nature neuroscience

SUPPLEMENTARY INFORMATION

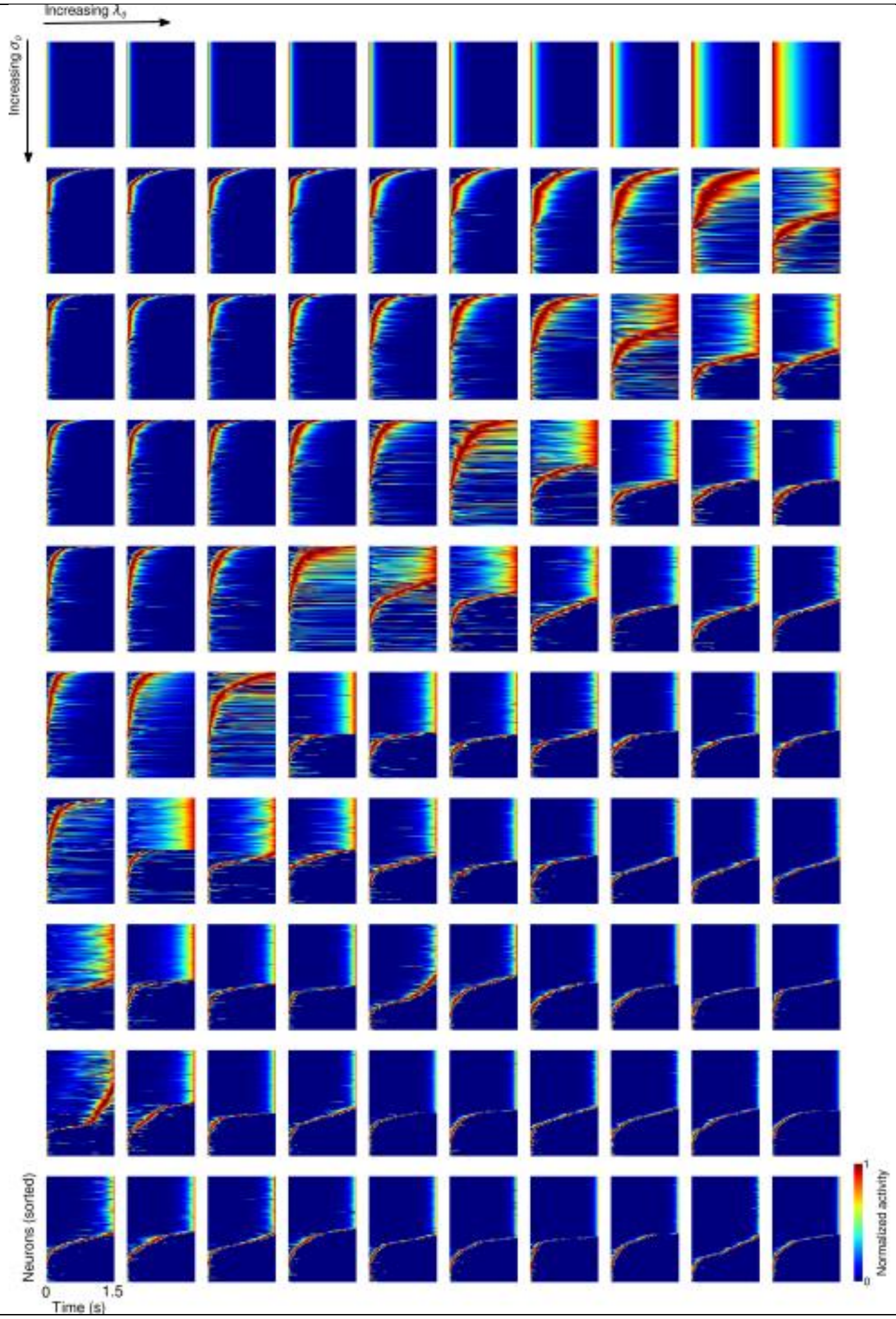
https://doi.org/10.1038/s41593-018-0314-y

In the format provided by the authors and unedited.

A diverse range of factors affect the nature of neural representations underlying short-term memory

A. Emin Orhan[1]* and Wei Ji Ma[2,3]

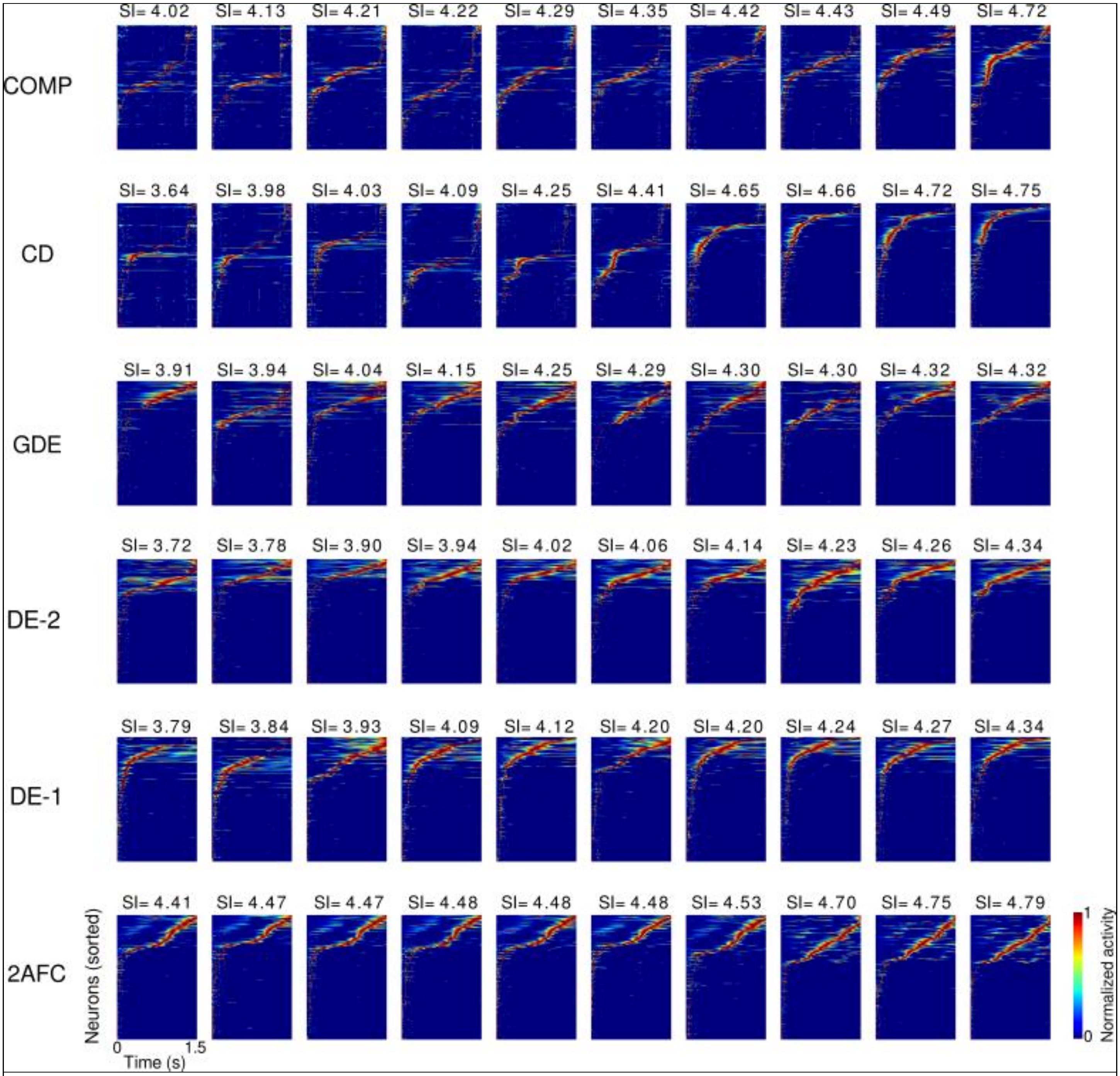
[1]Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA. [2]Center for Neural Science, New York University, New York, NY, USA. [3]Department of Psychology, New York University, New York, NY, USA. *e-mail: aeminorhan@gmail.com



Supplementary Figure 1

Initial, untrained network dynamics for different (λ_0,σ_0) values.

