Transition to chaos separates learning regimes and relates to measure of consciousness in recurrent neural networks

Dana Mastrovito, Yuhan Helena Liu, Lukasz Kusmierz, Eric Shea-Brown, Christof Koch, Stefan Mihalas

May 15, 2024

Abstract

Recurrent neural networks exhibit chaotic dynamics when the variance in their connection strengths exceed a critical value. Recent work indicates connection variance also modulates learning strategies; networks learn "rich" representations when initialized with low coupling and "lazier"solutions with larger variance. Using Watts-Strogatz networks of varying sparsity, structure, and hidden weight variance, we find that the critical coupling strength dividing chaotic from ordered dynamics also differentiates rich and lazy learning strategies. Training moves both stable and chaotic networks closer to the edge of chaos, with networks learning richer representations before the transition to chaos. In contrast, biologically realistic connectivity structures foster stability over a wide range of variances. The transition to chaos is also reflected in a measure that clinically discriminates levels of consciousness, the perturbational complexity index (PCIst). Networks with high values of PCIst exhibit stable dynamics and rich learning, suggesting a consciousness prior may promote rich learning. The results suggest a clear relationship between critical dynamics, learning regimes and complexity-based measures of consciousness.

As learning in artificial networks continues to amass practical successes, theorists have been making significant strides in rigorously characterizing the behavior of these models and explaining why they perform so well [1–8]. These theoretical tools present an opportunity to unravel mysteries in biological neural networks [9], such as how the learning rule and/or initial priors of a network could alter its learning dynamics and representations [10–13]. Among various theoretical developments that contribute to this progress, a popular theme is that networks can be successfully trained to learn a task using two distinct strategies: rich learning and lazy learning [14–23]. Interpolation between the two learning regimes can be achieved through adjusting the variance in connection strengths at initialization [14, 15, 24]. This adjustment tunes the extent to which the network alters its internal representation to fit the task statistics, leading to various degrees of task-specific representations. Consequently, whether learning occurs in the rich or lazy regime can have a profound effect on the nature of what the network learns and its performance in unseen situations post-training [15, 16].

Connection strength in artificial networks has also been studied in the context of dynamical systems theory, where it has been shown that networks exhibit a transition between ordered and chaotic dynamics when the variance of their connection strengths exceed a critical value [25]. The transition point between ordered and chaotic regimes can be identified mathematically using the maximum Lyapunov exponent λ, which when positive, indicates chaotic dynamics. At intermediate connection strengths, at a transition point in phase space known as the edge of chaos, systems are known to have optimal computational performance, exhibiting maximal information transfer [26, 27], and memory capacity [28]. Although, more recent work has suggested that networks can continue to perform well in the weakly chaotic regime, as degradation in autocorrelation of activity occurs more slowly in the chaotic than non-chaotic regime [29]. Building on these theoretical insights, we vary initial connection strengths of recurrent neural networks (RNNs) and characterize their learning properties, finding a direct correspondence between the transition to chaos, when the largest Lyapunov exponent prior to training becomes positive, and a shift between rich and lazy learning strategies.

The brain has long been theorized to operate at the edge of chaos due in part to the aforementioned optimal properties associated with this dynamical regime. Beyond arguments of optimality, however, it is reasonable to theorize that the highly recurrent brain operates within this regime, as self-organized criticality is observed in many natural systems and recent theory shows that suppression of chaos may be inherent in systems utilizing integrative feedback [30]. In fact, several recent studies have uncovered evidence to suggest that edge of chaos dynamics may underlie the capacity for consciousness itself [31–34]. Although consciousness is difficult to define and measure, one recently developed metric called the perturbational complexity index (PCI) has emerged as a reliable correlate of consciousness where it has been demonstrated to distinguish between brain states (awake, anesthetized, under the influence of psychedelics), and to reflect the potential for recovery in patients with disorders of consciousness [35–37]. The PCI metric is predicated on the theory that the capacity for consciousness relies on the ability to integrate information, and that this ability is achieved through the complex patterns of causal interactions between neurons. However, the original PCI metric is only applicable to systems for which one can employ transcranial magnetic stimulation in combination with electroencephalograpy (EEG). Therefore, in this study, we make use of an estimate of PCI (PCIst) [38] which is more broadly applicable. The metric similarly quantifies the spatiotemporal complexity of the propagation of evoked activity in response to an externally driven perturbation above that of baseline activity. Intuitively, the metric combines spatial principal component analysis (sPCA) with recurrence quantification analysis (RQA) [31, 39], which quantifies the temporal complexity as the recurrences of the evoked dynamics. RQA is commonly used in the analysis of dynamical systems [40] to identify state transitions and has properties that are directly related to Lyapunov exponents [40, 41]. Such methods have successfully been applied to the analysis of biological systems [42–44] where the direct computation of the Lyapunov spectrum is impractical. Nevertheless, no previous study has examined the relationship between PCIst and Lyapunov exponents. We therefore computed PCIst on RNNs and found that it increases as a function of initial connection strength up to the edge of chaos, where it is maximal, and subsequently sharply decreases.

Finally, we compare Lyapunov exponents, learning regime and PCIst on network models initialized with Gaussian weight distributions to those with biologically-realistic connectivity structures at two different scales: that of a cortical column within mouse primary visual cortex and of the mouse whole-brain mesoscopic connectivity. We find that biologically realistic connectivity yields non-chaotic dynamics, where PCIst is high and rich learning is favored, over a wider range of connectivity strengths.

Summary of results and contributions:

Building on three lines of literature — rich versus lazy learning, dynamical systems theory and consciousness — we find and characterize two regimes associated with the initial hidden weight gains below and above the critical point at which networks begin to exhibit chaotic dynamics. Below the critical point, in the ordered learning regime, networks learn rich low-dimensional representations. Beyond the critical point, in the chaotic learning regime, networks gradually transition towards a high-dimensional, lazy learning strategy that is more sensitive to noise. Importantly, we find that models in the chaotic regime, with gains close to the transition, still perform the task with high accuracy, and converge more quickly than in the ordered regime. Further increases in initial connection strength variance results in chaotic dynamics, drastically reducing learning ability. Connection sparsity allows stable dynamics with larger variances, and training moves networks closer to the edge of chaos. Intriguingly, we find that PCIst increases as a function of gain below the transition point, is maximal at the edge of chaos and subsequently sharply decreases. Finally, we show that several key features of biologically-realistic connectivity strongly bias networks towards ordered dynamics and rich learning. We focus on RNNs due to their widespread usage in brain modeling [24, 45–64]. Although in artificial learning networks, our results show a strong relationship between critical dynamics, learning and measures of consciousness.

1 Results

In a series of RNNs with systematically varied sparsity and degree of small-world structure, we characterize 1) their dynamical regime (ordered vs chaotic) 2) their PCIst scores and 3) their learning regime (rich vs lazy) as a function of an initial scaling parameter (g) on the strength of their hidden weight connectivity (H). Specifically, we trained a series of RNNs to perform the ten digit (0-9) sequential MNIST task, in

which the network receives a row of pixels from a digit image at each time step. All networks consisted of an input layer, a single recurrent layer with 198 neurons, and a linear readout layer. The hidden layer structure of each network was initialized according to a Watts-Strogatz [65] connectivity rule over a range of nearest neighbors (nNN: 4, 8, 16, 28, 32, 64, 128, 198 = fully connected), and rewiring probabilities (p: 0.0, 0.2, 0.5, 0.8, 1.0), while self-connections were prohibited. Thus, the resulting networks systematically varied in both their degree of sparsity with nNN = 4 most sparse; nNN = 198 least sparse and their degree of small-world structure with p = 0.0 highly structured; p = 1.0 Erdös-Rényi random connectivity (Figure 1). (see Table 1 for number of parameters in each model). Initial hidden layer weights were sampled from a Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = \frac{g}{\sqrt{N_{nz}}}$, where gain (g : 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 5.0) and N_{nz} is the number of non-zero elements, such that variance is scaled in accordance with sparsity. Network sparsity was maintained over training by restricting training to only the non-zero elements of the hidden weight matrices.

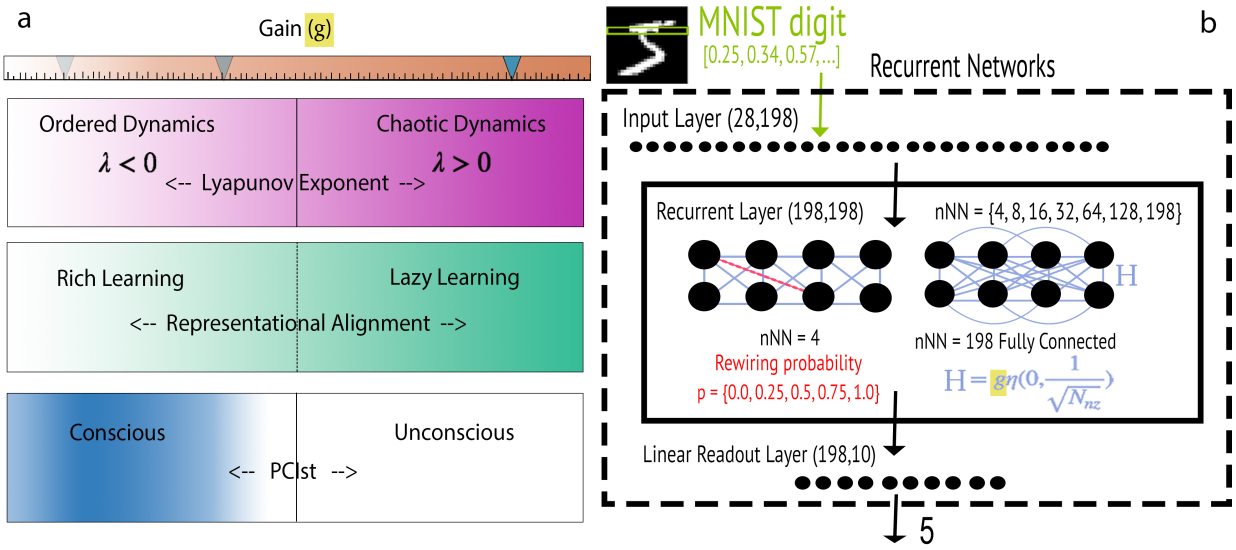

Figure 1: Summary of Results. **(a)** A single gain parameter (g) scaling initial hidden weights modulates ordered and chaotic dynamical regimes, rich and lazy learning strategies as well as a measure associated with consciousness. **(b)** Model Construction. Models consist of an input layer of size 28, a single recurrent layer of 198 hidden units. The hidden layer connectivity matrix (H) is initialized as a Watts Strogatz network with number of nearest neighbor connections nNN = {4, 8, 16, 28, 32, 64, 128, 198 = fully connected} and rewiring probability p = {0.0, 0.2, 0.5, 0.8, 1 = Erdös-Rényi}. Initial hidden layer connection strengths are sampled from a normal distribution $\mathcal{N}(0, \frac{g}{\sqrt{N_{nz}}})$, where N_{nz} is the number of nonzero elements and gain g = {0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 5.0}. The location of zero/non-zero elements are maintained over training on the sMNIST task.

1.1 Chaotic and Ordered Learning

We computed finite time Lyapunov spectra for each RNN model as the eigenvalues of the product of the Jacobians along the data-driven network trajectory, averaged over a batch of sMNIST test data [66] (see section 2.3.1 Lyapunov Exponents in Methods for details). Chaotic dynamics are indicated when the largest Lyapunov exponent $\lambda > 0$. For models at each sparsity level (nNN), we define the transition point between ordered and chaotic learning regimes in terms of a single multiplicative parameter g, the gain in connection strength of the initial (pre-training) hidden layer connectivity. Specifically, we define the critical transition point associated with each model's degree of sparsity as the gain ($g_{c_{nNN}}$) at which the initialized (prior to training) model's largest Lyapunov exponent λ becomes positive. We find that the transition from ordered to chaotic learning regimes shifts towards lower values of gain (g) as models become less sparse (nNN

Model	Number of Hidden Layer Parameters
nNN = 4	792
nNN = 8	1584
nNN = 16	3168
nNN = 28	5544
EM V1 Cortical Column	5573
nNN = 32	6336
nNN = 64	12672
nNN = 128	25344
Mesoscale Connectome (thresholded)	5573
Mesoscale Connectome	38984
nNN = 198 (fully connected)	39006

Table 1: Model Hidden Weight Sparsity and Parameter Count

increases)(Figure 2a,b). Rewiring probability, which shifts network structure away from modular small-world structures towards Erdös-Rényi random connectivity as it increases, had a much smaller impact on network dynamics than did the sparsity. In fact, we found that the transition point often did not vary substantially with rewiring probability (Table S1, Figure 2a). Therefore, unless otherwise noted, throughout the manuscript we generally report results for rewiring probability of 1.0 = Erdös-Rényi, as it is the most commonly used network initialization strategy, but see Figure S1 for similar results obtained with rewiring probabilities p: 0.0, 0.2, 0.5, 0.8. Consistent with prior work [27, 67, 68], we found that networks self-tuned towards criticality as a result of training with back-propagation, such that the maximum Lyapunov exponents after training shifts closer to zero for all models (Figure 3a). However, we find that changes in the magnitude of the maximum Lyapunov exponent over training decreases substantially as the models enter the chaotic regime (Figure 3b), suggesting that models are less able to tune towards the edge of chaos if the variance in their initial connectivity strength is too large.

Although models in the weakly chaotic regime are often able to obtain good accuracy (See section 1.2 for model performance), the trained model dynamics in the ordered and chaotic regimes are qualitatively very different. To illustrate the difference in model solution space, we project the hidden state space of the trained model responses to sMNIST test samples into 2-dimensions (for ease of visualization) using principle component analysis (PCA), and find qualitatively different solutions to the task are learned on either side of $g_{c_{nNN}}$. In the ordered learning regime, models exhibit a smooth trajectory from initial state (at the origin) towards a final state located in a cluster associated with digit identity. In contrast, in the chaotic learning regime, just past the transition, where models nevertheless still learn to perform the task, models exhibit a jagged trajectory (Figure S2, illustrating the resulting chaotic dynamics).

Finally, we tested for model sensitivity to noise at test time. Models, trained without noise were subjected to additional Gaussian noise during testing (See Methods for details). Because small perturbations should be amplified in the chaotic regime, we expected that models in the chaotic learning regime would be more sensitive to injected noise. Indeed we found that accuracy dropped more substantially for models in the chaotic than ordered regimes when exposed to additional Gaussian noise $\mathcal{N}(0, .01)$ at test time (Figure 4c).

1.2 Model Performance

Model performance was assessed via both the test accuracy achieved as well as speed of learning. Test accuracy was quantified as the average percent correct over all sMNIST test samples after 100 epochs of training, while speed of learning was assessed by the number of epochs required to reach at least 90% accuracy. For both metrics, the reported values are the mean over 10 model instantiations with the same model parameters: nNN, rewiring probability (p), gain (g). The least sparse models achieved higher accuracy after 100 epochs in comparison to sparse ones; while performance on models with equal sparsity, differing only in connectivity structure, performed nearly identically. After 100 epochs of training, accuracy increased as a function of gain up to $g_{c_{nNN}}$, where all models achieved their highest accuracy (Figure 4a). Of critical interest is the fact that models continued to have high accuracy after the transition into the chaotic learning

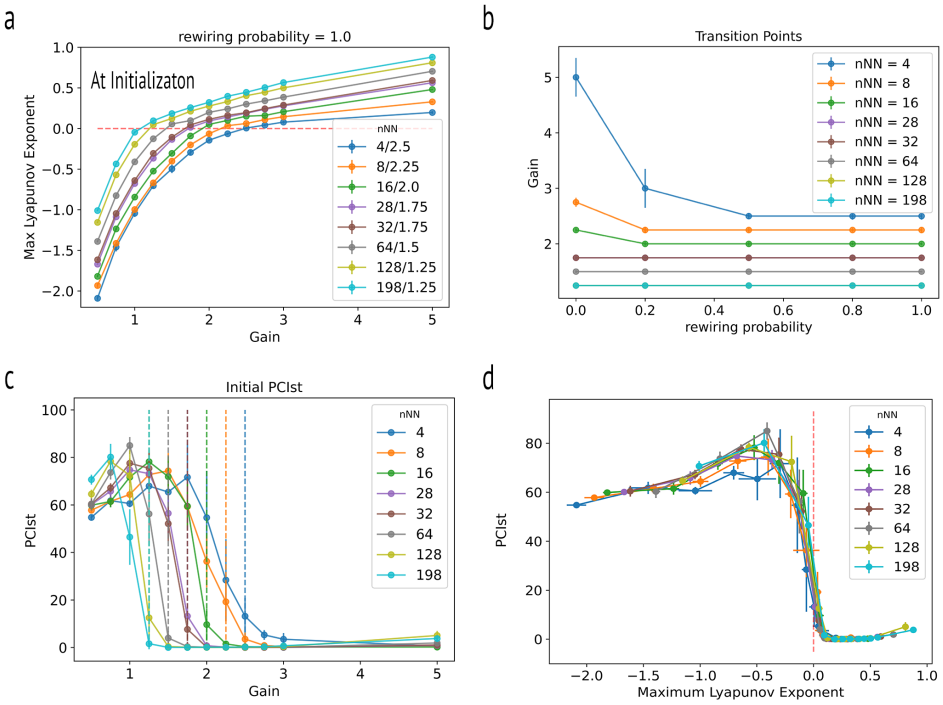

Figure 2: PCIst and Maximum Lyapunov Exponents Before Training **(a)** Maximum Lyapunov Exponent λ for driven models with rewiring probability $p = 1.0$ along trajectory. Gain at which each network transitions from ordered to chaotic regime $g_{c_{nNN}}$ is the gain at which λ becomes positive for each level of sparsity nNN. **(b)** Model Transition points. Transitions to chaos shift to smaller gains as hidden layer sparsity increases. **(c)** PCIst in pre-trained networks increases as a function of gain and begins to decrease just before $g_{c_{nNN}}$ where it decreases sharply. **(d)** PCIst in pre-trained networks begins to decrease as the maximum Lyapunov exponent λ approaches zero.

