Supplementary materials

Note on MISS:

We developed the Matrix Inversion and Subset Selection (MISS) algorithm to deconvolve spatially resolved gene expression data, such as those provided by the AGEA, into cell type densities using cell-type-specific gene expression signatures from scRNAseq. The fundamental problem can be stated in the following way:

$$E = C \times D$$

where E is the genes-by-voxels matrix from the AGEA, C is the genes-by-cell-types matrix from scRNAseq, and D is the unknown cell-types-by-voxels matrix to be determined. However, given that not all genes contribute information germane to the cell mapping problem, we first apply a novel feature selection algorithm, Minimum Redundancy Maximum Relevance Minimum Residual (MRx3), to create a maximally informative gene subset. Briefly, MRx3 is an extension of the popular mRMR approach[1], a greedy algorithm that uses the quotient between the F-statistic of each candidate gene across samples (in our case, cell types) and its average pairwise correlation to the other genes in the active set to iteratively choose maximally informative genes. MRx3 adds a prefiltering step to mRMR where the 10% of genes whose addition adds the most error to the reconstruction of E from C and D are excluded. Applying MRx3 results in row-decimated matrices E_{red} and C_{red}, which can then be used to infer D. Since we require that all entries of D be strictly greater than or equal to zero, we use the MATLAB unction lsqnonneg, which solves linear problems with a non-negativity constraint, to determine D voxel by voxel. A complete description of the algorithm can be found in the original publication[2].

Supplementary tables

Dataset Name	Description	Reprocessing Methods
AMBCA[3]	The Allen Mouse Brain Connectivity Atlas is a mesoscale connectome representing the whole-brain wiring diagram. Brain-wide, region-specific axonal projections were mapped into a common 3D space using viral tracing.	See **Methods**
Tasic, *et al.* scRNAseq dataset[4,5]	Single-cell RNA sequencing data obtained by the AIBS from three distinct brain regions (primary visual cortex, secondary motor cortex, and lateral geniculate complex), representing data from 25557 individual cells.	Grouped into 25 distinct cell types and mapped using MISS[2]
Zeisel, *et al.* scRNAseq dataset[6]	A single-cell RNA sequencing data consisting of 144147 individual cells sampled from twelve regions throughout the mouse brain.	200 cell types mapped using MISS[2]

S. Table 1. Dataset information for each of the three source datasets used.

Methods	AURUC_score	Accuracy
Random Forest	0.856	0.790
Ridge	0.5302	0.6474
Lasso	0.5781	0.6411
Support Vector Machine	0.673	0.738
MLP Classifier	0.712	0.759
Decision Tree Classifier	0.704	0.727
Gradient Boosting Classifier	0.605	0.694
Extra Trees Classifier	0.687	0.711
K-Neighbors Classifier	0.744	0.774

S. Table 2. Model comparison for classification.

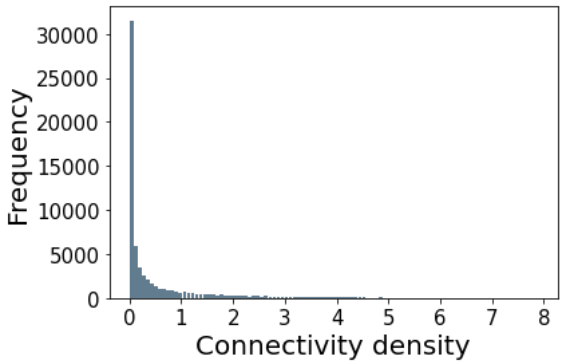
Methods	Adjusted R^2 score	RMSE
Random Forest	0.587	0.605
Ridge	0.060	0.911
Lasso	0.061	0.910
Support Vector Machine	0.363	0.750
MLP Regressor	0.515	0.654
Decision Tree Regressor	0.121	0.881
Gradient Boosting Regressor	0.328	0.770
Extra Trees Regressor	0.599	0.595
K-Neighbors Regressor	0.501	0.664

