Learning to cluster neuronal function

Nina S. Nellen (ninasophie.nellen@stud.uni-goettingen.de)
Institute of Computer Science and Campus Institute Data Science, University of Göttingen
Goldschmidtstr. 1, 37077 Göttingen, Germany

Polina Turishcheva (turishcheva@cs.uni-goettingen.de)
Institute of Computer Science and Campus Institute Data Science, University of Göttingen
Goldschmidtstr. 1, 37077 Göttingen, Germany

Alexander Ecker (ecker@cs.uni-goettingen.de)
Institute of Computer Science and Campus Institute Data Science, University of Göttingen
Goldschmidtstr. 1, 37077 Göttingen, Germany
Max Planck Institute for Dynamics and Self-Organization
Am Faßberg 17, 37077 Göttingen, Germany.

Abstract

Deep predictive models have recently shown great potential to create digital twins to predict neuronal activity in the visual cortex. These models provide per-neuron embeddings, which have been proposed as a basis to identify functional cell types. However, so far no clear clusters have been observed in the mouse visual cortex and the structure of the embedding space is not highly reproducible across independent model fits. To address these problems, we build upon state-of-the-art predictive networks and introduce an explicit inductive bias to enhance cluster separability. If functional cell types exist, such a clustering bias should improve model performance and consistency of clustering. Our approach is based on training a predictive model and adding an auxiliary loss function that encourages the per-neuron embeddings to be distributed according to a t mixture model. We jointly optimize both neuronal feature embeddings and clustering parameters. Our approach improves consistency of clusters and therefore leads to more consistent embedding spaces across models.

Keywords: neuronal response modeling, visual cortex, clustering, cell types

Introduction

Understanding whether neurons form discrete types or lie on a continuum is a fundamental question in neuroscience (Zeng, 2022). While discrete anatomical and transcriptomic classifications have been proposed (DeFelipe et al., 2013; Oberlaender et al., 2012; Markram et al., 2015), recent work in the mouse brain suggests a more continuous organization (Scala et al., 2019; Weis et al., 2025). Functional types are well established in the retina (Baden et al., 2016) but remain unclear in the visual cortex. Recent deep learning models improve neural activity prediction (Willeke et al., 2022; Turishcheva, Fahey, et al., 2024) and provide data-driven representations of neuronal function via per-neuron embeddings. These embeddings have been used in unsupervised clustering to map the functional organization (Ustyuzhaninov et al., 2022; Turishcheva, Burg, et al., 2024), yet results have been inconclusive: clear clusters rarely emerge, particularly among excitatory neurons in the mouse visual cortex. This raises the question: Do discrete functional types exist, or is neuronal diversity more continuous (F. Xie et al., 2025)? Here we explicitly encourage clustered structure in learned embeddings to test for functional cell type separability. We treat this as a form of model-driven hypothesis testing: if distinct types exist, a clustering bias should improve model performance, embedding structure, and/or cluster consistency. Inspired by Deep Embedding Clustering (DEC; J. Xie et al., 2016), we propose a modified loss that promotes clustering of neuronal embeddings during training of deep predictive models.

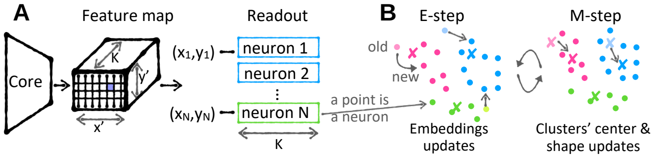

Figure 1: **A: Model architecture:** The model consists of a shared convolutional core outputting a (height × width × feature channels K) dimensional feature map and neuron-specific readouts, each defined by a receptive field position and a weight vector. We refer to the weight vector as neuronal embedding. The response is computed by taking the feature at the RF position and applying a dot product with the weight vector. **B: Clustering procedure:** We add a loss to incorporate a clustering bias into the weight vector and update clustering parameters (centers and scale matrices) using an EM step of a t mixture model, as shown in Algorithm 1.

Methods

Predictive model for visual cortex. We extend a state-of-the-art neural predictive model of neuronal responses to natural images (Willeke et al., 2022). It consists of a shared convolutional core and neuron-specific Gaussian readout (Figure 1A). The core extracts nonlinear features shared across neurons (Klindt et al., 2017), while each neuron-specific readout selects (x, y) location of the neuron's receptive field (RF) and computes the dot product between the neuron's weight vector (per-neuron embedding) and the feature map at that location to obtain the predicted neuronal response (Lurz et al., 2021).