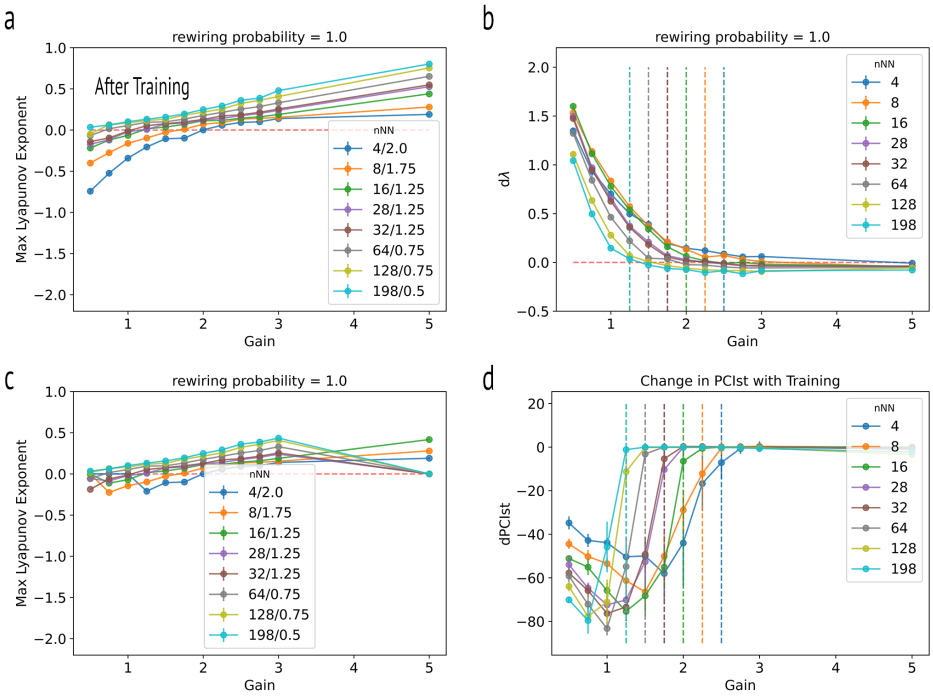

Figure 3: PCIst and Maximum Lyapunov Exponents After Training **(a)** Maximum Lyapunov Exponent λ after training for 100 epochs. Models tune towards the edge of chaos with training. **(b)** Change in maximum Lyapunov exponent λ with training (100 epochs). Models trained in a highly chaotic regime exhibit smaller changes in λ with training. **(c)** Maximum Lyapunov exponent in fully-trained models, trained till 90% accuracy. **(d)** Changes in PCIst over training decrease in the chaotic regime.

regime, while also learning faster in this regime (Figure 4b), with small increases in initial weight gain above $g_{c_{nNN}}$. We are not the first to report that models can perform well in the weakly chaotic regime. This result is consistent with previously reported numerical experiments on the lazy training regime [14, 69] where it was also found that lazy models converged faster. As we will show in the subsequent sections, however, learning in this regime occurs using a "lazier" strategy consistent with previous work indicating that dimensionality expansion in the weakly chaotic regime allows for linearly separable representations at the readout layer [70]. This is can be seen in Figure 5a,e, where the dimensionality of the trained networks at higher gains increases and is reflected by greater weight changes in the readout layer in successfully trained chaotic networks. However, further increases in gain result in sharp decreases in accuracy (Figure 4a). Again, analogously to the dynamical regime, we found that hidden layer sparsity had a much greater effect on model performance than connectivity structure varied through rewiring probability.

1.3 Rich vs Lazy Learning Regime

In the following sections, we report several metrics used to characterize the learning regime. First, we computed the Kaplan-Yorke (KY) dimension of the trained models derived from the Lyapunov spectrum. It has previously been demonstrated that, with random initialization, lazy learning leads to higher dimensional task-agnostic representation, whereas rich learning results in lower dimensional task-specific representations [12, 69]. Second, lazier learning results in less modification to the hidden weight parameters [14, 24, 69], and we report the vector norm of the magnitude weight changes at the input, hidden, and output layers. Finally, we compute representation alignment, which quantifies the directional change in a representational similarity matrix before and after training; lazier learning should result in higher representation alignment [16]. See also Supplementary Results where we also quantify the directional shift in the neural tangent kernel (NTK) pre- and post-training (Figure S4). This measure, referred to as tangent kernel alignment, provides another method for quantifying the richness/laziness of learning [16]. We emphasize that in this study, laziness and richness are quantified on a continuum rather than categorically, with lazier learning corresponding to smaller network changes to achieve learning of a task. In other words, we adopt the notion of an effective learning regime used in [71], which gauges effective richness or laziness by post-training changes rather than on initialization.

1.3.1 Dimensionality

We computed the Kaplan-Yorke (KY) dimension of the trained models derived from the Lyapunov spectrum (see section 2.6 in Methods for details). The KY dimensionality increased as a function of gain, accompanied by a shift towards lazier learning. Notably, models learned very low-dimensional solutions at the smallest gains (Figure 5a). Interestingly, as the task consists of 10 classes, the dimensionality of the trained models exceeded 10 at $g_{c_{nNN}}$, for all but the sparsest and most small-world models. KY dimensionality continued to increase with further increases in gain. We trained an additional series of models for rewiring probability = 1.0 only, on the 2-digit sMNIST task, using digits [2,5] and found that models exceeded KY dimension of 2 rather than 10 at $g_{c_{nNN}}$ (Figure S3).

1.3.2 Weight Changes

We found that the norm of weight changes over training in both the input and hidden layers decrease smoothly and monotonically as a function of gain for all models (Figure 5c-d). However, in the readout layer, changes in weights increase as models approach $g_{c_{nNN}}$, after which it gradually begins to decrease. This is consistent with previous work that found lazy learning can result from expanding the dimensionality of input signals. Specifically, random projections to the hidden layer create a representation that facilitates linear separability. Consequently, learning primarily occurs in the readout weights [72]. So as gain increases, learning smoothly shifts away from rich hidden layer representations, as evidenced by large magnitude changes in the hidden layer and when models engage in lazy learning, learning is confined to the readout layer. As the gain increases far into the chaotic regime where models fail to learn the task, presumably due to numeric instability, the normed weight changes of all layers approach zero.

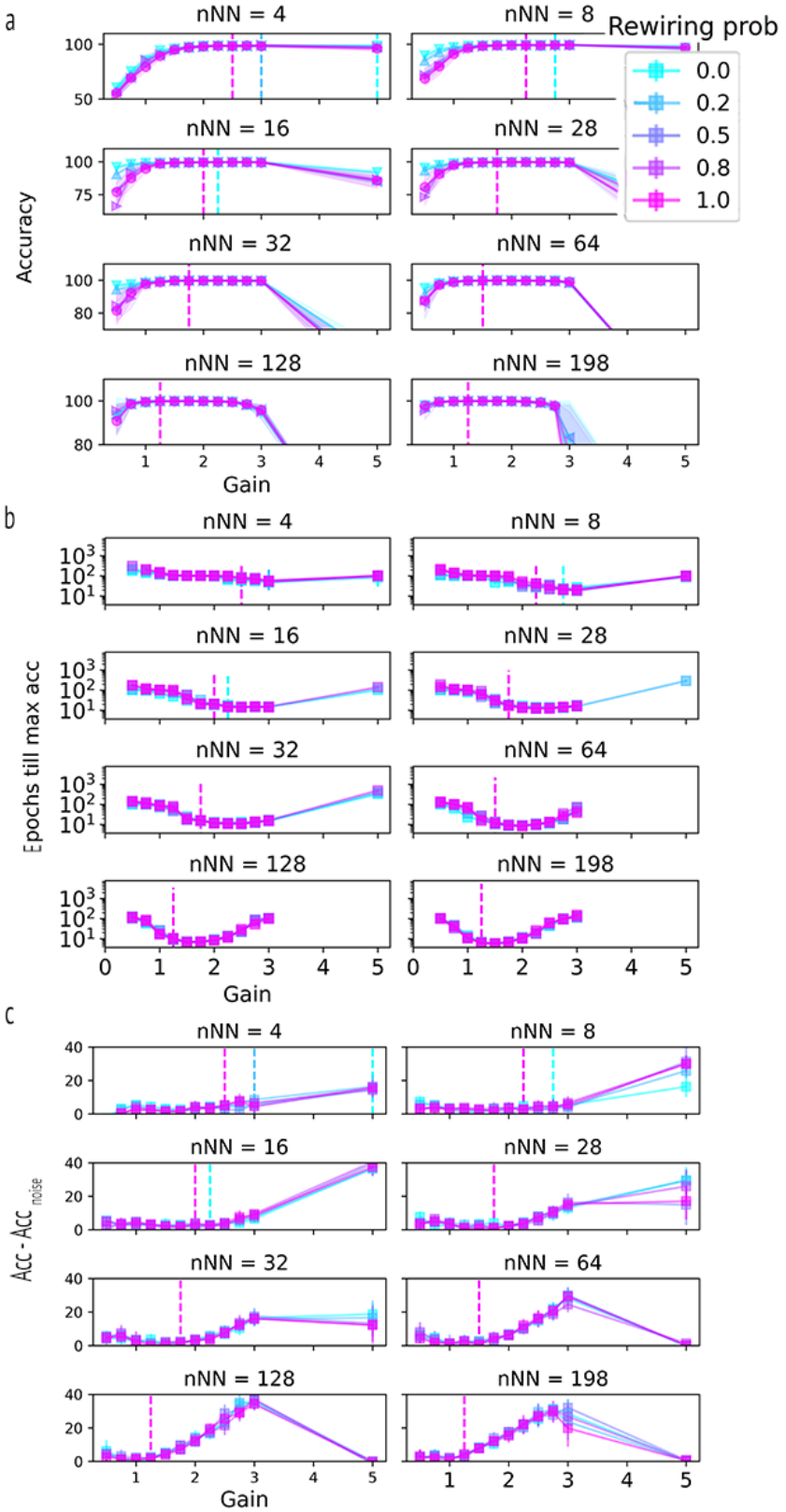

Figure 4: Model Performance. **(a)** Accuracy after 100 epochs. Dashed lines indicates $g_{c_{nNN}}$, where models achieve their highest accuracy. As sparsity decreases ability to learn the task falls sharply as gain increases beyond ≥ 3.0. **(b)** Number of Epochs till 90% accuracy. **(c)** Difference in accuracy with additional noise during testing. Networks are more sensitive to noise in the chaotic regime.

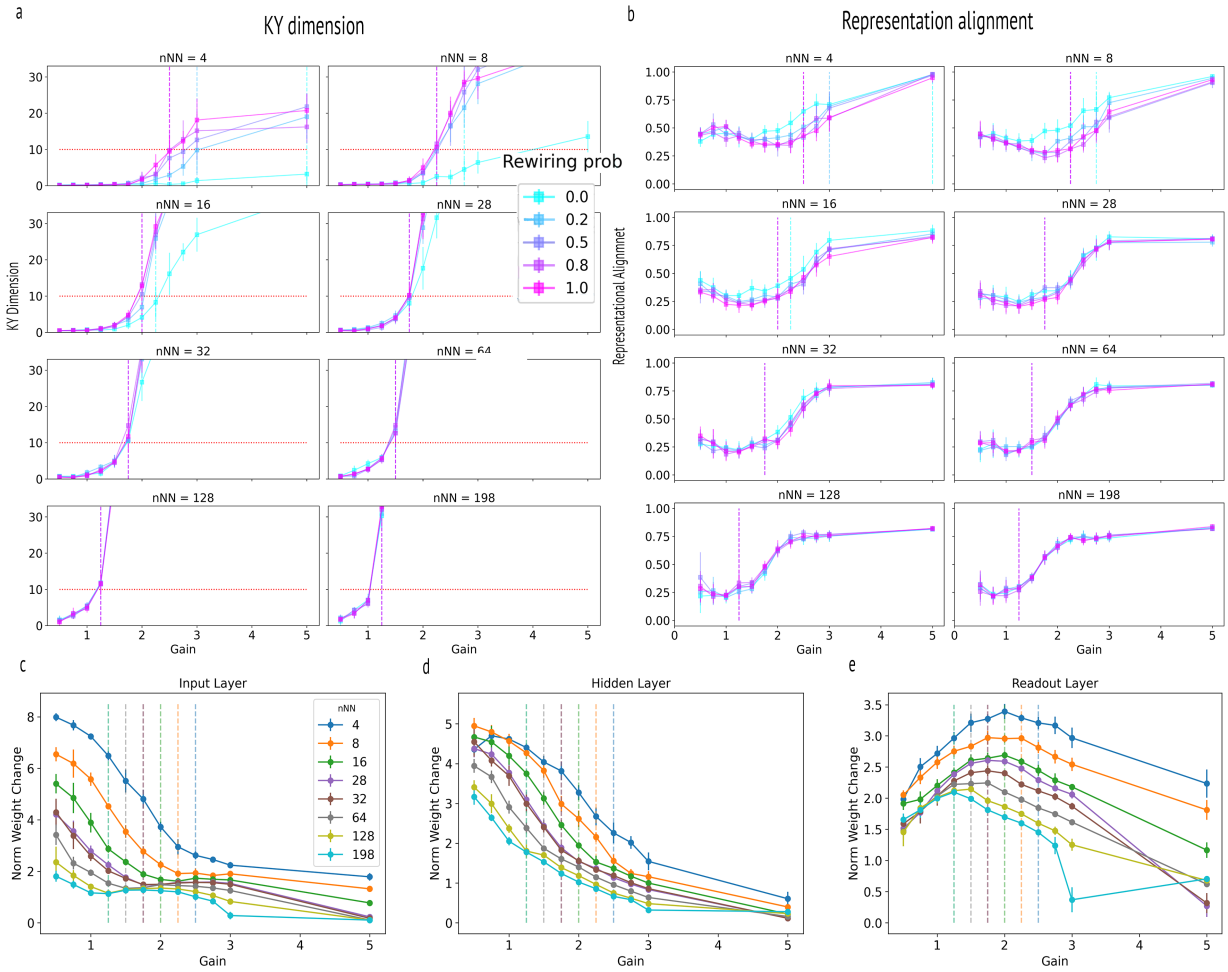

Figure 5: Rich vs. Lazy Learning **(a)** Kaplan-Yorke (KY) dimensionality. Dashed lines indicate $g_{c_{nNN}}$. **(b)** Representation alignment for different sparsity and rewiring probabilities. **(c)** Norm Weight Change of input layer weights over training. **(d)** Norm weight change of hidden layer. **(e)** Norm weight change of readout layer. Weight changes in the input and hidden layers decrease smoothly with gain. In the readout layer, however, weight changes are non-monotonic, increasing as a function of gain up to $g_{c_{nNN}}$ and then decreasing.

1.3.3 Representation Alignment

As expected, we find that representation alignment (see Methods section 2.7.2) between trained and untrained models largely increases with increased gain, indicating greater laziness as gain increases (Figure 5b). Although the curves are not always monotonic, the representation alignment typically increases just before the $g_{c_{nNN}}$ transition.

1.4 PCIst

PCIst was assessed in all networks before and after training on the sMNIST task. PCIst at initialization increased as a function of gain across all network architectures with g ≤ 1, however initial PCIst begins to decrease as the gain approaches $g_{c_{nNN}}$ for all models and nears zero when models enter the chaotic regime (Figure 2c). In section 1.1 we found that the gain at which the maximum Lyapunov exponent becomes positive shifts to ever smaller gains as sparsity decreases. Consistent with the maximum Lyapunov exponent itself, the gain at which PCIst begins to decrease also shifts to smaller gains as sparsity decreases. The metric hits a maximum prior to the critical point, as models near the edge of chaos transition (Figure 2d). Critically, PCIst reaches its minimum at $g_{c_{nNN}}$ or, in the sparser models, just beyond the point at which models become chaotic. Also consistent with the maximum Lyapunov exponent, PCIst decreases with learning in models initialized in a non-chaotic regime, as the models tune towards towards criticality as a result of training with back-propagation. For both the maximum Lyapunov exponent and PCIst, we do not observe decreases as a result of training when initialized in the chaotic regime (Figure 3b,d). All models considered here were of equal size. To test for finite size effects, we created two larger models with 500 and 1000 hidden units (Figure S12) and found that the point at which PCIst begins to decrease moves closer to the edge of chaos as networks increase in size. In all cases, we find PCIst is non-zero in the ordered regime where richer learning strategies are favored and maximal just before the transition to chaos.

