*Journal:* NETWORK NEUROSCIENCE

RESEARCH

# Modeling the cell-type specific murine connectome

**Samson Koelle** [1,2], **Jennifer Whitesell** [1], **Karla Hirokawa** [1], **Hongkui Zeng**[1], **Marina Meila**[2], **Julie Harris**[1], **Stefan Mihalas**[1]

[1]Allen Institute for Brain Science, Seattle, WA, USA

[2]Department of Statistics, University of Washington, Seattle, WA, USA

**Keywords:** [a series of capitalized words, separated with commas]

## ABSTRACT

The Allen Brain Connectivity Atlas consists of thousands of labelling experiments targeting interrogating diverse structures and classes of projecting neurons. This paper describes the conversion of these experiments into class-specific connectivity matrices representing the connection between source and target structures. We introduce and validate a novel statistical model for creation of connectivity matrices that combines spatial and categorical smoothing to share information between similar neuron classes. We then investigate patterns in the resultant connectivities, and show that our connectivities display expected cell-type specific structures.

## AUTHOR SUMMARY

## INTRODUCTION

The animal nervous system enables an extraordinary range of natural behaviors, and has inspired much of modern artifical intelligence. Neural connectivities - axon-dendrite connections from one region to another - form the architecture underlying this capability. These connectivities vary by neuron type, as well as axonic source and dendritic target structure. Thus, characterization of the

relationship between neuron type and source and target structure is an important step to understanding the nervous system.

Viral tracing experiments - in which a viral vector expressing GFP is transduced into neural cells through stereotaxic injection - are a useful tool for understanding these connections on the mesoscale (**???**). The GFP protein moves from axon to dendrite through the process of anterograde projection, so neurons 'downstream' of the injection site will also fluoresce. Two-photon tomography imaging can then determine the location and strength of the fluorescent signals in two-dimensional slices. These locations can then be mapped back into three-dimensional space, and the signal is partitioned into the transduced source and merely transfected target regions.

The conversion of such experiment-specific signals into an overall estimate of the connectivity strength of two regions is accomplished by a statistical model. **?** and **?** describe two such methods. Intuitively, both of these models provide some improvement over simply averaging the projection signals of injections in a given region.  is another. These models are evaluated based off of their ability to predict held-out experiments in leave-one-out cross validation. A model that performs well in such validation experiments is then assumed to generate the most accurate connectivity.

Both **?** and **?** develop models for mostly wild-type mice using a standardized vector over all experiments. However, recent work (**?**) has extended these datasets to include viral tracing experiments inducing cell-type specific fluorescence. This is accomplished by injecting vectors with Cre-recombinase triggered GFP promoters into transgenic mice with cell-type specific Cre-recombinase expression Thus, the this paper extends the methodology of **?** and **?** to deal with the diverse set of cre-lines described in **?**.

This extension relies on a to our knowledge novel estimator that takes into account both the spatial position of the labelled source, as well as the categorical cre-label. This model outperforms the model of **?**, even for wild-type experiments.

The resulting cell-type specific connectivity matrices form a multi-way *neural connection tensor* of information about neural structure. We do not attempt an exhaustive analysis of this data, but do demonstrate several basic phenomena. First, we verify several cell-type specific patterns found elsewhere in the literature. Second, we discover cell-type specific signals in the neural connection

tensor. Finally, we decompose the overall (wild-type) connectivity matrix into factors representing archetypal connective patterns.

## METHODS

We create cell-type specific connectivity matrices using a model trained on murine viral-tracing experiments. This model predicts projection patterns of different neuron classes at different locations within the brain that are more accurate than simple averages over nearby experiments in cross-validation. This section describes the data used to generate the model, the model itself, the evaluation of the model, and the use of the model in creation of the connectivity matrices. We then give exploratory analyses of the resulting connectivities that illustrate their key features. Additional information on our methods is given in Supplemental Section .

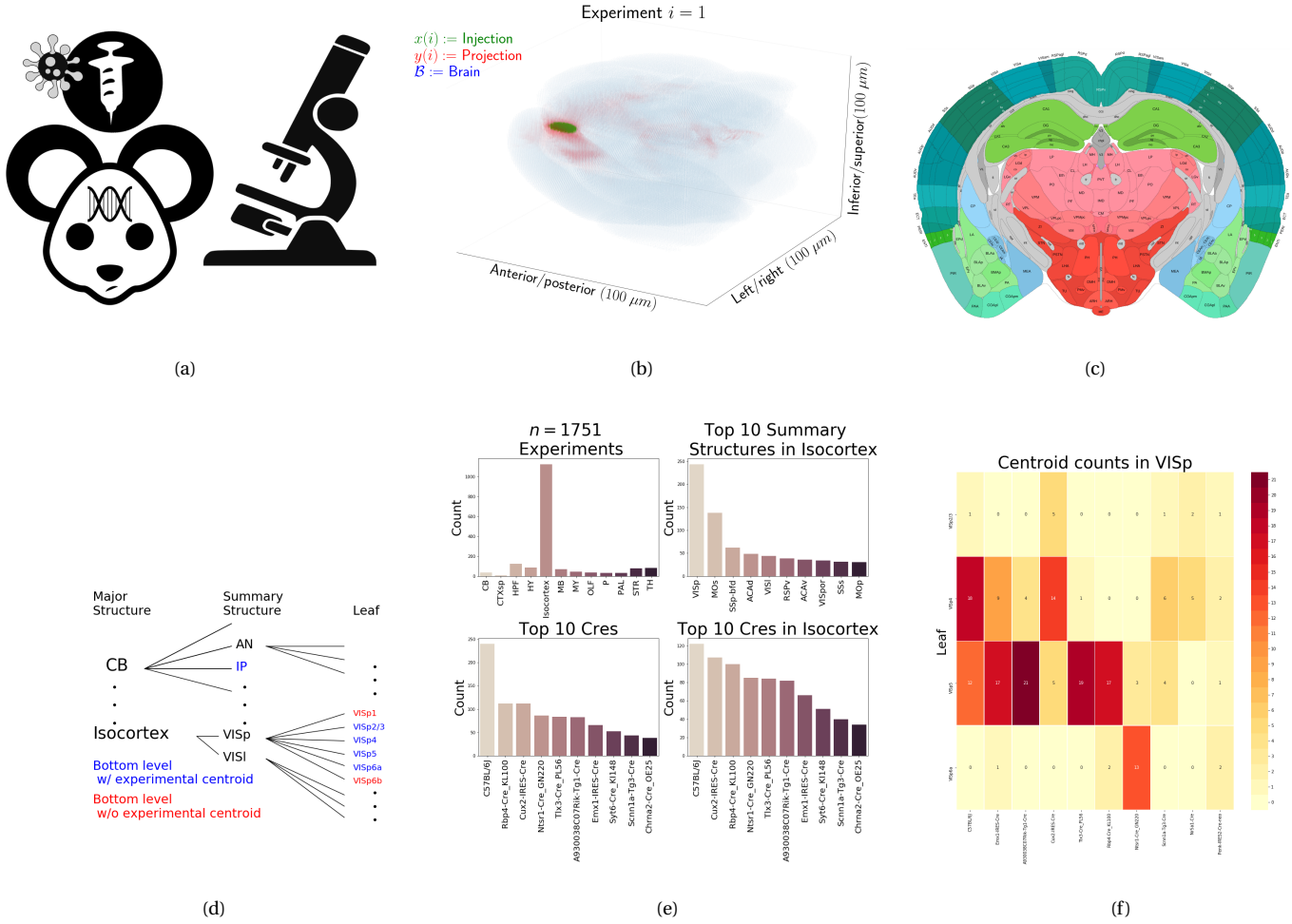



Figure 1: Experimental setting. a) Within the brain (blue), injection (green) and projection (red) areas are determined via histological analysis and alignment to the Allen Common Coordinate Framework (CCF). b) An example of the segmentation of projection and injection for a single experiment. c) Example of structural segmentation within a horizontal plane. d) Explanation of nested structural ontology highlighting lowest-level and data-relevant structures. e) Abundances of crelines and structural injections. f) Cooccurence of layer-specific centroids and creline within VISp

### *Mice*

(SK's comment:**Experiments involving mice were approved by the Institutional Animal Care and Use Committees of the Allen Institute for Brain Science in accordance with NIH guidelines.**)

### *Data*

Our dataset $\mathscr{D}$ consists of $n = 1751$ experiments from the Allen Mouse Connectivity Atlas. Figures 1a summarizes the multistage experimental process used to generate this data. In each experiment, a GFP-labelled transgene casette with a potentially cre-specific promoter is injected into a particular location in a cre-driver mouse. This causes fluorescence thats depends on the localization of cre-recombinase expression within the mouse. While frequently this localization corresponds to a specific cell-type, it can also correspond to a combination of cell-types. For example, in wild-type mice injected with non-cre specific promoters, fluorescence is observed in all areas projected to from the injection site, regardless of cell-type. Thus, we use the term *neuron class* to describe the neurons targeted by a specific combination of transgene and mouse-line. This is the notion of cell-type specificity that we model.

After injection, the resultant fluorescent signal is imaged, and aligned into the Allen Common Coordinate Framework (CCF) v3, a three-dimensional idealized model of the brain that is consistent between animals. This imaging and alignment procedure (described in detail in (**?**)) records fluourescent intensity discretized at the 100 $\mu$m *voxel* level. The image is histologically segmented into *injection* and *projection* areas corresponding to areas of transduction and transduction/transfection, respectively. An example for a single experiment is given in Figure 1b.

Our goal is the estimation of *s*tructural connectivity from one structure to another. Thus, a visual depiction of this structural regionalization for a slice of the brain is given in Figure 1c. For different areas of the brain, the Allen Atlas contains different depths of discretization. We denote these levels as Major Structures, Summary Structures, and Leafs. As indicated in Figure 1d, the dataset used to generate the connectivity model reported in this paper contains certain combinations of structure and neuron class $(S, V)$ frequently, and others not at all. A summary of the most frequently assayed neuron classes and structures is given in Figures 1e and 1f. Since users of the connectivity matrices may be interested in particular combinations, or interested in the amount of data used to generate a

particular connectivity estimate, we exhaustively present this information about all experiments in Appendix .

We can preprocess this publically available data in several ways. A detailed mathematical description of these is given Appendix .

***Connectivity***

At an essential level, cell-class specific neural connectivity is a function $f : \mathcal{V} \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^+$ giving the directed connection of a particular neuron class from a one position in the brain to another. However, what we actually create is, given a set of source regions $\mathcal{S} = \{S\}$, target regions $\mathcal{T} = \{T\}$, and neuron classes $V$,

$$\begin{aligned}
&\textit{connectivity strength } \mathscr{C} \in \mathcal{V} \times \mathcal{S} \times \mathcal{T} \times \mathbb{R}_{\geq 0} \text{ with } \mathscr{C}(V,S,T) = \sum_{s \in S} \sum_{t \in T} f(v,s,t),\\
&\textit{normalized connectivity strength } \mathscr{C}^S \in \mathcal{V} \times \mathcal{S} \times \mathcal{T} \times \mathbb{R}_{\geq 0} \text{ with } \mathscr{C}^S(V,S,T) = \frac{1}{|S|}\mathscr{C}(V,S,T),\\
&\textit{normalized projection density } \mathscr{C}^D \in \mathcal{V} \times \mathcal{S} \times \mathcal{T} \times \mathbb{R}_{\geq 0} \text{ with } \mathscr{C}^D(V,S,T) = \frac{1}{|S||T|}\mathscr{C}(V,S,T)).
\end{aligned}$$

These represent the strength of the connection from source to target regions for each class. Since the normalized strength and density are computable from the strength via a fixed normalization, our main statistical goal is to estimate $\mathscr{C}(V,S,T)$. We call this estimator $\widehat{\mathscr{C}}$.

Construction of such an estimator raises the questions of what data to use for estimating which connectivity, how to featurize the dataset, what statistical estimator to use, and how to reconstruct the connectivity using the chosen estimator. Mathematically, we represent these considerations as

$$\widehat{\mathscr{C}}(V,S,T) = e^*(\widehat{e}(e_*(\mathscr{I}(\mathscr{D})))). \tag{1}$$

This makes explicit the data featurization $e_*$, statistical estimator $\widehat{e}$, and any potential subsequent transformation $e^*$ such as averaging over the source region, as well as the fact that different data $\mathscr{D}$ may be used to estimate different connectivities. Table 1 reviews estimators used for this data-type. Additional information is given in Appendix .

| Model | $e^*$ | $\widehat{e}$ | $e_*$ | Training Data |
|---|---|---|---|---|
| (?) | $\widehat{e}(S)$ | NNLS(X,Y) | $X = r(x(I)), Y = r(y(I))$ | $I = I_M$ |
| (?) | $\sum_{s \in S} \widehat{e}(s)$ | NW(X,Y) | $X = c(x(I)), Y = r(y(I))$ | $I = I_M$ |
| Cre-NW | $\sum_{s \in S} \widehat{e}(s)$ | NW(X,Y) | $X = c(x(I)), Y = n(r(y(I)))$ | $I = I_S \cap I_V$ |
| Expected Loss (EL) | $\sum_{s \in S} \widehat{e}(s)$ | $\mathrm{EL}_S(X, Y, V)$ | $X = c(x(I)), Y = n(r(y(I))), V = v(I)$ | $I = I_S$ |

Table 1: Estimation of $\mathscr{C}$ using connectivity data. The regionalization, estimation, and featurization steps are denoted by $e^*, \widehat{e}$, and $e_*$, respectively. The training data used to fit the model is given by $I$. We generically denote the set of experiments used to train a particular model as $I$, and experiments from particular major brain divisions, summary structures, and leafs as $I_M$, $I_U$, and $I_L$, respectively.