S. Table 3. Model comparison for regressor.

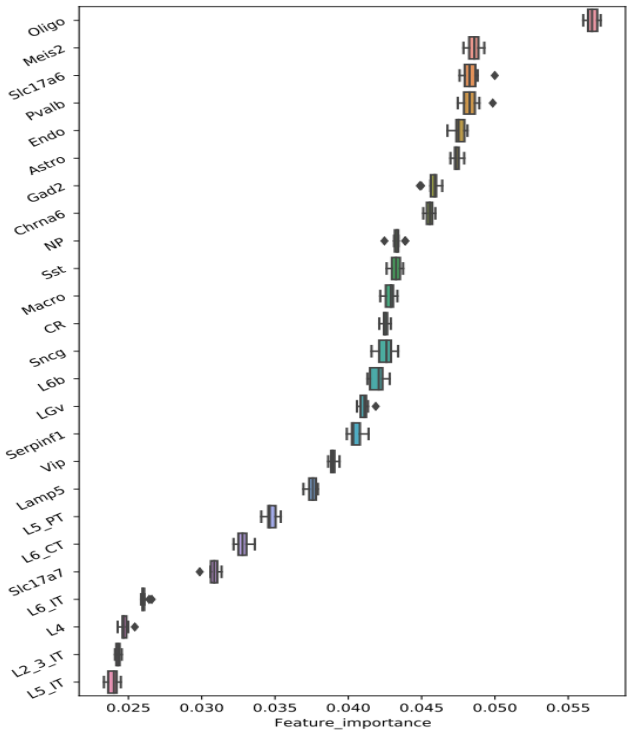
Cell Type Name	Major Cell Class	Derived From:	Description
L2/3 IT	Glutamatergic	VISp, ALM	L2/3 intratelencephalic projection neuron
L4	Glutamatergic	VISp, ALM	L4 projection neuron
L5 IT	Glutamatergic	VISp, ALM	L5 intratelencephalic projection neuron
L5 PT	Glutamatergic	VISp, ALM	L5 pyramidal-tract projection neuron
L6 CT	Glutamatergic	VISp, ALM	L6 corticothalamic projection neuron
L6 IT	Glutamatergic	VISp, ALM	L6 intratelencephalic projection neuron
L6b	Glutamatergic	VISp, ALM	L6b projection neuron
NP	Glutamatergic	VISp, ALM	Near-projecting neuron (?)
CR	Glutamatergic	VISp, ALM	Cajal-Retzius neuron (?)
Slc17a6	Glutamatergic	LGd	(?)
Slc17a7	Glutamatergic	LGd	(?)
Pvalb	GABAergic	VISp, ALM	Pv+ interneuron
Sst	GABAergic	VISp, ALM	Sst+ interneuon
Vip	GABAergic	VISp, ALM	Vip+ interneuron
Serpinf1	GABAergic	VISp, ALM	Serpinf1+ inhibitory neuron (?)
Sncg	GABAergic	VISp, ALM	Sncg+ inhibitory neuron (?)
Chrna6	GABAergic	VISp, ALM	Chrna6+ inhibitory neuron (?)
Meis2	GABAergic	VISp, ALM	Meis2+ inhibitory neuron (?)
Lamp5	GABAergic	VISp, ALM	Lamp5+ inhibitory neuron (?)
LGv	GABAergic	LGd	(?)
Gad2	GABAergic	LGd	(?)
Astro	Glial	VISp, ALM, LGd	Astrocyte
Macro/Micro	Glial	VISp, ALM, LGd	Microglia + macrophage
Oligo	Glial	VISp, ALM, LGd	Oligodendrocyte + OPC
Endo	Endothelial	VISp, ALM, LGd	Endothelia + other vessel-associated cells

S. Table 4. We provide a list of major classes of neuronal and glial cell types from the Tasic, et al. dataset with a metric of whether the type of cell is definitively a canonically well-characterized cell type, with types lacking characterization denoted with a (?). We note that *Pvalb*+, *Sst*+, and *Vip*+ GABAergic interneurons, the various laminar neocortical glutamatergic neurons, and the glia are well-characterized types by their major class, but other neuronal subtypes are not, necessarily. VISp – primary visual cortex; ALM – anterior lateral motor cortex; LGd – lateral geniculate complex, dorsal part. Adapted with permission from Mezias, et al., 2021.