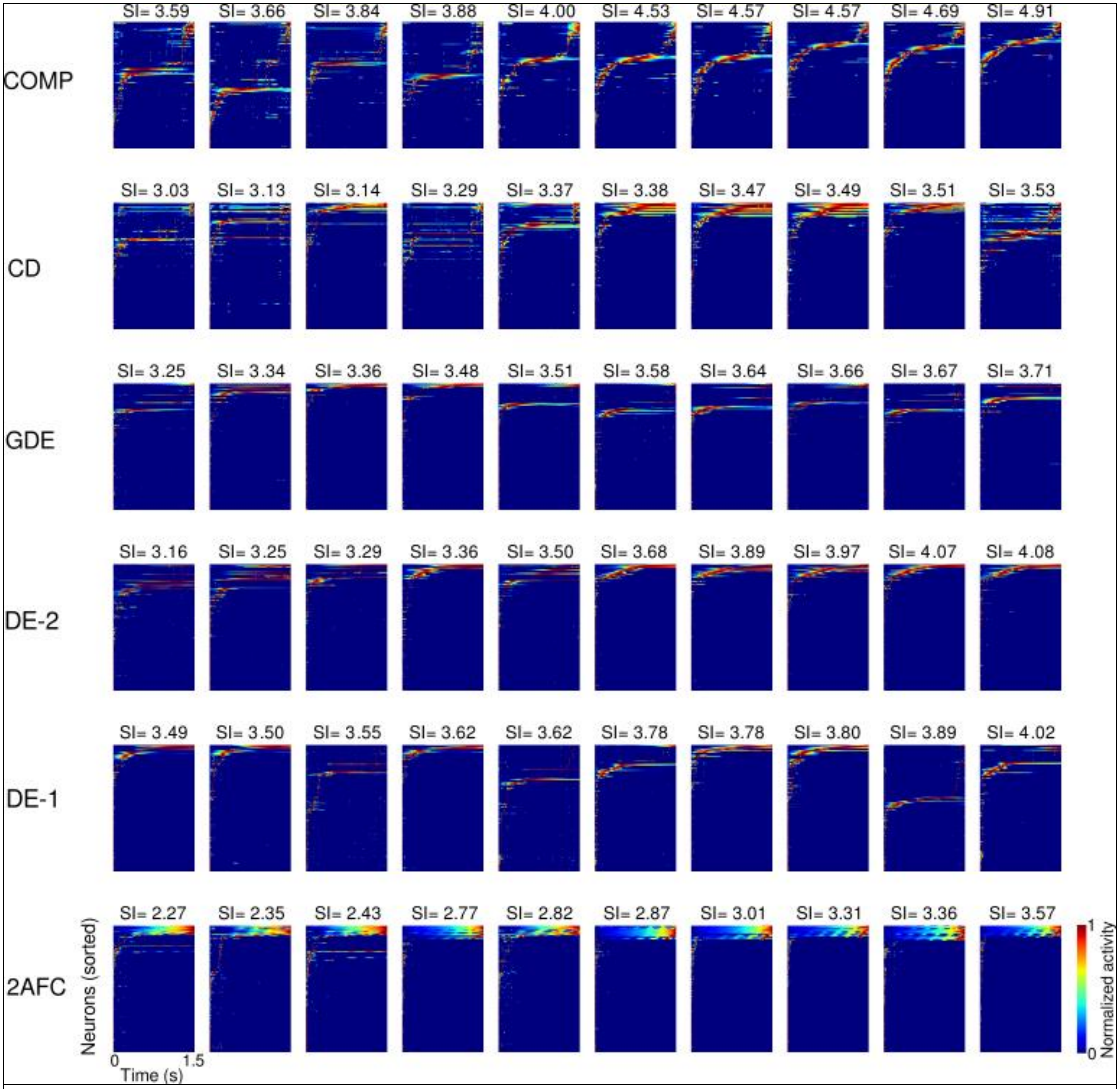
The heat maps show the normalized responses of the recurrent units to a unit pulse delivered at time *t*=0 to all units. Here, λ_0 takes 10 uniformly-spaced values between 0.8 and 0.98 (columns) and σ_0 takes 10 uniformly-spaced values between 0 and 0.4025 (rows).



Supplementary Figure 2

Normalized responses of the recurrent units in networks trained with strong initial network coupling and no regularization.

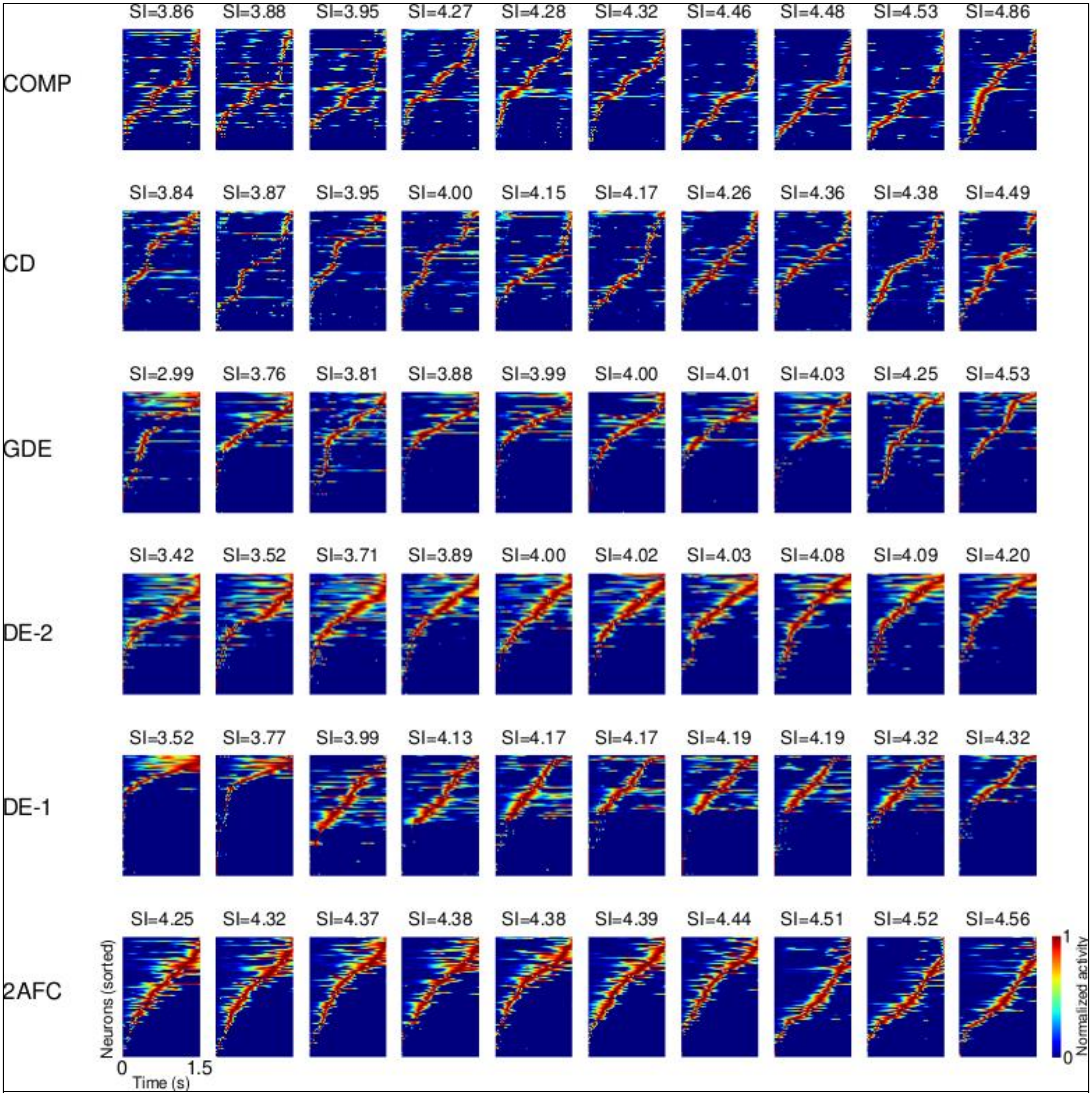
Each plot corresponds to an example trial from one of the six basic tasks. The SIs of the trials are indicated at the top of the plots. Trials are ordered by increasing SI from left to right. All trials shown here are from networks trained with λ_0=0.96, σ_0=0.313, ρ=0. After training, all networks shown here achieved a test set performance within 25% of the optimal performance. In Supplementary Figures 2-5, only the active recurrent units are shown.