Our contributions - the Cre-NW and Expected Loss (EL) models - have several differences from the previous methods. In contrast to the **?** non-negative least squares and **?** Nadaraya-Watson estimators that take into account $s$ and $t$, but not $v$, our new estimators specifically account for neural class. The Cre-NW estimator only uses experiments from a particular neural class to predict connectivity for that class, while the EL estimator shares information between classes.

100 ***Model evaluation***

We select optimum functions from within and between our estimator classes using empirical risk minimization. Equation 1 includes a deterministic step $e^*$ included without input by the data. The performance of $\widehat{\mathscr{C}}$ is therefore determined by performance of the model $\widehat{f}(v,s,t) = \widehat{e}(e_*(\mathscr{I}(\mathscr{D})))$. We can then evaluate $\widehat{f}(v,s,t)$ using *l*eave-one-out cross validation, in which the accuracy of the model is assessed by its ability to predict experiments excluded from the training data. In order to compare between methods, we necessarily restrict to the smallest set of evaluation experiments suggested by any of our models. Since the number of parameters fit is quite low relative to the size of the evaluation set, we do not make use of a formal validation-test split. We use $l2$-loss and weighted $l2$-loss to evaluate these predictions.

$$\text{l2-loss } l(\hat{f}) = \frac{1}{|I_M|} \sum_{i \in I_M} \| r(y(i)) - \hat{f}(c(i)) \|_2^2.$$

$$\text{weighted l2-loss } l(\hat{f}) = \frac{1}{|\{S,V\}|} \sum_{s,v \in \{S,V\}} \frac{1}{|I_{s,v}|} \sum_{i \in I_{s,v}} l(r(y(i)), \hat{f}(\mathbb{D} \setminus i)).$$

101 As a final modeling step, we establish a lower limit of detection. This is covered in Appendix

***Connectivity analyses***

We show neuronal processes underlying our estimated connectome using a variety of types of undersupervised learning. Clustering projection patterns by class and source structure. This shows that cell-class has a dominating effect on projection in certain regions. We flatten the connectivity tensor $\mathscr{C} \in \mathbb{R}^{c \times s \times t}$ to $\mathscr{C}_\flat \in \mathbb{R}^{cs \times t}$ and cluster the $cs$ sources by their $t$-dimensional projections.

Second, we extend the characterization of **?** on structural differences in short-range projections. These are primarily assumed to be due to diffusion, and the diffusion-rate helps to characterize the basic structural anatomy. Third, since the overall wild-type connectome results from the combination of underlying cell-classes, we apply non-negative matrix factorization (NMF) to decompose the observed long-range connectivity into *connectivity archetypes* that linearly combine to reproduce the observed connectivity. These methods identify structures with both known and plausible biological meaning, and simplistically exemplify useful posthoc analyses for data of this type. Technical details of these approaches are given in Appendix.

## RESULTS

Our results include evaluation of model fit, the cre-specific connectivity matrices themselves, and retrospective analyses of these matrices for patterns related to cre-type and source and target regions.

### *Model evaluation*

Table 2 contains the sizes of these evaluation sets in each major structure. This information may be cross-referenced visually with the figures in Our two-stage model generally performs better than the cre-line specific NW estimator.

| Structure | Estimator<br>Smoothing<br>Target<br># Eval exps | EL<br>SS<br>SS | NW<br>Cre-SS<br>SS | Average<br>Cre-SS<br>SS | NW<br>SS<br>SS | NW-wt<br>M<br>SS |
|---|---|---|---|---|---|---|
| CB | 10 | 0.044 | 0.081 | 0.081 | 0.058 | 0.439 |
| CTXsp | 2 | 0.497 | 0.497 | 0.497 | 0.497 | 0.000 |
| HPF | 79 | 0.122 | 0.140 | 0.143 | 0.155 | 0.471 |
| HY | 41 | 0.241 | 0.266 | 0.269 | 0.244 | 1.019 |
| Isocortex | 838 | 0.173 | 0.195 | 0.202 | 0.234 | 0.404 |
| MB | 23 | 0.151 | 0.151 | 0.166 | 0.139 | 0.759 |
| MY | 7 | 0.186 | 0.233 | 0.233 | 0.184 | 0.452 |
| OLF | 17 | 0.069 | 0.095 | 0.100 | 0.073 | 0.110 |
| P | 8 | 0.236 | 0.239 | 0.239 | 0.264 | 0.984 |
| PAL | 11 | 0.190 | 0.198 | 0.198 | 0.260 | 1.401 |
| STR | 45 | 0.084 | 0.088 | 0.089 | 0.097 | 0.265 |
| TH | 29 | 0.351 | 0.678 | 0.678 | 0.365 | 1.088 |

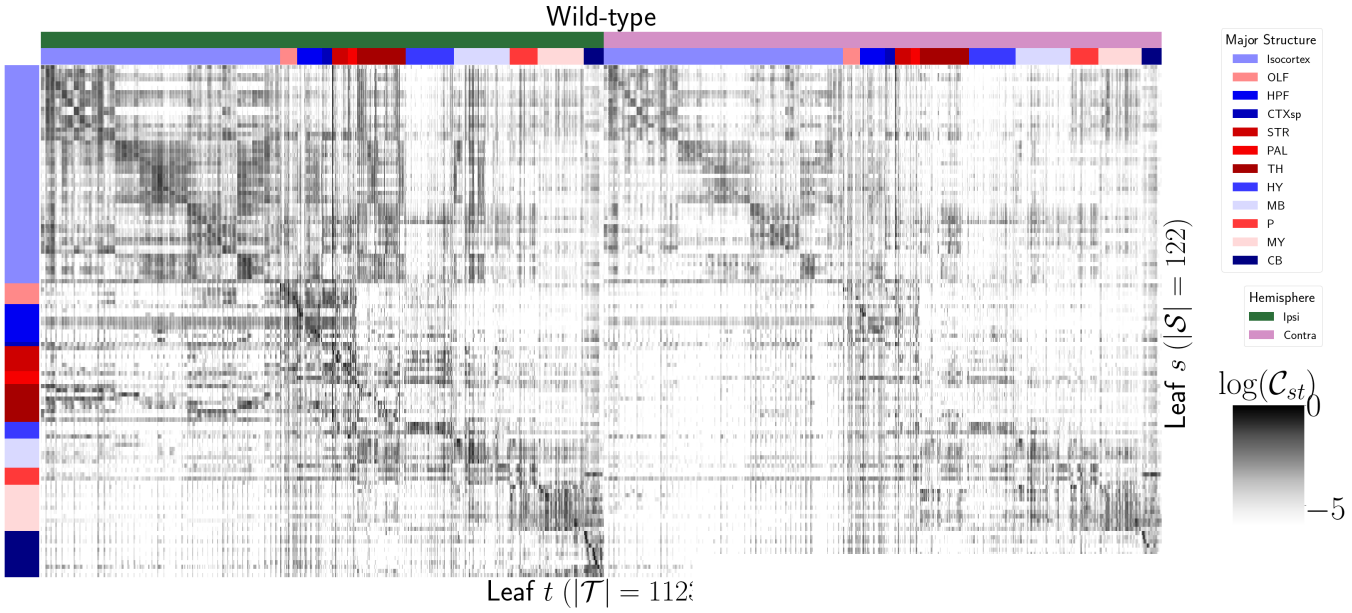
Table 2: Weighted losses with summary structure targets.

Additional information on model evaluation, including class and structure specific performance, is given in Appendix

### *Connectivities*

Our main result is the estimation of matrices $\hat{\mathscr{C}}_v$ representing connections of source structures to target structures for particular cre-lines $v$. We exhibit several characteristics of interest, and confirm the detection of several well-established connectivities within our tensor. Many additional interesting biological processes are visible within this matrix - more than we can report in this paper - and it is our expectation that these will be identified by users of our results. The connectivity tensor and code to reproduce it are available at
https://github.com/AllenInstitute/mouse_connectivity_models/tree/2020.

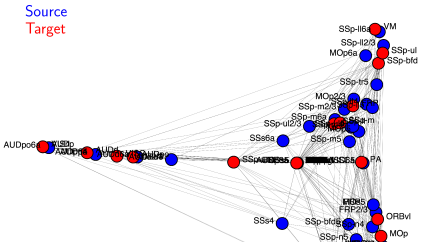
*Overall connectivity* The connectivity matrix for wild-type connectivities from leaf sources to summary structure targets is illustrated in Figure **??**. The clear intraareal connectivities mirror previous estimates in **?** and **?** and descriptive depictions of individual experiments in **?**. Compared with **?**, our more discritized source smoothing and greater number of experiments leads to a significantly more discritized connectivity matrix. This is generally expected - for example, different cortical layers have more substantially different connectivities.




(a)

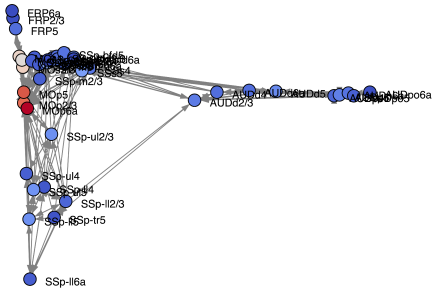
| | # Ipsilateral Leaf Targets | Top Entropy | Bottom Sparsity | Bottom Entropy | Top Sparsity |
|---|---|---|---|---|---|
| Isocortex | 51 | CP | BAC | BAC | ENTl |
| OLF | 11 | TMv | III | III | NaN |
| HPF | 15 | IG | EPv | PA | NaN |
| CTXsp | 7 | TT | FC | APr | TT |
| STR | 14 | RPA | ISN | PYR | TU |
| PAL | 9 | PG | ACVII | GR | MG |
| TH | 44 | NOD | DN | SSp-ll | SCm |
| HY | 44 | CLA | SH | LSc | DG |
| MB | 39 | NDB | SubG | SGN | SUB |
| P | 26 | MT | Acs5 | SOC | NDB |
| MY | 43 | RT | NaN | OV | EPd |
| CB | 18 | ECT | AOB | MOB | GU |

(b)



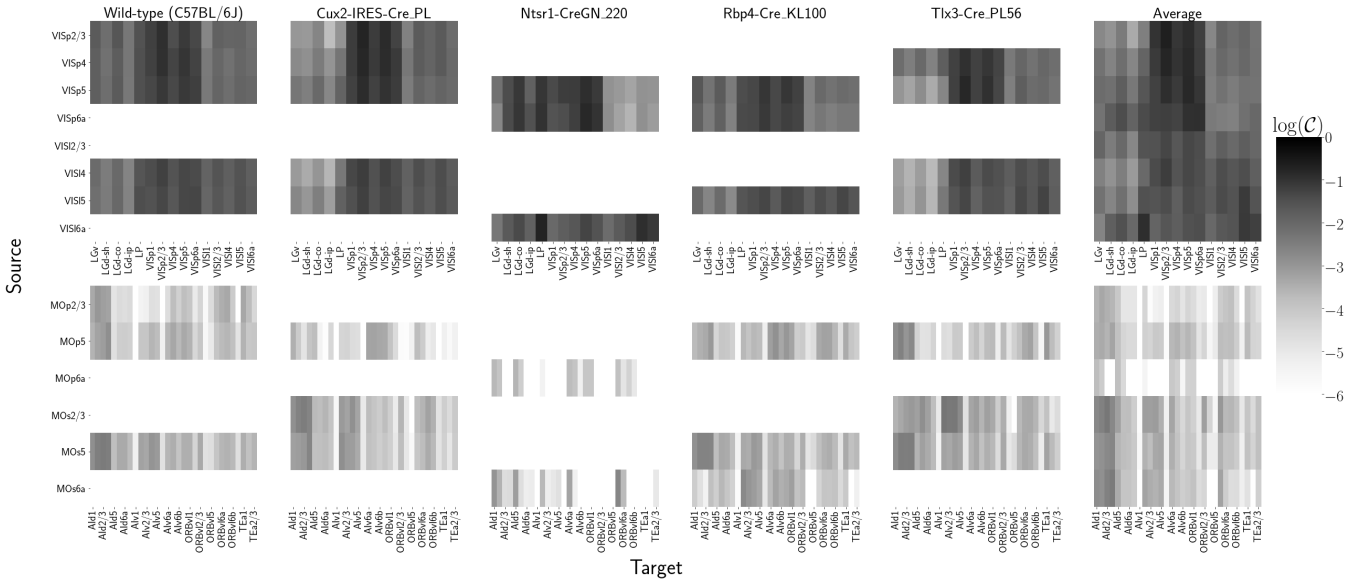

| | # Ipsilateral Leaf Targets | Top Entropy | Bottom Sparsity | Bottom Entropy | Top Sparsity |
|---|---|---|---|---|---|
| Isocortex | 51 | CP | BAC | BAC | ENTl |
| OLF | 11 | TMv | III | III | NaN |
| HPF | 15 | IG | EPv | PA | NaN |
| CTXsp | 7 | TT | FC | APr | TT |
| STR | 14 | RPA | ISN | PYR | TU |
| PAL | 9 | PG | ACVII | GR | MG |
| TH | 44 | NOD | DN | SSp-ll | SCm |
| HY | 44 | CLA | SH | LSc | DG |
| MB | 39 | NDB | SubG | SGN | SUB |
| P | 26 | MT | Acs5 | SOC | NDB |
| MY | 43 | RT | NaN | OV | EPd |
| CB | 18 | ECT | AOB | MOB | GU |