1.5 Biologically Realistic Connectivity

Biological brain networks are known to have small-world architectures [65, 73], but differ from the models previously explored in important ways including the distribution of their weights (Figure 7a, the degree distribution of each neuron, and adherence to Dale's law. We therefore trained a series of models with hidden weight matrices defined by biologically realistic connectivity structures at two different scales. The first connectivity structure was defined by the normalized whole-brain mesoscopic connectivity density of the mouse connectome [74, 75] as measured from hundreds of anterograde tracing experiments. Normalized connection density is defined as the directional inter-regional connection strength divided by the product of the size of the regions. The mesoscopic connectivity model has a similar number of parameters as the nNN = 198 (fully connected) model, but the distribution of connection strengths has a longer-tailed distribution (Figure 7a). As connection strengths are by definition all positive, for these models, an equal number of positive and negative weights were assigned randomly. The second model, at the scale of a single micro-column, was derived from electron microscopy of the mouse primary visual cortex (V1)[]. In this model, the connection strength H_{ij} from neuron j to neuron i was defined as the sum of the volume of synaptic densities at neuron i coming from neuron j. For comparison, this cortical column model has approximately the same number of parameters as the nNN = 28 models (See Table 1). The sign of each connection weight is cell-type specific as dictated by Dale's Law, such that all post-synaptic connections of excitatory neurons are positive, and visa-versa. We now describe the differences we observed in model performance, dynamical and learning regimes in biologically realistic connectivity structures in contrast to models initialized with Gaussian hidden weight distributions previously described.

1.5.1 Chaotic and Ordered Dynamics

In the case of the mesoscopic connectivity model, we find that the relationship between initial hidden weight gain and the maximum Lyapunov exponent pre-training has a shallower slope, such that it enters the chaotic regime at a higher gain than the fully connected Gaussian model (Figure 6a:green dash-dot/cyan). Additionally, the magnitude of the largest Lyapunov exponent remains below that of the fully connected model at higher gain values. For comparison to the cortical column model, we also trained a modified

mesocopic model with a thresholded connectivity structure, such that sparsity was matched to the cortical column model. Here again, relative to the nNN = 28 Gaussian model, we see that the mesoscopic model has a shallower slope and the magnitude of the largest Lyapunov exponent remains below that of the Gaussian model at high gains (Figure 6a:dark blue dashed/brown), suggesting that the long-tailed distribution of the weights of the mesoscopic connectivity structure affords the model greater stability than a Gaussian model of equal sparsity. See also Figures S6, S7 for comparison of eigenspectrum and Lypaunov spectrums for Gaussian and biologically realistic connectivity.

For models based on the cortical column connectivity, the effect of initial gain on dynamical regime is even more pronounced (Figure 7b (V1 23 4 Dales - purple dash-dot), where the maximum Lyapunov exponent is less than 1 for all gains except gains of 1 and 1.25. Moreover, as the gain increases, we observe the maximum Lyapunov exponent decreases rather than increases, deviating substantially from the Gaussian models as well as that of the mesoscopic model.

We endeavored to explore which of the characteristics of the cortical connectivity structure (weight distribution, degree distribution, topological structure, Dale's law, relative balance of excitation and inhibition) enables such stability. To this end we created a series of models with altered structures and found that the observed stability required the combination of block structured connectivity, multiple cell types and adherence to Dale's law. Below we describe each of the altered structures tested.

1) to test whether Dale's law alone was sufficient, we created a network in which Dale's Law was artificially applied to a connectivity structure with Gaussian weights of equal sparsity (Figure 7b - nNN=28D: cyan solid). This model did not exhibit stability at high gains, indicating that Dale's law alone was not sufficient.

2) to test whether the distribution of weights alone was sufficient, we permuted the cortical column weights such that all topological structure including Dale's law is disrupted (Figure 7b - V1 2/3 4 permuted: peach). This model also did not exhibit stability at high gains, indicating that weight distribution alone was insufficient.

3) to test whether the degree distribution alone could account for the observed stability, we created a model with Gaussian weights but matched the degree distribution (DD) of the cortical column (Figure S5b - DD green). Degree-distribution alone was not sufficient to maintain stability.

4) to test whether the specific topology of connections was sufficient to reproduce the stability observed in the cortical column, we created a model with Gaussian weights, maintaining the topology (location of non-zero weights) of the cortical column model. Topology alone was also insufficient to maintain stability (Figure S5b - (V1 T magenta solid).

5) To test whether stability could be explained by the combination of weight distribution and topology, we created a network in which the weights are fully permuted but the topological structure is maintained, such that the location of non-zero weights are preserved but Dales law was not preserved (Figure S5b - V1 2/3 4 permuted T, pink dash-dot). This model did result in lower slope as a function of gain, suggesting that the combination of topological structure and weight distribution contributes to stability. However this model did not reproduce the shape of the curve we observed in the cortical column model, where the maximum Lyapunov exponent decreased with increasing gain.

6) To test the possibility that having multiple populations of neurons (E - excitatory, I- inhibitory) with different distributions could account for our observations, we created a model in which cortical column weights were permuted within blocks (E-E, I-I, E-I, I-E), such that block structure and Dale's law were preserved, but the topological structure was not (Figure 7b - V1 Dales BP: magenta solid). This model did reproduce the stability of the cortical column and the shape of the curve as a function of gain, suggesting that block structure contributes importantly to the stability of the cortical column.

7) a model with weights permuted within block while maintaining topological structure and Dales law (Figure S5b - V1 dales BPT: magenta dash-dot) also maintained stability.

8) Finally, to test whether the overall balance of excitation and inhibition was critical to stability, we created a model in which the signs of all connections are flipped, such that the topology, block structure and Dale's law are all maintained but the balance of excitation and inhibition is reversed (Figure 7c V1 Dales +-: fuchsia-dotted). Surprisingly, we found that this model was equally as stable as the true connectivity at high gains.

From these models we can conjecture that none of the weight distribution, topology, degree distribution nor balance of excitation and inhibition *alone* is sufficient to reproduce the stability of the cortical column model. Rather, models that matched the observed pattern of stability in the cortical column model featured

the combination of a block structured connectivity matrix with multiple cell types and adherence to Dale's law. We further explored this possibility through a series of simulations (See Supplementary Section 5.4.1), which further suggested that the dynamical stability we observed can be achieved whenever the connectivity contains at least 2 populations of cells, with differing means, such that for each, the population mean is large enough compared to the variance. The simulations also reveal that strict adherence to Dale's law is not required; approximate adherence to Dale's law is sufficient. In this case, the mean activity drives the overall system dynamics. Stability is achieved by interaction between excitatory and inhibitory populations in combination with a saturating non-linearity, such that strong oscillations dominate, effectively quenching the chaotic dynamics.

1.5.2 Model Performance

The mesoscopic connectivity models performed similarly to the Gaussian nNN = 198 models to which they are most similar, in that models achieved similar accuracy as a function of gain after 100 epochs (Figure 6c). This is perhaps unsurprising given the matched degree of sparsity and balance of positive and negative weights in both sets of models. Models initialized with connectivity of the cortical column, however, did not perform as well as their nearest Gaussian comparator in terms of sparsity (nNN = 28) (Figure 7d - V1 2/3 4 purple dash-dot trace). Instead, these models only achieved equivalent accuracy to their Gaussian counterparts at much lower gains ($g < 0.5$). Note that Dale's Law was enforced during training such that the sign of each connection was not allowed to change with back-propagation. Both outcomes are consistent with previous work on networks with cell-type specific connectivity. Specifically, it is known that networks that obey Dale's law perform more poorly [76], presumably because the constraint that each row of the hidden weight matrix must be either positive or negative, adversely restricts the space of possible solutions. Furthermore, at least one study found that the effective gain of a network with block connectivity structure is greater than the average gain, leading to larger learning capacity at lower gains [77].

1.5.3 PCIst

PCIst, when computed on biologically realistic connectivity models, was similar to that of the previously described Gaussian models. That is, the value increased as a function of initial weight gain, is maximal at the edge of chaos and decreases sharply as the maximum Lyapunov exponent nears zero. However, unlike the Gaussian models (Fig 2d), as the gain is further increased, the maximum Lyapunov exponent becomes negative again for the cortical column model (Figure S13)a. In this case, the PCIst metric reflects the fact that there are nevertheless outliers beyond the unit circle in the eigenspectrum of the connectivity matrix at higher gains (See Supplementary Section 5.4.1). The oscillatory dynamics that dominate at larger gains are amplified enough in the baseline condition as to make the relative response to perturbation negligible.

1.5.4 Learning Regime

In comparison to the Gaussian models, those with mesoscopic connectivity have similar respresentational alignment curves, transitioning towards lazier learning in the chaotic regime. Accordingly, the pattern of weight changes as a function of gain were also similar to the Gaussian models (Figure 6e-h). One exception is the thresholded mesoscopic model, which has larger changes in the hidden weight layer than the other models and higher representational alignment. The difference for this model is likely due to having thresholded the smallest weights while leaving the tails of the distribution, with larger weights. The cortical column models were unique in that they consistently found low-dimensional solutions, despite having notably higher representational alignment for all gains. The apparently lazier strategy employed to reach the solutions in this case likely reflects the restricted solution space of biologically unrealistic training with back-propagation under the constraint of Dale's law (Figure 7f-i).

2 Methods

2.1 Model Construction

We trained a series of recurrent neural networks (RNNs) to perform the ten digit (0-9) sequential MNIST task, in which networks sequentially receive 28 rows of 28 pixels from a handwritten digit. The model's task is to learn to correctly identify the digit after receiving the last row of pixels. All models consisted of an input layer with 28 units (U), a single recurrent layer with 198 hidden units ($H = [198, 198]$ matrix) , and a linear readout layer with input size 198 and output size (o_t) of 10.

$$h_t = tanh(Hh_t + Ux_t + b) \tag{1}$$

$$o_t = h_t A + c \tag{2}$$

The size of the hidden layer was chosen for consistency with the size of the experimentally derived mesoscopic connectivity of the mouse brain, which reflects the biological connectivity density between 198 major brain areas [74, 75]. The hidden layers of each network were initialized according to a Watts-Strogatz [65] connectivity rule over a range of nearest neighbors (nNN: 4, 8, 16, 28, 32, 64, 128, 198 = fully connected), and rewiring probabilities (p: 0.0, 0.2, 0.5, 0.8, 1.0), while self-connections were prohibited. The resulting networks systematically varied in both their degree of sparsity (nNN = 4 most sparse; nNN = 198 least sparse) and their degree of small-world structure (p = 0.0 highly structured small world architecture; p = 1.0 random Erdös Rényi connectivity). At initialization, non-zero weights of the hidden layer were sampled from a Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = \frac{g}{\sqrt{N_{nz}}}$, where N_{nz} is the number of non-zero elements and gain (g : 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 5.0), thus normalizing the variance across models at initialization. For each combination (nNN, p, g) we trained 10 randomly initialized models, 4800 models in total.

2.1.1 Biologically Realistic Connectivity

Mesoscopic Connectivity

We made use of a previously published whole-brain model of inter-regional mesoscopic connectivity of the mouse brain. The model is based on hundreds of anterograde tracing experiments in C57BL/6J mice across 12 major brain divisions including isocortex, olfactory areas, hippocampus, cortical subplate, striatum, pallidum, thalamus, hypothalamus, midbrain, pons, medulla and cerebellum, allowing for the creation of a whole-brain connectivity model at the scale of 100 μm voxels [74, 75]. Hidden layer connectivity between each of 198 regions are their connection densities. Connection density is defined as the sum of the connection strengths from all voxels in a source region to all voxels in the target region divided by the product of the region sizes, where voxel-wise connection strengths represent the fraction of voxel volume expressing fluorescence. Connection density values are positive by definition, but the signs of each connection in the matrix were assigned randomly, making approximately equal number of positive and negative weights. As above, network connection strengths were scaled at initialization by a range of gain values (g : 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 5.0). As we are interested in the potential computational advantage of biologically realistic connectivity structure, we did not explore rewiring these model connections (all models p = 0). A sparse mesoscopic connectivity model was subsequently derived from this connectivity matrix by removing connections below a value of .0188 in order to match its sparsity to that of the cortical column model described below.

Synaptic Connectivity of a Cortical Column

The hidden layer connectivity of the cortical column model was generated from a data set containing reconstructions of the dendritic trees of hundreds of thousands of neurons as well as their local axonal projections using electron microscopy, giving unprecedentedly accurate information on their 0.5 billion synaptic

connections []. The synapses are located within the binocular area of the primary visual cortex (V1) of a single mouse (192 days old). We made use of a subset of this data selecting 198 cells from the set of fully proofread neurons with the nearest euclidean distance from the center point between layers 2/3 and 4. For each neuron in the column, the connectivity strength is calculated as the sum over the volume of each post-synaptic density to each target cell, while noting cell types (excitatory vs inhibitory). For example, if cell a has 10 synapses on to cell b the connection strength of connection H_{ba} is the sum of the volume of those 10 synaptic densities at cell b. Connections from inhibitory neurons have sign -1, and those from excitatory have sign +1, adhering to Dale's law [78, 79]. The resulting connection matrix included 52 inhibitory cells and 146 excitatory cells. However, as the inhibitory cells tend to make a larger number of local connections. the ratio of inhibitory to excitatory weights was 1.60. The cortical column models differ substantially in both the distribution of their weights (Figure 7a) as well as their degree of sparsity . The sparsity of the cortical column-derived hidden layer is 85.71%, while the mescoscopic connectivity structure is fully connected, with the exception of self-connections. It is for this reason we included the nNN = 28 model as it matches the sparsity of the cortical column model (Table 1). Once again, networks connection strengths were scaled at initialization by a range of gain values (g : 0.5, 0.75, 1.0, 1.25 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 5.0) and for these models rewiring probability p = 0.

2.2 Training

All networks were trained using the Adam optimizer for 100 epochs with a learning rate of 1e-5, a batch size of 256 and cross-entropy loss. Note that while connection variance across models was normalized at initialization, it was not controlled over training. Model readout occurred at a single time point immediately following input of the last row of sMNIST digit input. Hidden states were initialized as zero. Sparsity was enforced over learning such that only non-zero weights were updated over the course of training. For cortical column models with cell-type specific connectivity, Dale's law was enforced over training, such that sign changes were prevented by clamping negative weights to a maximum of -.0001 and positive weights to a minimum of .0001.

2.3 Identifying Chaotic and Ordered Learning Regimes

2.3.1 Lyapunov Exponents

Gradients can explode or vanish exponentially over recurrent steps through a network, especially when the connection strengths between recurrent processing units are large, making models numerically unstable and difficult to train. As Lyapunov exponents represent the exponential growth rates of nearby trajectories in phase space of the model, we compute the finite time Lyapunov spectrum as a measure of model stability. Following the standard QR-decomposition technique [80] for computing the Lyapunov spectrum, we compute the average, over input samples, of the eigenvalues of the Jacobian of hidden state dynamics of each model using the python implementation in [66].

Specifically: A matrix Q is initialized as the identity matrix and hidden states are initialized as zero. At each of the T = 28 time points through the model, we compute the Jacobian matrix (first-order partial derivatives of equation 1 with respect to the hidden state dynamics h) and the product of the Jacobian matrix with the Q matrix. QR decomposition is applied to this product and used to update Q, which tracks the relative expansion or contraction of the model over time. The Lyapunov exponent of the i^{th} batch at timestep t r_t^i is the expansion of the i^{th} vector corresponding to the i^{th} diagonal element of R, and the i^{th} Lyapunov exponent λ_i is then given by

$$\lambda_i = \frac{1}{T}\sum_{t=1}^{T} log(r_i^t) \tag{3}$$

The exponents λ_i are computed in parallel for a batch of input samples (size 100) of sMNIST test data to drive the networks and are subsequently averaged over the batch. We compute Lyapunov exponents both at network initialization and after training. In Figure 2a,b, we report the maximum Lyapunov exponent resulting from this process after averaging over each of 10 model instantiations of a given model parameter

combination (nNN, p, g) for p = 1.0. (See Figure S1 for similar results obtained for rewiring probabilities [0, 0.2, 0.5, 0.75]). Chaotic dynamics are indicated by positive average largest Lyapunov exponent ($\lambda > 0$). Though we computed the spectrum both pre and post-training, we define the transition from ordered to chaotic $g_{c_{nNN}}$ dynamics as the gain at which the largest Lyapunov exponent $\lambda > 0$ in the network prior to training.

2.4 Assessing Sensitivity to Noise

Model sensitivity to noise was assessed by adding Gaussian noise $\mathcal{N}(0, 0.01)$ to sMNIST test data. We compare the accuracies of the trained model (trained in the absence of noise) to test data with and without noise. Models with chaotic dynamics are expected to exhibit greater sensitivity to noise and therefore greater decreases in accuracy.