Supplementary Figure 3

Normalized responses of the recurrent units in networks trained with weak initial network coupling and no regularization.

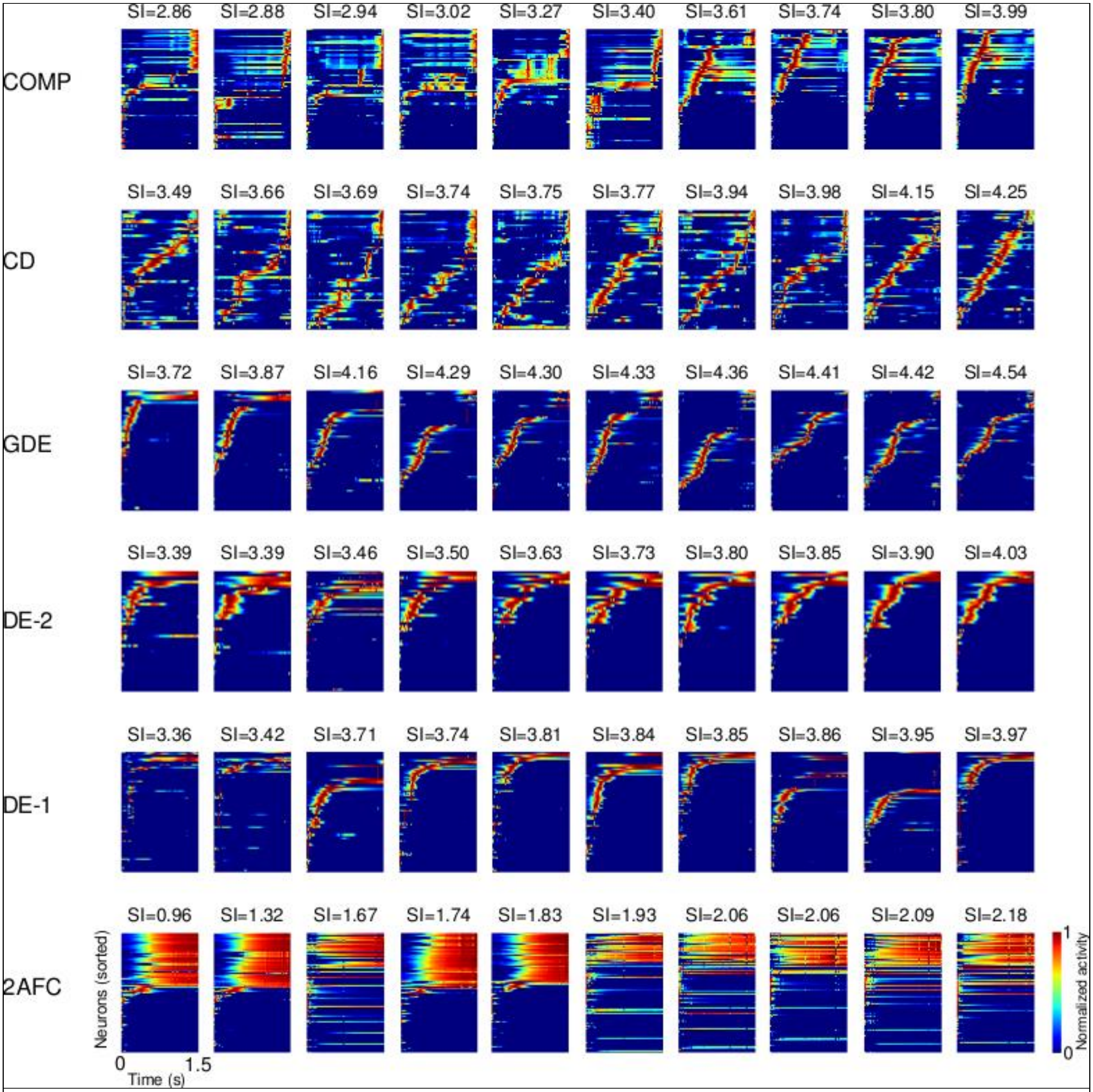
Each plot corresponds to an example trial from one of the six basic tasks. The SIs of the trials are indicated at the top of the plots. Trials are ordered by increasing SI from left to right. All trials shown here are from networks trained with λ_0=0.96, σ_0=0.134, ρ=0. After training, all networks shown here achieved a test set performance within 50% of the optimal performance.



Supplementary Figure 4

Normalized responses of the recurrent units in networks trained with strong initial network coupling and strong regularization.

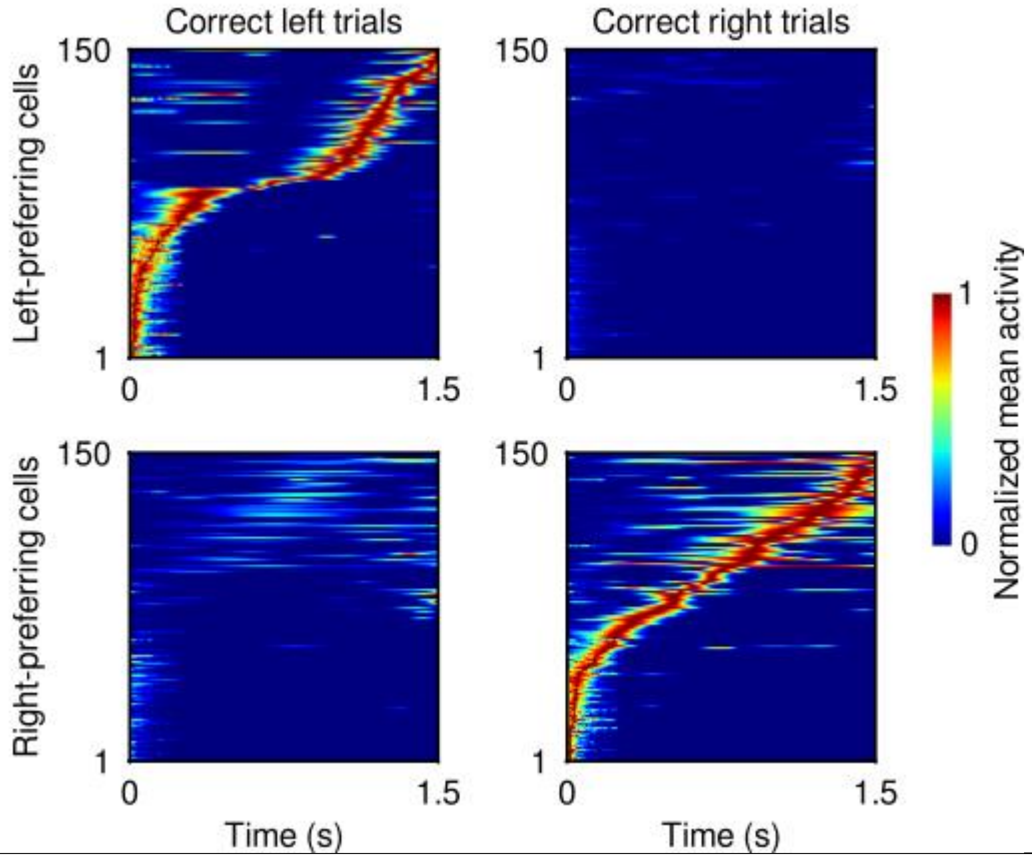
Each plot corresponds to an example trial from one of the six basic tasks. The SIs of the trials are indicated at the top of the plots. Trials are ordered by increasing SI from left to right. All trials shown here are from networks trained with λ_0=0.96, σ_0=0.313, $\rho=10^{-3}$. After training, all networks shown here achieved a test set performance within 50% of the optimal performance.



Supplementary Figure 5

Normalized responses of the recurrent units in networks trained with weak initial network coupling and strong regularization.