(d)

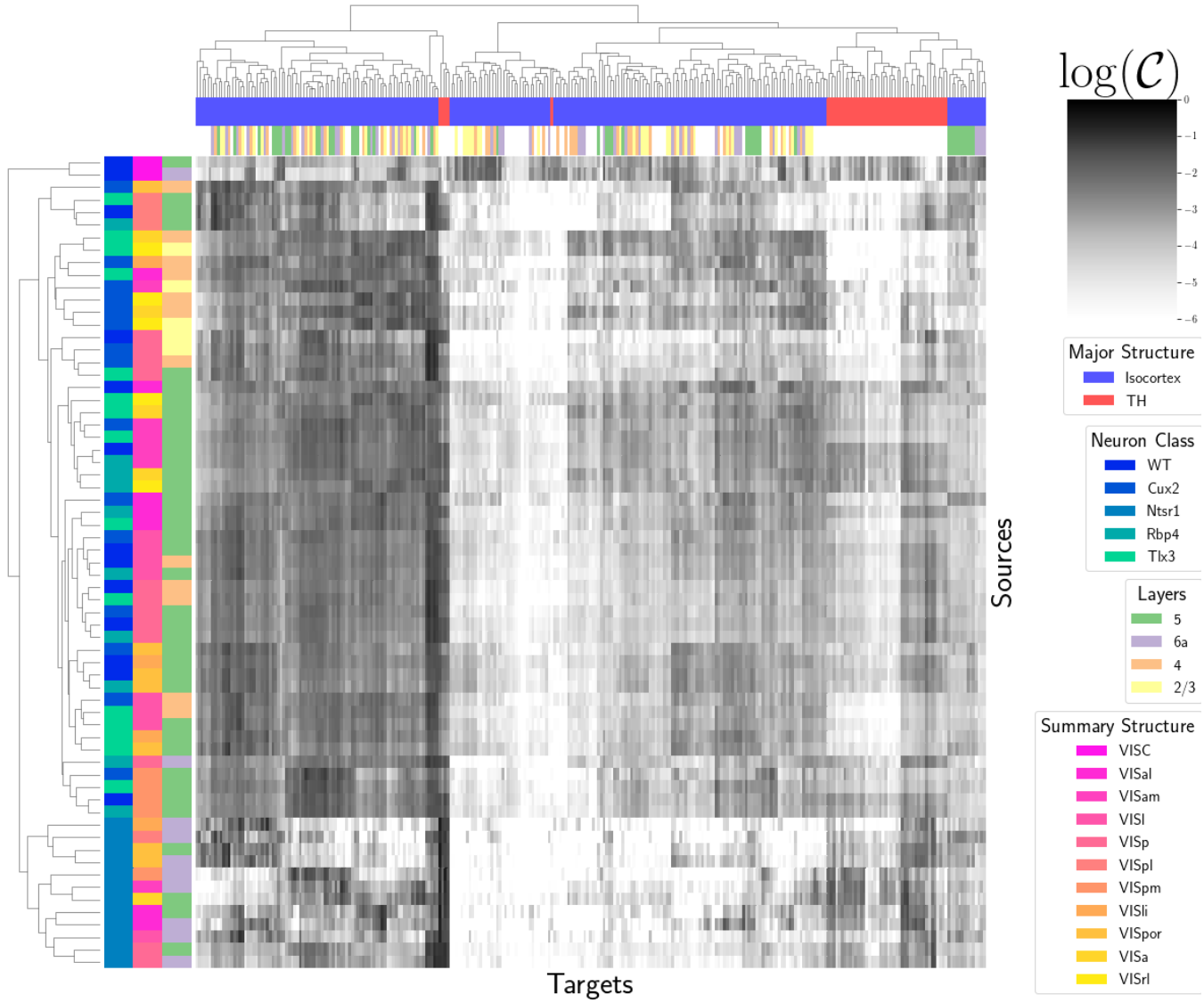



(e)

*Class-specific connectivities* The cell-type specific connectivities that we provide also conform to well-known behaviors. Examples from the visual processing and motor control regions of the cortex are given in Figure **??** for both wild type and several cre-lines. Rbp4-Cre and Ntsr1-Cre target layers 5 and 6, respectively. As in **?**, layer 5 projects to anterior basolateral amygdala (BLA) and capsular central amygdala (CEA), while layer 6 does not.


Wild-type (C57BL/6J)
Cux2-IRES-Cre_PL
Ntsr1-CreGN_220
Rbp4-Cre_KL100
Tlx3-Cre_PL56
Average
log(C)
Source
Target

(f)


log(C)
Major Structure
Isocortex
TH
Neuron Class
WT
Cux2
Ntsr1
Rbp4
Tlx3
Layers
5
6a
4
2/3
Summary Structure
VISC
VISal
VISam
VISl
VISp
VISpl
VISpm
VISli
VISpor
VISa
VISrl
Sources
Targets


(g)

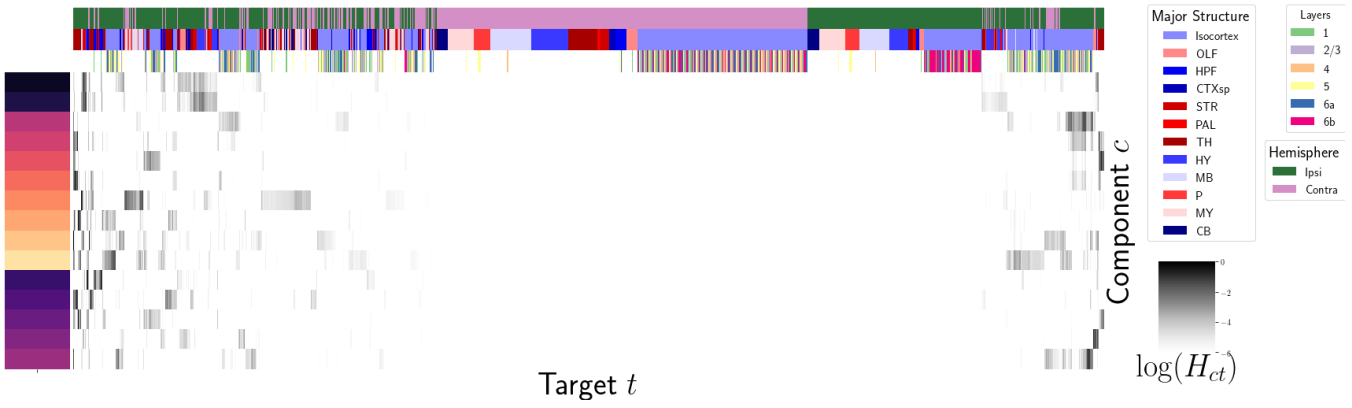
### *Connectivity Analyses*

The connectivity matrix represents a collection of relatively few biological processes. For example, certain cell-types and layers have a characteristic connectivity pattern, and structures tend to connect most strongly to the most proximal areas. We elucidate these patterns through two types of analyses. First, we demonstrate cell-type specific connectivity patterns by heirarchical clustering of connectivities from multiple cre-lines, and showing that cre-line is a key factor driving the observed behavior. Then, we perform a different unsupervised analysis - non-negative matrix factorization - of distal wild-type connectivities, to estimate underlying overall connectivity patterns.

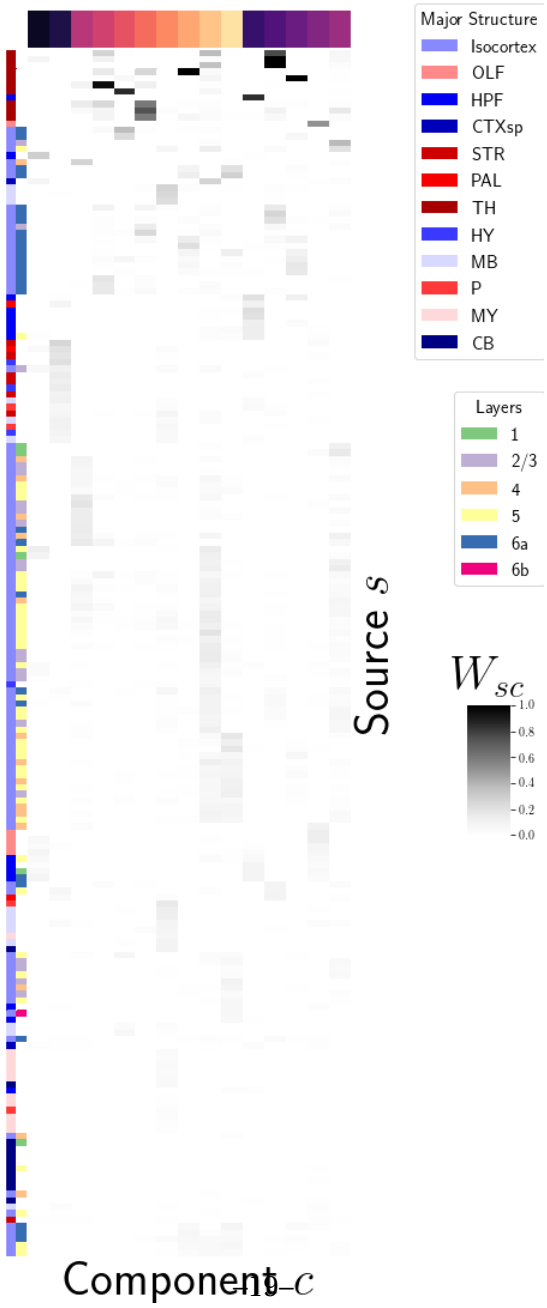
Figure **??** shows a collection of connectivity strengths generated using cre-specific models for wild-type, Cux2, Ntsr1, Rbp4, and Tlx3 cre-lines from visual signal processing leafs in the cortex to cortical and thalymic nucleii. Heirarchical clustering is applied to sort the different source/cre combinations by the similarity of their connectivities to summary-structure targets. This analysis shows that Ntsr1 cre-lines tend to target thalymic nucleii, in particular LP and LD **?**. However, with this exception, for the other plotted cre-lines, connectivity tends to cluster by source structure. That the tendency for structures to connect to themselves is quite strong emphasizes the special nature of the Ntsr1-Thalymic connection in this analysis.

The overall wild-type connectivity strength matrix also displays an underlying modellable structure. As discussed in **?**, one of the most basic processes underlying the observed connectivity is the tendency of each source region to predominantly project to proximal regions. The heatmap in **??**a) shows intraregion distances clearly contains an overall pattern reminscent of the connectivity matrix in **??**. This relationship is plotted in **??** b), showing that there exists substantial variability that would be impossible to model with low-error in a univariate model, even using the diffusion model suggested in **?**. These connections are biologically meaningful, but also unsurprising, and their relative strength biases learned latent coordinate representations away from long-range structures. For this reason, we establish a 1500$\mu m$ 'distal' threshold within which to exclude connections for our analysis. We then apply non-negative matrix factorization (NMF) to decompose the remaining censored matrix into a relatively small number of distinct projection signals, and apply an

unsupervised cross-validation method to select the optimum number of signals (SK's comment:**Percent error... show reconstruction? log scale?**).


Major Structure
Isocortex
OLF
HPF
CTXsp
STR
PAL
TH
HY
MB
P
MY
CB
Layers
1
2/3
4
5
6a
6b
Hemisphere
Ipsi
Contra
Component $c$
$\log(H_{ct})$
Target $t$

(h)


Major Structure
Isocortex
OLF
HPF
CTXsp
STR
PAL
TH
HY
MB
P
MY
CB
Layers
1
2/3
4
5
6a
6b
Source $s$
$W_{sc}$
1.0
0.8
0.6
0.4
0.2
0.0
Component $c$

(i)

## DISCUSSION

Flattening $\mathscr{C}$ prior to unsupervised analysis is not necessarily recommended, but provides an easy solution for this problem.

With respect to the model, a Wasserstein-based measure of injection similarity per structure would combine both the physical simplicity of the centroid model while also incorpating structural knowledge.

The Nadaraya-Watson weighting procedure introduced here is, to our knowledge, novel. In particular, our method of utilizing the expected loss to weight points differs from the minimization task of fitting data to weighted sums of neighbors (**?**). We make a key assumption: that the additional statistical accuracy of including more samples makes up for the fact that their expected accuracy is lower. Note that this assumption can be easily violated, if, for example, the data is distributed on a circle without error, and only nearest neighbors are most predictive.

Model averaging based off of cross-validation has been implemented in **?**, but we note that our approach makes use of a non-parametric estimator, rather than an optimization method for selecting the weights. (SK's comment:**CITE METHOD THAT SELECTS WEIGHTS IN KERNEL (has catchy name)**)

## ACKNOWLEDGMENTS

The Funder and award ID information you input at submission will be introduced by the publisher under a Funding Information head during production. Please use this space for any additional acknowledgements and verbiage required by your funders.