Clustering loss. We want to cluster the N per-neuron feature embeddings $z_i \in \mathbb{R}^K$ into J distinct clusters to improve functional cell type separation. To enforce this structure, we adapt DEC (J. Xie et al., 2016) and introduce a clustering loss during training, which is the Kullback-Leibler (KL) divergence $KL(Q||P)$ between the soft cluster assignments Q under a t mixture model and a sharpened target distribution P.

EM-based cluster updates. We train the mixture of t model simultaneously with the core and readout of the predictive model, using Expectation Maximization (EM). We use the shape-rate form of the t-distribution (McLachlan & Peel, 2000), which represents it as a scale mixture of Gaussians, whose (latent) scaling factor u follows a Gamma distribution. We iterate over three steps (Algorithm 1): (1) E-Step: Compute soft cluster assignment probabilities q_{ij} and the expectation of u_{ij}. (2). M-Step: Update cluster means μ_j and (diagonal) scale matrices Σ_j. (3) Update the parameters of core and readout via one iteration of stochastic gradient descent.

The probability density of the multivariate t-distribution is:

$$f_t(z_i;\mu_j,\Sigma_j,\nu) = \frac{\Gamma\left(\frac{\nu+K}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{K}{2}}\pi^{\frac{K}{2}}|\Sigma_j|^{\frac{1}{2}}}\left(1+\frac{1}{\nu}\delta_{ij}\right)^{-\frac{\nu+K}{2}}$$

where $\delta_{ij} = (z_i-\mu_j)^{\mathrm{T}}\Sigma_j^{-1}(z_i-\mu_j)$. We fix the degrees of freedom to $\nu = 2.1$ to ensure the variance is defined ($\nu > 2$), and use diagonal Σ_j to balance flexibility and overfitting.

```
Algorithm 1 Model Training with clustering loss
  Inputs: Degrees of freedom $\nu$, clustering weight $\beta$, core parameters $\theta$, neuronal embeddings (readout) $Z$
  Output: Parameters $\mu_j, \Sigma_j$, $\theta$ and $Z$
  Pretraining: Train the predictive model by optimizing $L_{\text{model}}$ w.r.t. $\theta$ and $Z$ for $m$ epochs
  Initialize: Cluster centers $\mu_j$ with $k$-means and diagonal scale matrix $\Sigma_j$ as within-cluster variance
  for epoch $t = 1$ to $T$ do
      for minibatch $b$ in dataset do
          (1) E-step (Expectation):
              1.1 Soft assignments $q_{ij} = \frac{f_t(z_i;\mu_j,\Sigma_j,\nu)}{\sum_{j'=1}^{J} f_t(z_i;\mu_{j'},\Sigma_{j'},\nu)}$
              1.2 Latent scales $u_{ij} = \frac{\nu+K}{\nu+(z_i-\mu_j)'\Sigma_j^{-1}(z_i-\mu_j)}$
          (2) M-step (Maximization): Update parameters
              2.1 Update $\mu_j = \frac{\sum_{i=1}^{N} q_{ij}u_{ij}z_i}{\sum_{i=1}^{N} q_{ij}u_{ij}}$
              2.2 Update $\Sigma_j = \frac{\sum_{i=1}^{N} q_{ij}u_{ij}(z_i-\mu_j)(z_i-\mu_j)'}{\sum_{i=1}^{N} q_{ij}}$
          (3) Optimize predictive model parameters
              3.1 Minimize $L = L_{\text{model}} + \beta KL(Q||P)$ w.r.t $\theta, Z$
              with $p_{ij} = \frac{q_{ij}^2/f_j}{\sum_k q_{ik}/f_k}$ and $f_j = \sum_i q_{ij}$
  return $\mu, \Sigma, \theta, Z$
```

Cluster initialization and loss. We pretrain the predictive model for m epochs (pretrain epochs PE) following Willeke et al. (2022) and Turishcheva, Burg, et al. (2024) before turning on the KL loss. After pretraining, we initialize cluster centers μ_j with k-means (MacQueen, 1967), and scale matrix Σ_j as the within-cluster variances. We scale the KL term with β to match the magnitude of the model loss.

