Deep learning spatial augmentation of single-cell transcriptomic data

Djuna von Maydell [1,2,*], Na Sun[1,*], Guillaume Leclerc[1,*], Fatima Gunter-Rahman[3], Manolis Kellis[1,4,5]

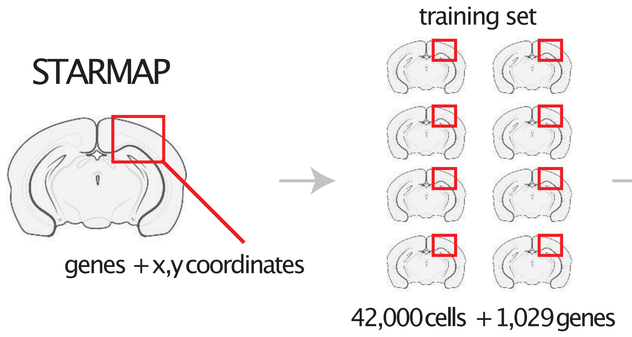
1 Computer Science and Artificial Intelligence Lab, MIT, Cambridge, MA,
2 The Picower Institute for Learning and Memory, MIT, Cambridge, MA,
3 Harvard-MIT Program in Health Sciences and Technology, Harvard/MIT, Cambridge, MA,
4 Broad Institute of Harvard and MIT, Harvard/MIT, Cambridge, MA, 5these authors contributed equally, MA,
5 Correspondence to: manoli@mit.edu
* contributed equally

ABSTRACT

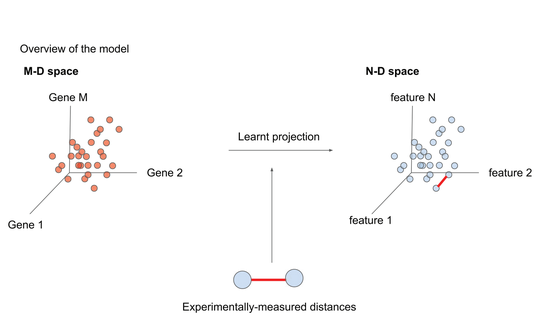
Spatial localization is not explicitly preserved in single-cell RNA-sequencing data and current techniques to preserve this information are low-throughput. To augment single-cell transcriptomic data analysis with spatial information, we apply a deep-learning approach to predict relative spatial localization of single-cells from gene expression data alone. We trained a regularized feed-forward neural network on multiple independent STARMAP datasets from the mouse cortex (> 42,000 cells with experimentally determined spatial coordinates and gene expression profiles from Wang et al, Science, 2018), to project per-cell gene expression vectors into a new N-dimensional space embedding, with the objective function of minimizing the difference between learned pairwise distances and experimentally measured ones. On a held-out STARMAP mouse brain dataset, the model estimates relative pairwise distances with Pearson correlation coefficients of up to 0.8 per cell type. Importantly, the model was trained jointly across cell types, enabling us to recognize inter-cell type spatial relationships, something that was previously not possible with expression only based similarity metrics. Interestingly, we found that the first principal axis of the learned space resolves cellular distributions along the cortical layer axis, indicating that the embedding dimensions of the model capture spatial axes in the original space, despite training constraints being on pairwise distances, not absolute positions. The cortical layer axis is resolved best for excitatory neurons (R=0.9). Importantly, the model captures this distribution at sub-layer resolution (e.g. R=0.7 for layer 4 neurons), suggesting that the model explains residuals of current marker-based excitatory-layer annotation methods. Surprisingly, we find that the model also resolves the layer axis for glial cells (e.g. R=0.8 for oligodendrocytes and R=0.7 for astrocytes). Together, these data suggest that the transcriptome alone provides sufficient information to learn pairwise distances and interpretable tissue structures for both neuronal and glial cells.