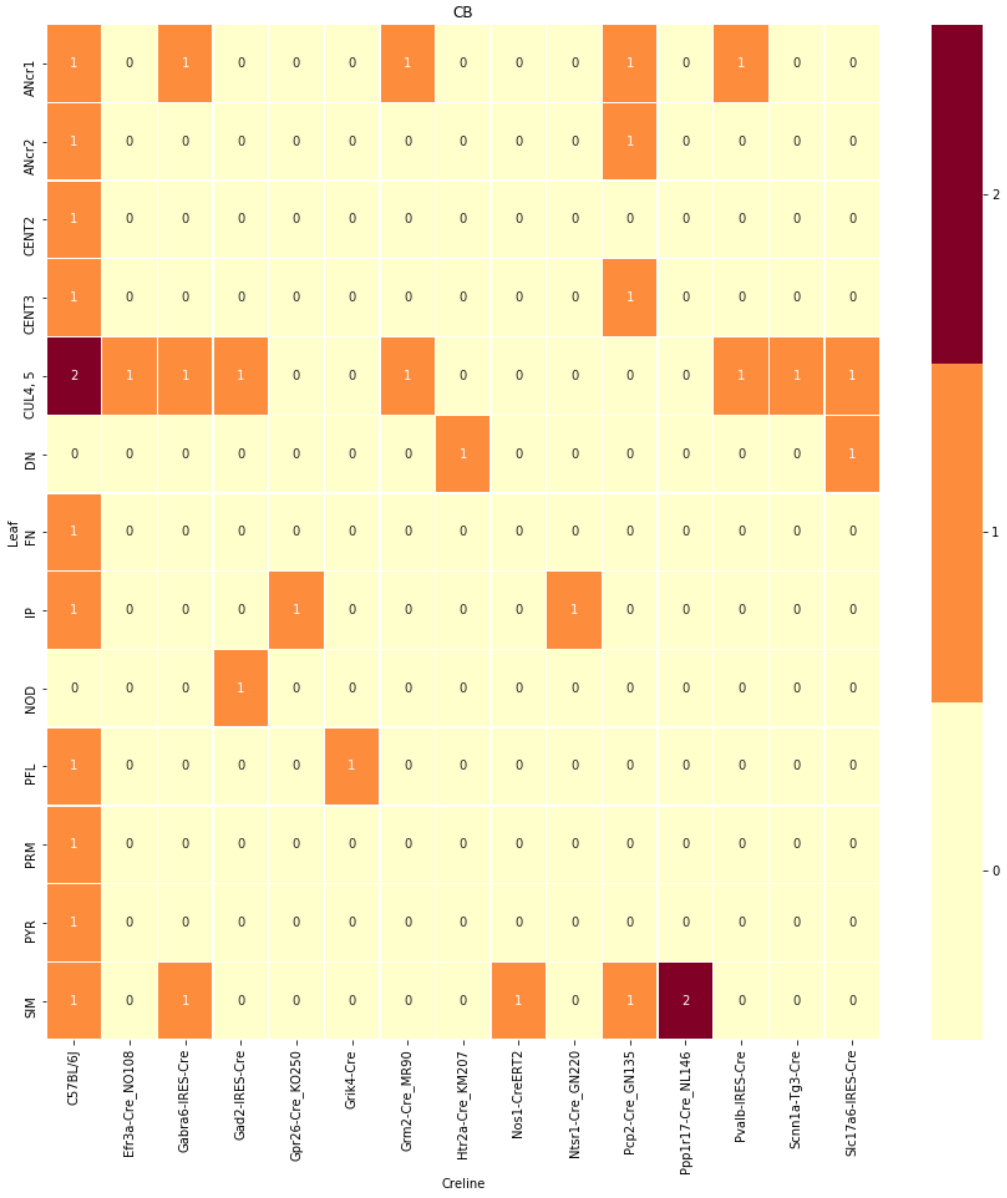
## SUPPORTING INFORMATION

## SUPPLEMENTAL INFORMATION

***Data***

This section describes the set of leaf and cre-line experimental combinations.

centroid densityoct12.png



centroid densityoct12.png

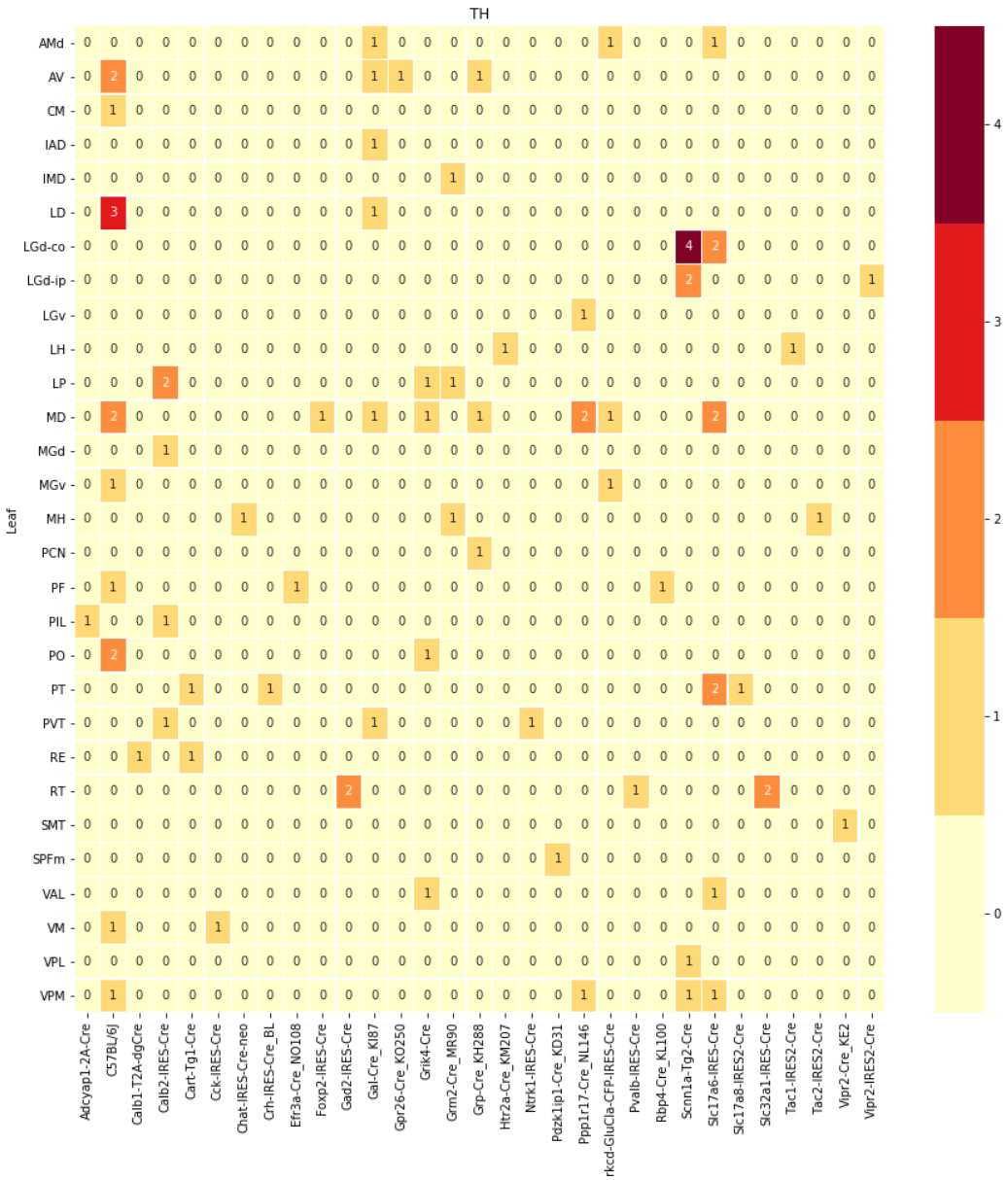



Figure 2: Caption

centroid densityoct12.png

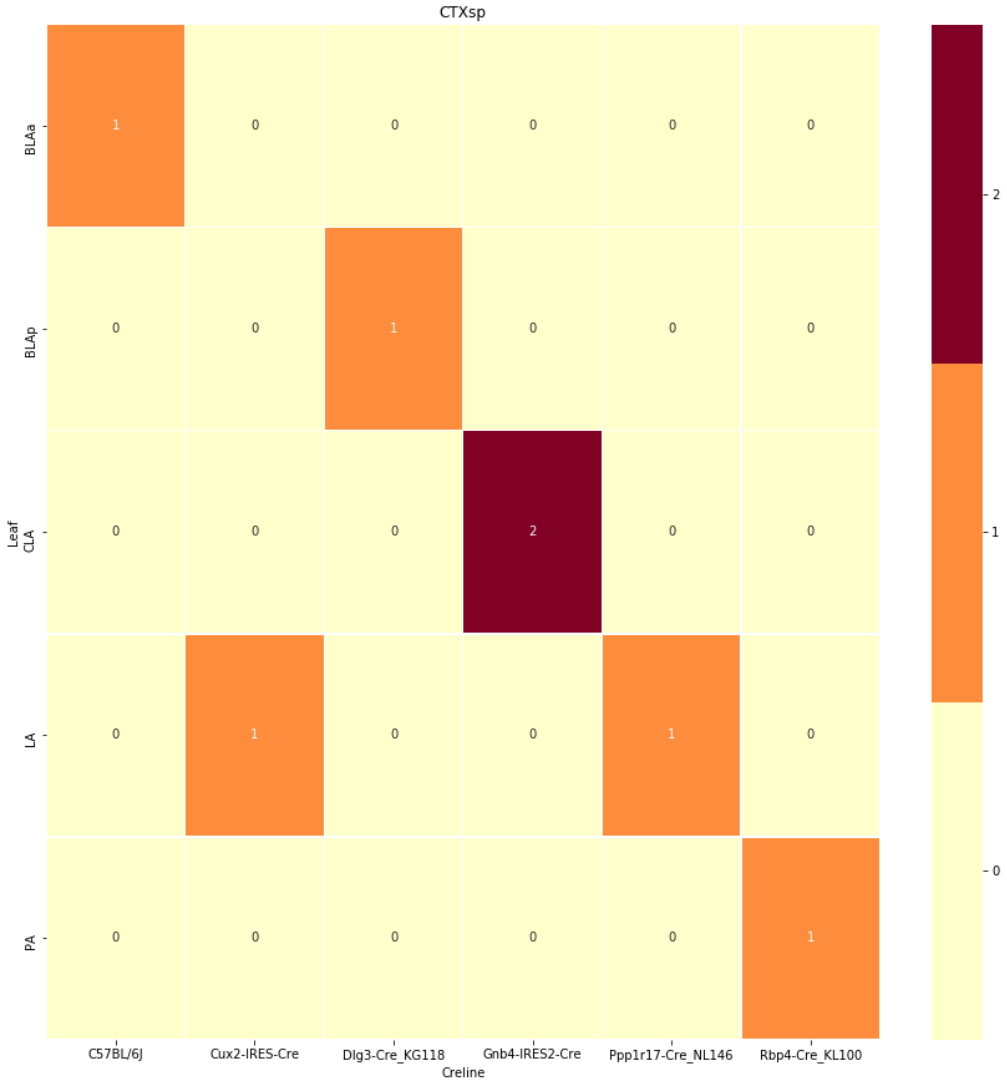
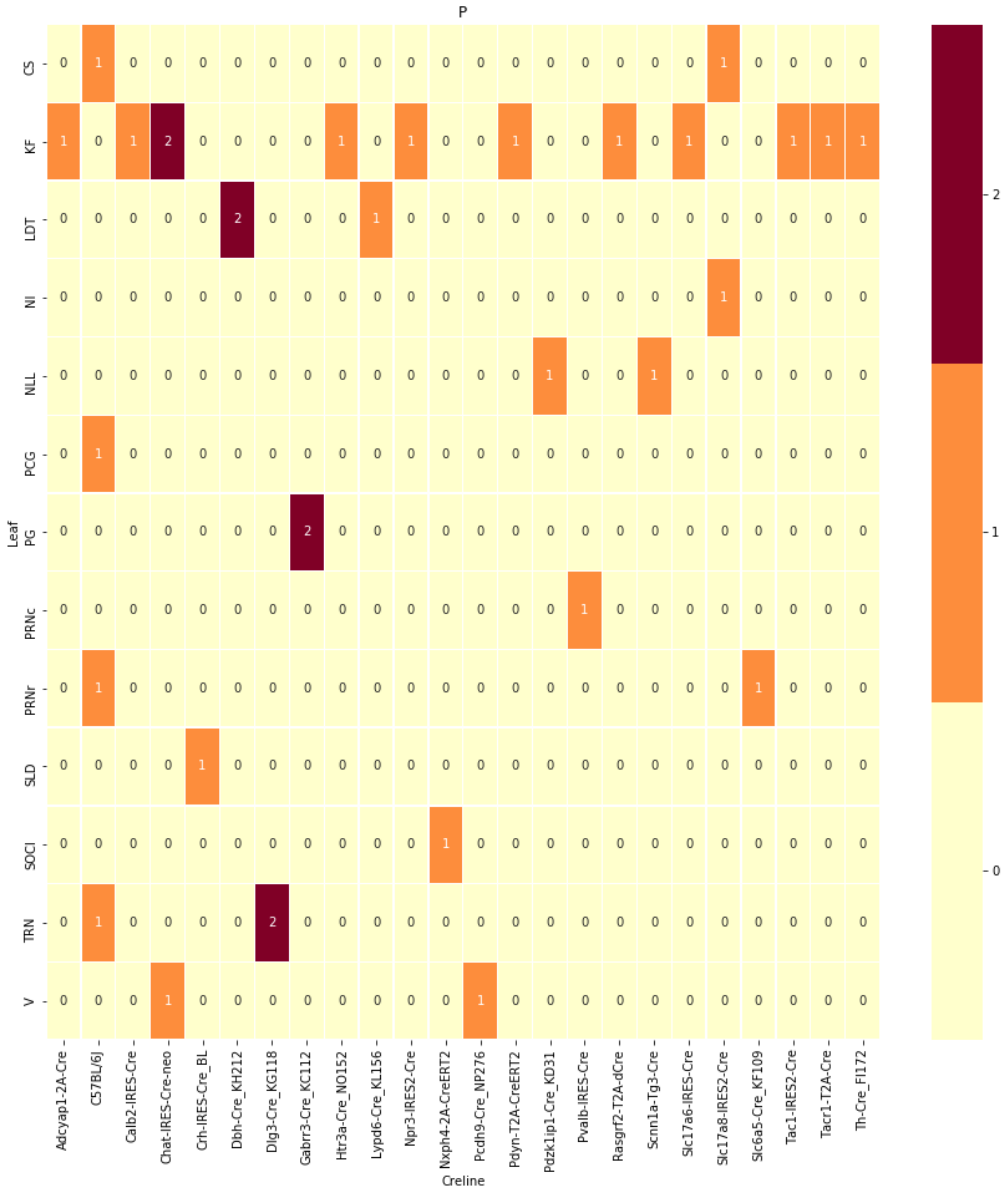
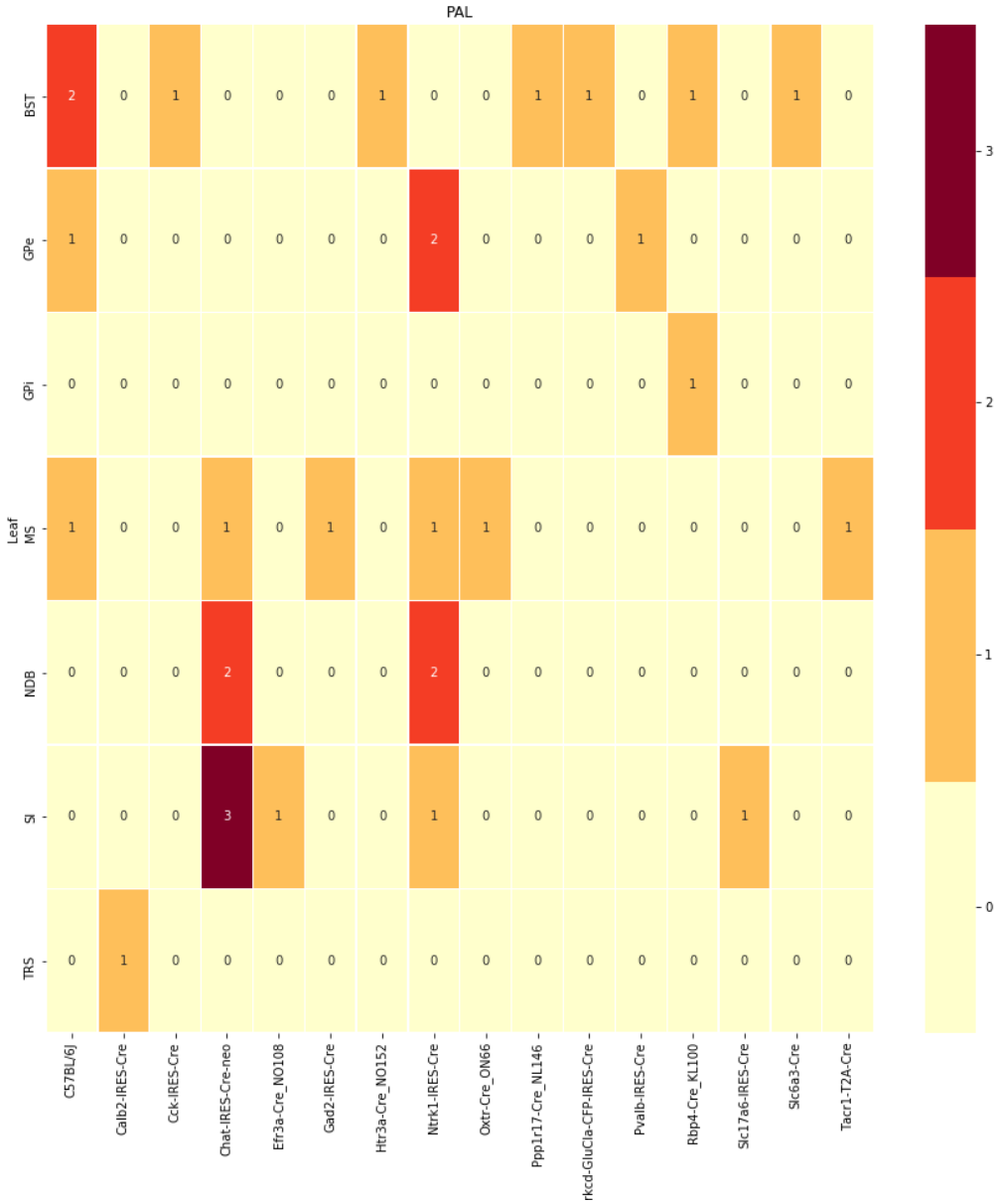