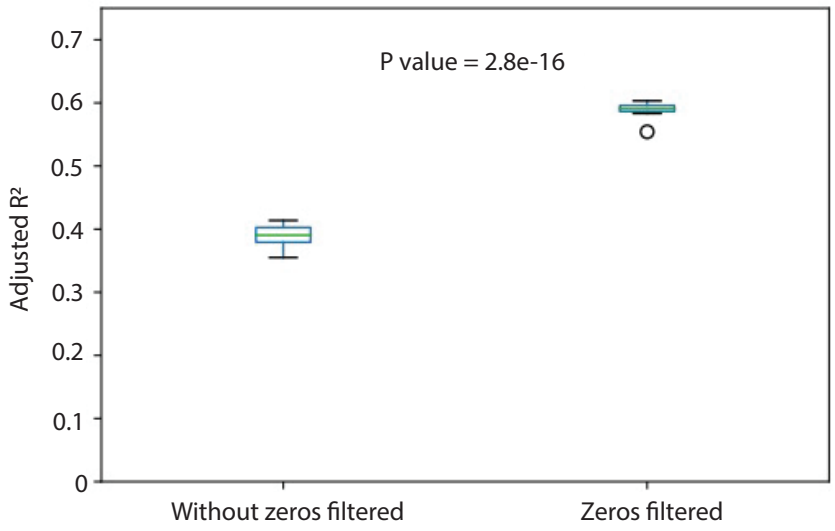
Supplementary figures



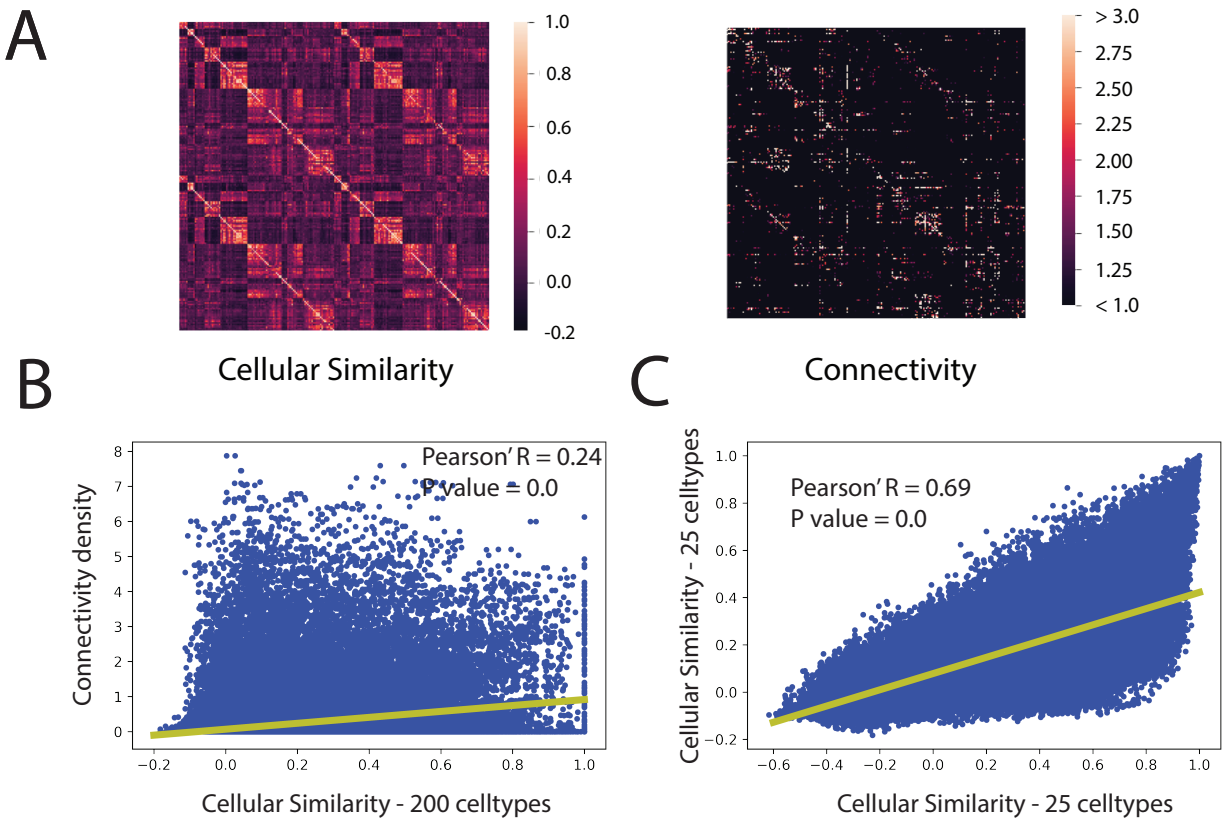
S. Figure 1. Histogram showing of the log-transformed connectivity distribution