2.5 PCIst

PCIst is a measure of the spatio-temporal complexity of the evoked response of a system to perturbation [35, 38]. It was developed for use with EEG data in response to a perturbation delivered via transcranial magnetic stimulation (TMS). However, the metric is applicable more generally to any evoked signal composed of a baseline state and a well-defined response period to a perturbation in a network of causally interacting units; thus, importantly it can be applied to artificial networks. Computing PCIst involves several steps: 1) Baseline and response periods are defined. 2) Singular value decomposition (SVD) is performed on the matrix consisting of the response time-series of the hidden states over time 3) Principal components are selected from the eigenvalues of the decomposed matrix so as to account for a user-specified amount of variance. 4) components are then selected in terms of their signal-to-noise ratio (SNR), calculated as the square root of the ratio of average response power 5) recurrence quantification proceeds on the remaining components by computing a distance matrix between all time points 6) This matrix is thresholded at several values and transition matrices are computed as the number of times the state crosses each threshold. The "st" in PCIst is derived from quantifying the number of state transitions in this segment of the analysis, where state transitions are measured. 6) An optimal threshold ϵ is determined such that the number of state transitions in the response relative to baseline is maximized. 7) The average number of state transitions in the baseline and response period of the n^{th} component is the difference in the number of state transitions in the ϵ-thresholded matrix of the response relative to the baseline scaled by the number of response samples. 8) PCIst is then the sum of these differences over components. Therefore, PCIst is simply a product of the number of retained components (reflecting the spatial differentiation of the response across the network), and the average number of state transitions across components (which reflect the temporal complexity present in each component.) See [38] for more details.

In our study, the time series used to compute PCIst comes from the hidden state space of the model. We defined the response period as the 28 steps corresponding to the trajectory of the network to a single sMNIST test digit (the perturbation). Since the network's response unfolds over 28 time steps, we defined a baseline period of equal length during which the network receives a small Gaussian noise input sampled from $\mathcal{N}(0, .01)$. The perturbation was a batch (size 256) from the sMNIST test data set of a given digit. The metric was computed on the mean over batches of the hidden layer states for baseline period and perturbations. PCIst was computed on all networks both at initialization (IPCIst) and after training on the sMNIST task.

2.6 Dimensionality

Dimensionality was computed from the Lyapunov spectrum of the trained models, as the attractor dimension using the Kaplan-Yorke conjecture [81]. First the Lyapunov spectrum was computed as in 2.3.1 on the trained rather than pre-trained models. The resulting average values for a particular model parameter set (nNN, p, g) are sorted from largest to smallest. Then let j be the largest index for which the sum of the cumulative sum of the exponents is greater than zero.

$$\sum_{i=1}^{j} \lambda_i \geq 0 \tag{4}$$

and

$$\sum_{i=1}^{j+1} \lambda_i < 0 \tag{5}$$

Then the dimension

$$D = j + \frac{\sum_{i=1}^{j} \lambda_i}{|\lambda_{j+1}|} \tag{6}$$

2.7 Characterizing Learning Regime

2.7.1 Weight Change

To assess the magnitude of weight changes at each layer of the network, we computed the L^2 or euclidean norm of the difference between the final weights W^f and the initial weights W^0.

$$|W^f - W^0| = \sqrt{\sum_i \sum_j (W_{ij}^f - W_{ij}^0)^2} \tag{7}$$

2.7.2 Representation Alignment

Representation alignment (RA) quantifies the directional change in the representational similarity matrix (RSM) due to training. Instead of focusing on the dissimilarity as used in representational similarity analysis [82], RSM focuses on the similarity between how two pairs of input are represented by computing the Gram matrix of last step hidden activity.

An increased level of representation alignment indicates a higher degree of lazy learning in the network, and it is obtained by [16]:

$$\begin{aligned} RA(R^{(T)}, R^{(0)}) &:= \frac{Tr(R^{(T)} R^{(0)})}{\|R^{(T)}\| \|R^{(0)}\|}, \\ R &:= H^T H, \end{aligned} \tag{8}$$

where H is the hidden activity, and $R^{(0)}$ and $R^{(T)}$ are the initial and final RSM, respectively.

2.7.3 Tangent kernel alignment

Tangent kernel alignment provides a measure of the directional change in the neural tangent kernel (NTK). The NTK is a mathematical tool that calculates the inner product of gradients for each input pair. Like representational alignment (RA), tangent kernel alignment calculates the Gram matrix between input pairs. However, unlike RA, it does so based on the gradient rather than the final hidden activity, thereby quantifying network similarity in terms of the gradients. In more specific terms, the NTK, for any given pair of inputs, determines the covariance between the gradients of the neural network's output with respect to its parameters.

A heightened degree of tangent kernel alignment points to a greater extent of lazy learning within the network. As outlined in [16], this is calculated as follows:

$$\begin{aligned} KA(K^{(T)}, K^{(0)}) &:= \frac{Tr(K^{(T)} K^{(0)})}{\|K^{(T)}\| \|K^{(0)}\|}, \\ K &:= \nabla y^T \nabla y \end{aligned} \tag{9}$$

where $K^{(0)}$ and $K^{(T)}$ denote the initial and final NTK, respectively.

3 Discussion

In this study, we centered our attention on recurrent neural networks (RNNs) and revealed that the manipulation of initial hidden weight gain results in "lazier" or "richer" learning. In the latter, which is brought about by smaller initialization gains, we observe more significant network changes (including substantial weight alteration and rotation of representation) as well as lower-dimensional representation. These findings are in line with existing literature on feedforward networks [14, 69]. To the best of our knowledge, prior research specifically characterizing the dichotomy of rich and lazy learning within the context of RNNs is scant [83]. Expanding on this exploration, we ventured further into the interplay between the rich and lazy learning transition and its connection to the transition to chaos. Our investigation characterizes solution attributes and learning behavior on both sides of this transition and suggests that rich representations can be biased into a network's learning strategy by tuning the Lyapunov spectrum towards ordered dynamics.

In the context of infinite-depth feedforward networks, the correlation between the rich/lazy transition and the chaotic/ordered transition has been previously documented [84, 85]. However, this exploration has yet to be extended to RNNs. Given the divergent behavior at the infinite-depth limit, and if such an infinite-depth limit in feedforward networks translates to an infinite sequence length limit in RNNs, our study admittedly has some limitations. Notably, we only considered finite sequence lengths in this work, leaving the exploration of infinite sequence limits as an area for future inquiry. Additionally, more nuanced categorization of learning regimes is left for future work [83]. Further, we manipulated the initial weight scale to tune between the learning regimes, but we recognize there are other parameters — such as network width, scaling of readout weights (α parameter) [14, 16] and initial weight rank [71] — that could also be adjusted to affect the transition. Hence, future research will explore these additional dimensions.

Additionally, in this work, we explored differences in network dynamics and learning regime on biologically realistic connectivity structures at two different scales. These structures differ in their weight distributions, degree distributions, inclusion of Dale's law, and the balance of positive/negative weights. In the mesoscale model, we found that longer-tailed distributions with only a single population did afford greater stability over a wider range of gains than Gaussian models, while learning equally well. Importantly, using the cortical column model, we found that having multiple populations from different distributions changed network dynamics dramatically. Theoretical work has yet to fully describe the expected dynamical transitions for connectivity matrices with multiple populations drawn from separate distributions, each with different non-zero means and standard deviations. Our results are nevertheless consistent with previous theoretical work showing that block connectivity structure with more than one population will have outliers in their eigenspectrum, when the sum of synaptic weights into each neuron is non-zero, while noting that such networks can have non-intuitive dynamics [86].

Finally, we computed PCIst, a clinically relevant measure of consciousness, for these models. To our knowledge, this is the first time that the metric has been characterized in artificial neural networks in relation to their associated Lyapunov exponents. Interestingly, we found that the metric can be predicted by the value of the maximum Lyapunov exponent, peaking at the edge of chaos, with larger values in the non-chaotic regime where we also observed a tendency towards rich learning. It is important to note that the biological connectivity structures used in our study result from fine-tuning both prenatally and over the course of the first months of life. As a result, their ordered dynamics may reflect the impact of prior tuning through biological feedback mechanisms to have this desirable characteristic. The result suggests that 1. biological systems are biased towards rich learning. 2. consciousness as measured by PCIst may be an evolutionary consequence of favoring rich learning strategies.

Although beyond the scope of this work, our observation that networks shift towards lazier learning strategies as dynamical complexity increases, raises interesting questions about brain states associated with more complex neural activity, such as when under the influence of psychedelics. The use of psychedelics in the treatment of psychiatric disorders has become increasingly commonplace. Although their therapeutic mechanism is not understood, recent research has found that psychedelics robustly increase brain complexity [87–89]. It has been suggested that such treatment may increase brain flexibility [90, 91]. Thus, this work raises an intriguing, testable hypothesis that presumed increased flexibility may include a shift in learning strategy under the influence of these drugs.

Overall, our work connects critical network dynamics, learning regimes, and measures of consciousness and characterizes the influence of network sparsity, structure, and weight variance on network dynamics and

learning strategy. We show that both learning regime and measures of consciousness undergo a transition at the coupling strength that delineates chaotic from ordered dynamics. As we continue to investigate, these findings promise to unlock deeper understanding and more robust applications within the field of artificial intelligence and neuroscience.

4 Acknowledgements

We wish to thank the Tiny Blue Dot Foundation, the NSF and the NIH for their support as this work was funded in part by grants from the Tiny Blue Dot Foundation, NSF 2223725, and NIH R01EB029813 and RF1DA055669. We thank Guillaume Lajoie and Stefano Recanatesi for their valuable discussion and feedback on this work. We also wish to thank the Allen Institute founder, P. G. Allen, for his vision, encouragement and support.

References

[1] Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

[2] Yiding Jiang, Pierre Foret, Scott Yak, Daniel M Roy, Hossein Mobahi, Gintare Karolina Dziugaite, Samy Bengio, Suriya Gunasekar, Isabelle Guyon, and Behnam Neyshabur. Neurips 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*, 2020.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Can sgd learn recurrent neural networks with provable generalization? *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[5] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[6] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

[8] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

[9] 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.

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

[11] Yuhan Helena Liu, Arna Ghosh, Blake Richards, Eric Shea-Brown, and Guillaume Lajoie. Beyond accuracy: generalization properties of bio-plausible temporal credit assignment rules. *Advances in Neural Information Processing Systems*, 35:23077–23097, 2022.

[12] 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.

[13] 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.

[14] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

[15] Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Rich and lazy learning of task representations in brains and neural networks. *BioRxiv*, pages 2021–04, 2021.

[16] Thomas George, Guillaume Lajoie, and Aristide Baratin. Lazy vs hasty: linearization in deep networks impacts learning schedule based on example difficulty. *arXiv preprint arXiv:2209.09658*, 2022.

[17] 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.

[18] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

[19] Alexander Atanasov, Blake Bordelon, Sabarish Sainathan, and Cengiz Pehlevan. The onset of variance-limited behavior for networks in the lazy and rich regimes. *arXiv preprint arXiv:2212.12147*, 2022.

[20] Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *arXiv preprint arXiv:2304.03408*, 2023.

[21] Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

[22] Edward Moroshko, Blake E Woodworth, Suriya Gunasekar, Jason D Lee, Nati Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *Advances in neural information processing systems*, 33:22182–22193, 2020.

[23] 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.

[24] 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.

[25] H. Sompolinsky, A. Crisanti, and H. J. Sommers. Chaos in Random Neural Networks. *Physical Review Letters*, 61(3):259–262, July 1988.

[26] Chris G. Langton. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: Nonlinear Phenomena*, 42(1-3):12–37, June 1990.

[27] Ling Feng, Lin Zhang, and Choy Heng Lai. Optimal Machine Intelligence at the Edge of Chaos. *arXiv*, October 2020. arXiv:1909.05176 [nlin, stat].

[28] Jannis Schuecker, Sven Goedeke, and Moritz Helias. Optimal Sequence Memory in Driven Random Networks. *Physical Review X*, 8(4):041029, November 2018.

[29] T. Toyoizumi and L. F. Abbott. Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime. *Physical Review E*, 84(5):051908, November 2011.

[30] Hübler Alfred Wotherspoon, Timothy. Adaptation to the edge of chaos with random-wavelet feedback. *J. Phys. Chem A*, 113:19–22, 2009.

[31] C. L. Webber and J. P. Zbilut. Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, 76(2):965–973, February 1994.

[32] N. Pradhan and P.K. Sadasivan. The nature of dominant lyapunov exponent and attractor dimension curves of eeg in sleep. *Computers in Biology and Medicine*, 26(5):419–428, September 1996.

[33] Daniel Toker, Ioannis Pappas, Janna D. Lendner, Joel Frohlich, Diego M. Mateos, Suresh Muthukumaraswamy, Robin Carhart-Harris, Michelle Paff, Paul M. Vespa, Martin M. Monti, Friedrich T. Sommer, Robert T. Knight, and Mark D'Esposito. Consciousness is supported by near-critical slow cortical electrodynamics. *Proceedings of the National Academy of Sciences*, 119(7):e2024455119, February 2022.

[34] Brandon R. Munn, Eli J. Müller, Jaan Aru, Christopher J. Whyte, Albert Gidon, Matthew E. Larkum, and James M. Shine. A thalamocortical substrate for integrated information via critical synchronous bursting. *Proceedings of the National Academy of Sciences*, 120(46):e2308670120, November 2023.

[35] Adenauer G. Casali, Olivia Gosseries, Mario Rosanova, Mélanie Boly, Simone Sarasso, Karina R. Casali, Silvia Casarotto, Marie-Aurélie Bruno, Steven Laureys, Giulio Tononi, and Marcello Massimini. A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Science Translational Medicine*, 5(198), August 2013.

[36] Dmitry O. Sinitsyn, Alexandra G. Poydasheva, Ilya S. Bakulin, Liudmila A. Legostaeva, Elizaveta G. Iazeva, Dmitry V. Sergeev, Anastasia N. Sergeeva, Elena I. Kremneva, Sofya N. Morozova, Dmitry Yu. Lagoda, Silvia Casarotto, Angela Comanducci, Yulia V. Ryabinkina, Natalia A. Suponeva, and Michael A. Piradov. Detecting the Potential for Consciousness in Unresponsive Patients Using the Perturbational Complexity Index. *Brain Sciences*, 10(12):917, November 2020.

[37] Mario Rosanova, Silvia Casarotto, Camilla Derchi, Gabriel Hassan, Simone Russo, Simone Sarasso, Alessandro Viganò, Marcello Massimini, and Angela Comanducci. The perturbational complexity index detects capacity for consciousness earlier than the recovery of behavioral responsiveness in subacute brain-injured patients. *Brain Stimulation*, 16(1):371, January 2023.

[38] Renzo Comolatti, Andrea Pigorini, Silvia Casarotto, Matteo Fecchio, Guilherme Faria, Simone Sarasso, Mario Rosanova, Olivia Gosseries, Mélanie Boly, Olivier Bodart, Didier Ledoux, Jean-François Brichant, Lino Nobili, Steven Laureys, Giulio Tononi, Marcello Massimini, and Adenauer G. Casali. A fast and general method to empirically estimate the complexity of brain responses to transcranial and intracranial stimulations. *Brain Stimulation*, 12(5):1280–1289, September 2019.

[39] Joseph P. Zbilut and Charles L. Webber. Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*, 171(3-4):199–203, December 1992.

[40] J.-P. Eckmann, S. Oliffson Kamphorst, and D. Ruelle. Recurrence Plots of Dynamical Systems. *Europhysics Letters*, 4(9):973, November 1987.

[41] Joseph P. Zbilut, José-Manuel Zaldivar-Comenges, and Fernanda Strozzi. Recurrence quantification based Liapunov exponents for monitoring divergence in experimental data. *Physics Letters A*, 297(3-4):173–181, May 2002.

[42] Pawel Kałużny and Remigiusz Tarnecki. Recurrence plots of neuronal spike trains. *Biological Cybernetics*, 68(6):527–534, April 1993.

[43] Nitza Thomasson, Thomas J. Hoeppner, Charles L. Webber, and Joseph P. Zbilut. Recurrence quantification in epileptic eegs. *Physics Letters A*, 279(1):94–101, 2001.

[44] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, and J. Kurths. Recurrence Plot Based Measures of Complexity and its Application to Heart Rate Variability Data. *Physical Review E*, 66(2):026702, August 2002. arXiv:physics/0201064.

[45] 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.

[46] 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.

[47] Christoph Stöckl, Dominik Lang, and Wolfgang Maass. Probabilistic skeletons endow brain-like neural networks with innate computing capabilities. *bioRxiv*, 2021.

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

[49] 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.

[50] 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.

[51] Danil Tyulmankov, Guangyu Robert Yang, and LF Abbott. Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. *Neuron*, 110(3):544–557, 2022.

[52] Michael Kleinman, Chandramouli Chandrasekaran, and Jonathan Kao. A mechanistic multi-area recurrent network model of decision-making. *Advances in Neural Information Processing Systems*, 34, 2021.

[53] Jimmy Smith, Scott Linderman, and David Sussillo. Reverse engineering recurrent neural networks with jacobian switching linear dynamical systems. *Advances in Neural Information Processing Systems*, 34, 2021.

[54] Joshua Glaser, Matthew Whiteway, John P Cunningham, Liam Paninski, and Scott Linderman. Recurrent switching dynamical systems models for multiple interacting neural populations. *Advances in neural information processing systems*, 33:14867–14878, 2020.

[55] Jonathan Kadmon, Jonathan Timcheck, and Surya Ganguli. Predictive coding in balanced neural networks with noise, chaos and delays. *Advances in neural information processing systems*, 33:16677–16688, 2020.