Figure 3: Caption

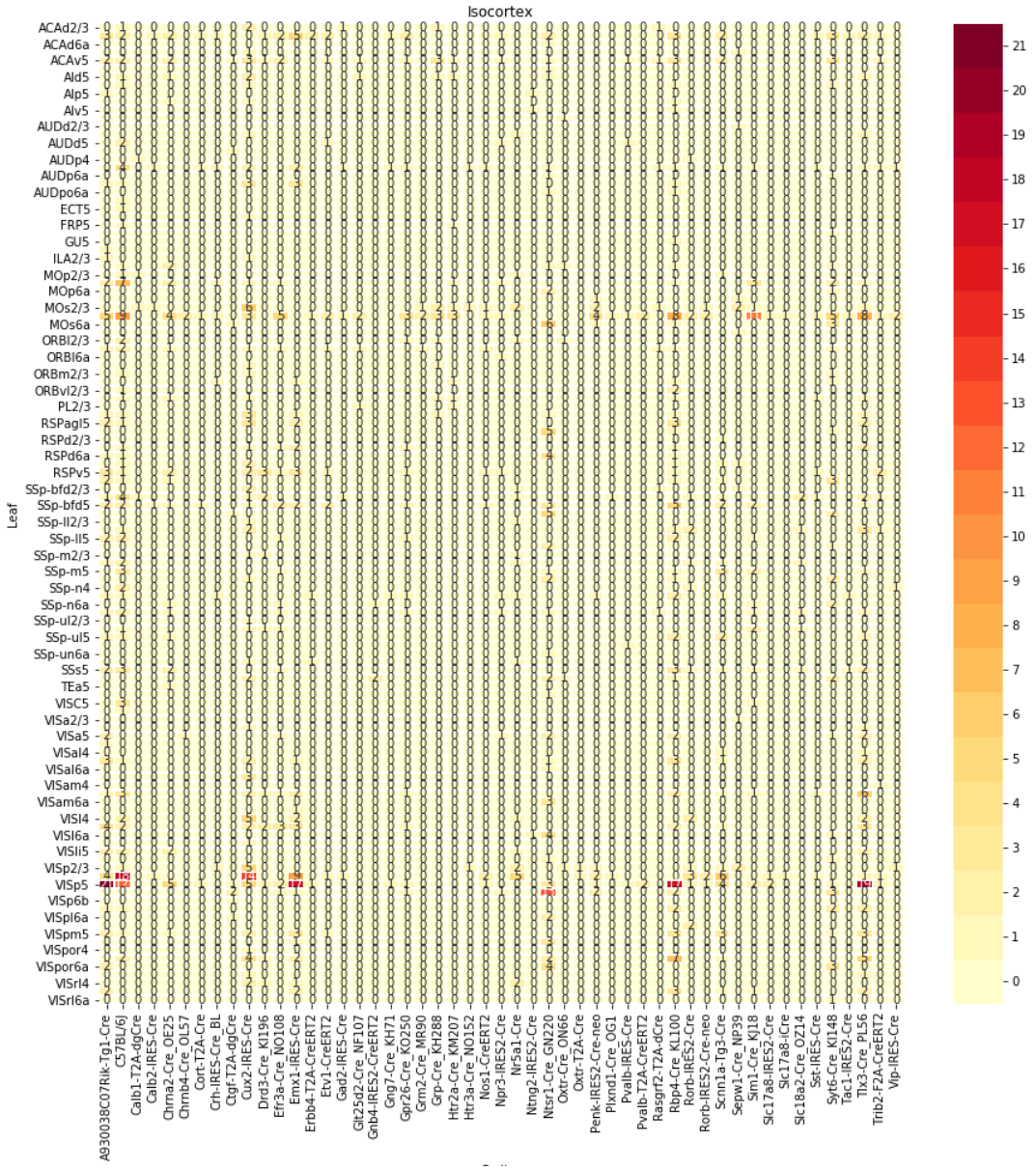
centroid densityoct12.png



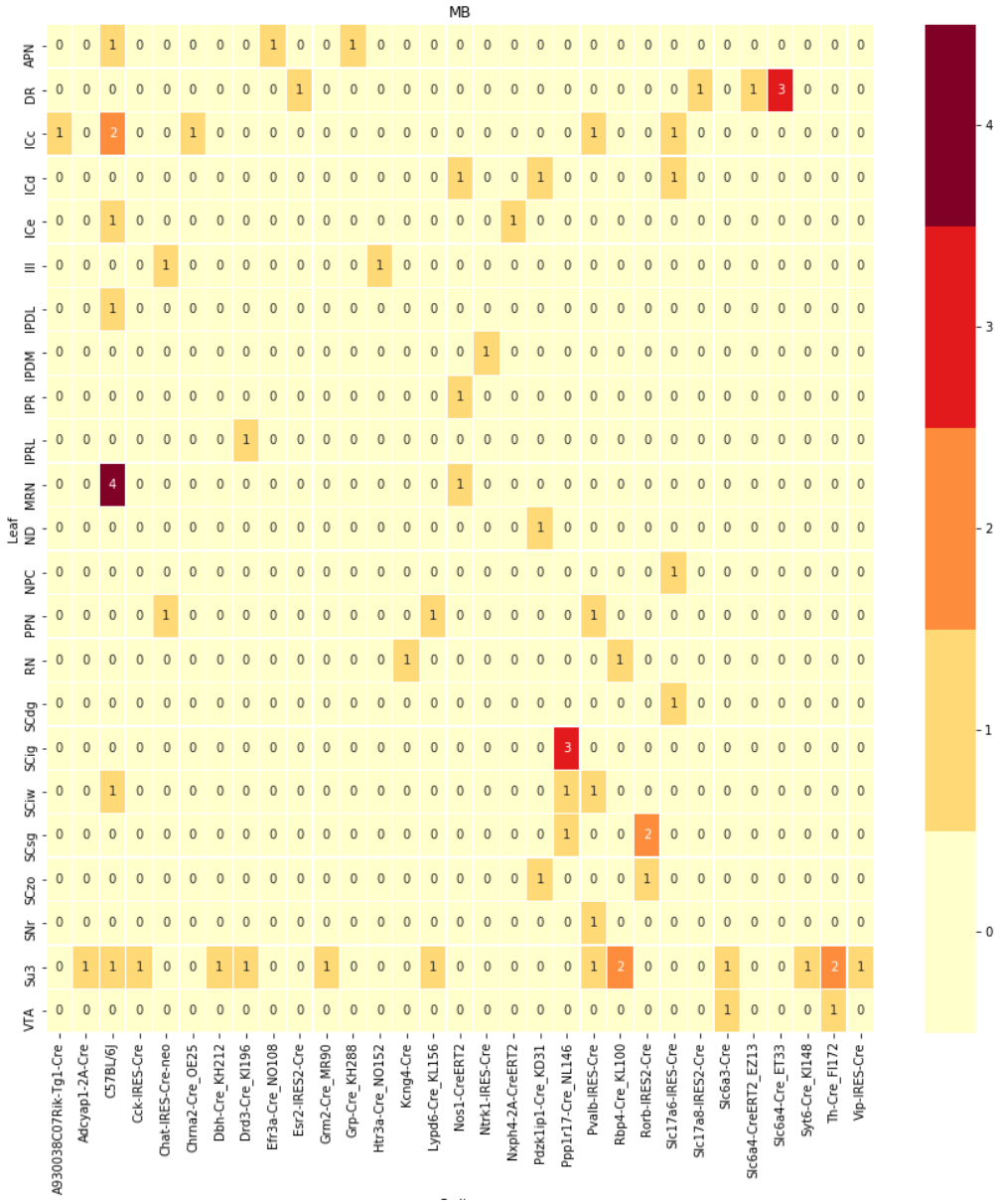
centroid densityoct12.png



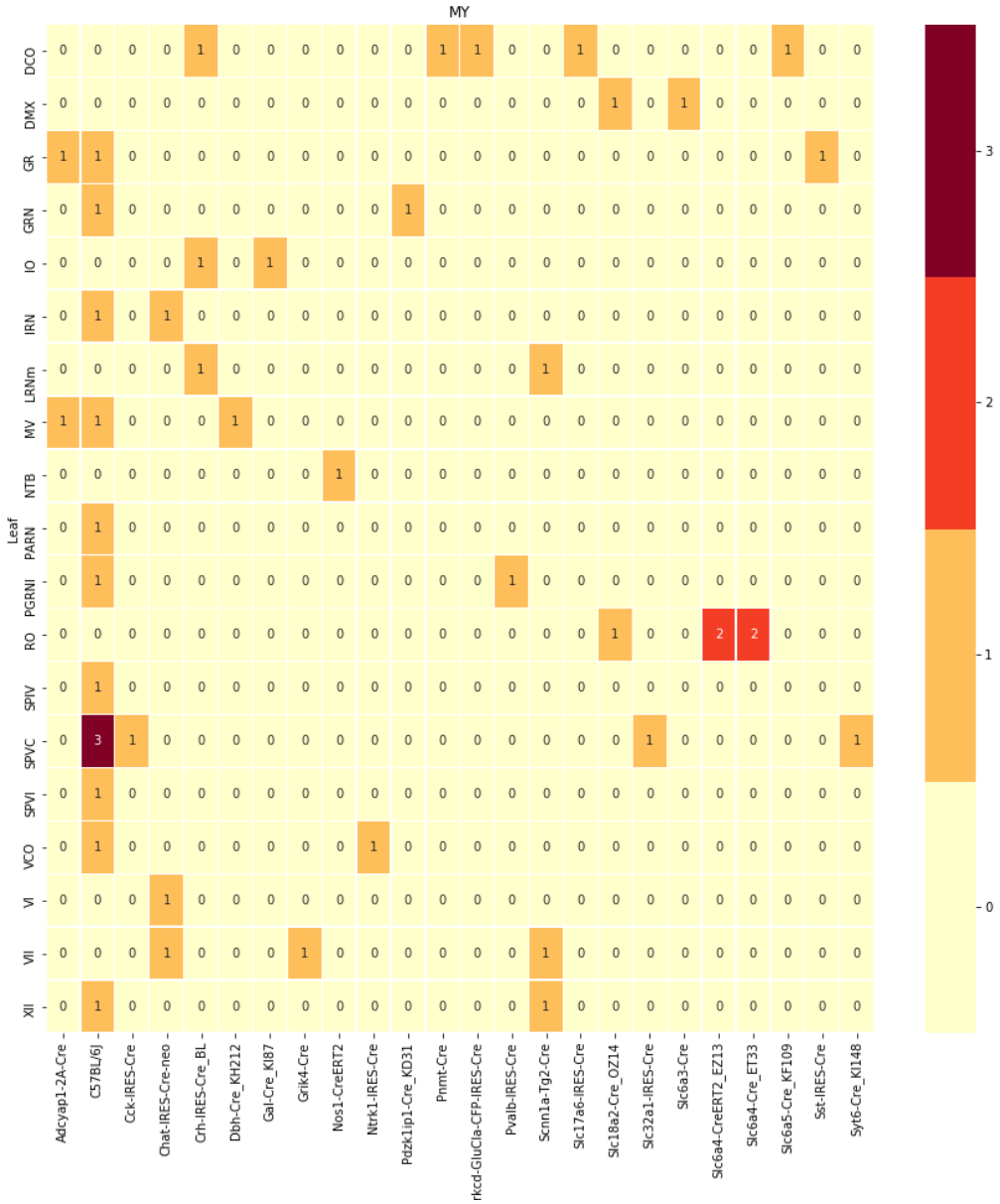
centroid densityoct12.png



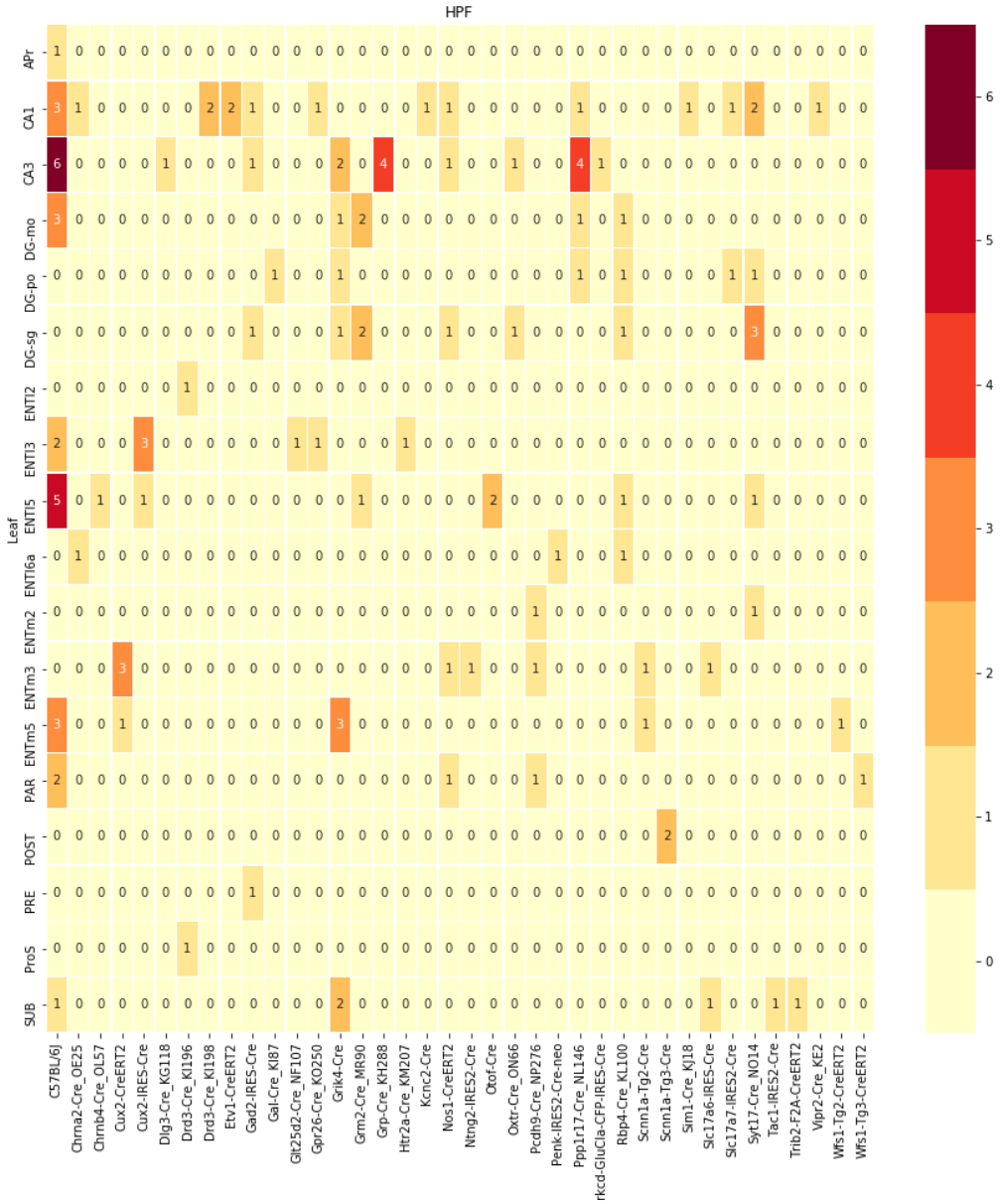
centroid densityoct12.png



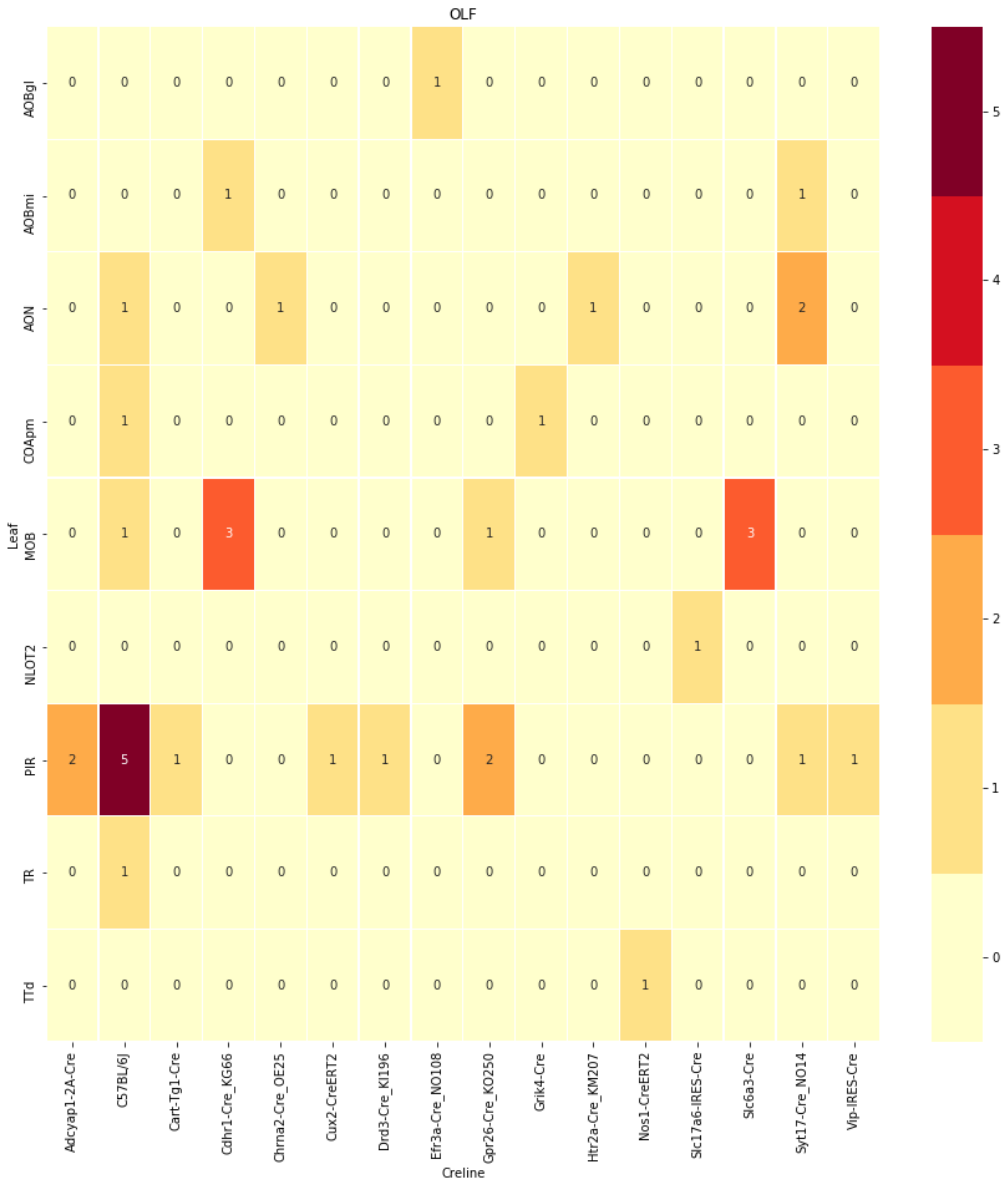
centroid densityoct12.png



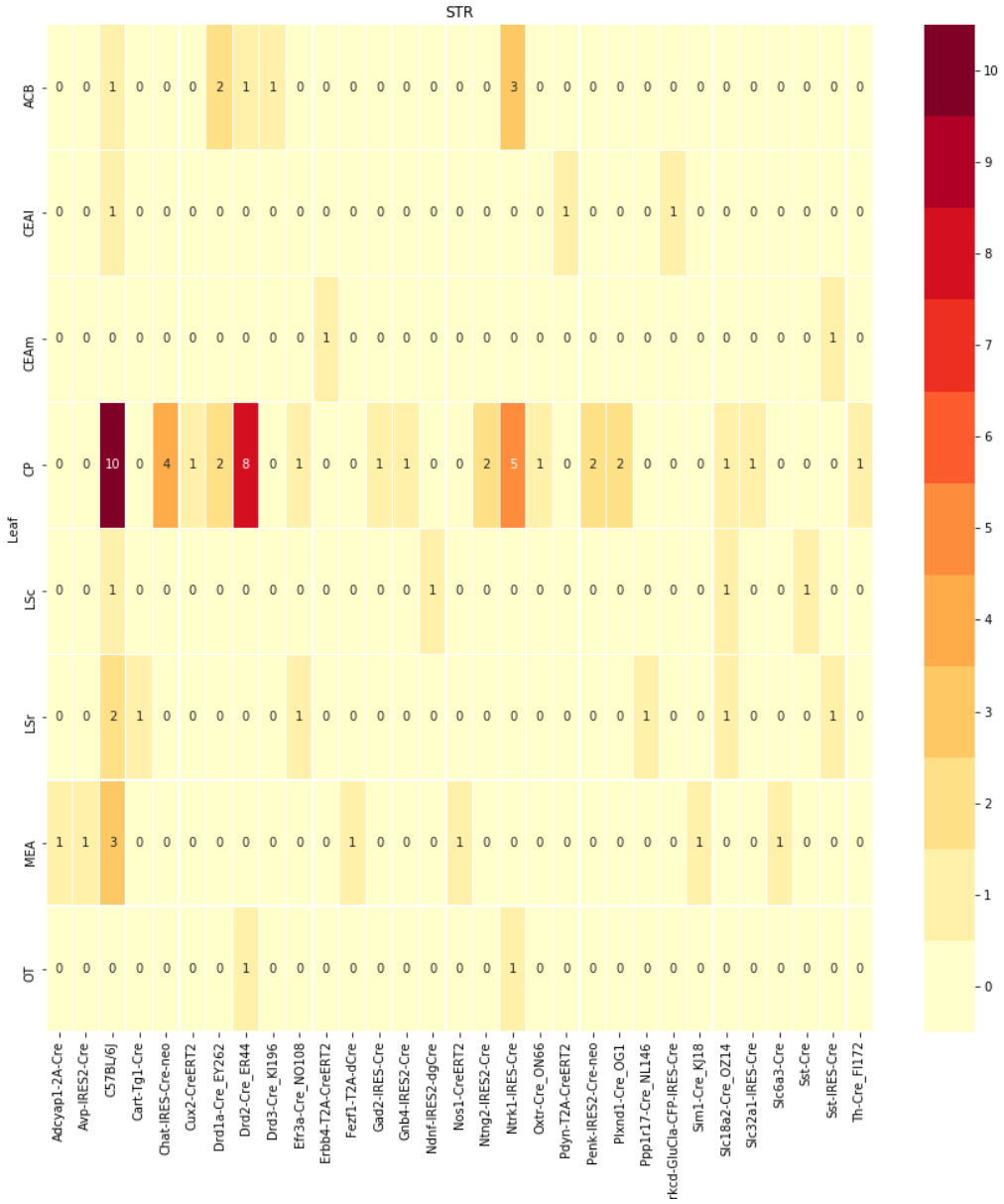
centroid densityoct12.png



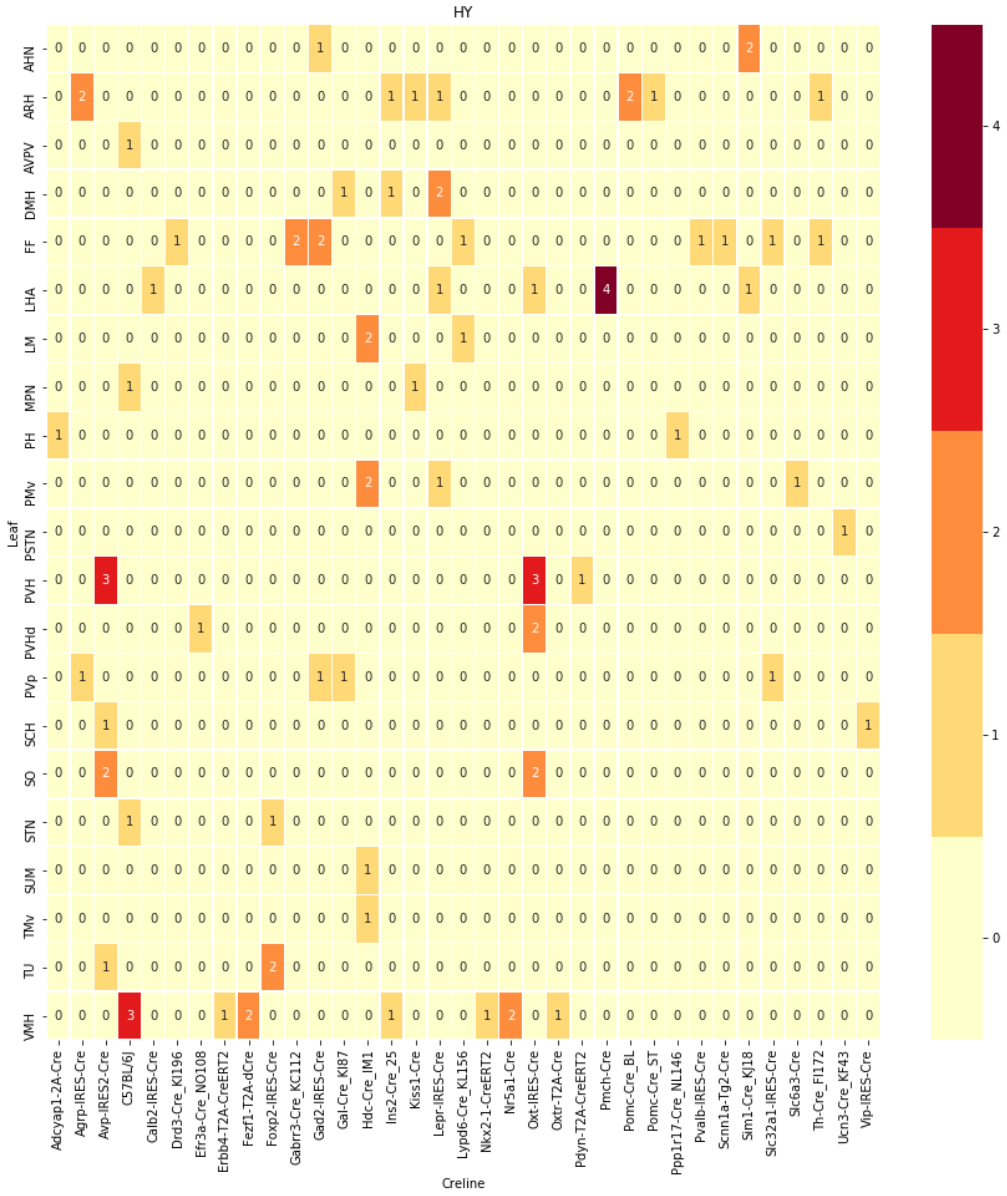
centroid densityoct12.png



centroid densityoct12.png



centroid densityoct12.png



191 ***Structure information***

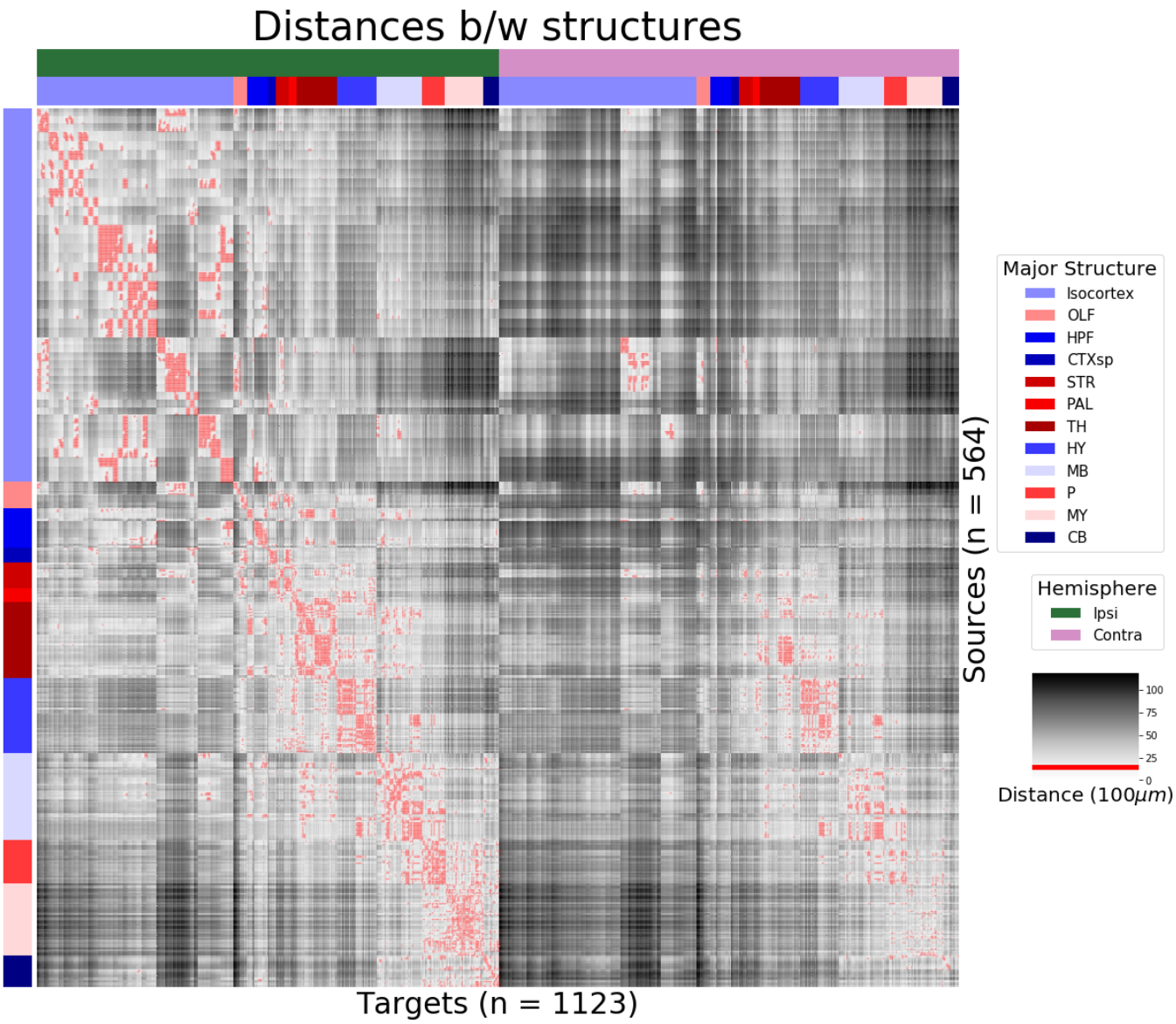



Figure 4: Distance between structures. Short-range connections are masked in red

192 ***Model evaluation***

| | Total | Cre-Summary | Cre-Summary, Leaf | Cre-Leaf |
|---|---|---|---|---|
| 0 | 36 | 10 | 9 | 4 |
| 1 | 7 | 2 | 2 | 2 |
| 2 | 122 | 79 | 79 | 62 |
| 3 | 85 | 41 | 41 | 41 |
| 4 | 1128 | 838 | 829 | 732 |
| 5 | 68 | 23 | 23 | 18 |
| 6 | 46 | 7 | 7 | 7 |
| 7 | 35 | 17 | 17 | 17 |
| 8 | 33 | 8 | 8 | 8 |
| 9 | 30 | 11 | 11 | 11 |
| 10 | 78 | 45 | 45 | 45 |
| 11 | 83 | 29 | 29 | 29 |

Table 3: Number of experiments available to evaluate models in leave-one-out cross validation. Models that rely on a finer granularity of modeling have less data available to validate with.

## SUPPLEMENTAL METHODS

This section consists of additional information on preprocessing of the neural connectivity data, estimation of connectivity, and matrix factorization.