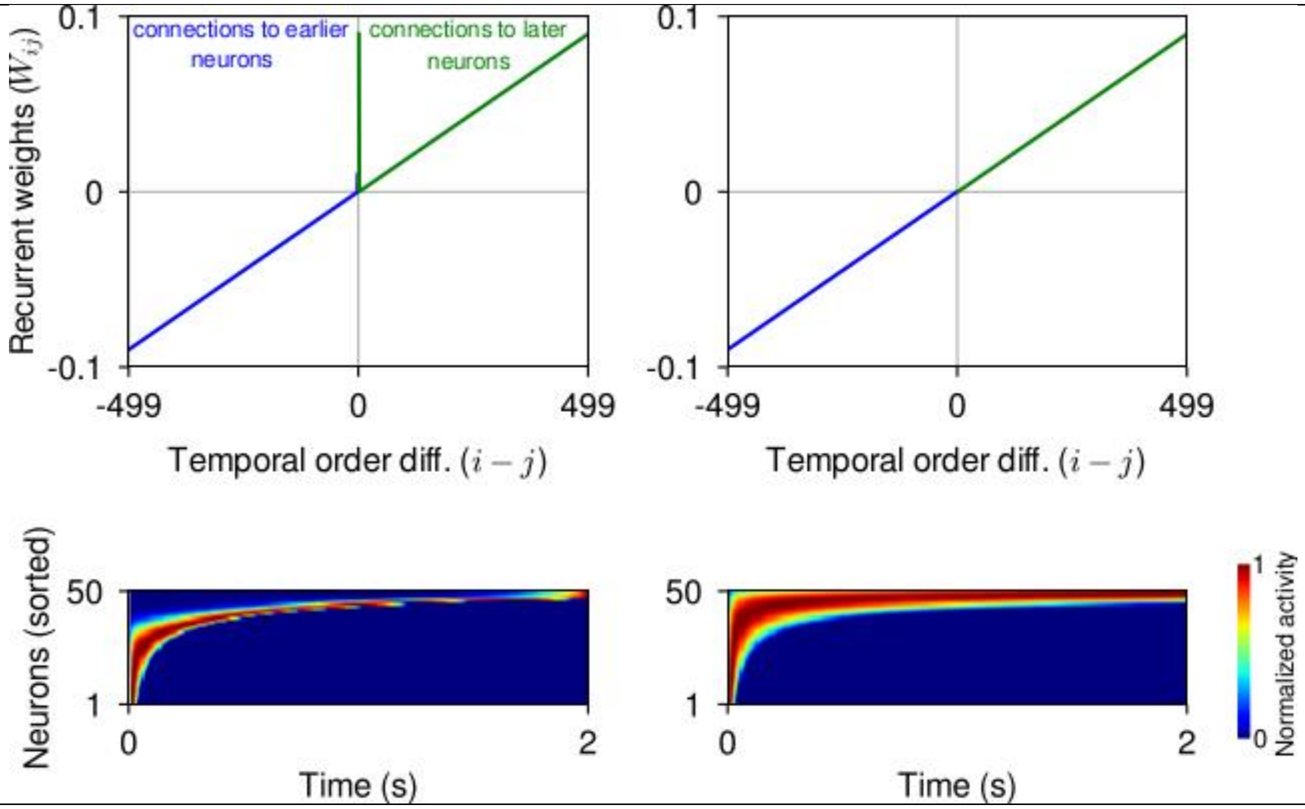
Each plot corresponds to an example trial from one of the six basic tasks. The SIs of the trials are indicated at the top of the plots. Trials are ordered by increasing SI from left to right. All trials shown here are from networks trained with λ_0=0.96, σ_0=0.134, ρ=10^{-3}. After training, all networks shown here achieved a test set performance within 50% of the optimal performance.



Supplementary Figure 6

Average normalized activity of recurrent units in an example network trained in the 2AFC task.

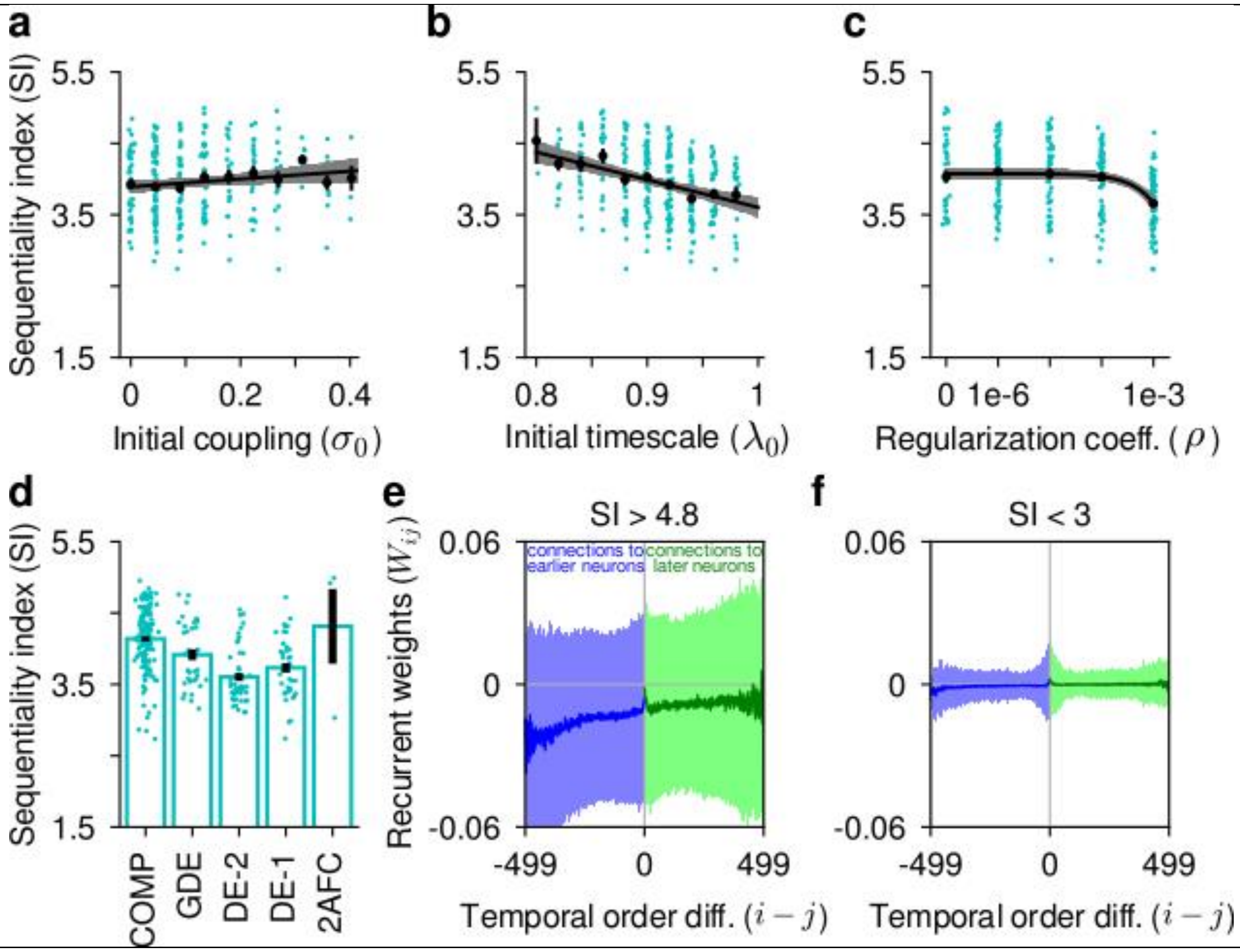
The network shown here was trained with λ_0=0.96, σ_0=0.313, ρ=0. After training, the network achieved a test set performance within 0.1% of the optimal performance. As in ref. 16, we divided the recurrent units into left-preferring and right-preferring ones based on whether they responded more strongly during correct left choices or during correct right choices. The upper panel shows the average normalized responses of the left-preferring units in the correct left and correct right trials, respectively. Similarly, the lower panel shows the average normalized responses of the right-preferring units in the correct left and correct right trials. As reported in ref. 16, the trained network developed choice-specific sequences in the 2AFC task (cf. Figure 2c in ref. 16). Only the most active 150 units from each group are shown in this figure; as always, the original network contained 500 recurrent units. This figure also demonstrates that the sequences are consistent from trial to trial, since the sequential activity pattern does not disappear when the responses are averaged over multiple trials.



