A Appendix

A.1 Training details

We perform Bayesian hyperparameter optimization to obtain 120 candidates on each dataset for subsequent model ensembling. Training was done on RTX 2080 Ti GPUs. The sweep ranges for hyperparameters optimization are shown in Table 1.

Table 1: Training details

	MC_Maze	MC_RTT	Area2_Bump	DMFC_RSG
Dropout ratio	0 – 0.4	0 – 0.4	0 – 0.6	0 – 0.4
Temporal backward context	1 – 100	1 – 100	1 – 100	1 – 240
Temporal forward context	1 – 100	1 – 100	1 – 100	1 – 240
Initial learning rate	1e-5 – 1e-2	1e-4 – 1e-1	1e-5 – 1e-2	1e-5 – 1e-2
Learning rate warmup	0 – 7000	0 – 7000	0 – 7000	0 – 2000
Mask ratio	0 – 0.4	0 – 0.4	0 – 0.6	0 – 0.4
Zero mask ratio	0.5 – 1.0	0.5 – 1.0	0.5 – 1.0	0.5 – 1.0
Random mask ratio	0.3 – 1.0	0.6 – 1.0	0.9 – 1.0	0.9 – 1.0
Training time	∼65 hrs, 6 GPUs	∼71 hrs, 4 GPUs	∼19 hrs, 5 GPUs	∼91 hrs, 4 GPUs
Ensemble size	20	40	50	77

A.2 Model robustness across random initializations

To assess the robustness of STNDT against random initializations, we trained our best STNDT model and best AESMTE model with five different random seeds and report the mean as well as the standard error in Tables 2-5 below. For AESMTE, we used the same public code and the same set of hyperparameters of the best performing model they provided to ensure a fair comparison. All the results are obtained on the hidden test set held by NLB. The results indicate that STNDT maintains a gap over AESMTE and is more robust across initializations. The effect is observed on all four datasets and is most notable on the primary metric co-bps.

A.3 Correlations of evaluation metrics

We show in Figure 1 the correlation between evaluation metrics and validation mask loss obtained at the final training epoch where the best model is checkpointed. The mask loss is still a good objective to guide the training in the early episodes. However, after reaching certain goodness of fit, it is no longer indicative of the model performance as measured by the four metrics. Therefore we chose to optimize co-bps during Bayesian hyperparameter optimization.

A.4 Visualization of temporal and spatial attention maps

STNDT employs two attention modules over the temporal dimension and the spatial dimension. In Results section, we showed the spatial attention weights of two layers for four example trials in MC_Maze dataset. For completeness, we visualize the accompanying temporal attention weights in Figure 2. In addition, we also show spatial attention weights from layer 1 to layer 4 for four example trials of all datasets in Figure 3-6. While temporal attention weights are pretty uniformly distributed and have minimal interpretability, spatial attention weights in the first layer - which correspond directly to physical neurons - delineate a subset of neurons that are heavily attended to by most of other neurons in the population, suggesting these neurons might have an important role in inferring underlying firing rates of the entire population.

A.5 Impacts of heavily-attended neurons to performance of latent variable models

To assess how excluding heavily-attended neurons identified by STNDT's spatial attention affects STNDT's performance and whether that effect generalizes to other modeling methods, we show in Figures 7- 10 the performance of STNDT and other modeling methods (NDT [1], Smoothing [2], GPFA [3]) as neurons are dropped from the spike train input. For NDT, Smoothing and GPFA methods, we use the optimal hyperparameters reported in [4] and [2]. In general, performance of all

Table 2: Performance (mean±SEM) of STNDT with and without contrastive loss (CL) across 5 random seeds on MC_Maze dataset.

	MC_Maze			
Methods	co-bps↑	vel R^2↑	psth R^2↑	fp-bps↑
AESMTE1 (single)	0.3476 ± 0.0035	0.9057 ± 0.0006	0.6320 ± 0.0071	0.2365 ± 0.0031
STNDT single w/o CL	$\mathbf{0.3659} \pm 0.0003$	0.8937 ± 0.0013	$\mathbf{0.6562} \pm 0.0029$	$\mathbf{0.2446} \pm 0.0014$
STNDT single w/ CL	$\mathbf{0.3668} \pm 0.0005$	0.8932 ± 0.0012	$\mathbf{0.6534} \pm 0.0046$	$\mathbf{0.2447} \pm 0.0009$

Table 3: Performance (mean±SEM) of STNDT with and without contrastive loss (CL) across 5 random seeds on MC_RTT dataset.