[56] 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.

[57] Rylan Schaeffer, Mikail Khona, Leenoy Meshulam, Ila Rani Fiete, et al. Reverse-engineering recurrent neural network solutions to a hierarchical inference task for mice. *bioRxiv*, 2020.

[58] Luke Y Prince, Ellen Boven, Roy Henha Eyono, Arna Ghosh, Joe Pemberton, Franz Scherr, Claudia Clopath, Rui Ponte Costa, Wolfgang Maass, Blake A Richards, et al. Ccn gac workshop: Issues with learning in biological recurrent neural networks. *arXiv preprint arXiv:2105.05382*, 2021.

[59] Timothy P Lillicrap and Adam Santoro. Backpropagation through time and the brain. *Current opinion in neurobiology*, 55:82–89, 2019.

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

[61] 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), 2021.

[62] 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.

[63] Owen Marschall, Kyunghyun Cho, and Cristina Savin. A unified framework of online learning algorithms for training recurrent neural networks. *Journal of Machine Learning Research*, 21(135):1–34, 2020.

[64] 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.

[65] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[66] Ryan Vogt, Maximilian Puelma Touzel, Eli Shlizerman, and Guillaume Lajoie. On Lyapunov Exponents for RNNs: Understanding Information Propagation Using Dynamical Systems Tools. *Frontiers in Applied Mathematics and Statistics*, 8:818799, March 2022.

[67] Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability, November 2022. arXiv:2103.00065 [cs, stat].

[68] Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability, April 2023. arXiv:2209.15594 [cs, math, stat].

[69] Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270.e11, April 2022.

[70] Matthew Farrell, Stefano Recanatesi, Timothy Moore, Guillaume Lajoie, and Eric Shea-Brown. Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion. *Nature Machine Intelligence*, 4(6):564–573, June 2022.

[71] 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*, 2023.

[72] Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Rich and lazy learning of task representations in brains and neural networks. preprint, Neuroscience, April 2021.

[73] Danielle Smith Bassett and Ed Bullmore. Small-World Brain Networks. *The Neuroscientist*, 12(6):512–523, December 2006.

[74] Joseph E. Knox, Kameron Decker Harris, Nile Graddis, Jennifer D. Whitesell, Hongkui Zeng, Julie A. Harris, Eric Shea-Brown, and Stefan Mihalas. High-resolution data-driven model of the mouse connectome. *Network Neuroscience*, 3(1):217–236, January 2019.

[75] Samson Koelle, Dana Mastrovito, Jennifer D Whitesell, Karla E Hirokawa, Hongkui Zeng, Marina Meila, Julie A Harris, and Stefan Mihalas. Modeling the cell-type specific mesoscale murine connectome with anterograde tracing experiments. preprint, Neuroscience, May 2023.

[76] Jonathan Cornford, Damjan Kalajdzievski, Marco Leite, Amélie Lamarquette, Dimitri M. Kullmann, and Blake Richards. Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. preprint, Neuroscience, November 2020.

[77] Johnatan Aljadeff, Merav Stern, and Tatyana Sharpee. Transition to Chaos in Random Networks with Cell-Type-Specific Connectivity. *Physical Review Letters*, 114(8):088101, February 2015.

[78] H. Dale. Pharmacology and nerve-endings (walter ernest dixon memorial lecture). *Theraputics and Pharmacology Section of Proceedings of the Royal Society of Medicine*, 28(3):319–332, 1935.

[79] Fat-P. Koketsu K. Eccles, J.C. Cholinergic and inhibitory synapses in a pathway from motor-axon collaterals to motoneurones. *The Journal of Physiolog*, 126(3):524–562, 1954.

[80] Giancarlo Benettin, Luigi Galgani, Antonio Giorgilli, and Jean-Marie Strelcyn. Lyapunov Characteristic Exponents for smooth dynamical systems and for hamiltonian systems; A method for computing all of them. Part 2: Numerical application. *Meccanica*, 15(1):21–30, March 1980.

[81] Nikolay Kuznetsov and Volker Reitmann. *Attractor Dimension Estimates for Dynamical Systems: Theory and Computation: Dedicated to Gennady Leonov*, volume 38 of *Emergence, Complexity and Computation*. Springer International Publishing, Cham, 2021.

[82] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4, 2008.

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

[84] Lechao Xiao, Jeffrey Pennington, and Samuel Schoenholz. Disentangling trainability and generalization in deep neural networks. In *International Conference on Machine Learning*, pages 10462–10472. PMLR, 2020.

[85] Mariia Seleznova and Gitta Kutyniok. Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization. In *International Conference on Machine Learning*, pages 19522–19560. PMLR, 2022.

[86] Kanaka Rajan and L. F. Abbott. Eigenvalue Spectra of Random Matrices for Neural Networks. *Physical Review Letters*, 97(18):188104, November 2006.

[87] A. Viol, Fernanda Palhano-Fontes, Heloisa Onias, Draulio B. de Araujo, and G. M. Viswanathan. Shannon entropy of brain functional complex networks under the influence of the psychedelic Ayahuasca. *Scientific Reports*, 7(1):7388, August 2017.

[88] Robin L. Carhart-Harris. The entropic brain - revisited. *Psychedelics: New Doors, Altered Perceptions*, 142:167–178, November 2018.

[89] Andres Ort, John W. Smallridge, Simone Sarasso, Silvia Casarotto, Robin Von Rotz, Andrea Casanova, Erich Seifritz, Katrin H. Preller, Giulio Tononi, and Franz X. Vollenweider. TMS-EEG and resting-state EEG applied to altered states of consciousness: oscillations, complexity, and phenomenology. *iScience*, 26(5):106589, May 2023.

[90] David Papo. Commentary: The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience*, 10, 2016.

[91] R. L. Carhart-Harris and K. J. Friston. REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics. *Pharmacological Reviews*, 71(3):316, July 2019.

[92] Terence Tao. Outliers in the spectrum of iid matrices with bounded rank perturbations. *Probability Theory and Related Fields*, 155(1-2):231–263, 2013.

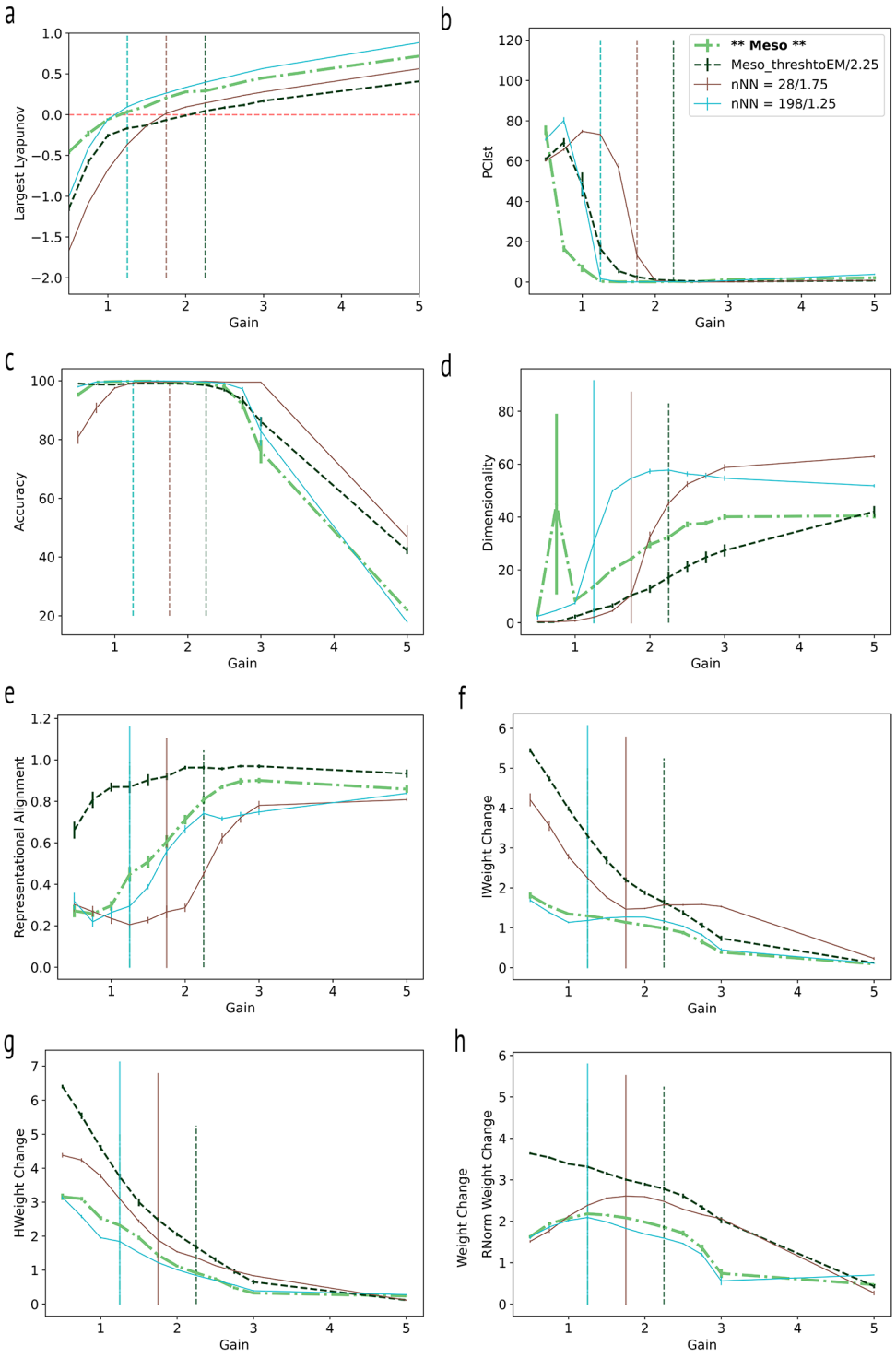

Figure 6: Mesoscopic Connectivity. **(a)** Largest lyapunov exponent λ of pre-trained models. True mesoscopic connectivity (indicated as **Meso** in green dash-dot) has a shallower slope than the Gaussian nNN = 198 model(brown). The same holds true for the thresholded mesoscopic model (dark blue - dashed) with respect to the sparsity-matched Gaussian nNN = 28 model (brown) **(b)** Initial PCIst in pre-trained models. **(c)** Model accuracy after 100 epochs of training **(d)** Kaplan-York (KY) dimensionality of trained models **(e)** Representational Alignment. **(f-h)** Norm Weight change in input, hidden and readout layers.

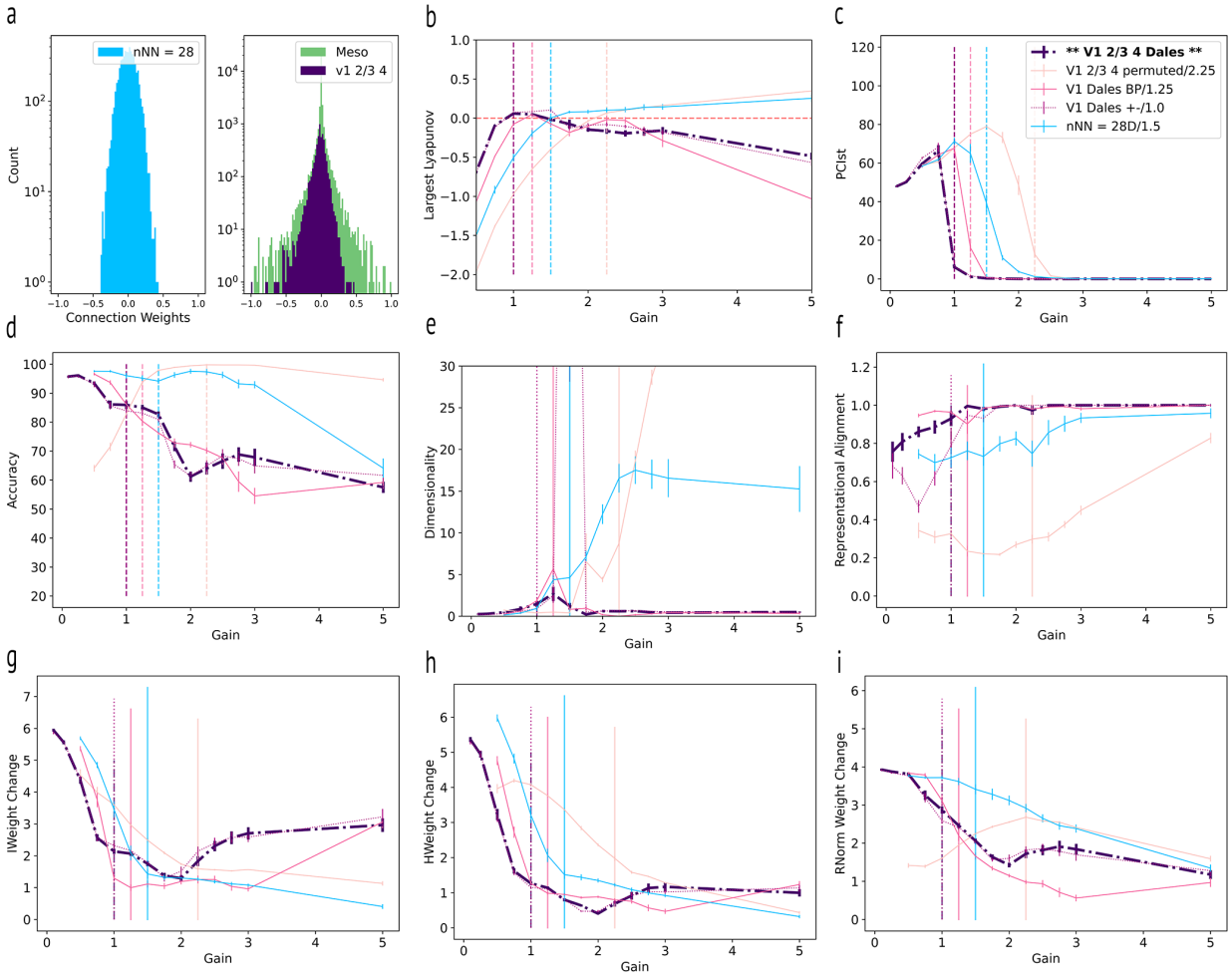

Figure 7: Cortical Column Connectivity. **(a)** Initial distributions of hidden layer connectivity with g = 1.0 for models with nNN = 28 drawn from $\mathcal{N}(0, \frac{1}{\sqrt{(N_{nz})}})$ (blue) in comparison to two experimentally derived biological connectivity structures: mouse mesocopic connectiome and synaptic connections within a cortical column of mouse V1. **(b)** Largest lyapunov exponent λ of pre-trained models: cortical column connectivity (indicated as ** V1 2/3 4 Dales ** in purple dash-dot), a permuted version of the cortical column in which the distribution is the same but the connection patterns are randomly altered, a block-permuted model of the cortical column where connections are permuted but Dale's law is preserved, and a sparsity-matched Gaussian nNN = 28 model with Dale's Law imposed. Biologically realistic connectivity distributions biases these models away from highly chaotic dynamics and lazy learning regime even at high gains. **(c)** PCIst in pre-trained models. **(d)** Model accuracy after 100 epochs of training **(e)** Kaplan-York (KY) dimensionality of trained models **(f)** Representational Alignment. **(g-i)** Norm Weight change in input, hidden and readout layers.

5 Supplementary

5.1 Maximum Lyapunov Exponents

We reported the maximum Lyapunov Exponents for models with rewiring probability = 1.0 (Erdös-Rényi) in the main text. Here we report similar results for all other rewiring probabilities [0.0, 0.2, 0.5, 0.8]. The overall pattern of transitions into the chaotic regime prior to training, with largest Lyapunov exponent $\lambda > 0$ are consistent with the Erdös-Rényi connectivity models, such that the sparsest models transition at higher gains and the least sparse models transition at lower gains (Figure S1 - left column). Additionally, for all models, λ tunes closer to the critical point ($\lambda = 0$) with training, regardless of whether λ is positive or negative prior to training (Figure S1 - right column).

5.2 Dimensionality sMNIST 2-digits

We trained an additional series of models for rewiring probability = 1.0 only, on the 2-digit sMNIST task, using digits [2,5] and found that models exceeded KY dimension of 2 rather than 10 at $g_{c_{nNN}}$ (Figure S3).

5.3 Neural Tangent Kernel

The NTK provides another method for quantifying the richness/laziness of learning [15]. The NTK increases towards a maximum value of one as network models begin to use lazier learning strategies. Consistent with the other metrics used to quantify the richness/laziness of learning, for most models, the NTK increases as a function of gain as networks approach the transition to chaos (Figure S4). However, because it is based on the gradients of the network's output with respect to its parameters, values can become numerically unstable as the model dynamics become chaotic. We nevertheless, include them for completeness.

5.4 Biologically Realistic Connectivity

We endeavored to identify the characteristics of the cortical connectivity structure (weight distribution, degree distribution, topological structure, Dale's law, relative balance of excitation and inhibition) that led to negative maximum lyapunov exponents at high gain. As described in the main text, we tested many altered connectivity structures to see which aspect of connectivity was necessary to achieve this. Here we show several additional models described but not shown in the main text (Figure S5.)

