seeds (except for 10 seeds for Figure 3B), and results for each seed were plotted as separate lines in our figures. Training involved 1000 SGD iterations for Sussillo 2015 and 3000 for Mante 2013, with input, recurrent, and readout weights all trainable. Local learning rule approximations were specifically applied to input and recurrent weights, due to the locality issues discussed in Section B.4. Unlike these weights, readout weights do not encounter such issues; hence, by default, the same readout weights were used for both forward and backward computations. However, as verified in Appendix Figure 10, employing random feedback readout weights for training (i.e., feedback alignment [90]) resulted in comparable distances.

By default, zero-mean Gaussian noise with a standard deviation of 0.1 was added to the hidden activity, except in cases where the noise was removed to assess its impact. Typically, no connectivity constraints were applied, except for settings in Figure 7B where only 25% of recurrent weights were set as nonzero and trainable, and in Figure 7C where 80% of the neurons were enforced as strictly excitatory and 20% as inhibitory. To enforce Dale's law, we used the same masking procedure in [18]. To initialize the weights, we initialized with random Gaussian distributions where each weight element $W_{h,ij} \sim \mathcal{N}(0, g^2/N)$, with an initial weight variance of g; unless otherwise mentioned, we set $g = 1.0$. Input and readout weights were initialized similarly as in [18] (see their $EIRNN.ipynb$ notebook).

Normalized accuracy, which appears as the x-axis in several plots, is defined such that a value of 1 corresponds to perfect performance. For Sussillo 2015, normalized accuracy is calculated as $1 -$ normalized mean squared error, as used in [94] In the case of Mante 2013, which involves a classification task where mean squared error is not applicable, normalized accuracy is computed as $1 -$ cross entropy loss to maintain consistency with the definition where 1 indicates the best performance. We also applied x-axis limits to constrain the range between 0 and 1 for uniformity.

We detail the data-splitting procedure used for generating the noise floor, i.e. the baseline, in Figure 4. We split the neural data into nonoverlapping groups each containing N_{sample} neurons ($ineurons1$, $ineurons2$). We sample N_{sample} units from the RNN model ($iunits$). We compute the distance between two samples of neural data $d1 = D(ineurons2, ineurons1)$. $d1$ is the lowest we can hope to get given the variability in the neurons that were recorded. We compute the distance between samples of the model and neural data $d2 = D(iunits, ineurons1)$. For each iteration of this procedure we get a new estimate for the distance between the model and data, and the data-to-data distance.

C Additional simulations

In Appendix Figure 5, we examine the top demixed principle components between data and models. In Appendix Figure 6 displays the similarity among models in terms of their pairwise distances and their embeddings across different sampled training snapshots. In Appendix Figure 7, we demonstrate consistent patterns when recurrent noise is removed, sparsity constraints are applied, and Dale's law is enforced. We also explore ModProp [54], which incorporates cell-type-specific local modulatory signals to reintroduce terms omitted by e-prop; however, as ModProp is effective only under specific conditions (Dale's law and $ReLU$ activation), confining Appendix Figure 7C to these settings. Further analysis of post-training weight eigenspectrums and distances, conducted using Dynamical Similarity Analysis (DSA), reinforces the similarity between BPTT and e-prop, as shown in Appendix Figure 8.