	MC_RTT		
Methods	co-bps↑	vel R^2↑	fp-bps↑
AESMTE1 (single)	0.1729 ± 0.0090	0.5847 ± 0.0618	0.0974 ± 0.0044
STNDT single w/o CL	$\mathbf{0.1883} \pm 0.0019$	$\mathbf{0.6021} \pm 0.0051$	0.0958 ± 0.0039
STNDT single w/ CL	$\mathbf{0.1923} \pm 0.0009$	$\mathbf{0.5996} \pm 0.0060$	0.0932 ± 0.0030

Table 4: Performance (mean±SEM) of STNDT with and without contrastive loss (CL) across 5 random seeds on Area2_Bump dataset.

	Area2_Bump			
Methods	co-bps↑	vel R^2↑	psth R^2↑	fp-bps↑
AESMTE1 (single)	0.2483 ± 0.0096	0.8370 ± 0.0175	0.5628 ± 0.0423	0.1261 ± 0.0080
STNDT single w/o CL	$\mathbf{0.2717} \pm 0.0011$	$\mathbf{0.8730} \pm 0.0048$	$\mathbf{0.7145} \pm 0.0029$	$\mathbf{0.1435} \pm 0.0019$
STNDT single w/ CL	$\mathbf{0.2738} \pm 0.0009$	$\mathbf{0.8720} \pm 0.0020$	$\mathbf{0.7098} \pm 0.0038$	$\mathbf{0.1477} \pm 0.0025$

Table 5: Performance (mean±SEM) of STNDT with and without contrastive loss (CL) across 5 random seeds on DMFC_RSG dataset.

	DMFC_RSG			
Methods	co-bps↑	tp-corr ↓	psth R^2↑	fp-bps↑
AESMTE1 (single)	0.1795 ± 0.0008	-0.7297 ± 0.0104	0.5584 ± 0.0207	0.1597 ± 0.0041
STNDT single w/o CL	$\mathbf{0.1820} \pm 0.0011$	-0.5210 ± 0.0435	$\mathbf{0.6080} \pm 0.0015$	0.1429 ± 0.0059
STNDT single w/ CL	$\mathbf{0.1840} \pm 0.0008$	-0.5148 ± 0.0408	$\mathbf{0.6097} \pm 0.0071$	0.1444 ± 0.0095

modeling methods declines when more neurons are dropped from the population inputs. However, when heavily-attended neurons (important neurons) that were identified by STNDT's spatial attention module are dropped, the performance deteriorates more significantly compared to when the same number of random neurons are dropped. This is most conspicuous in MC_Maze and Area2_Bump datasets. This gap can be observed in all four modeling methods but is wider in the cases of STNDT and NDT as compared to Smoothing and GPFA.

References

[1] Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with neural data transformers. *arXiv preprint arXiv:2108.01210*, 2021.

[2] Felix Pei, Joel Ye, David Zoltowski, Anqi Wu, Raeed H Chowdhury, Hansem Sohn, Joseph E O'Doherty, Krishna V Shenoy, Matthew T Kaufman, Mark Churchland, et al. Neural latents benchmark'21: Evaluating latent variable models of neural population activity. *arXiv preprint arXiv:2109.04463*, 2021.

[3] Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Advances in neural information processing systems*, 21, 2008.

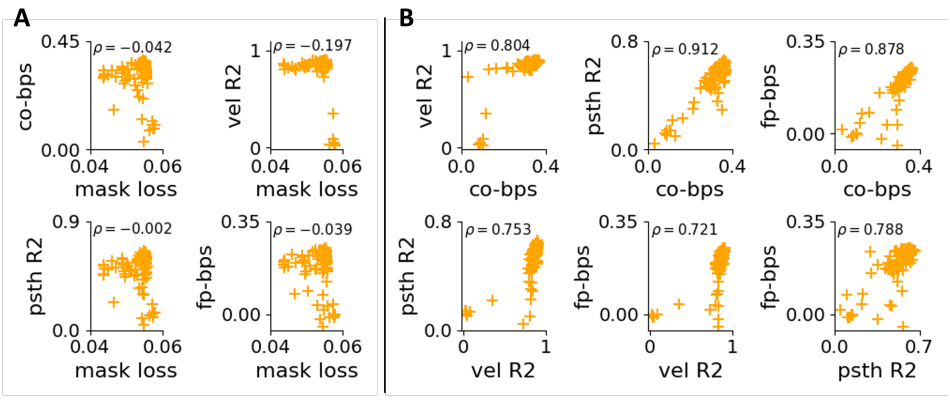

Figure 1: Correlations of evaluation metrics. **A**: Four evaluation metrics of 120 models obtained from Bayesian hyperparameter optimization on MC_Maze dataset are plotted against mask loss. The metrics evaluated at the end of the training do not correlate well with mask loss. **B**: The four metrics are more correlated with each other, therefore we opted for co-bps as the objective for Bayesian hyperparameter optimization.