5.4.1 Generative Model Reproducing Dynamical Stability of Cortical Column Model

Biological neural networks can be divided into various sub-populations of morphologically and functionally defined cell types. Here, we focus on the two most basic cell types: excitatory and inhibitory. We split neurons in our experimental data set into two sub-populations ((E)xcitatory and (I)nhibitory). Experimentally estimated synaptic weights can be rewritten as a block matrix of the form

$$\mathbf{W} = \begin{bmatrix} \mathbf{W^{EE}} & \mathbf{W^{EI}} \\ \mathbf{W^{IE}} & \mathbf{W^{II}} \end{bmatrix} \tag{10}$$

In order to generate random surrogate weights we compute statistics (means and variances) of weights between each possible pair of populations ($E \to E$, $E \to I$, $I \to E$, $I \to I$). More precisely, for a given pair of populations (α, β) of sizes (N^α, N^β), we take the set of non-zero synaptic weights from neurons in population β to neurons in population α and calculate the mean $\overline{w^{\alpha\beta}}$ and variance $\overline{(w^{\alpha\beta} - \overline{w^{\alpha\beta}})^2}$. We define $\mu^{\alpha\beta}$ as the sum of mean weights that, on average, any neuron from population α receives from neurons in population β. If there are $C^{\alpha\beta}$ non-zero $\beta \to \alpha$ weights, the average number of non-zero weights from population β per neuron in population α is $K^{\alpha\beta} = C^{\alpha\beta}/N^\alpha$ and we should have

$$\mu^{\alpha\beta} = \overline{w^{\alpha\beta}} C^{\alpha\beta}/N^\alpha = \overline{w^{\alpha\beta}} K^{\alpha\beta} \tag{11}$$

Similarly, we define $\sigma^{\alpha\beta}$ as

$$\sigma^{\alpha\beta} = \sqrt{\overline{(w^{\alpha\beta} - \overline{w^{\alpha\beta}})^2} K^{\alpha\beta}} \tag{12}$$

nNN	p	Transition point
4	0.00	5.0
4	0.20	3.0
4	0.50	2.5
4	0.75	2.5
4	1.00	2.5
8	0.00	2.75
8	0.20	2.25
8	0.50	2.25
8	0.75	2.25
8	1.00	2.25
16	0.00	2.25
16	0.20	2.0
16	0.50	2.0
16	0.75	2.0
16	1.00	2.0
28	0.00	1.75
28	0.20	1.75
28	0.50	1.75
28	0.75	1.75
28	1.00	1.75
32	0.00	1.75
32	0.20	1.75
32	0.50	1.75
32	0.75	1.75
32	1.00	1.75
64	0.00	1.5
64	0.20	1.5
64	0.50	1.5
64	0.75	1.5
64	1.00	1.5
128	0.00	1.25
128	0.20	1.25
128	0.50	1.25
128	0.75	1.25
128	1.00	1.25
198	0.00	1.25
198	0.20	1.25
198	0.50	1.25
198	0.75	1.25
198	1.00	1.25

Table S1: Model Critical Points

The surrogate network is fully connected and is not subject to strict Dale's law. Its weights form a block matrix with entries generated randomly from normal distribution with matched experimental statistics. The numbers of neurons in each population, $\tilde{N}^E$ and $\tilde{N}^I$, do not have to match those in experimental data. Means and variances are rescaled by the number of neurons in the presynaptic population. More precisely, the random surrogate weight matrix takes the form

$$\tilde{\mathbf{W}} = \begin{bmatrix} \tilde{\mathbf{W}}^{\mathbf{EE}} & \tilde{\mathbf{W}}^{\mathbf{EI}} \\ \tilde{\mathbf{W}}^{\mathbf{IE}} & \tilde{\mathbf{W}}^{\mathbf{II}} \end{bmatrix} \tag{13}$$

$\alpha\beta$	$K^{\alpha\beta}$	$\overline{w^{\alpha\beta}}$	$\mu^{\alpha\beta}$	$\sigma^{\alpha\beta}$
EE	4.8	0.061	0.290	0.117
	(0.25)	(0.002)	(0.018)	(0.005)
EI	17.0	-0.058	-0.992	0.248
	(0.56)	(0.0012)	(0.038)	(0.009)
IE	27.9	0.054	1.51	0.261
	(3.09)	(0.0013)	(0.17)	(0.019)
II	18.2	-0.083	-1.52	0.467
	(1.05)	(0.0036)	(0.011)	(0.033)

Table S2: Connectivity statistics of the subset of EM V1 Cortical Column data set used in our numerical experiments.

where entries of each block are generated i.i.d. as

$$\tilde{W}_{ij}^{\alpha\beta} \sim \mathcal{N}\left(\mu^{\alpha\beta}/\tilde{N}^{\beta}, \sigma^{\alpha\beta}/\sqrt{\tilde{N}^{\beta}}\right) \tag{14}$$

We simulated networks driven entirely by recurrent connections (i.e., without any inputs) with either experimental or surrogate weights (Figure S8). Although the eigenvalues of the resulting random matrix do not match eigenvalues of experimental weights (Figure S8b,f), the qualitative features of its dynamics, including suppression of chaos (Figure S8a,e) and low-dimensional, periodic or quasiperiodic attractors (Figure S8c,d,g,h) are reproduced. In the surrogate network, the emergence of oscillations is driven by the presence of a pair of extreme eigenvalues ("outliers"). The position of the outliers can be predicted (Figure S8f) from the eigenvalues of the 2×2 matrix:

$$\mathbf{M} = \begin{bmatrix} \mu^{EE} & \mu^{EI} \\ \mu^{IE} & \mu^{II} \end{bmatrix}. \tag{15}$$

The effect of chaos suppression disappeared when standard deviations of weight distributions were increased by a factor of 2 (Figure S8i-l). These results were not finite-size effects as confirmed in simulations of larger surrogate networks (Figure S9).

Overall, our findings suggest that the main drivers of the chaos-suppressing oscillations in the original network may be imbalanced excitatory and inhibitory input weights in tandem with relatively low variability of the weights around the mean. However, the experimental eigenspectrum is markedly more complex than the eigenspectrum of random surrogate weights (Figure S8b,f), indicating that our simple generative model does not capture other, potentially crucial features of the original weight matrix. Outliers could for example appear due to pairwise weight correlations or cell-type-specific connectivity

The estimated mean weights $\overline{w^{\alpha\beta}}$ had comparable magnitudes for all four types of connections and the differences in the values of $\mu^{\alpha\beta}$, although statistically significant, were driven mostly by large differences in the values of $K^{\alpha\beta}$ (see Table S2). In particular, K^{EE} was much smaller than K^{EI}, leading to $\mu^{EE} < |\mu^{EI}|$. At this point it is worth noting that the EM V1 Cortical Column data is focused on local circuits as it does not include projections to a given neuron from neurons from outside its near neighborhood. This raises the possibility that the lack of E/I balance in the original network may only reflect local circuit connectivity patterns. In this view, statistics at larger ("global") spatial scales may be significantly different and we may expect the overall connectivity to be closer to the balanced regime. Importantly, however, the presence of multiple complex-valued outliers in the mean-balanced regime is still possible, albeit only if mean weights strongly dominate over their standard deviations ($\mu^{\alpha\beta} \gg \sigma^{\alpha\beta}$) [92], see Figure S10. Indeed, due to the local circuit sampling, the experimental values of $K^{\alpha\beta}$ in the data set are much lower than the total average number of presynaptic partners per neuron. Since the ratio $\mu^{\alpha\beta}/\sigma^{\alpha\beta}$ scales like $\sqrt{K^{\alpha\beta}}$ (assuming that global and local connectivity statistics are comparable), this may be the regime the underlying biological networks are operating in.

5.4.2 NTK on Biologically Realistic Connectivity

We report on the NTK results of the mesoscopic and cortical column models in the main text. Additionally, for completeness, we have included the results for all other model variants tested (Figure S11).

5.4.3 Relationship between PCIst and Lyapunov Exponents

In all models we explored, we found a similar relationship between the maximum Lyapunov exponent and PCIst. However, the size of the models we explored were kept constant. We therefore explored two larger models with recurrent layers of size 500 and 1000. From these larger models we see that the point at which PCIst begins to decrease moves closer to the edge of chaos as the size increases (Figure S12).

As we observed for Gaussian networks, for biologically realistic connectivity, PCIst drops off sharply when the Lyapunov exponent becomes positive (Figure S13). As gain increases further, the maximum Lyapunov exponent becomes negative again without a corresponding increase in PCIst. In this regime, where (as described above) chaos is quenched by oscillatory dynamics, the metric reflects the outliers in the eigenspectrum beyond the unit circle and the fact that perturbations to such networks do not result in strong signal beyond the background response to a Gaussian noise input (See Methods).

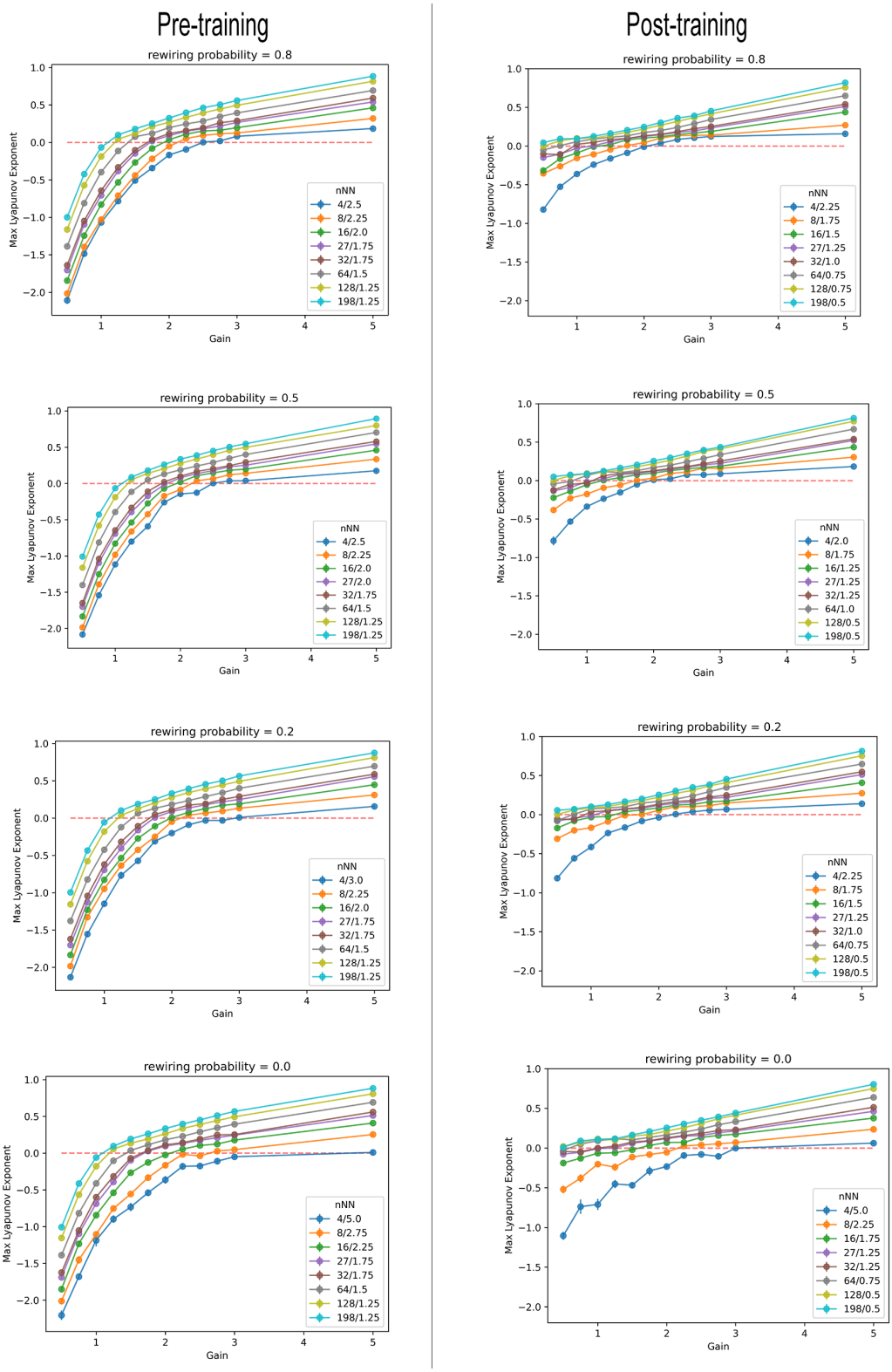

Figure S1: Maximum Lyapunov Exponents for rewiring probabilities [0.0, 0.2, 0.5, 0.8]. Pre-trained models (left column) Post-training (right column).

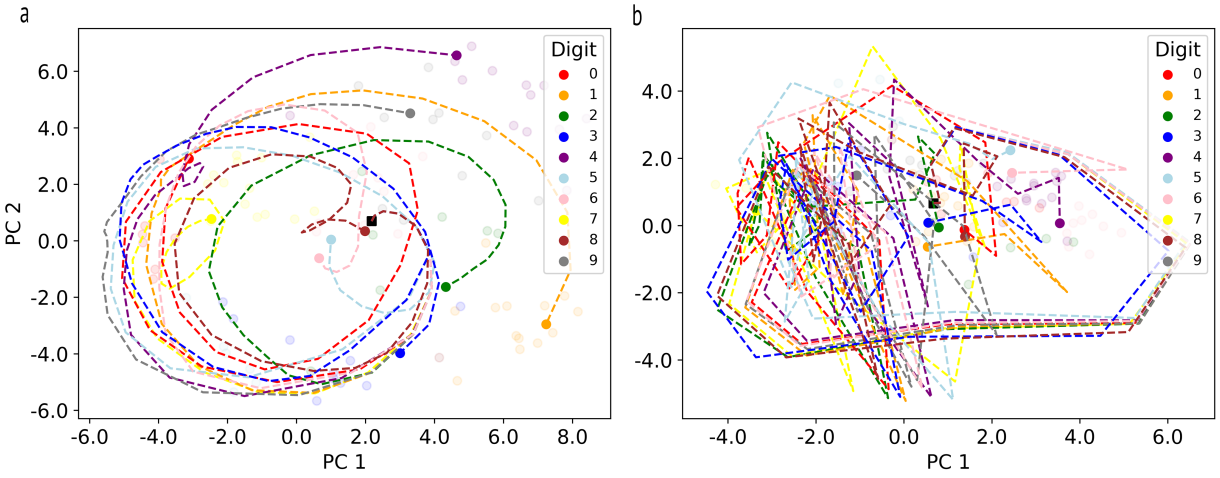

Figure S2: Example Trajectories for nNN = 28, p = 1.0 **(a)** Trajectory associated with ordered learning at gain = 1.0 < $g_{c_{nNN}}$ projected onto the first 2 principle components of the hidden states. Colors indicate MNIST digit class. **(b)** Trajectory associated with chaotic learning at gain = 2.5 > $g_{c_{nNN}}$

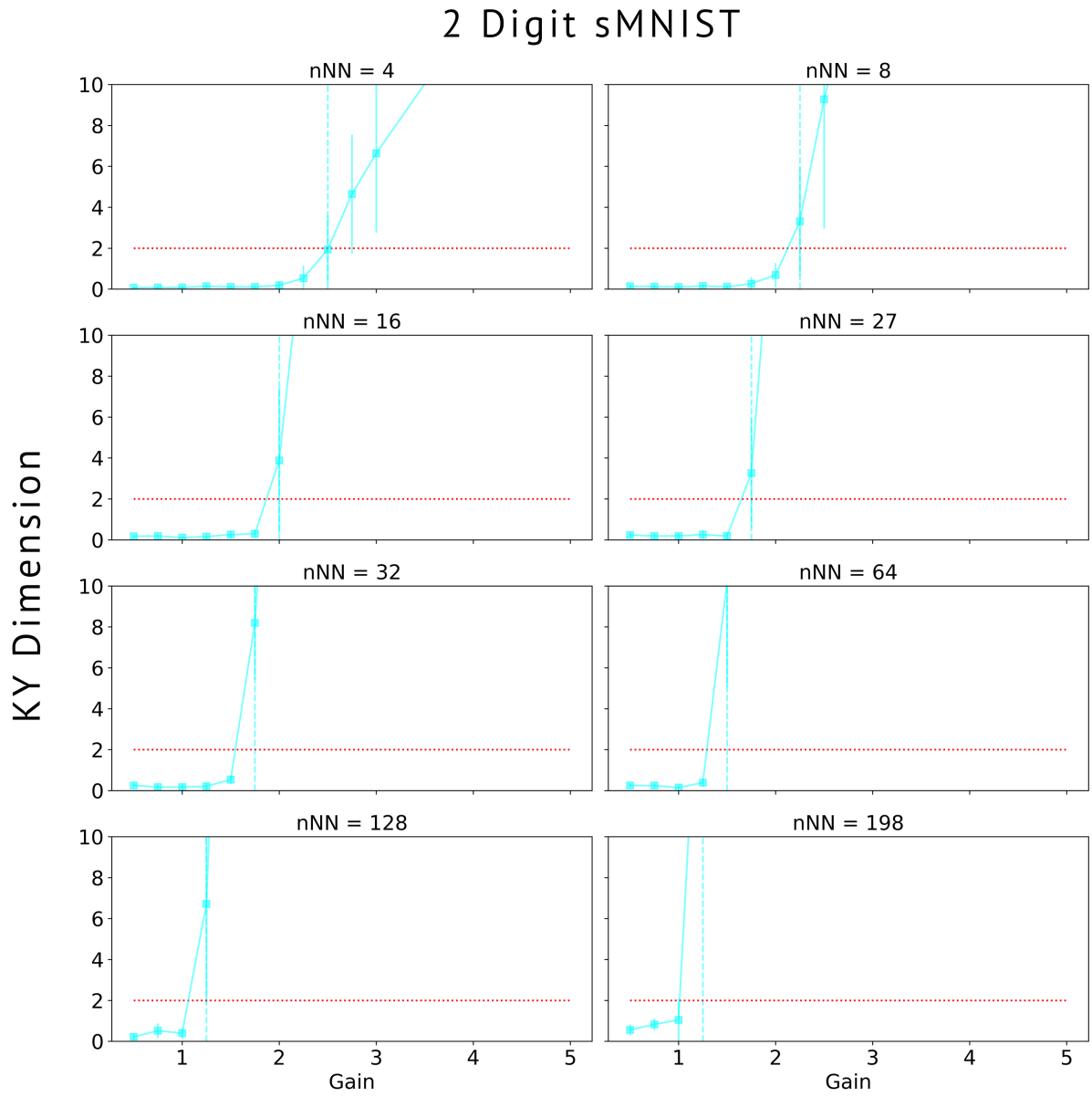

Figure S3: KY dimensionality for the 2-digit sMNIST task. Dotted red line = two.