Evaluation of embedding consistency. The number of excitatory cell types in mouse visual cortex remains unclear, with estimates ranging from 20 to 50 (Gouwens et al., 2019; Ustyuzhaninov et al., 2022) and some studies finding a high degree of continuous variation (Scala et al., 2019; Weis et al., 2025). To evaluate clustering structure, we compute the Adjusted Rand Index (ARI; Hubert & Arabie, 1985) between model runs for cluster counts between 5 and 60, measuring how consistently pairs of neurons are grouped across runs.

Training data. The model was trained on SENSORIUM 2022 (Willeke et al., 2022) responses to natural images of seven mice (54569 neurons total) using behavioral variables as well.

Model evaluation. Following prior work, we assess model performance using the Pearson correlation between predicted and observed neuronal responses averaged across neurons (Vintch et al., 2015; Sinz et al., 2018; Willeke et al., 2022).

Visualization. We visualize embeddings using t-SNE (Maaten & Hinton, 2008), with perplexity $N/100$, learning rate 1, and early exaggeration $N/10$, as in Linderman & Steinerberger (2019). We sample 2,000 neurons per mouse from seven mice, using the same subset across visualizations.

Results

Clustering loss improves consistency without revealing a clear number of clusters. Our method (Algorithm 1), achieves higher embedding consistency (ARI) than the baseline model (Figure 2), matching the previous state of the art (rotation-equivariant model; Turishcheva, Burg, et al., 2024) but with better predictive performance. The optimal schedule is pretraining for 10 epochs before turning on the clustering loss. To assess the impact of this loss, we varied its weight β while keeping other factors fixed (e. g., number of clusters, length of pretraining), tuning only the learning rate per β. Cluster consistency remains stable across β weights (Figure 3E). However, predictive performance drops as β increases (Figure 3D), suggesting a rather continuous variation of neuronal function.

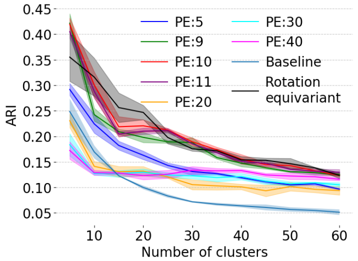

Figure 2: Clustering loss improves cluster consistency (ARI) across model fits for different numbers of clusters depending on length of pretraining (PE). Learning rate tuned for each model. $\beta = 10^5$.

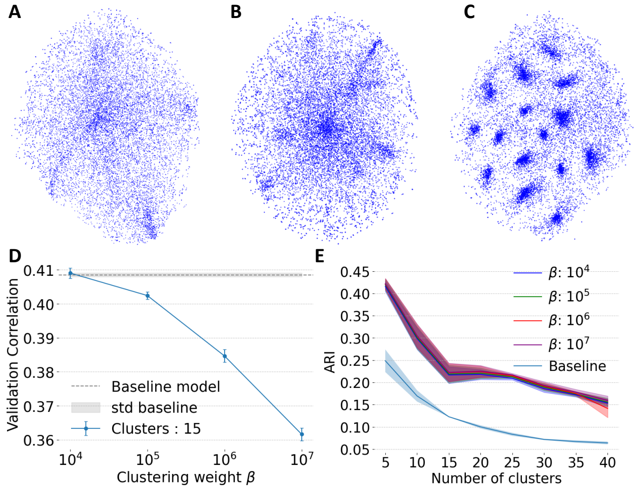

Figure 3: **A** *t*-SNE of baseline model. **B-C** *t*-SNE of our model with 15 clusters (PE = 10). **B.** Clustering weight $\beta = 10^5$. **C.** $\beta = 10^7$. **D.** Model performance for different β. **E.** ARI for different number of clusters and β.

Acknowledgments

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 101041669). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. We thanks Rasmus Steinkamp for technical support.

References

Baden, T., Berens, P., Franke, K., Román Rosón, M., Bethge, M., & Euler, T. (2016). The functional diversity of retinal ganglion cells in the mouse. *Nature*, *529*(7586), 345–350.

DeFelipe, J., López-Cruz, P. L., Benavides-Piccione, R., Bielza, C., Larrañaga, P., Anderson, S., ... others (2013). New insights into the classification and nomenclature of cortical gabaergic interneurons. *Nature Reviews Neuroscience*, *14*(3), 202–216.

Gouwens, N. W., Sorensen, S. A., Berg, J., Lee, C., Jarsky, T., Ting, J., ... others (2019). Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nature neuroscience*, *22*(7), 1182–1195.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*, 193–218.