### *Data preprocessing*

Several data prepreprocessing steps take place prior to evaluations of the connectivity matrices. Injections and projections were downloaded using the Allen SDK. These were originally annotated manually. The injection and projection vectors are Hadamard multiplied by a data quality matrix. We also have a map $A : \mathbb{R} \to \mathbb{R}^{|S|}$ where $S$ is the number of structures that takes the average value for voxels in that structure

---

**PREPROCESS 1 Input** Injection $x(i)$, Projection $y(i)$, Injection centroid $c(i) \in \mathbb{R}^3$, injection fraction $F(i)$, data mask $M(i)$

---

Data-quality censor $y_M(i) = M(i) \odot y(i)$

Average structures $y_S(i) = A y_M(i)$

Normalize $\tilde{y}(i) = \frac{y_S(i)}{\|y_S(i)\|}$

**Output** $\tilde{y}(i)$ )

---

The data-quality censor is established by (SK's comment:**fill**) The injection fraction accounts for the relatively coarse graining of the voxel grid compared with the histological analysis used to establish the injection region. In particular, certain voxels are only partially contained within the injection region.

One basic but significant methodological change from **?** is the normalization of projection vectors.

The loss function in **?** is

$$\frac{\|y - \hat{y}\|}{\|y\| \|\hat{y}\|}$$

***Estimators***

Our estimators span a range of training and featurization methods. One commonality is that they model a connectivity vector $f(\mathscr{D}, v, s) \in \mathbb{R}^T$, and so we may write