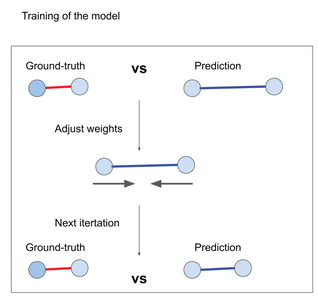
DESIGN

A. TRAINING DATA

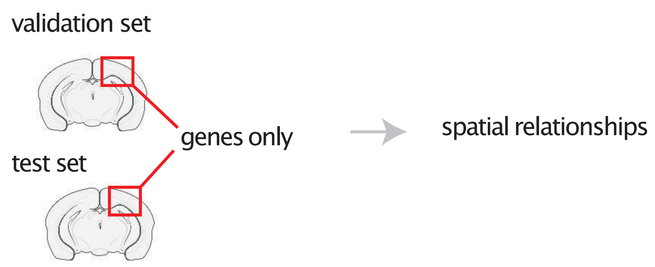


B. TRAINING PROCEDURE



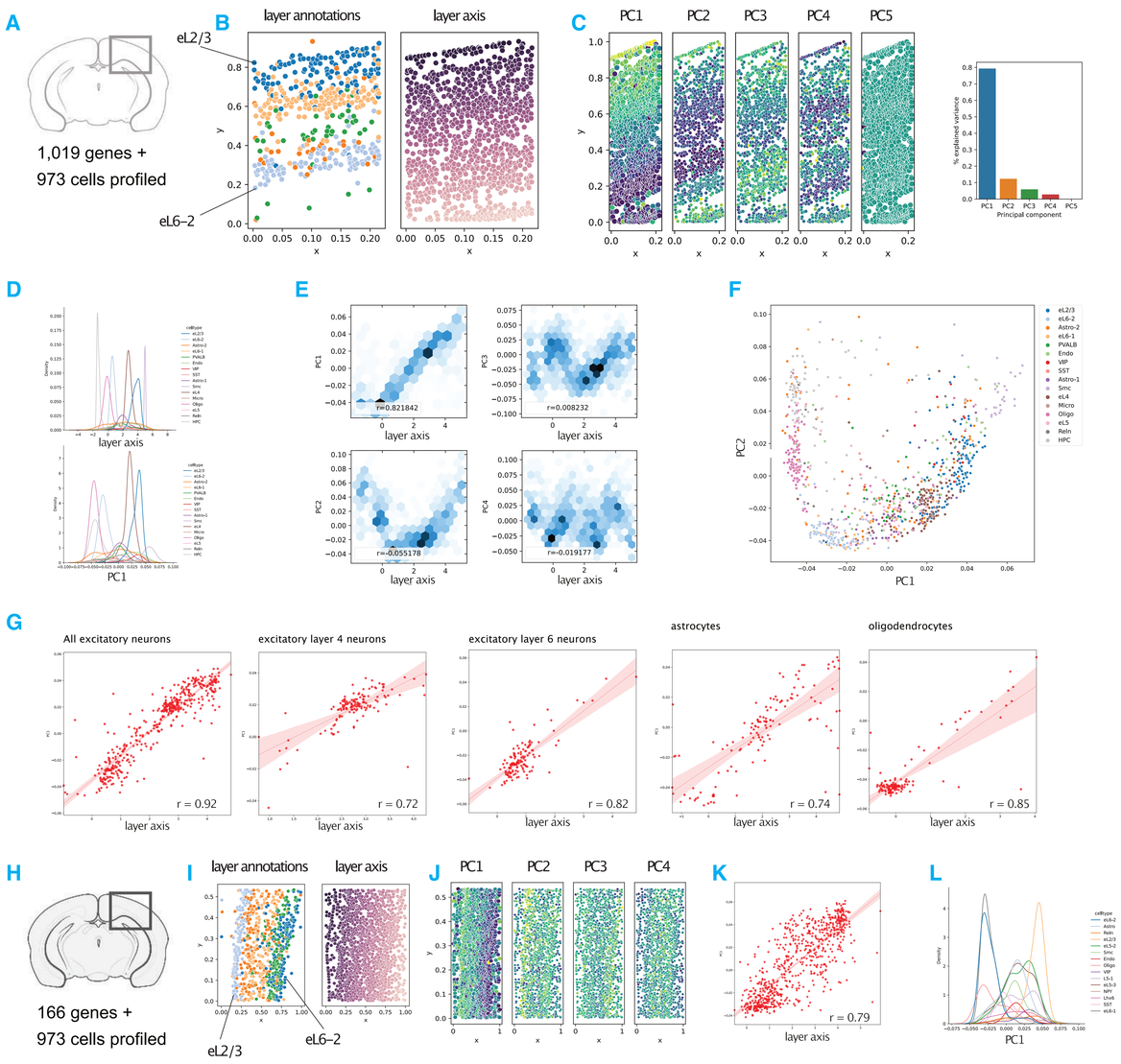


C. PREDICTION



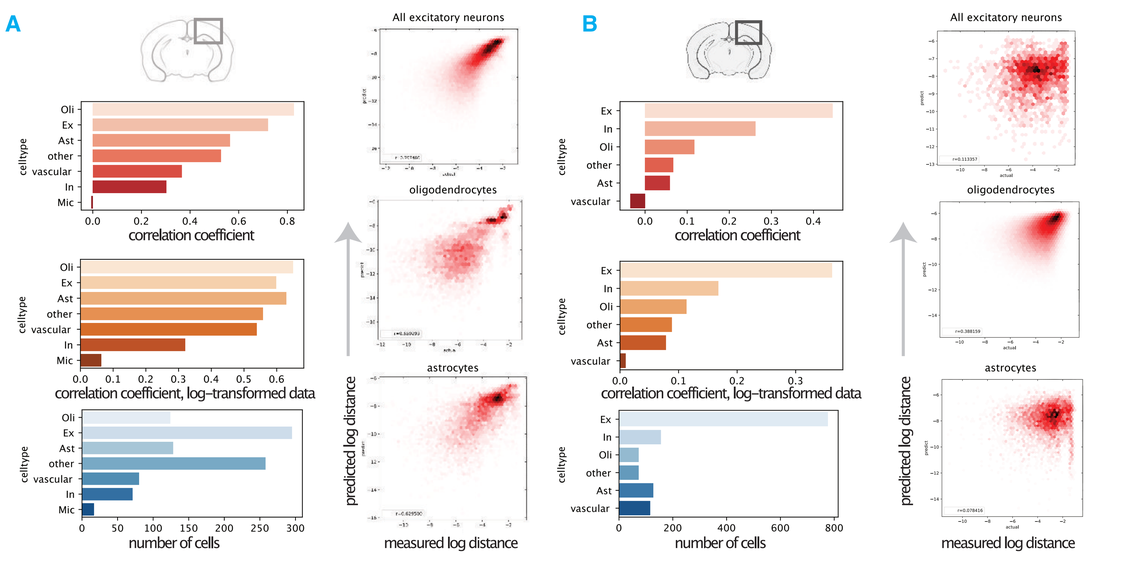
RESULTS

LEARNED DIMENSION CAPTURES CORTICAL LAYER AXIS



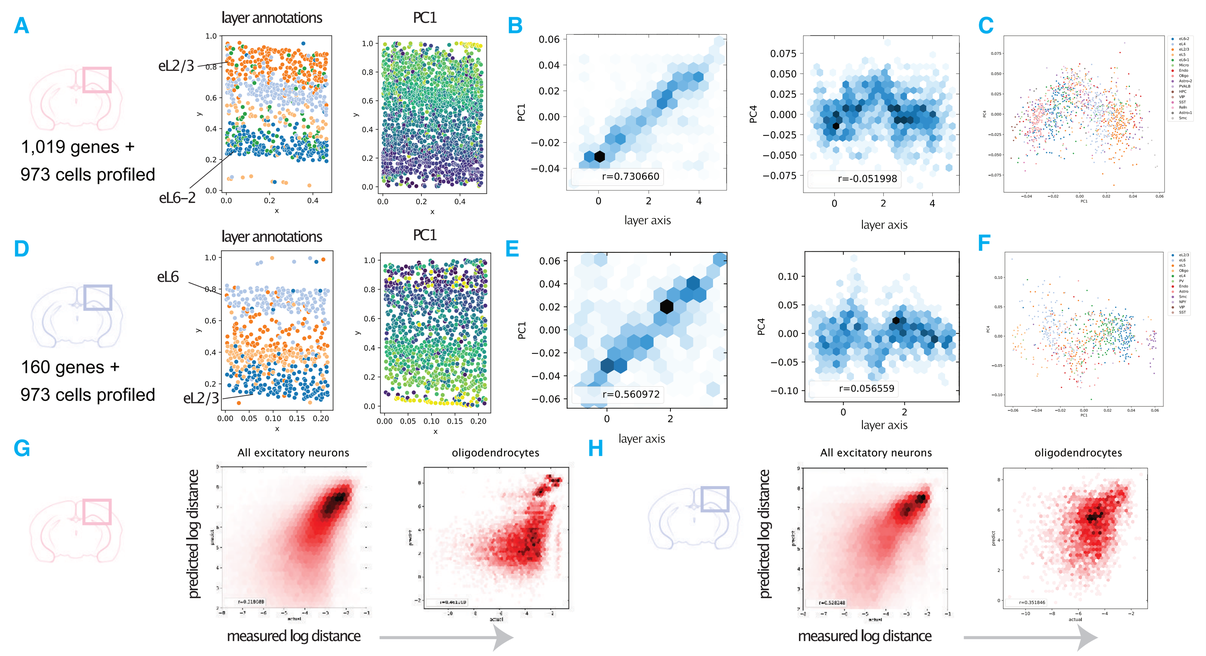
PANEL 1: A. Held-out dataset with ground-truth measurements. **B.** Neuronal labels provided by Wang *et al* (left) from which the layer axis was computed (colored purple). X and Y indicate measured coordinates of cells in tissue. **C.** Same tissue slice from B with cells colored by projection onto one of the five dimensions ("PCs") learnt by the model, while preserving pairwise distances. Yellow indicates a large positive value, purple a large negative value. **D.** Cell distributions along true layer axis and model PC1. **E.** Relationship between model PCs and true layer axis. **F.** Cells projected onto PCs1 & 2. **G.** Correlations between PC1 and true layer axis, stratified by cell type. **H.** Second held-out dataset. **I.** Same as B for second dataset. **J.** Same as C. **K.** Same as G. **L.** Same as D.

PREDICTION OF CELL-CELL PAIRWISE DISTANCES