[4] Darin Sleiter, Joshua Schoenfield, and Mike Vaiana. ae-nlb-2021. `https://github.com/agencyenterprise/ae-nlb-2021.git`, 2021.

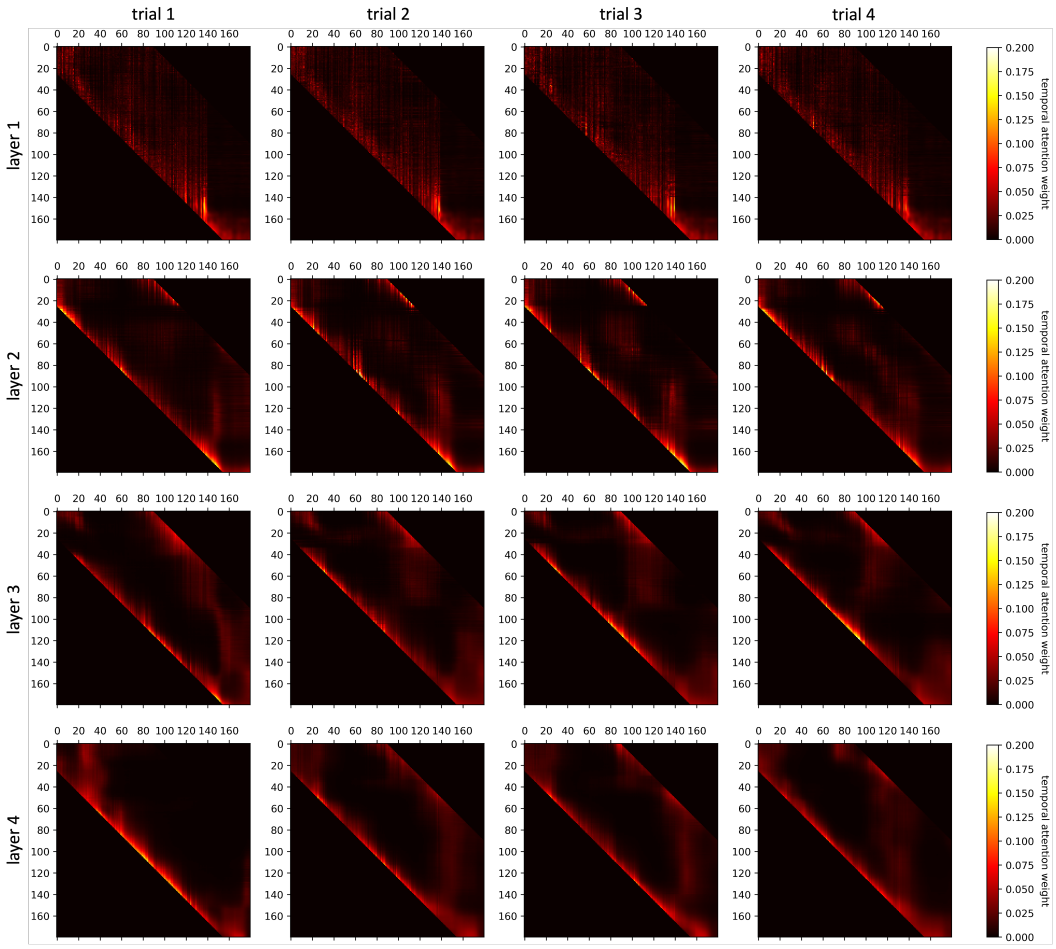

Figure 2: Temporal attention weights across four attention layers of four example trials in MC_Maze dataset.