Supplementary Figure 7

A simplified model of recurrent dynamics.

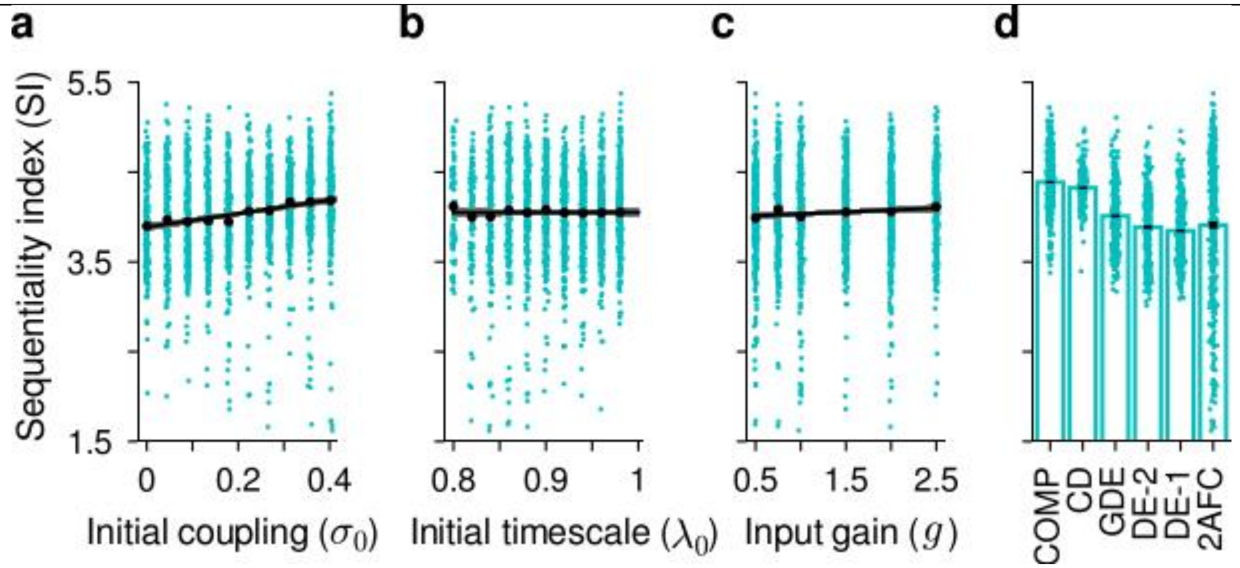
A simplified model that only incorporated the ReLU nonlinearity and the mean recurrent connection weight profiles shown in the upper panel (with no fluctuations around the mean) qualitatively captured the difference between the emergent sequential vs. persistent activity patterns (lower panel, left and right plots respectively). The networks simulated here had 500 recurrent units (only the most active 50 units are shown in the lower panel). All recurrent units received a unit pulse input at *t*=0. The self-recurrence term in the recurrent connectivity matrix (not shown in the upper panel for clarity) was set to 1 in both cases. In the sequential case, the off-diagonal band was set to 0.09 in the forward direction and 0.01 in the backward direction, i.e. $W_{i,i-1}$=0.09 and $W_{i-1,i}$=0.01. The recurrent units did not have a bias term and they did not receive any direct inputs during the trial other than the unit pulse injected at the beginning of the trial.



Supplementary Figure 8

Results for the clipped ReLU networks.

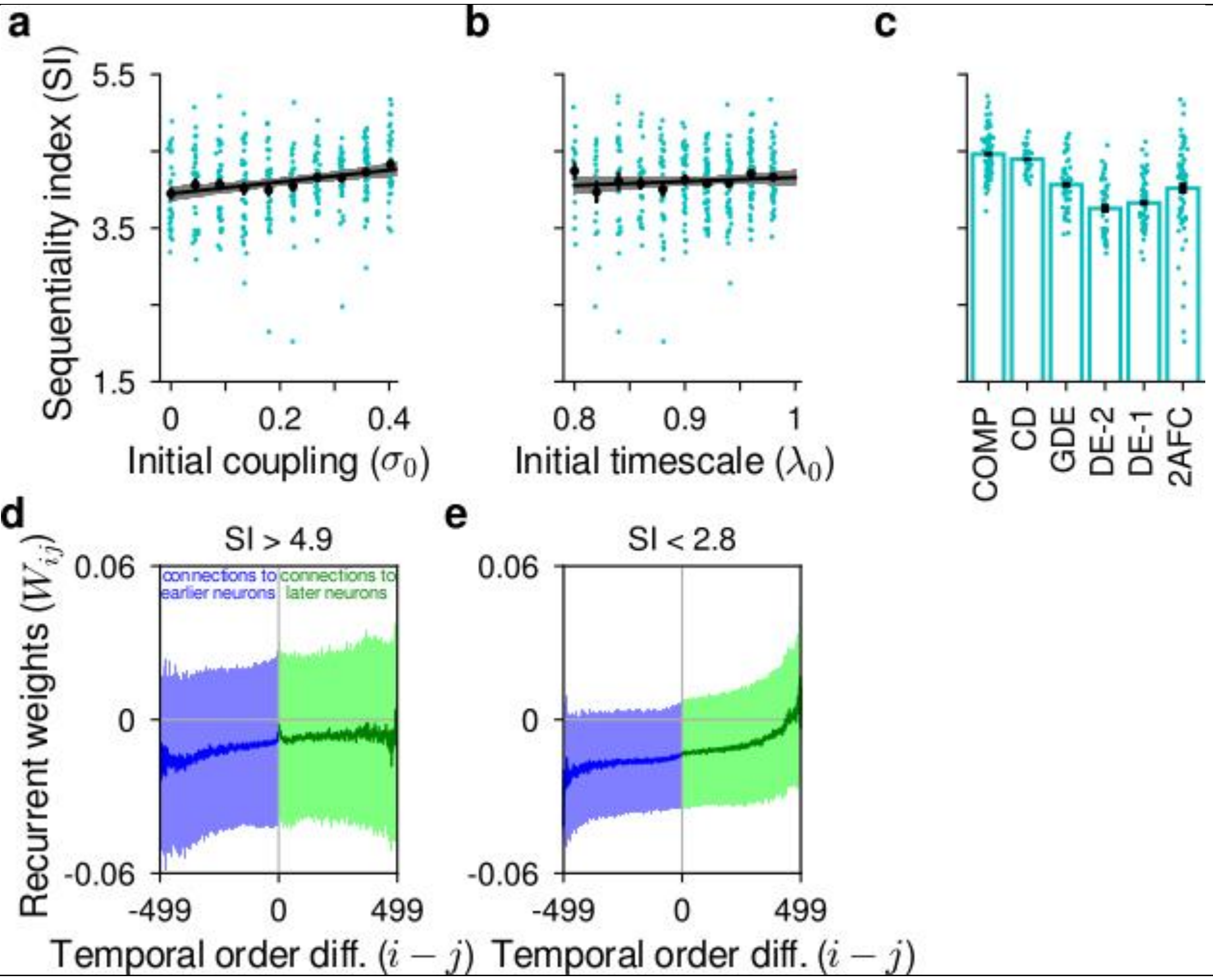
The clipped ReLU nonlinearity is similar to ReLU except that it is bounded above by a maximum value: i.e. $f(x)=clip(x, r_{min}, r_{max})$, where r_{min}=0 and r_{max}=100. **a** SI increased significantly with σ$_0$. Linear regression slope: 0.55∓0.28, R^2=0.01 (two-sided Wald test, n=280 experimental conditions, p=0.049). In **a-c**, solid black lines are the linear fits and shaded regions are 95% confidence intervals for the linear regression. **b** SI decreased significantly with λ$_0$. Linear regression slope: -3.87∓0.66, R^2=0.11 (two-sided Wald test, n=280 experimental conditions, p=0.000). Note that this result differs from the corresponding result in the case of ReLU networks, where λ$_0$ did not have a significant effect on the SI (Figure 2c). **c** SI decreased significantly with ρ. Linear regression slope: -418∓64, R^2=0.13 (two-sided Wald test, n=280 experimental conditions, p=0.000). **d** SI as a function of task. Overall, the ordering of the tasks by SI was similar to that obtained with the ReLU nonlinearity (Figure 3a). However, note that training was substantially more difficult with the clipped ReLU nonlinearity than with the ReLU nonlinearity. Across all tasks and all conditions, ReLU networks had a training success (defined as reaching within 50% of the optimal performance) of ~60%, whereas the clipped ReLU networks had a training success of only ~9.3%. In particular, we were not able to successfully train any networks in the CD task and very few in the 2AFC task. As a consequence, some of the differences between the tasks ended up not being significant in the clipped ReLU case. Error bars represent mean ∓ standard errors across different hyperparameter settings. Exact sample sizes for the derived statistics shown in **d** are reported in Supplementary Table 1. **e, f** Recurrent connection weight profiles (as in Figure 6a-c) in conditions where SI>4.8 and in conditions where SI<3, respectively. The weights were smaller in magnitude in **f**, because most of the low SI networks were trained under strong regularization. Solid lines represent mean weights and shaded regions represent standard deviations of weights. Both means and standard deviations are averages over multiple networks.