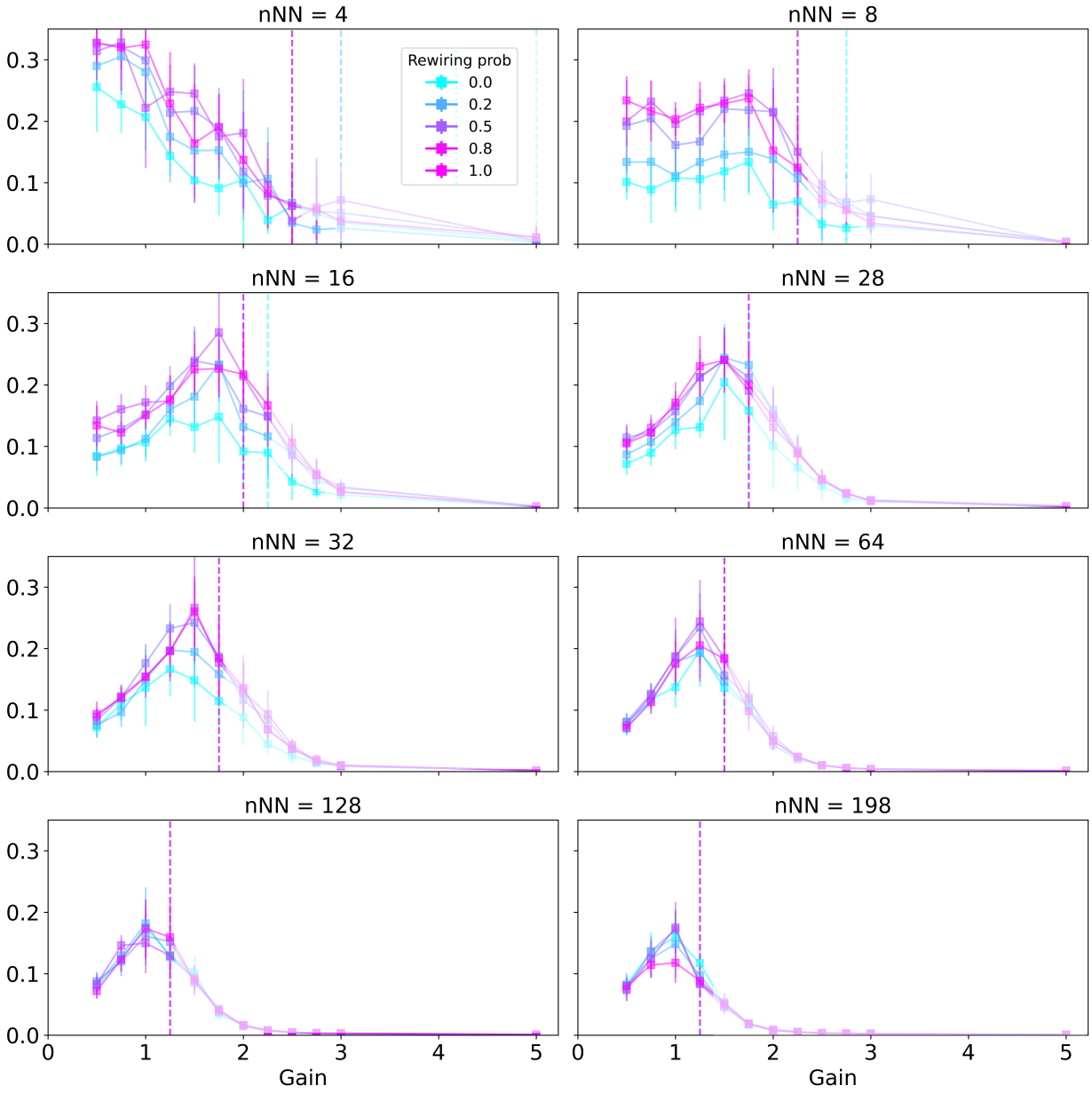

Figure S4: Neural Tangent Kernel Alignment. NTK values after the transition to chaos are greyed to indicate the fact that they may reflect numeric instabilities in their estimation.

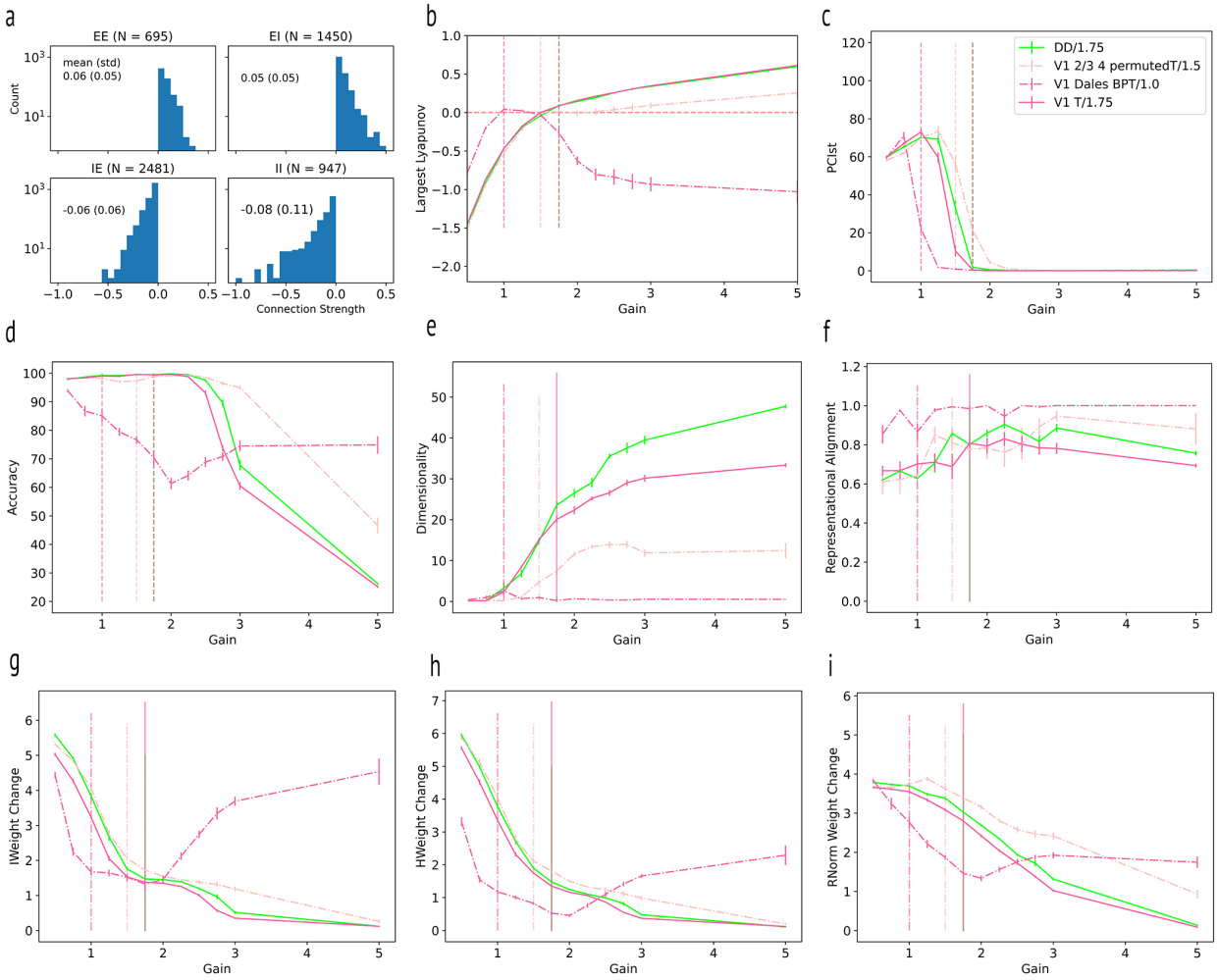

Figure S5: Cortical Column model and variants **(a)** Cortical column - distribution of weights for each cell type (E/I) block (E-E), (E-I), (I-E), (I-I) connectivity. **(b)** Maximum Lyapunov Exponents **(c)** PCIst pre-training **(d)** Accuracy achieved after 100 epochs of training on sMNIST task **(e)** KY Dimensionality of the final state. **(f)** Representational alignment **(g)** Input layer norm weight change **(h)** Hidden layer norm weight change **(i)** Output layer norm weight change

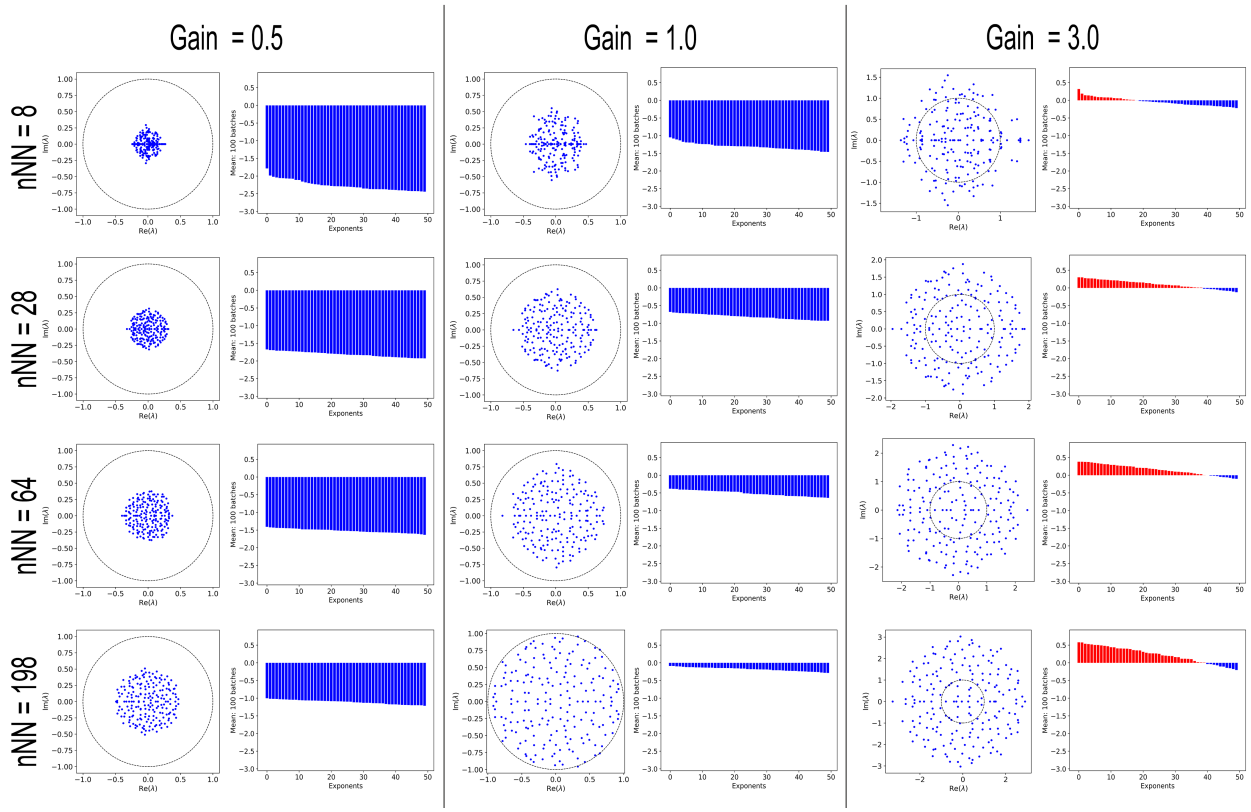

Figure S6: Lyapunov Exponents and Eigenvalues for Gaussian Models

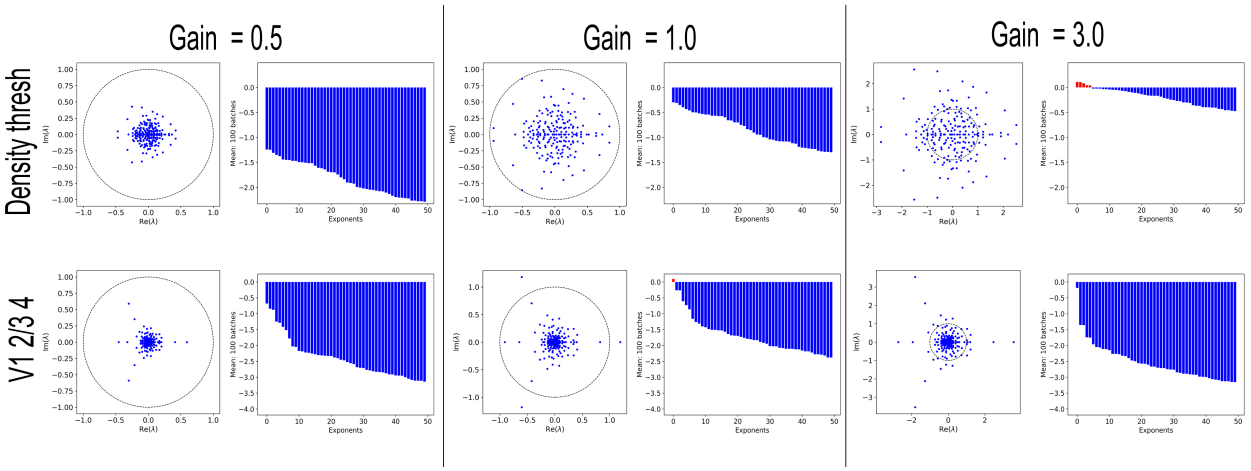

Figure S7: Lyapunov Exponents and EigenValues for Biologically Realistic Connectivity Models

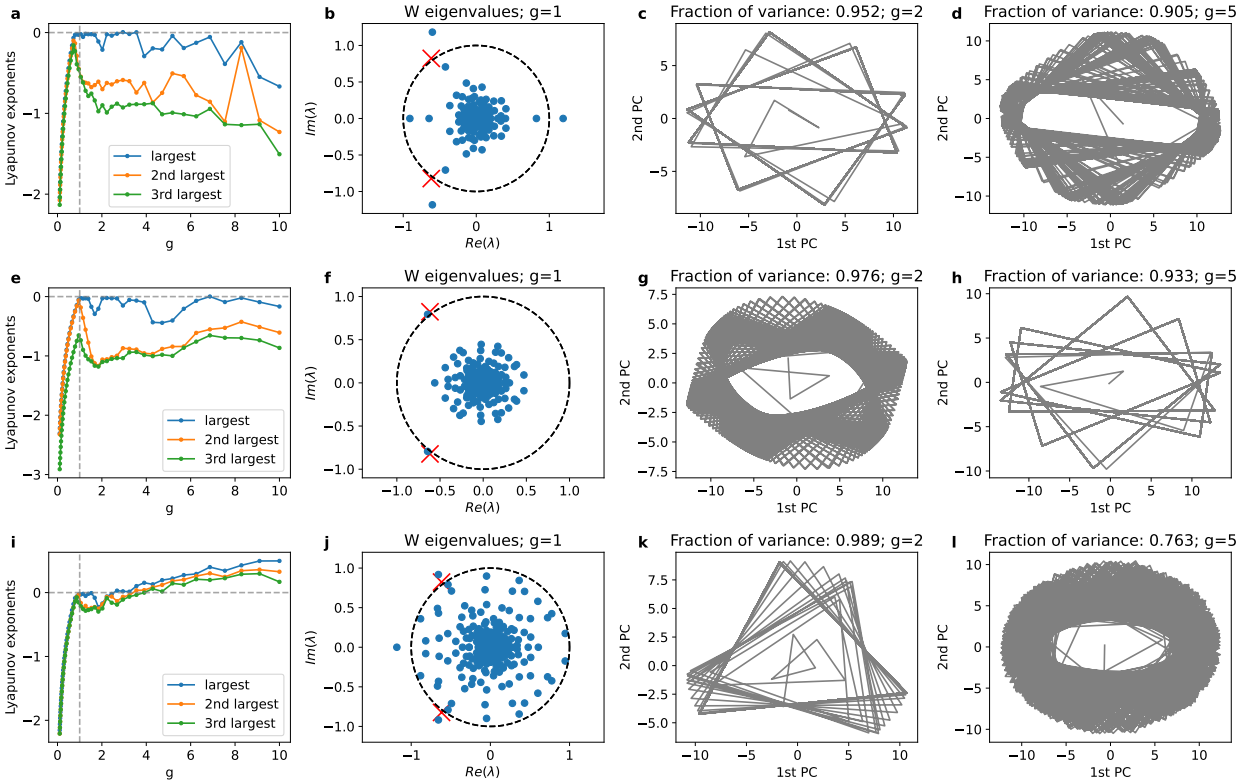

Figure S8: Random surrogate weights can reproduce qualitative features of neural dynamics driven by experimental weights. **(a)** Three largest Lyapunov exponents as functions of gain g using experimental weights. **(b)** Eigenvalues of the experimental weight matrix ($g = 1$). Most eigenvalues are contained within a relatively small circular-shaped core, but the presence of multiple outliers suggest non-random features of the connectivity. Red crosses correspond to a pair of eigenvalues of a 2×2 matrix $\mathbf{M}$ constructed from experimental mean weights. They do not match any of the outliers very well, but are close to the most extreme pair of complex-valued outliers. **(c)** Trajectory of the first two principal components for $g = 2$. **(d)** Same as (e) but with $g = 5$. **(e-h)** Same as (a-d) but with random surrogate weights with statistics matched to experimental data. The number of neurons in each population is the same as in the experiments, leading to a relatively large realization dependence (not shown), but the qualitative features of chaos suppression is robust. **(i-l)** Same as (e-h) but with $\sigma^{\alpha\beta}$ scaled up by a factor of 2. Here weight distributions are wide enough to diminish the influence of average inter-population structure. Thus, the classical scenario of transition of chaos is recovered.