PANEL 2: A. Correlation between actual and predicted pairwise distances between cells by cell type (top), computed for the first held-out dataset. Correlation between log-actual and log-predicted pairwise distances between cells by cell type (middle). Number of cells belonging to each cell type group (bottom). Hexagon plot showing log-actual *vs* log-predicted pairwise distances for select cell types (right). **B.** Same as in A but computed for second held-out dataset.

MODEL GENERALIZATION TO DATA WITH FEWER GENES BY FEATURE DROP-OUT REGULARIZATION



PANEL 3: Panels 1-2 suggest that the initial model performs well on a held-out dataset with ~ 1,000 predicting genes, but does not generalize as well to one with fewer genes (~160). After training with heavy regularization of the model - increasing the feature (gene) and internal node drop-out probability from ~0.02-0.2 to ~0.7 - the model performs significantly better on a held-out dataset with fewer genes, however, at the cost of top performance on the larger dataset. **A.** Performance of regularized model on first held-out dataset. Layer annotations provided by Wang *et al*. (left). Same tissue slice as on left with cells colored by projection onto model PC1 (right). B. Relationship between model PCs and true layer axis. **F.** Cells projected onto PCs1 & 4., colored by cell type. **D-F.** Same as A-C for second held-out dataset. **G-H.** Hexagon plot showing log-actual *vs* log-predicted pairwise distances for select cell types for first and second held-out datasets, respectively.

DATA AND MODEL DESCRIPTION

Supplemental to DESIGN panel:. A. Training data was generated by Wang *et al* (Science, 2018) and is publically available. Wang *et al* developed a novel *in situ* high-throughput sequencing method ("STARMAP"), which, for a given slice of tissue, provides single-cell level quantification of multiple genes and matched single-cell-level coordinates. From this data, we constructed a training set comprising 8 independent mouse cortical STARMAP datasets, for which between ~30 and ~1000 gene expression values were available per cell. **B.** We first train a simple feed forward neural network on the union of gene expression values, with the objective of learning a projection from the gene-space to a new space (Here, shown in 5 dimensions) that preserves the experimentally-derived pairwise distance matrix between cells in the new space. **B.** Finally, we use this model to predict pairwise distances onn held-out datasets from single-cell gene expression alone.