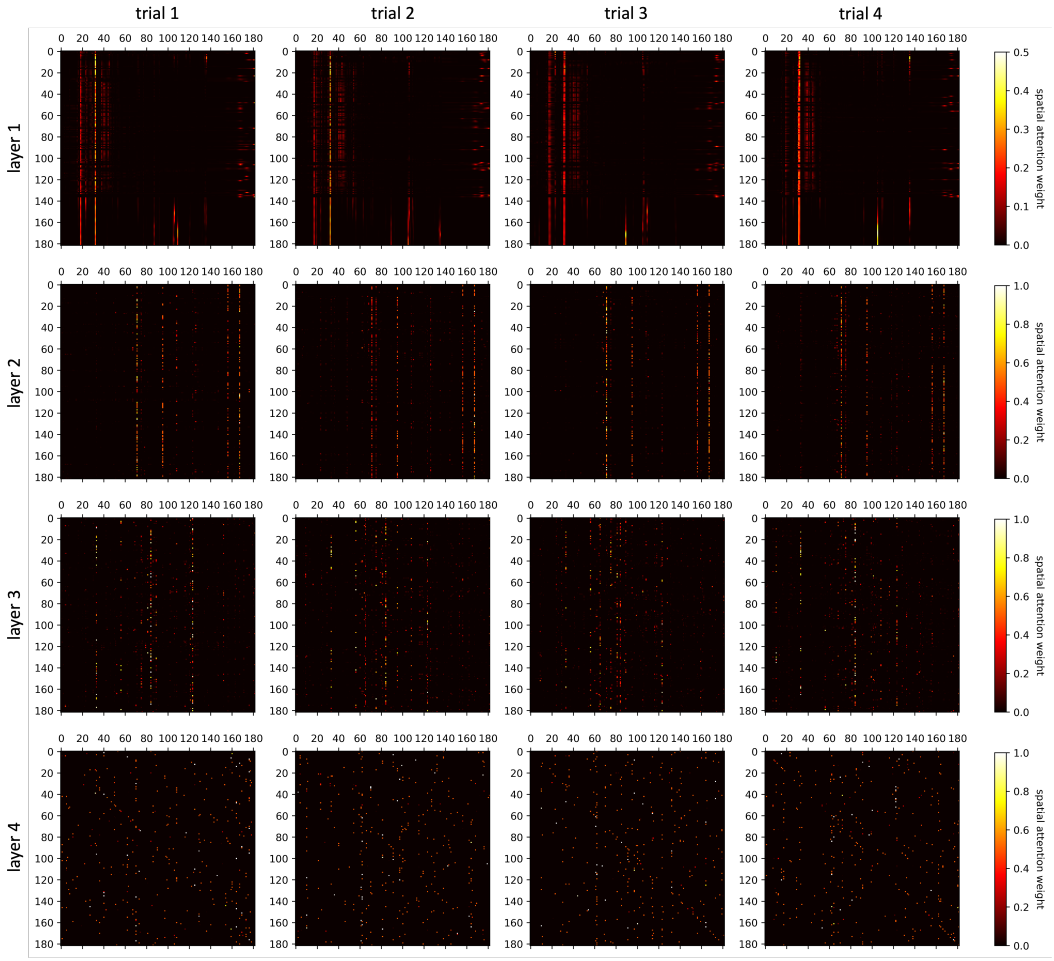

Figure 3: Spatial attention weights across four layers of STNDT on four example trials in MC_Maze dataset.