$$f(v, s, t) = f(v, t)[t].$$

Thus, for the remainder of this section, we will discuss only $f(s, v)$.

*Centroid-based Nadaraya-Watson* In the Nadaraya-Watson approach of **?**, the injection is considered only through its centroid, while the projection is considered regionalized. That is,

$$f_*(\mathscr{D}_i) = \{c(x_i), r(y_i)\}.$$

Since the injection is considered only by its centroid, this model only generates predictions for particular locations $c$, and the prediction for a structure $s$ is given by integrating over locations within the structure

$$f^*(\hat{f}(f_*(\mathscr{D})))(v, s) = \sum_{c \in s} \hat{f}(f_*(\mathscr{D}))(v, c),$$

This $\hat{f}$ is the Nadaraya-Watson estimator

$$\hat{f}_{NW}(c(x_{1:n}), r(y_{1:n}))(c, v) := \sum_{i \in I} \frac{\omega_{c(x_i)c}}{\sum_{i \in I} \omega_{c(x_i),c}} r(y_i)$$

where $\omega_{c(x_i)c} = \exp(-\gamma d(c, c(x_i))^2)$ and $d$ is the Euclidean distance between centroid $c(x_i)$ and voxel $c$. Several facets of the estimator are visible here. A smaller $\gamma$ corresponds to a greater amount of smoothing, and index set $I \subseteq \{1 : n\}$ indicates which experiments to use to generate the prediction. Fitting $\gamma$ via empirical risk minimization therefore bridges between 1-nearest neighbor prediction and averaging of all experiments in $I$. In **?**, $I$ consisted of experiments sharing the same brain division. Restricting of index set to only include experiments with the same neuron class gives the class-specific model.