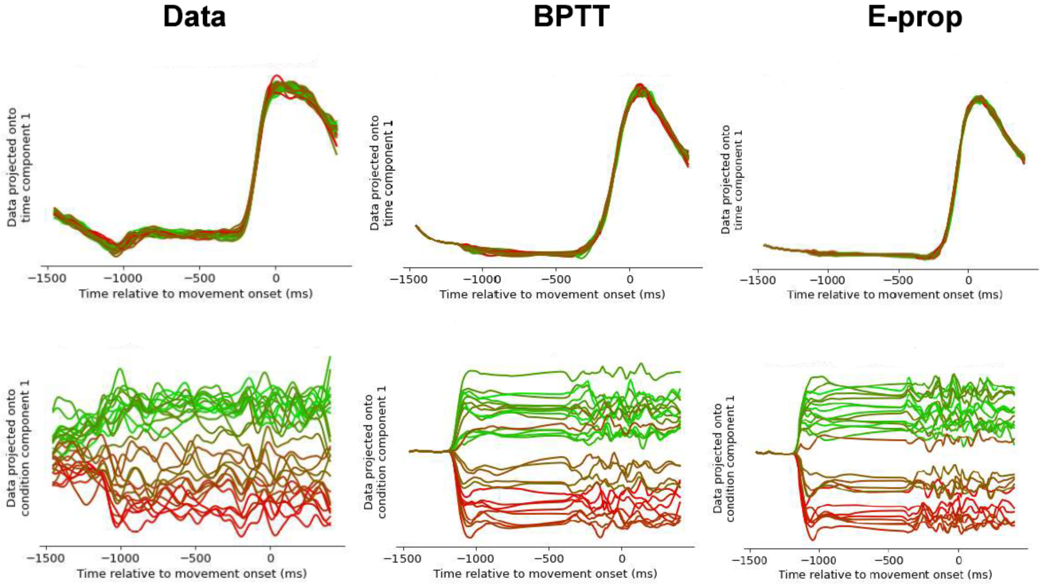

Figure 5: Demixed principle component analysis (dPCA) show qualitative match between model and data when projected onto the time component 1 and condition component 1. Here the Sussillo 2015 dataset is illustrated. Each color represents a different reach condition.