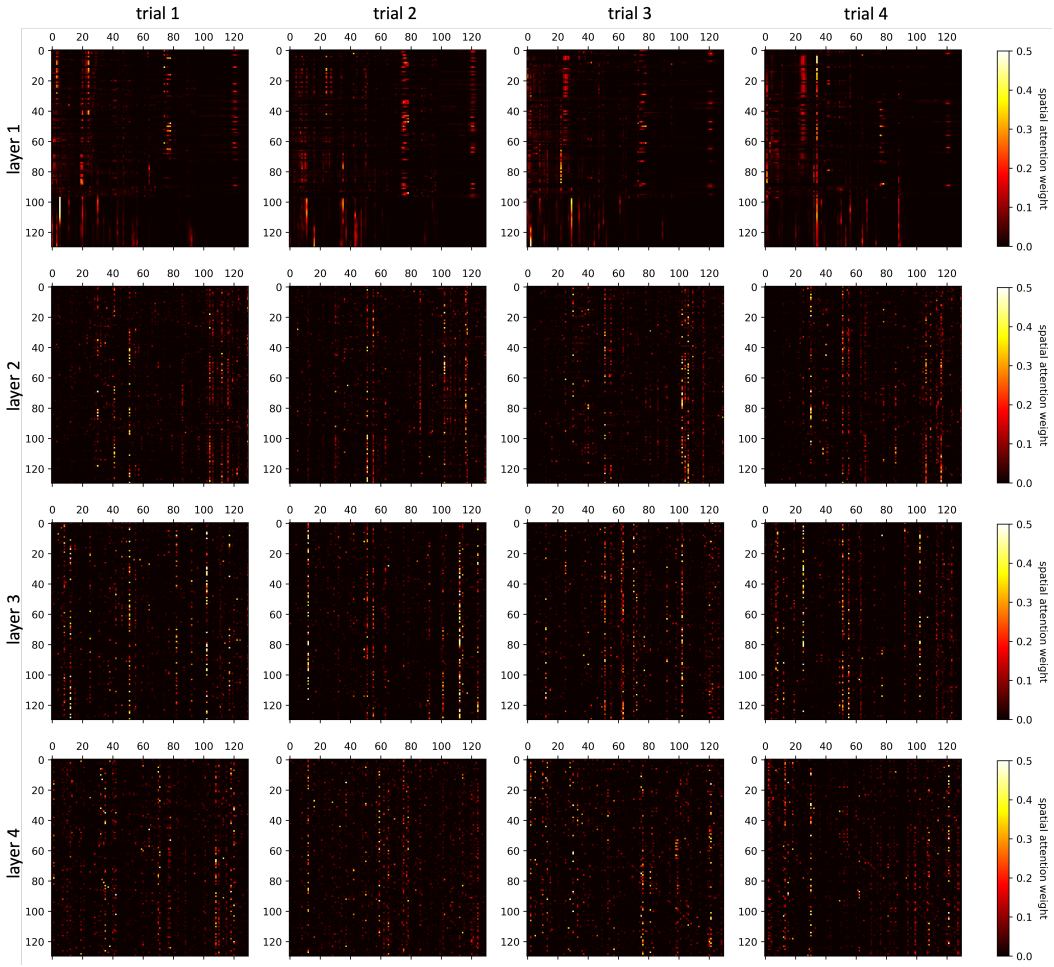

Figure 4: Spatial attention weights across four layers of STNDT on four example trials in MC_RTT dataset.

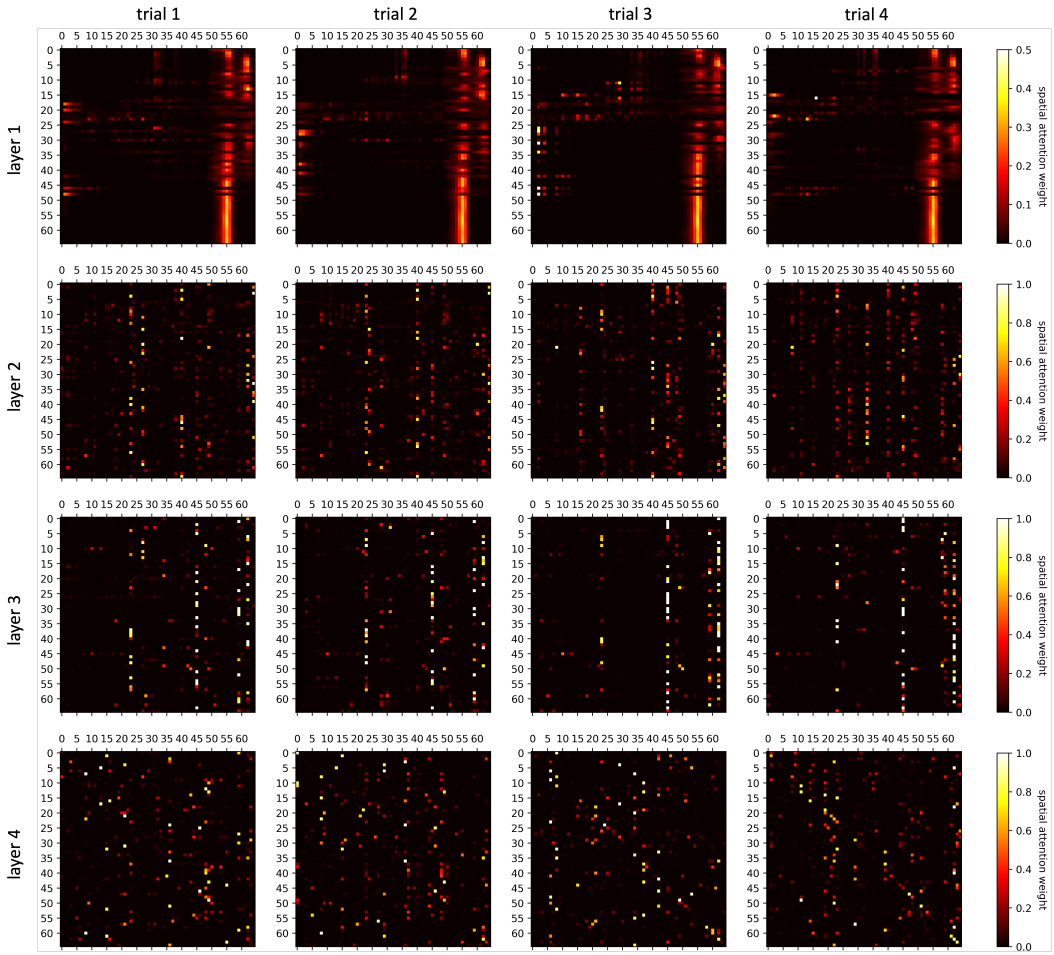

Figure 5: Spatial attention weights across four layers of STNDT on four example trials in Area2_Bump dataset.