Supplementary Figure 9

Changing the amount of input noise.

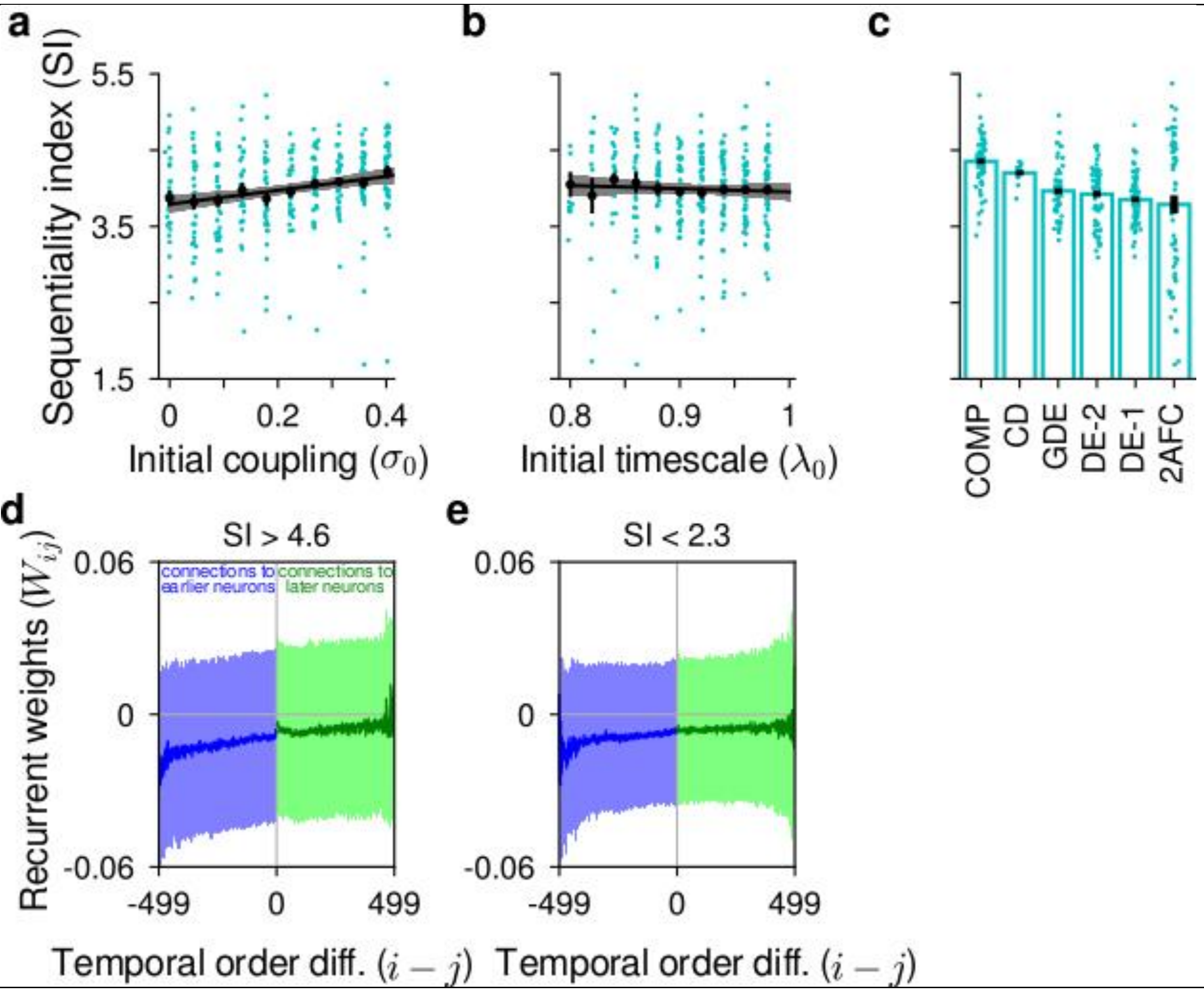
In these simulations, we set ρ=0 and varied the gain of the input population(s), *g*. *g*=1 corresponds to the original case reported in the main text; lower and higher values of *g* correspond to higher and lower amounts of input noise, respectively. **a** Combined across all noise conditions, SI increased significantly with σ_0. Linear regression slope: 0.76∓0.08, R^2=0.04 (two-sided Wald test, *n*=2239 experimental conditions, *p*=0.000). In **a-c**, solid black lines are the linear fits and shaded regions are 95% confidence intervals for the linear regression. **b** λ_0 did not have a significant effect on SI (two-sided Wald test, *n*=2239 experimental conditions, *p*=0.958). **c** The input gain *g* slightly increased the SI. Linear regression slope: 0.04∓0.02, R^2=0.003 (two-sided Wald test, *n*=2239 experimental conditions, *p*=0.003). **d** Again, combined across all input noise levels, the ordering of the tasks by SI was similar to that obtained in the main set of experiments, where *g*=1 (Figure 3a). Error bars represent mean ∓ standard errors across different hyperparameter settings and noise levels. Exact sample sizes for the derived statistics shown in **d** are reported in Supplementary Table 1.



Supplementary Figure 10

Results for the lowest level of input noise (*g* = 2.5).

