We thank the reviewers for their time and the many positive comments, in particular their appreciation of the novelty, originality, and clarity of our work. As a result of the thoughtful reviews, we have strengthened our work by 1) Re-running all quantification on training data in addition to validation data, showing that M-PHATE can evaluate performance without validation data; 2) Showing via running M-PHATE on white noise that generalization results are not purely due to MNIST being predominantly zeros; 3) Defining the results of the task-switch experiment quantitatively; and 4) Repeating experiments on CIFAR10, showing that our results are robust. Specific points below will be incorporated into the manuscript.

Validation set (R1): The new runs on training data show that our method identifies generalizability without requiring access to validation data. As opposed to differences in loss/accuracy, the dynamics of units revealed through M-PHATE retain key geometric features of the network when evaluated with both training and validation data. Thus M-PHATE is applicable even when a validation loss cannot be computed. We applied this to the generalization experiment using only training data and obtained $\rho = 0.93$, consistent with prior results.

Other datasets (R1,R2,R3): We added a generalization experiment on both CIFAR10, ($\rho = 0.95$; consistent with MNIST results) and white noise (finding homogeneous structure similar to scrambled classes, see Fig. 1b).

Quantitative evaluation of task-switch (R1): We added a metric to measure loss of structure by computing Adjusted Rand Index averaged over clusterings of the visualized units pre- and post-task switch (4 task switches $\times$ 6 parameters (3–8 clusters) $\times$ 20 repetitions). High structure preservation is strongly associated with low validation loss ($\rho = 0.90$).

Weaknesses and future directions (R2,R3): *Other architectures*: Extending M-PHATE to CNNs/RNNs is a very interesting future direction. Defining appropriate similarity measures to enable comparison of units across layers for these networks will require non-trivial extensions beyond the scope of this work. Also, using M-PHATE in its current form on such networks will be slow due to the $O(n^2)$ complexity. We plan to explore such extensions, both in designing appropriate metrics and developing computationally efficient models (e.g., online computation). *Overcrowding*: There is always information loss in low-dimensional visualizations of all units at all time points, especially in large networks over many epochs. Despite this, we show in specific tasks that local structure of the units can still be informative.

Higher dimensions (R3): We will add supplementary figures and quantitative measures for 2D vs 3D embeddings. Beyond 3D, dimensionality reduction methods become difficult to visually interpret (the initial intended use).

User studies (R2): We agree feedback is invaluable and will publish our code to get such feedback from the community.

Kernel details (R3): The "standard" kernel differs for each algorithm: t-SNE uses the perplexity kernel, ISOMAP the k-NN kernel, and PHATE and DM use the adaptive bandwidth Gaussian kernel (defined in the paper).

t-SNE/neighborhood preservation (R2, R3): It is well known that t-SNE does not preserve global structure, and that PHATE does [Moon et al. 2017, Linderman and Steinerberger 2019]. Here, preserving interslice vs. intraslice neighbors has an inherent trade-off. t-SNE guarantees almost perfect interslice neighborhood preservation by clustering each hidden unit with only itself, at the cost of losing all of the intraslice neighbors. We further show t-SNE's lack of utility in Figure 1a, in which you see no useful difference in the embeddings between the dropout and scrambled networks from the generalization experiment; M-PHATE showed significant, obvious, and interpretable differences. t-SNE (as all other methods) was run with default parameters in scikit-learn.

Visualization quality metrics (R3): Since standard algorithms do not provide any useful visualization of the data (see Fig 1a as an example), an algorithm that faithfully represents the structure of the raw data (e.g. as measured by coranking) is not expected to perform well. We will extend the notion of coranking to interslice and intraslice neighborhoods and report the result of this to give a notion of global structure in the multislice context.

Neighboring unit interactions (R3): Kernel construction over all node similarities and ignoring *a priori* knowledge of the data yields standard PHATE (paper, Figure 2), which is a poor visualization of the data. In addition, the construction of the multislice kernel such that $K((\tau, i), (\nu, j)) = 0$ for $\tau \neq \nu, i \neq j$ is standard in prior work in multislice graph construction (Mucha et al., 2010).

Deeper insights/conclusions from M-PHATE (R1,R2): The collapse in Adam is likely due to the extremely large gradients causing a rapid build-up of momentum and pushing all units in the same direction much farther than their current spread. The lack of heterogeneity of the activity regularizers is a direct consequence of their formulation: forcing all activities to be small induces similar activations. For continual learning, in appendix Fig. S4, S5 we show separate plots for the layers of the task switch network. These visualizations depict that the significant changes to the network due to the task switch are happening in layer 2, which seems to undergo rapid learning after each task change and then stays static. Layer 1 is highly heterogeneous, indicating a more universal representation which is evolving more smoothly throughout training when compared to layer 2. These layer-specific insights are not accessible from plotting global validation metrics.

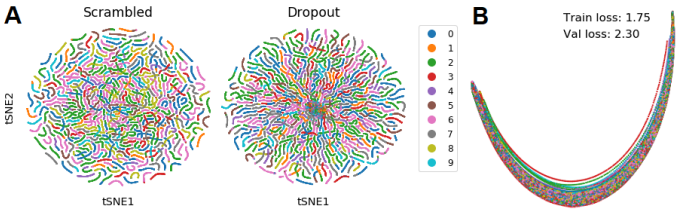

Figure 1: A) tSNE on FFNNs with (left) scrambled labels (right) dropout; B) M-PHATE on an FFNN trained on white noise.