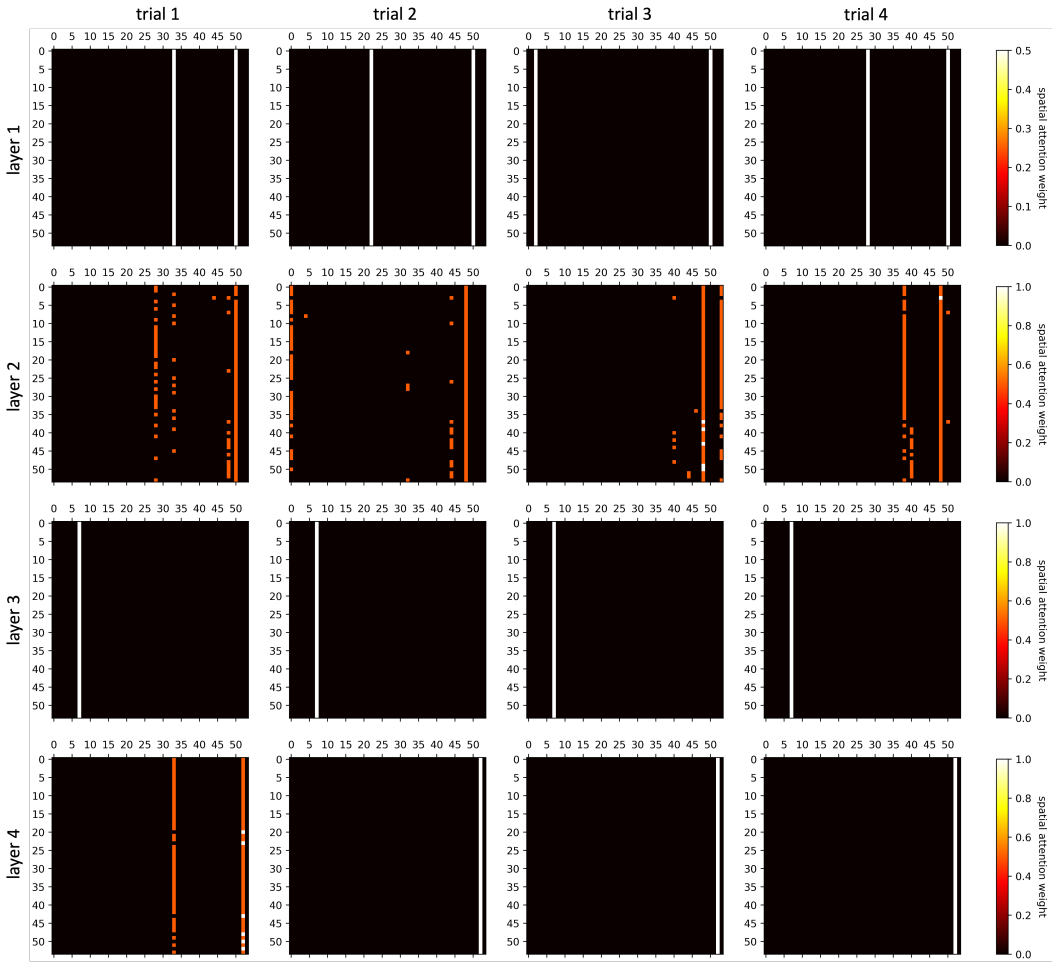

Figure 6: Spatial attention weights across four layers of STNDT on four example trials in DMFC_RSG dataset.