Klindt, D., Ecker, A. S., Euler, T., & Bethge, M. (2017). *Neural system identification for large populations separating "what"and "where"* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8c249675aea6c3cbd91661bbae767ff1-Paper.pdf

Linderman, G. C., & Steinerberger, S. (2019). *Clustering with t-sne, provably* (Vol. 1) (No. 2). Retrieved from https://doi.org/10.1137/18M1216134 doi: 10.1137/18M1216134

Lurz, K., Bashiri, M., Willeke, K., Jagadish, A. K., Wang, E., Walker, E. Y., ... Sinz, F. H. (2021). Generalization in data-driven models of primary visual cortex. Retrieved from https://iclr.cc/virtual/2021/poster/3042

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (pp. 281–297).

Markram, H., Muller, E., Ramaswamy, S., Reimann, M. W., Abdellah, M., Sanchez, C. A., ... others (2015). Reconstruction and simulation of neocortical microcircuitry. *Cell*, *163*(2), 456–492.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.

Oberlaender, M., De Kock, C. P., Bruno, R. M., Ramirez, A., Meyer, H. S., Dercksen, V. J., ... Sakmann, B. (2012). Cell type–specific three-dimensional structure of thalamocortical circuits in a column of rat vibrissal cortex. *Cerebral cortex*, *22*(10), 2375–2391.

Scala, F., Kobak, D., Shan, S., Bernaerts, Y., Laturnus, S., Cadwell, C. R., ... others (2019). Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas. *Nature communications*, *10*(1), 4174.

Sinz, F., Ecker, A. S., Fahey, P., Walker, E., Cobos, E., Froudarakis, E., ... Tolias, A. (2018). Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *Advances in neural information processing systems*, *31*.

Turishcheva, P., Burg, M. F., Sinz, F. H., & Ecker, A. S. (2024). *Reproducibility of predictive networks for mouse visual cortex* (Vol. 37). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2024/file/0f293260c9e3e9527c06920316326114-Paper-Conference.pdf

Turishcheva, P., Fahey, P. G., Vystrčilová, M., Hansel, L., Froebe, R., Ponder, K., ... Ecker, A. S. (2024). Retrospective for the dynamic sensorium competition for predicting large-scale mouse primary visual cortex activity from videos. , *37*, 118907–118929. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2024/file/d758d7c0a88d741c8ca4637579c9df87-Paper-Datasets_and_Benchmarks_Track.pdf

Ustyuzhaninov, I., Burg, M. F., Cadena, S. A., Fu, J., Muhammad, T., Ponder, K., ... others (2022). Digital twin reveals combinatorial code of non-linear computations in the mouse primary visual cortex. *bioRxiv*, 2022–02.

Vintch, B., Movshon, J. A., & Simoncelli, E. P. (2015). A convolutional subunit model for neuronal responses in macaque v1. *Journal of Neuroscience*, *35*(44), 14829–14841.

Weis, M. A., Papadopoulos, S., Hansel, L., Lüddecke, T., Celii, B., Fahey, P. G., . . . Ecker, A. S. (2025, April). An unsupervised map of excitatory neuron dendritic morphology in the mouse visual cortex. *Nature Communications*, *16*(1), 3361. doi: 10.1038/s41467-025-58763-w

Willeke, K. F., Fahey, P. G., Bashiri, M., Hansel, L., Blessing, C., Lurz, K.-K., . . . Sinz, F. H. (2022, 28 Nov–09 Dec). *Retrospective on the sensorium 2022 competition* (Vol. 220). PMLR. Retrieved from https://proceedings.mlr.press/v220/willeke23a.html

Xie, F., Jain, S., Xu, R., Butrus, S., Tan, Z., Xu, X., . . . Zipursky, S. L. (2025). Spatial profiling of the interplay between cell type- and vision-dependent transcriptomic programs in the visual cortex. *Proceedings of the National Academy of Sciences*, *122*(7), e2421022122. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2421022122 doi: 10.1073/pnas.2421022122

Xie, J., Girshick, R., & Farhadi, A. (2016, 20–22 Jun). *Unsupervised deep embedding for clustering analysis* (Vol. 48). New York, New York, USA: PMLR. Retrieved from https://proceedings.mlr.press/v48/xieb16.html

Zeng, H. (2022). What is a cell type and how to define it? *Cell*, *185*(15), 2739–2755.