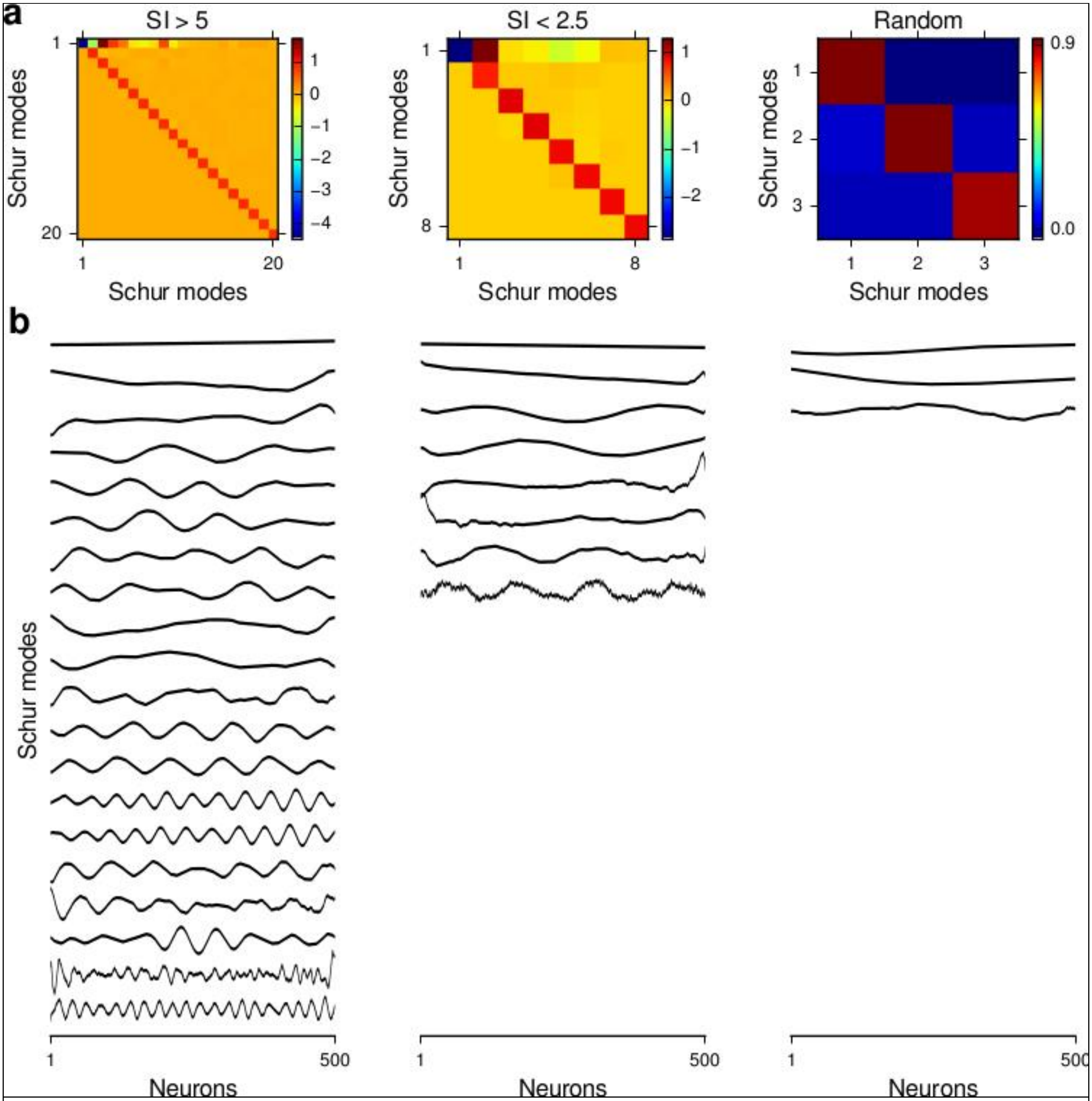
a SI increased significantly with σ$_0$. Linear regression slope: 0.76∓0.18, R^2=0.05 (two-sided Wald test, *n*=365 experimental conditions, *p*=0.000). In **a-b**, solid black lines are the linear fits and shaded regions are 95% confidence intervals for the linear regression. **b** λ$_0$ did not have a significant effect on SI (two-sided Wald test, *n*=365 experimental conditions, *p*=0.253). **c** The ordering of the tasks by SI was similar to that obtained in the main set of experiments. Error bars represent mean ∓ standard errors across different hyperparameter settings. Exact sample sizes for the derived statistics shown in **c** are reported in Supplementary Table 1. **d, e** Recurrent connection weight profiles (as in Figure 6a-c) in conditions where SI>4.9 and in conditions where SI<2.8, respectively. Solid lines represent mean weights and shaded regions represent standard deviations of weights. Both means and standard deviations are averages over multiple networks.



Supplementary Figure 11

Results for the highest level of input noise (g = 0.5).

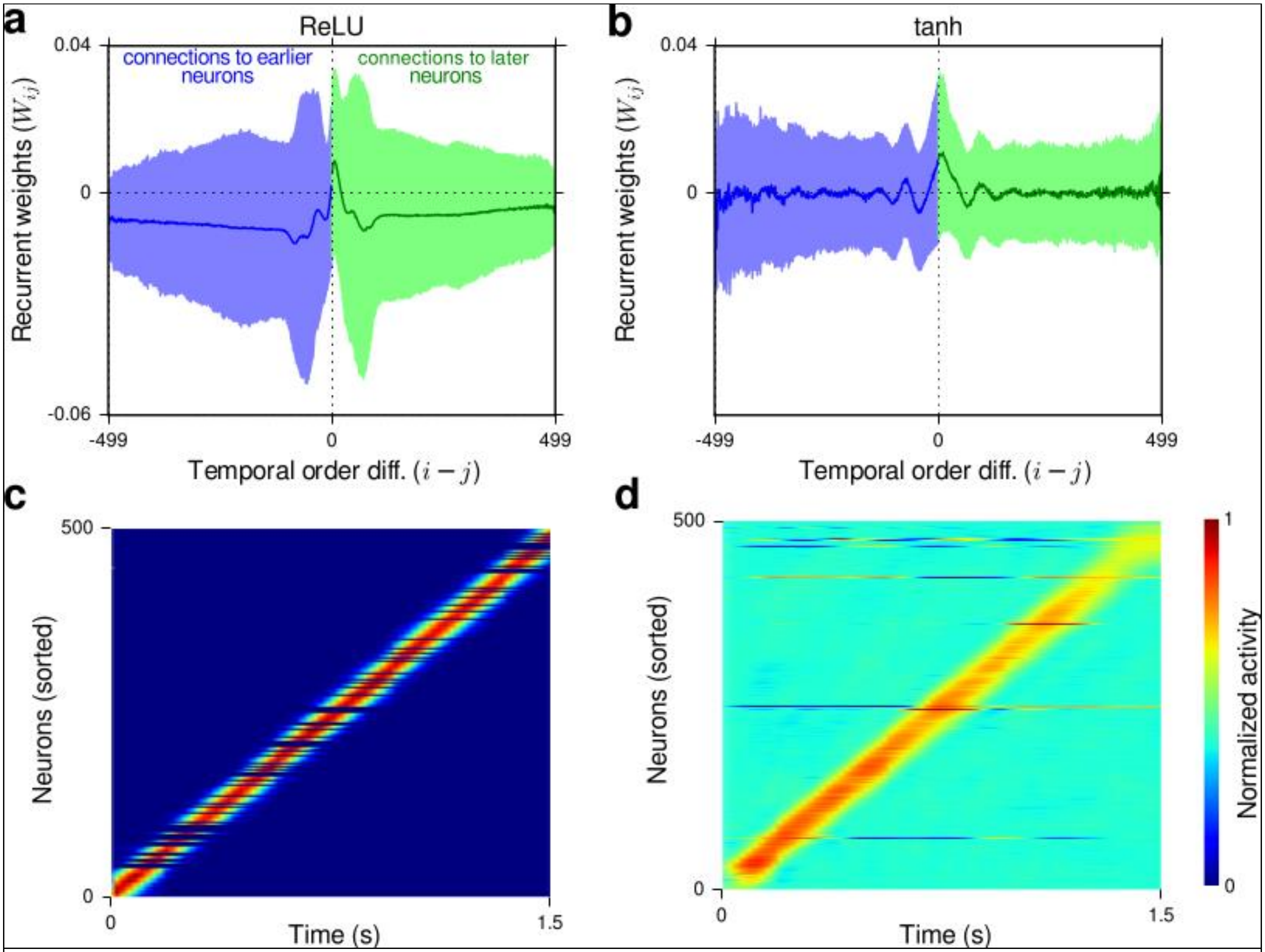
a SI increased significantly with σ_0. Linear regression slope: 0.91∓0.21, R^2=0.05 (two-sided Wald test, n=361 experimental conditions, p=0.000). In **a-b**, solid black lines are the linear fits and shaded regions are 95% confidence intervals for the linear regression. **b** λ_0 did not have a significant effect on SI (two-sided Wald test, n=361 experimental conditions, p=0.457). **c** The ordering of the tasks by SI was similar to that obtained in the main set of experiments. Error bars represent mean ∓ standard errors across different hyperparameter settings. Exact sample sizes for the derived statistics shown in **c** are reported in Supplementary Table 1. **d, e** Recurrent connection weight profiles (as in Figure 6a-c) in conditions where SI>4.6 and in conditions where SI<2.3, respectively. Solid lines represent mean weights and shaded regions represent standard deviations of weights. Both means and standard deviations are averages over multiple networks.



Supplementary Figure 12

Schur decomposition of trained and random-connectivity matrices.