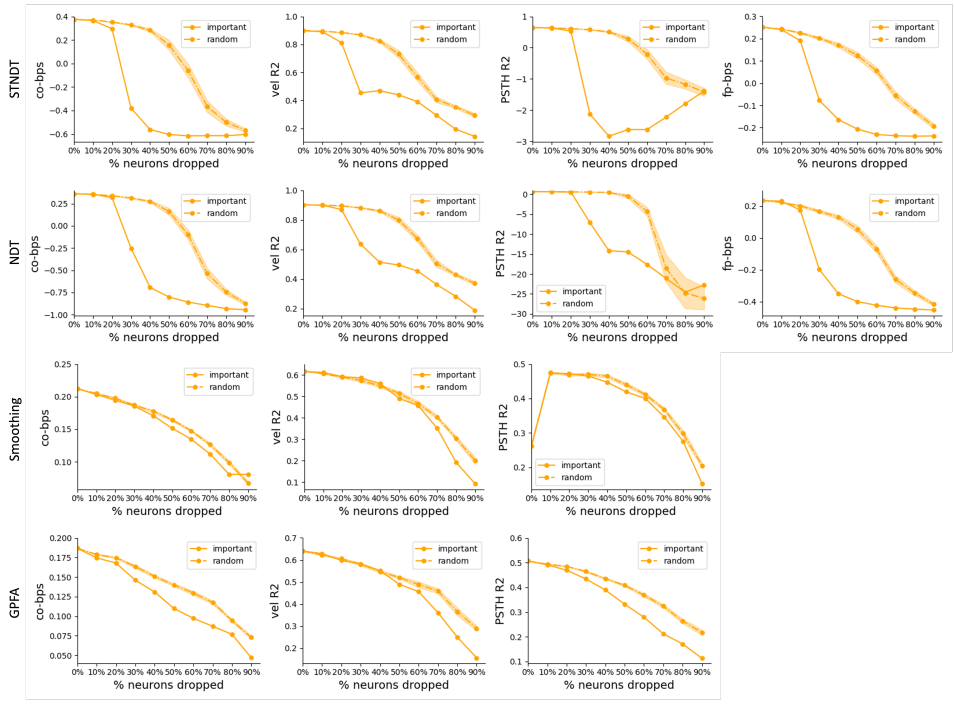

Figure 7: Performance of STNDT, NDT, Smoothing and GPFA models as neurons are dropped randomly vs in descending order of the average attention weights identified by STNDT's spatial attention. Shaded region represents 2 standard error of the mean. Results shown for MC_Maze dataset.

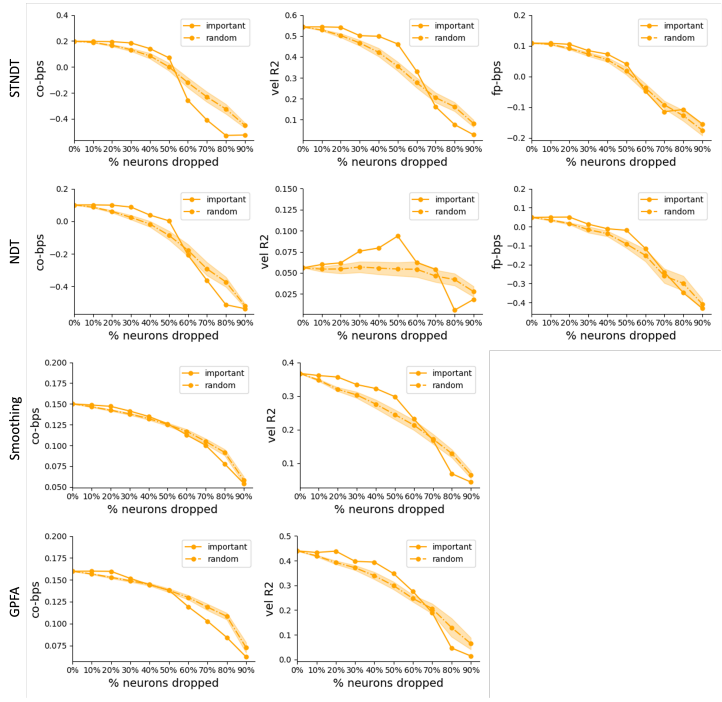

Figure 8: Performance of STNDT, NDT, Smoothing and GPFA models as neurons are dropped randomly vs in descending order of the average attention weights identified by STNDT's spatial attention. Shaded region represents 2 standard error of the mean. Results shown for MC_RTT dataset.