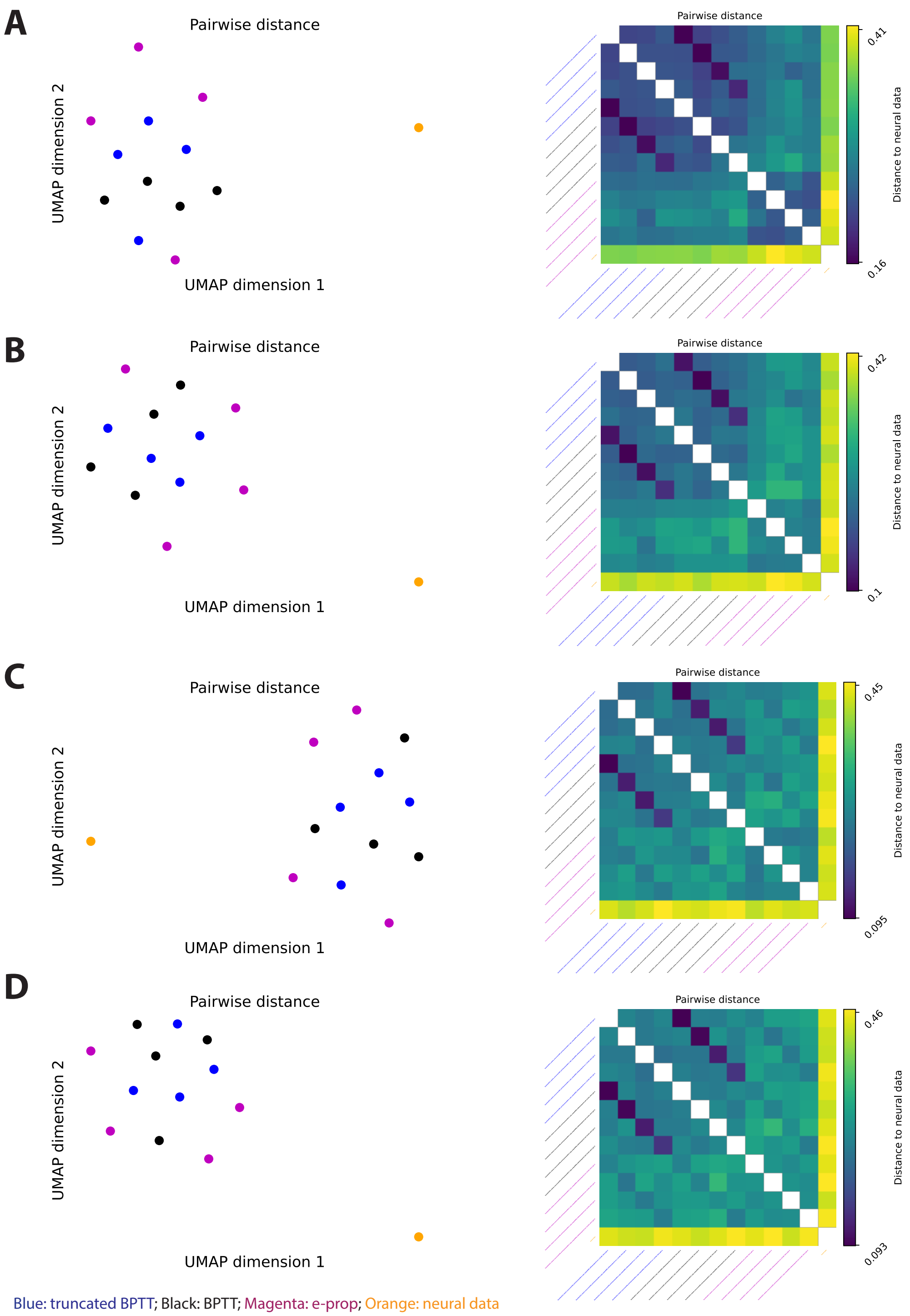

Figure 6: UMAP embedding and pairwise distance matrix heatmap for different models when (A) best e-prop accuracy, (B) 80%, (C) 60%, and (D) 40% accuracies are reached. Here, the Sussillo 2015 dataset is illustrated. Black: BPTT, blue: truncated BPTT, magenta: e-prop, orange: neural data. The pairwise distances show similarities across learning rules relative to data, indicated by lower distances between models as compared to model-data distance.

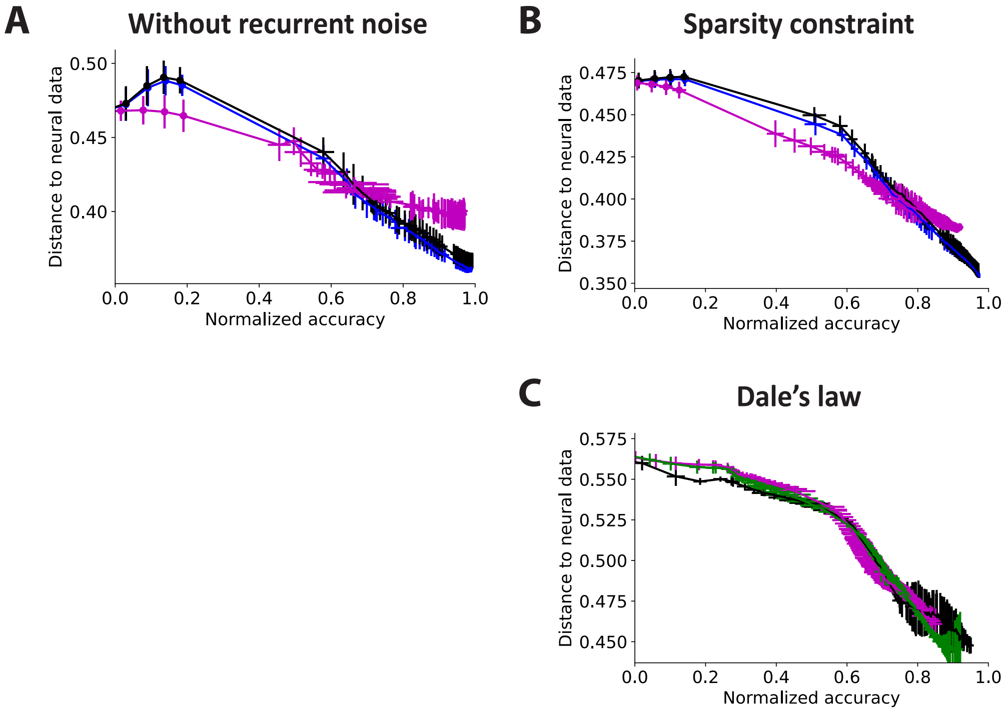

Figure 7: This plot compares Procrustes distances versus accuracy for three learning rules: BPTT (black), e-prop (magenta), and ModProp (green) — the latter functioning exclusively under Dale's law constraint and $ReLU$ activation. Consistent with trends observed in Figure 1, variations include: (A) removal of RNN hidden activity noise, (B) application of a sparsity constraint (limiting to only 25% of the recurrent weights as nonzero and trainable), and (C) enforcement of Dale's law. The results pertain to the Sussillo 2015 task, with plotting conventions mirroring those in Figure 1.