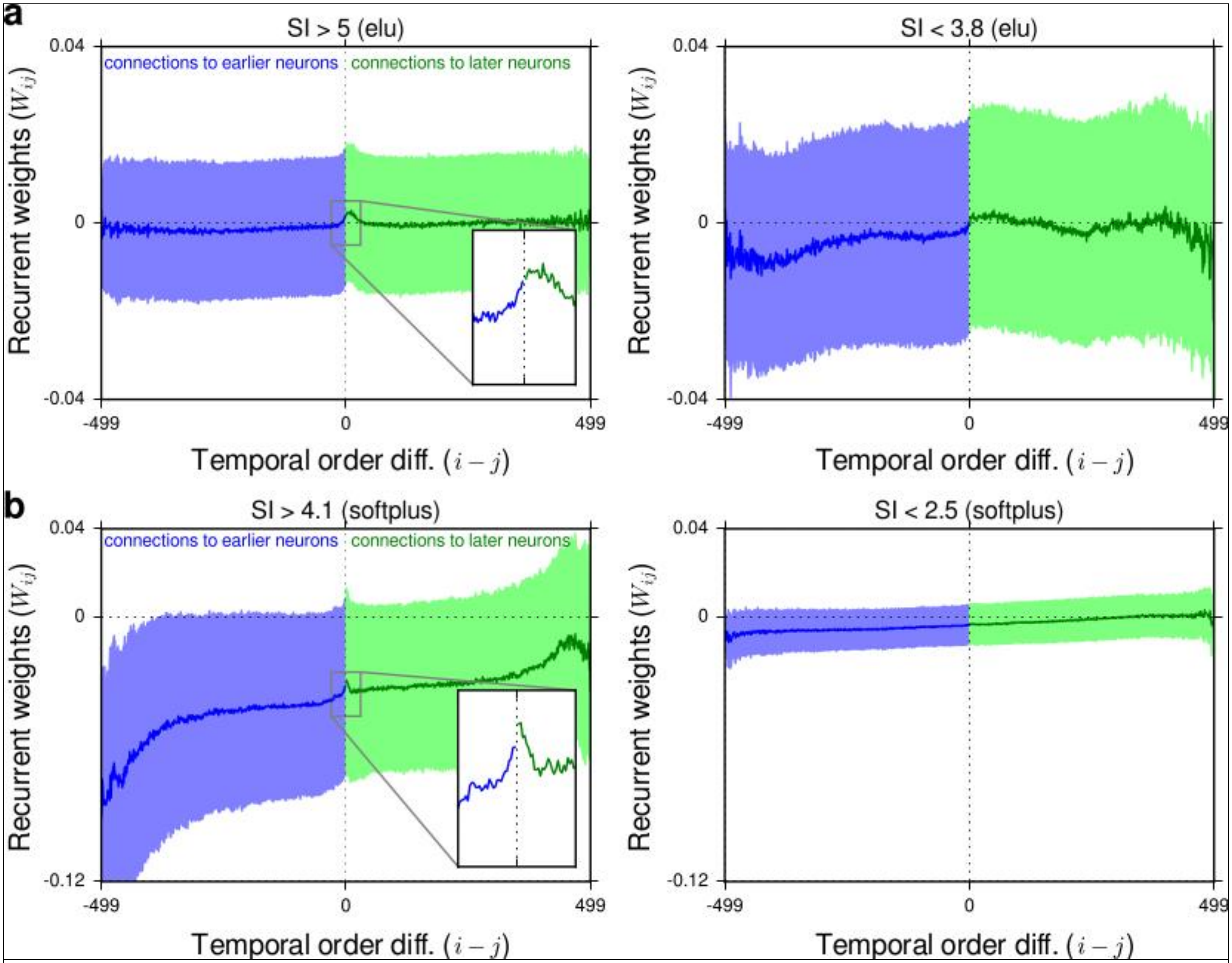
a Schur mode interaction matrices for the mean recurrent connectivity patterns shown in Figure 6a-c. Only significant Schur modes with at least one interaction of magnitude greater than 0.04 with another Schur mode are shown here. **b** The corresponding significant Schur modes. Networks with more sequential activity (SI>5) have more high-frequency Schur modes than networks with less sequential activity (SI<2.5). The random networks are close to normal.



Supplementary Figure 13

Results from networks explicitly trained to generate sequential activity **[AU: We can't support reference citations in titles. Okay to delete the citation to ref. 35 here? Or cite it in the legend.]**

a-b are analogous to Figure 6a-b and show the recurrent weight profiles obtained in trained networks with ReLU and tanh nonlinearities, respectively. **c-d** show example trials for the corresponding networks (trained with the same initial condition). Only networks with sequentiality index larger than 5.45 were included in the results shown here.



Supplementary Figure 14

Circuit mechanism that generates sequential vs. persistent activity in networks with alternative activation functions.

This figure is analogous to Figure 6a-b, but the results shown are for networks with the exponential linear (elu) activation function (**a**) and networks with the softplus activation function (**b**). Note that the elu activation function typically produced larger SIs than softplus, hence slightly different SI thresholds were used in the two cases to determine low and high SI networks.

Figure	Exact sample sizes (*n*)	Statistical tests
Figure 3a	COMP: 590 CD: 252 GDE: 546 DE-2: 537 DE-1: 590 2AFC: 391	None
Figure 3c	*f*=0.25: 27 *f*=1: 19 *f*=2: 8	Reported in Figure 3 legend.
Figure 3d	COMP: 67 (red & cyan each) CD: 27 (each) GDE: 54 (each) DE-2: 48 (each) DE-1: 64 (each) 2AFC: 48 (each)	*p*=0.140, *t*=1.484 *p*=0.137, *t*=-1.511 *p*=0.006, *t*=-2.963 *p*=0.486, *t*=-0.700 *p*=0.000, *t*=-3.866 *p*=0.000, *t*=-3.769 All tests are two-sided Welch.
Figure 3e	Combined: 308 (each)	*p*=0.000, *t*=-4.259 (two-sided Welch)
Figure 4a	COMP: 56 (red & cyan each) CD: 12 (each) GDE: 30 (each) DE-2: 23 (each) DE-1: 44 (each) 2AFC: 46 (each) Combined: 211 (each)	*p*=0.000, *t*=7.349 *p*=0.019, *t*=2.624 *p*=0.084, *t*=1.761 *p*=0.497, *t*=0.684 *p*=0.528, *t*=0.634 *p*=0.986, *t*=0.018 *p*=0.000, *t*=3.589 All tests are two-sided Welch.
Figure 4b	COMP: 76 (red & cyan each) CD: 20 (each) GDE: 57 (each) DE-2: 57 (each) DE-1: 66 (each) 2AFC: 64 (each) Combined: 340 (each)	*p*=0.000, *t*=16.096 *p*=0.000, *t*=20.023 *p*=0.000, *t*=8.825 *p*=0.000, *t*=8.226 *p*=0.000, *t*=8.462 *p*=0.189, *t*=1.322 *p*=0.000, *t*=13.802 All tests are two-sided Welch.
Figure 4c	COMP: 71 (red & cyan each) CD: 16 (each) GDE: 41 (each) DE-2: 35 (each) DE-1: 39 (each)	*p*=0.960, *t*=0.050 *p*=0.427, *t*=-0.805 *p*=0.000, *t*=-6.996 *p*=0.000, *t*=-9.835 *p*=0.000, *t*=-14.816

	2AFC: 43 (each) Combined: 245 (each)	*p*=0.028, *t*=-2.233 *p*=0.000, *t*=9.021 All tests are two-sided Welch.
Figure 5a	COMP->2AFC: 56 (each) 2AFC->COMP: 50 (each)	*p*=0.003, *t*=-3.084 *p*=0.015, *t*=-2.470 All tests are two-sided Welch.
Figure 5b	CD->2AFC: 28 (each) 2AFC->CD: 20 (each)	*p*=0.200, *t*=-1.305 *p*=0.002, *t*=-3.268 All tests are two-sided Welch.
Supplementary Figure S8d	COMP: 151 GDE: 36 DE-2: 48 DE-1: 42 2AFC: 3	None
Supplementary Figure S9d	COMP: 481 CD: 191 GDE: 398 DE-2: 354 DE-1: 419 2AFC: 396	None
Supplementary Figure S10c	COMP: 88 CD: 41 GDE: 63 DE-2: 41 DE-1: 64 2AFC: 68	None
Supplementary Figure S11c	COMP: 74 CD: 12 GDE: 66 DE-2: 71 DE-1: 75 2AFC: 63	None

Supplementary Table 1: Exact sample sizes and test statistics for all derived statistics and statistical tests reported in each figure, including the supplementary figures.