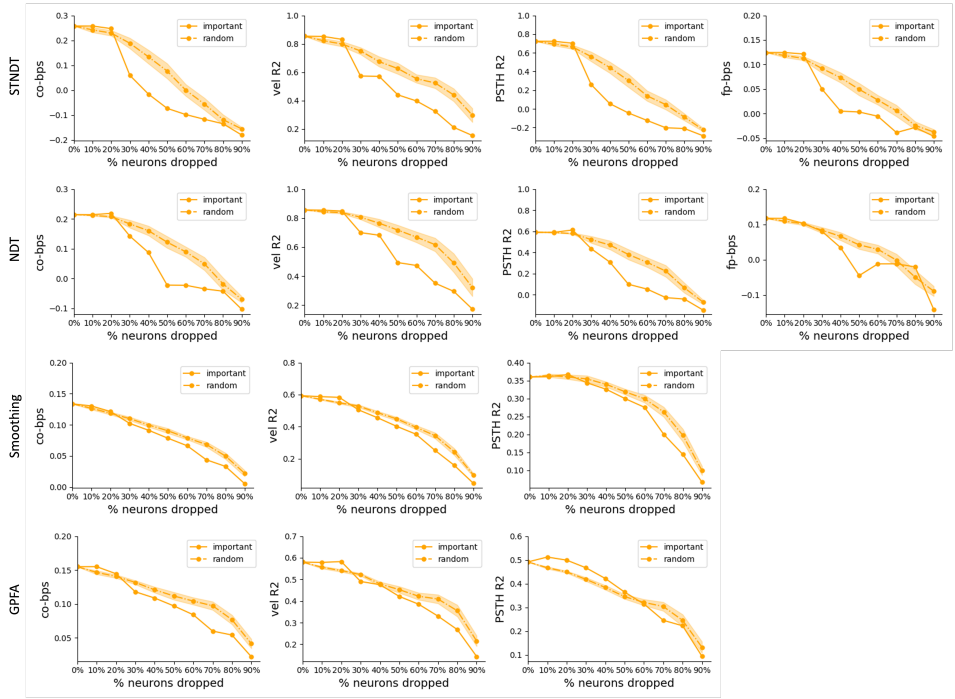

Figure 9: Performance of STNDT, NDT, Smoothing and GPFA models as neurons are dropped randomly vs in descending order of the average attention weights identified by STNDT's spatial attention. Shaded region represents 2 standard error of the mean. Results shown for Area2_Bump dataset.

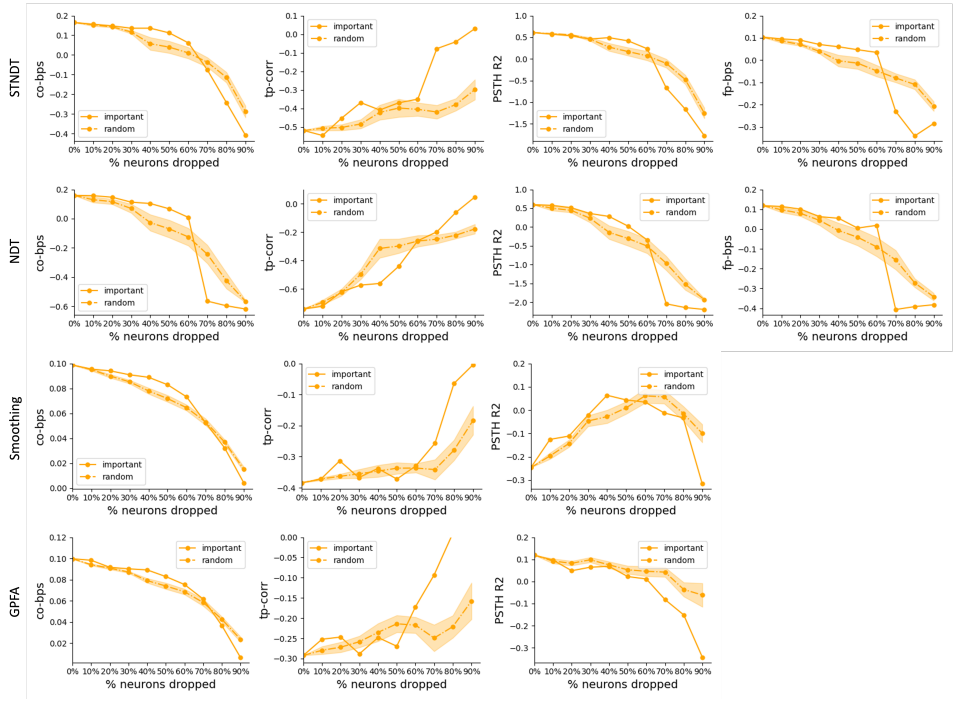

Figure 10: Performance of STNDT, NDT, Smoothing and GPFA models as neurons are dropped randomly vs in descending order of the average attention weights identified by STNDT's spatial attention. Shaded region represents 2 standard error of the mean. Results shown for DMFC_RSG dataset.