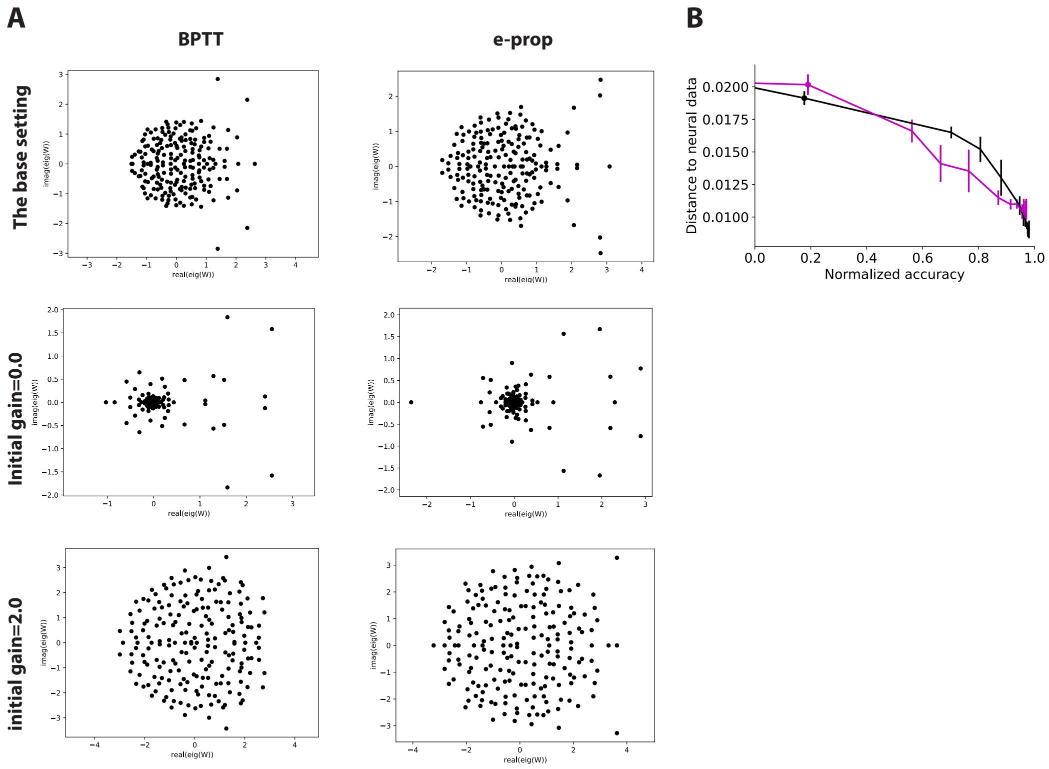

Figure 8: (A) presents the eigenvalues of the recurrent weight matrix post-training, with columns representing BPTT and e-prop respectively. Each row displays a different training setting: the base setting (referenced in Figure 1), initial weight standard deviation set to 0, and initial weight standard deviation set to $2/\sqrt{N}$. Notably, eigenvalue distributions appear more similar within each setting across learning rules (BPTT vs. e-prop) than across different settings for the same learning rule, further highlighting the similarity between BPTT and e-prop. B) The Dynamical Similarity Analysis (DSA), which evaluates systems based on their dynamical characteristics, is also unable to distinguish between learning rules when considering their proximity to neural data.