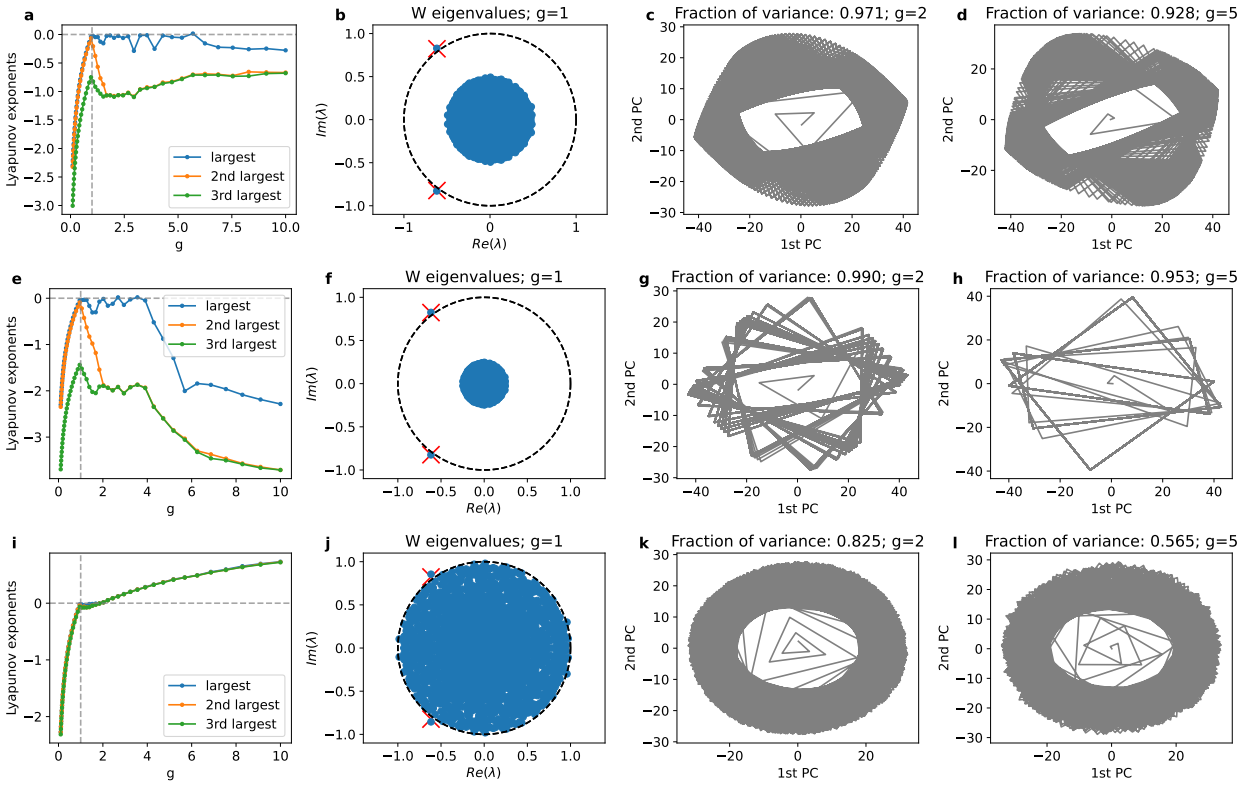

Figure S9: Results obtained with random surrogate weights are qualitatively the same as we increase the number of simulated neurons. **(a-d)** Same as Figure S8(b-c) but with larger populations ($N^E = N^I = 1000$) and appropriately rescaled moments otherwise matched to experimental data. **(e-h)** Same as (a-d) but with $\sigma^{\alpha\beta}$ scaled down by a factor of 0.5. **(i-l)** Same as (a-d) but with $\sigma^{\alpha\beta}$ scaled up by a factor of 2.

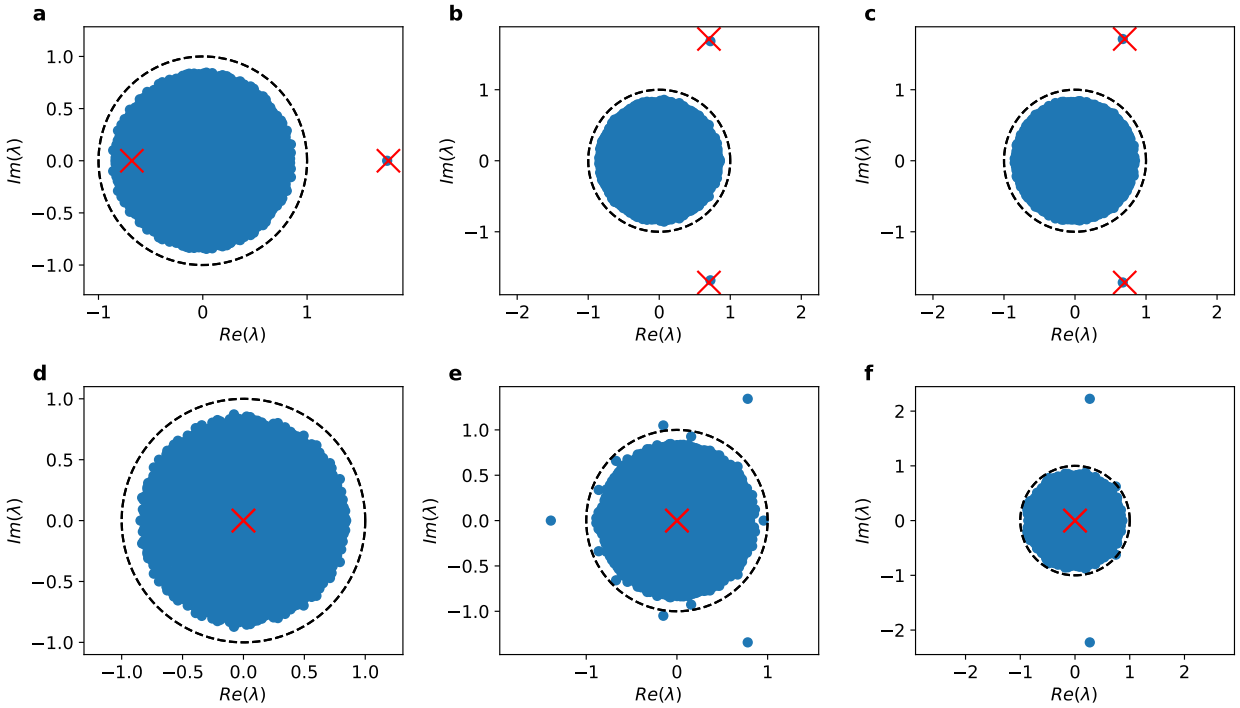

Figure S10: Eigenvalues of the weight matrix $\mathbf{W}$ depending on the matrix of means $\mathbf{M}$. Here, $N^E = N^I = 1000$ and $\sigma^{\alpha\beta} = 0.6$. **(a)** $\mathbf{M} = \begin{bmatrix} 2.2 & -1.1 \\ 1.1 & -1.1 \end{bmatrix}$. There are two real eigenvalues of $\mathbf{M}$ but only one lies outside of the disk defined by the circular law, so matrix W features a single deterministic outlier. **(b-c)** $\mathbf{M} = \begin{bmatrix} 0.7 & -1.4 \\ 2.1 & 0.7 \end{bmatrix}$. A pair of complex outliers is well-predicted by the eigenvalues of matrix $\mathbf{M}$. (b) and (c) correspond to two independent realizations of the weight matrix; the locations of the outliers are subject to small fluctuations that are expected to disappear with $N^E = N^I \to \infty$. **(d)** $\mathbf{M} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$. In the balanced regime, $\mathbf{M}$ has no non-zero eigenvalues and, as a result, no clear outliers are produced. **(e-f)** $\mathbf{M} = \begin{bmatrix} 100 & -100 \\ 100 & -100 \end{bmatrix}$. Although $\mathbf{M}$ has no non-zero eigenvalues, clear outliers are produced as a result of the large magnitude of the low-rank perturbation. The positions of the outliers are not deterministic, as confirmed by comparing two independent realizations in (e) and (f), and as such cannot be directly predicted based solely on $\mathbf{M}$.

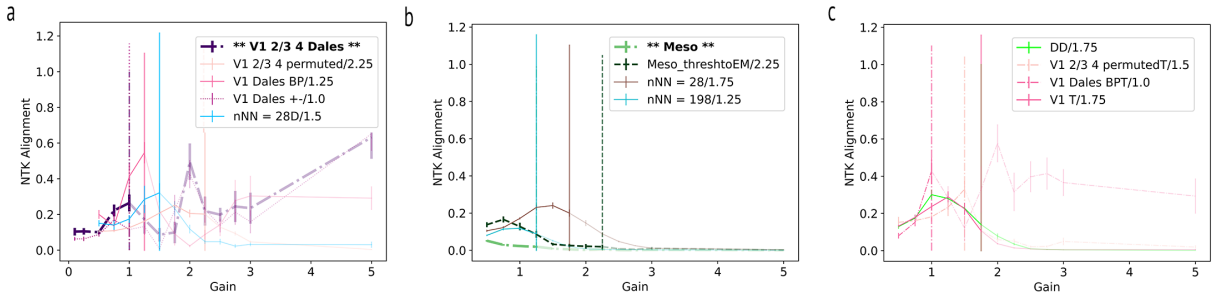

Figure S11: Neural tangent kernel alignment in networks with biologically realistic connectivity.

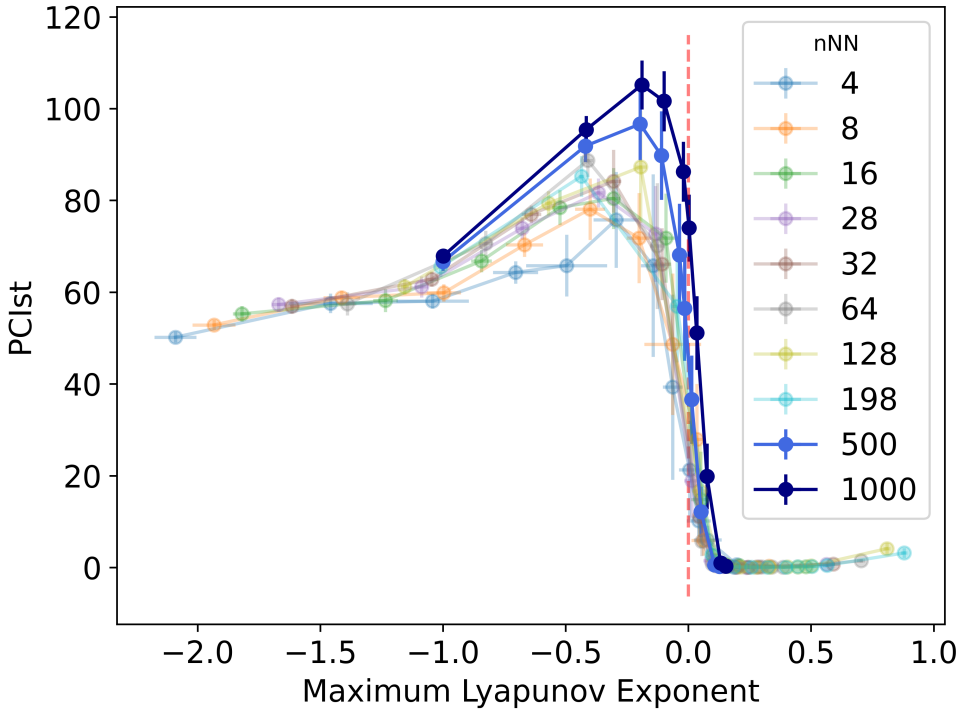

Figure S12: Relationship between PCIst and largest Lyapunov Exponent: Finite Size Effects. The point at which PCIst begins to decrease moves closer to the edge of chaos as networks increase in size.

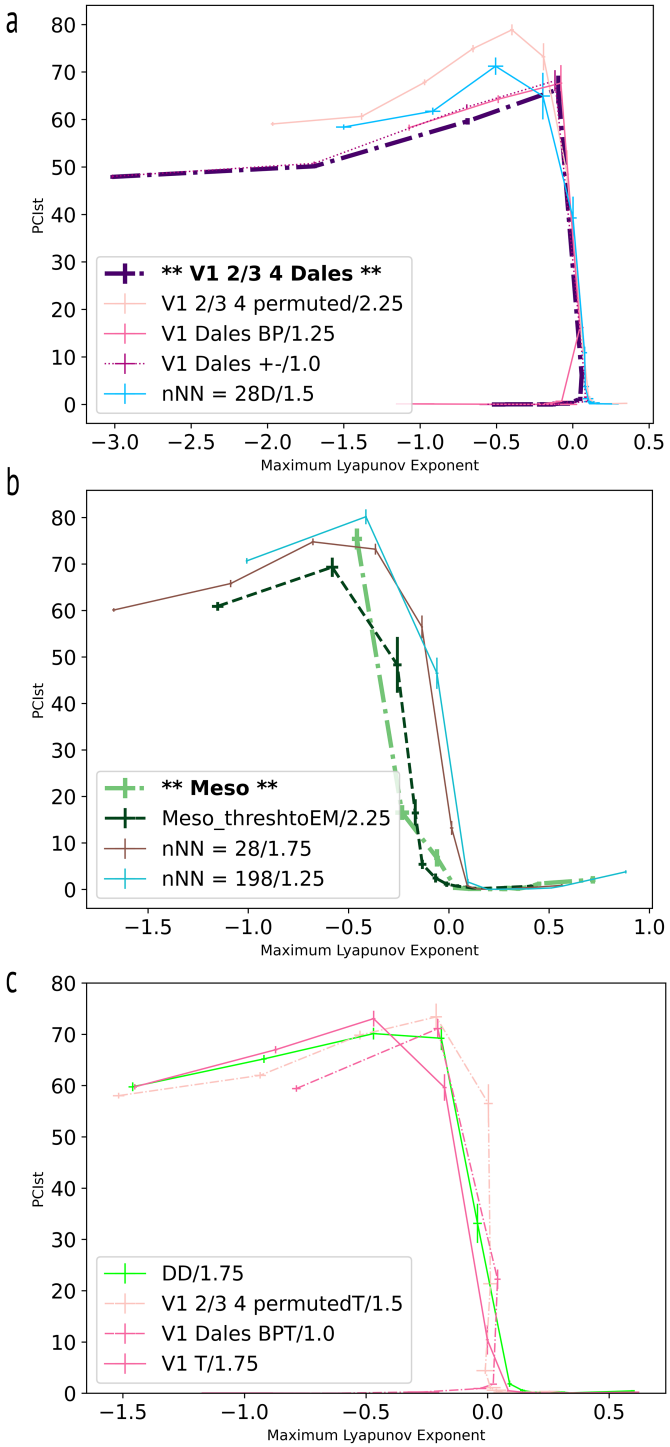

Figure S13: Relationship between PCIst and Lyapunov Exponents in all other model variants tested. Biologically realistic connectivity structure indicated in bold. PCIst decreases towards zero in chaotic regime as well as in models with Dale's Law with eigenspectrum outliers leading to oscillatory quenched chaos.