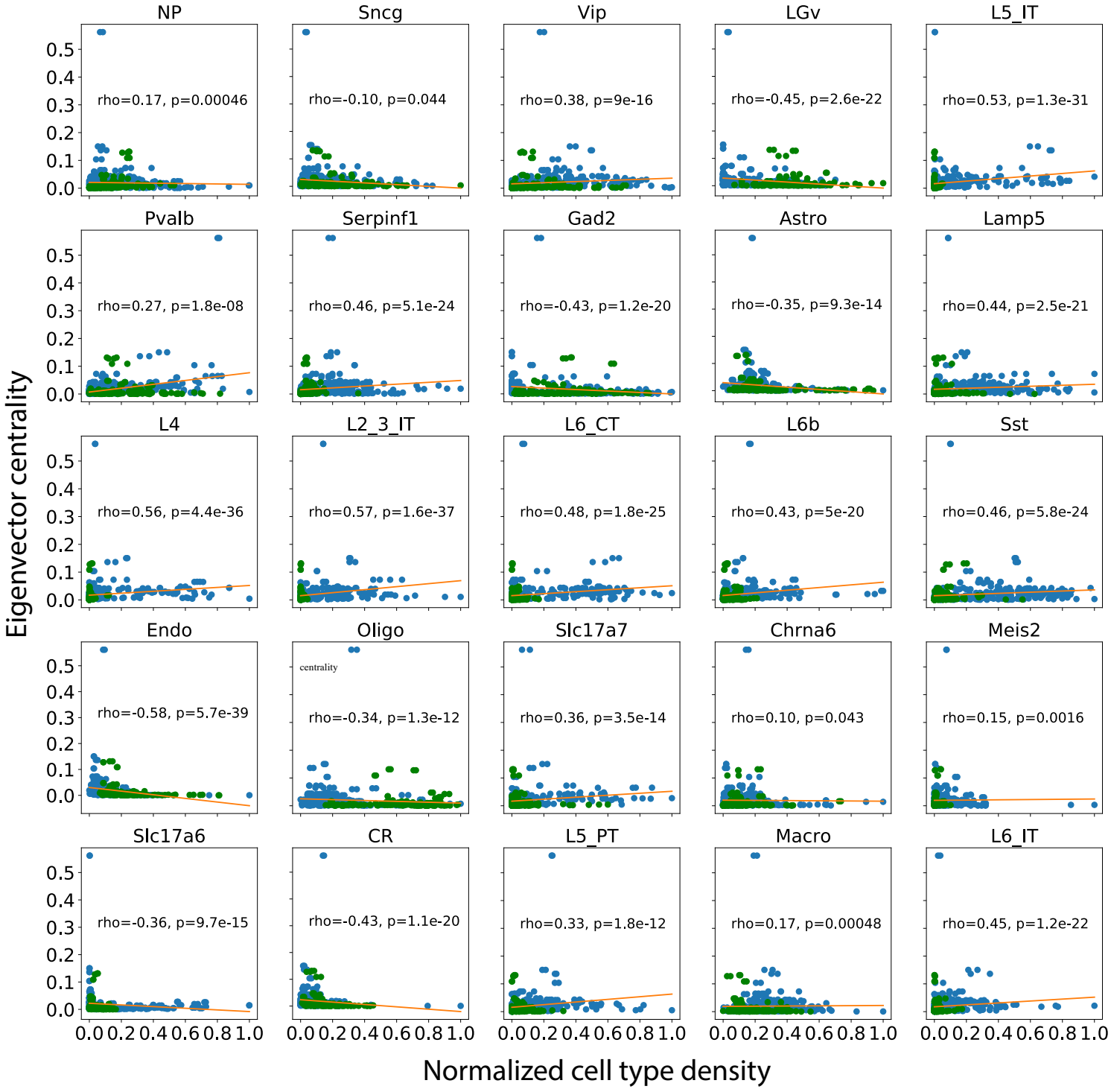
S. Figure 2. Feature importance for examing the existence of connectivity.



S. Figure 3. Comparison between the performance of predicting connectivity density without first filtering out unconnected region pairs (left) and with zero-filtering (right), tenfold cross validation. Independent student's t-test was performed to assess statistical significance.



S. Figure 4. Correlation between regional cell-type similarity matrix and connectivity matrix. **A**. Pearson correlation matrix of spatial cell-type enrichment quantification across brain regions (using the Zeisel, *et al*. scRNAseq dataset[6]). **B.** Scatterplot of regional cell-type similarity versus the log-transformed connectivity strength between the two regions. Pearson's R = 0.24. **C.** Scatterplot of regional cell-type similarity computed by 25 features versus that computed by 200 features. Pearson's correlation gave: r = 0.69.



S. Figure 5. Scatterplot of normalized cell type density versus eigenvector centrality. Blue color indicates non-hindbrain regions; green color indicates hindbrain regions. Spearman's correlation is reported in each subplot.

References

1. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238 (2005).
2. Mezias, C., Torok, J., Maia, P. D., Markley, E. & Raj, A. Matrix Inversion and Subset Selection (MISS): A novel pipeline for quantitative mapping of diverse cell types across the murine brain. *bioRxiv* (2019). doi:10.1101/833566
3. Oh, S. W. *et al*. A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
4. Tasic, B. *et al*. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
5. AIBS. Allen Cell Types Database - Technical White Paper: Transcriptomics. 1–14 (2018). Available at: http://help.brain-map.org/display/celltypes/Documentation.
6. Zeisel, A. *et al*. Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999–1014 (2018).