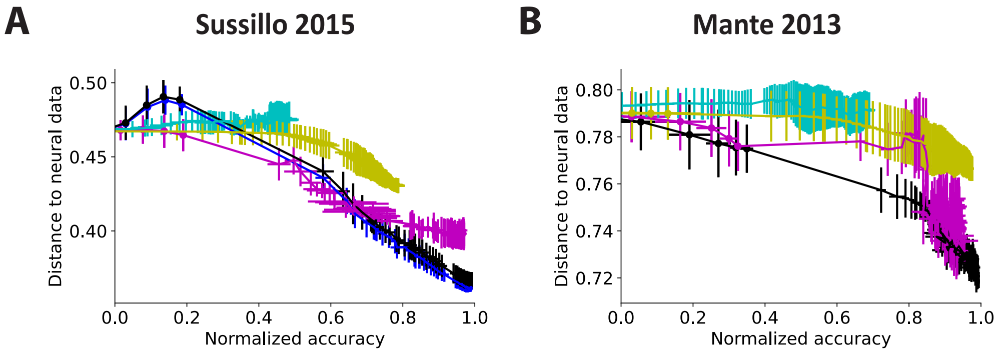

Figure 9: Node perturbation (cyan) and evolutionary strategies (yellow) lead to higher Procrustes distances from the neural data compared to BPTT (black) and e-prop (magenta) when accuracies are equivalent. This figure presents the Procrustes distance versus accuracy plots, adhering to the plotting conventions established in Figure 1, for (A) the Sussillo 2015 task and (B) the Mante 2013 task.

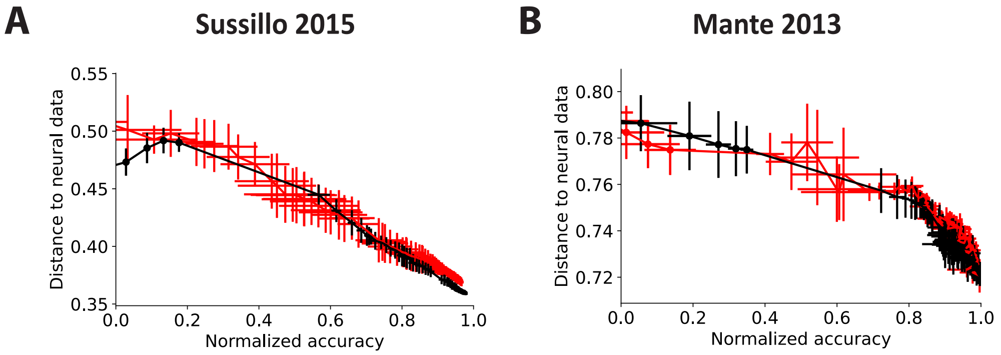

Figure 10: The use of random feedback readout weights for gradient computation (red) resulted in distances comparable to those achieved using exact readout weights (black). Plotting conventions are consistent with those used in previous figures.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: To make this easier for the readers, we have referred to the pertinent figures and sections under "Main contributions" in Introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Details on limitations and future work are discussed in our Discussion section.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [NA]

Justification: This paper does not introduce new theorems or lemmas.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: Training details are provided in Appendix B.5. Moreover, our code is available at `https://anonymous.4open.science/r/XYZ2442-860A/`. However, as explained in Appendix B.5 and the *readme.txt* file for our code, it only contains the code to reproduce our Mante 2013 results, as we do not have the permission to redistribute the Sussillo 2015 datasets.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: Our code is available at `https://anonymous.4open.science/r/XYZ2442-860A/`. However, as explained in the *readme.txt* file for our code, it only contains the code to reproduce our Mante 2013 results, as we do not have permission to redistribute the Sussillo 2015 datasets.

   Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: Simulation details are provided in Appendix B.5.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: We tried to provide this information in all applicable figures. This is stated as "The mean is plotted with error bars denoting the standard deviations across four different seeds" in the figure legends.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

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

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: Information pertaining to computing resources and simulation time can be found in Appendix B.5.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: We have carefully read the NeurIPS Code of Ethics and attest that the research conforms.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

Justification: This research advances our understanding of biologically plausible learning models in recurrent neural networks, with no immediate ethical or societal impacts expected. Over time, the findings could influence related fields like neuroscience and deep learning, potentially affecting society based on how these disciplines evolve.

Guidelines:

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

11. **Safeguards**

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

Answer: [NA]

Justification: This research advances our understanding of biologically plausible learning models in recurrent neural networks, with no immediate ethical or societal impacts expected. Over time, the findings could influence related fields like neuroscience and deep learning, potentially affecting society based on how these disciplines evolve.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: Please see Appendix B.5.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

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

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

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

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

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

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

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

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.