*The expected-loss estimator* The response induced by each of the cre-lines is effected by both the injection location and the targeted cell types. Cre-lines that target similar cell types are therefore expected to induce similar projections, and including similar cre-lines in our estimator thus increases the effective sample size. In order to leverage this fact in a data-driven way, we introduce an estimator that assigns a predictive weight to each training point that depends both on its centroid-distance and cre-line. This weight is determined by the expected prediction error of each of the two feature types, as determined by cross-validation. These weights are then utilized in a Nadaraya-Watson estimator in a final prediction step.

We formalize cre-line behavior as the average regionalized projection of a cre-line in a given leaf. This vectorization of categorical information is known as target encoding. We define a *c*re-distance in a leaf to be the distance between the target-encoded projections of two cre-lines. The relative predictive accuracy of cre-distance and centroid distance is determined by fitting a surface of projection distance as a function of cre-distance and centroid distance.

In mathematical terms, our full feature set consists of the centroid coordinates and the target-encoded means of the combinations of virus type and injection-centroid structure. That is,

$$f_*(\mathscr{D}_i) = \{c(x_i), \bar{r}(y_{I_v}), r(y_i)\}.$$

$f^*$ is defined as in (2). The expected loss estimator is then

$$\hat{f}_{EL}(c, c(x_i), v, r(y_{I_v})) = \sum_{i \in I} \frac{\nu(c(x_i), c, v_i, v)}{\sum_{i \in I} \nu(c(x_i), c, v_i, v)} r(y_i)$$

where

$$\nu_i = \exp(-\gamma g(d(c, c(x_i))^2, d(\bar{r}(v), \bar{r}(v_i))^2))$$

Note that $g$ must be a concave, non-decreasing function of its arguments with with $g(0,0) = 0$, then $g$ defines a metric on the product of the metric spaces defined by experiment centroid and target-encoded cre-line, and $\hat{f}_{EL}$ is a Nadaraya-Watson estimator. A derivation of this fact is given in Appendix , and we therefore use shape-constrained B-splines to estimate $g$.

This contrasts with the model is **?**, where $\hat{f}(c)$ does not depend on $v$, and **?**, where connectivity was directly estimated by $\hat{f}$ a function of $S$ without an integral. Estimating $\hat{f}(v, c)$ shares the advantage of

fine-scale spatial resolution with **?**, but in addition enables us to model a particular virus-type $v$, and, as we will see, make use of experimental data in our estimator.

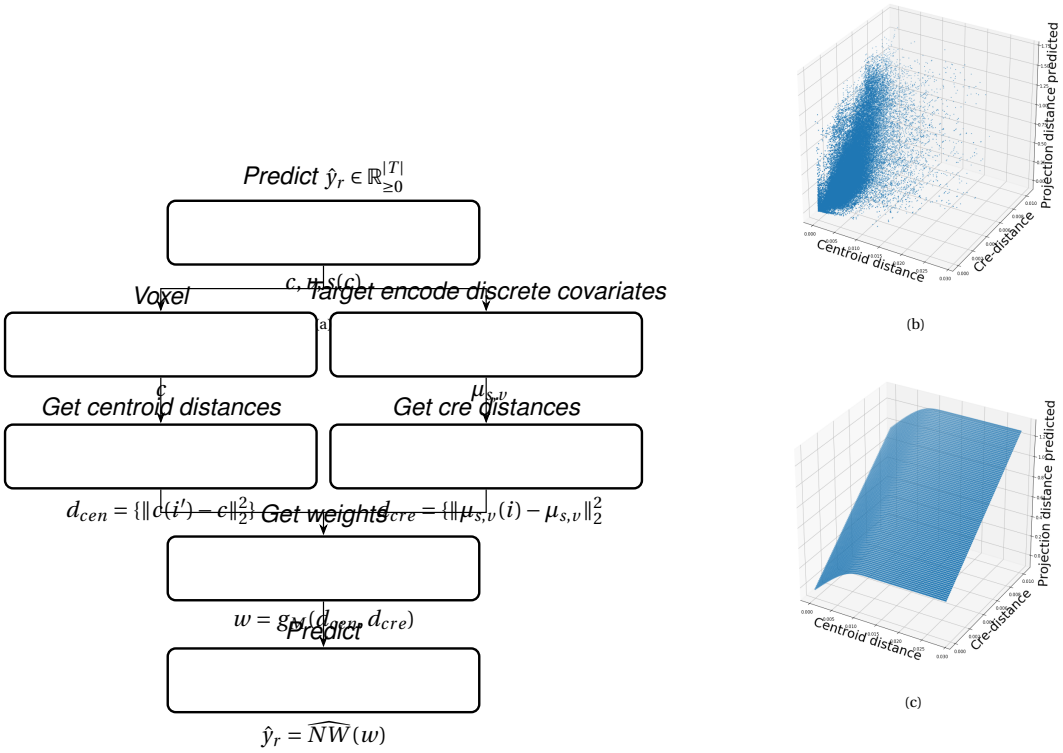



Figure 5: The Expected-Loss estimator

JUSTIFICATION OF SHAPE CONSTRAINT The shape-constrained expected-loss estimator introduced in this paper is, to our knowledge, novel. It should be considered an alternative method to the classic weighted kernel method. While we do not attempt a detailed theoretical study of this estimator, we do establish the need for the shape constraint in our spline estimator. Though this fact is probably well known, we prove a (slightly stronger) version here for completeness.

Given a collection of metric spaces $X_1, ...X_n$ with metrics $d_1...d_n$ (e.g. $d_{centroid}, d_cre$), and a function $f : (X_1 \times X_1)... \times (X_n \times X_n) = g(d_1(X_1 \times X_1), ...d_n(X_n \times X_n))$, then then $f$ is a metric iff $g$ is concave, non-decreasing and $g(d) = 0 \iff d = 0$.

We first show $g$ satisfying the above properties implies that $f$ is a metric.

- The first property of a metric is that $f(x,x') = 0 \iff x = x'$. The left implication: $x = x' \implies f(x_1, x_1', ...x_n, x_n') = g(0, ...., 0)$, since $d$ are metrics. Then, since $g(0) = 0$, we have that $f(x,x') = 0$. The right implication: $f(x,x') = 0 \implies d = 0 \implies x = x'$ since $d$ are metrics.
- The second property of a metric is that $f(x,x') = f(x',x)$. This follows immediately from the symmetry of the $d_i$, i.e. $f(x,x') = f(x_1, x_1', ...x_n, x_n') = g(d_1(x_1, x_1'), ...d_n(x_n, x_n')) = g(d_1(x_1', x_1), ...d_n(x_n', x_n)) = f(x_1', x_1, ...x_n', x_n) = f(x',x)$.
- The third property of a metric is the triangle inequality: $f(x,x') \leq f(x,x^*) + f(x^*,x')$. To show this is satisfied for such a $g$, we first note that $f(x,x') = g(d(x,x')) \leq g(d(x,x^*) + d(x^*,x'))$ since g is non-decreasing and by the triangle inequality of $d$. Then, since $g$ is concave, $g(d(x,x^*) + d(x^*,x')) \leq g(d(x,x^*)) + g(d(x^*,x')) = f(x,x^*) + f(x^*,x')$.

We then show that $f$ being a metric implies that $g$ satisfies the above properties.

- The first property is that $g(d) = 0 \iff d = 0$. We first show the right implication: $g(d) = 0$, and $g(d) = f(x,x')$, so $x = x'$ (since $f$ is a metric), so $d = 0$. We then show the left implication: $d = 0 \implies x = x'$, since $d$ is a metric, so $f(x,x') = 0$, since $f$ is a metric, and thus $g(d) = 0$.
- The second property is that $g$ is non-decreasing. We proceed by contradiction. Suppose g is decreasing in argument $d_1$ in some region $[l,u]$ with $0 < l < u$. Then $g(d_1(0,l),0) \geq g(d_1(0,0),0) + g(d_1(0,u),0) = g(d_1(0,u),0)$, which violates the triangle inequality on f. Thus, decreasing $g$ means that $f$ is not a metric, so $f$ a metric implies non-decreasing $g$.
- The final property is that $g$ is concave. We proceed by contradiction. Suppose $g$ is strictly convex. Then there exist vectors $d, d'$ such that $g(d+d') < g(d) + g(d')$. Assume that $d$ and $d'$ only are non-zero in the first position, and $d = d(0,x), d' = d(0,x')$. Then, $f(0,x) + f(0,x') < f(0,x+x')$, which violates the triangle inequality on $f$. Therefore, $g$ must be concave.

*Establishing a lower detection limit* The lower detection limit of our approach is a complicated consequence of our experimental and analytical protocols. For example, the Nadaraya-Watson estimator is likely to generate many small false positive connections, since the projection of even a single experiment within the source region to a target will cause a non-zero connectivity in the Nadaraya-Watson weighted average. On the other hand, the complexities of the experimental

protocol itself and the image analysis and alignment can also cause spurious signals. Therefore, it is of interest to establish a lower-detection threshold below which we have very little power-to-predict, and set estimated connectivities below this threshold to zero. This should make our estimated connectivities more accurate, especially in the biologically-important sense of sparsity.

We establish this limit with respect to the sum of Type 1 and Type 2 errors

$$\tau = \sum 1_{f(s,t,c)=0} 1_{\hat{f}(s,t,c)>0} + 1_{f(s,t,c)>0} 1_{\hat{f}(s,t,c)=0}.$$

***Decomposing the connectivity matrix***

We utilize non-negative matrix factorization (NMF) to analyze the principal signals in our connectivity matrix. Here, we review this approach as applied to decomposition of the distal elements of the estimated connectivity matrix $\widehat{\mathscr{C}}$ to identify $q$ connectivity archetypes. Aside from the NMF program itself, the key elements are selection of the number of archetypes $q$ and stabilization of the tendency of NMF to give random results over different initialization.

*Non-negative matrix factorization* Given a matrix $X \in \mathbb{R}^{a\times b}_{\geq 0}$ and a desired latent space dimension $q$, the non-negative matrix factorization is

$$NMF(X,q) = \arg \min_{W\in\mathbb{R}^{a\times q}_{\geq 0}, H^{q\times b}_{\geq 0}} \|(X - WH)\|_2^2.$$

NMF creates a useful decomposition since $X$ is in the positive orthant, and PCA cannot not apply. There is no orthogonality without sparsity.

We note the existence of NMF with alternative norms for certain marginal distributions, but leave utilization of this approach for future work (**?**). We can also apply a mask $1_M \in \mathbb{R}^{S\times T}$ of ones and zeros and solve

$$\arg \min_{W\in\mathbb{R}_{\geq 0}, H_{\geq 0}} \|1_M \odot ((\hat{\mathscr{C}} - WH))\|_2^2$$

For us, such a mask serves for two purposes. First, it enables computation of the NMF objective while excluding self and nearby connections. These connections are both strong and linearly independent, and so would dominate the $NMF$ reconstruction error. Long range connections are more biologically

interesting or cell-type dependent. Second, it enables cross-validation based selection of the number of retained components.

*Cross-validating NMF* Perhaps surprisingly, cross-validation techniques may also be applied to unsupervised learning problems. These techniques are somewhat standard, but not entirely well-known, so we review them here, in particular as they apply to the NMF problem. A NMF model is first fit on a reduced data set, and an evaluation set is held out. After random masking of the evaluation set, the loss of the learned model is then evaluated on the basis of successful reconstruction of the held-out values. This procedure is performed repeatedly, with different held out regions and random mask at different dimensionalities $l$, to determine to point past which additional hidden units provide no reconstructive value.

That is, given a matrix $X \in \mathbb{R}^{S\times T}$ we can decompose $X \sim d(e(X))$ where $e(X)$ is some map that encodes $X$ in a learned representation, and $d$ is the decoding reconstruction map. In our case, $d$ is simply left multiplication by $W$, and $e$ is the solution of a regularized non-negative least squares optimization problem

$$H := e_W(X) = \arg\min_{\beta} \|X - W\beta\|_2^2.$$

The form of this solution particularly motivates our cross-validation estimator.

Recall that in supervised learning, the learned model is $Y \sim f(X)$. Standard cross-validation removes elements of $X$, fits $f$, and then uses the $f$ learned from part of the data to predict $Y$. A good $f$ will have low error on the training data, and also low error on the test data, indicating that it has not overfit. Although there is no assumed dichotomy between $X$ and $Y$ in unsupervised learning, for techniques like autoencoders, the above paradigm still applies, i.e., one can still hold out values of $X$. We can then estimate

$$\arg\min_{d,e} \widehat{E}(l(X, d_{X^C}(e_{X^C}(X)))) = \sum_{r=1}^{R} l(X_r, d_{X^{C_r}}(e_{X^{C_r}}(X_r)))$$

over $R$ random samples of rows of $X$. However, in our setting, since computing $e(X)$ on the test rows amounts to fitting a non-negative least squares w.r.t. $W$, so the negative effects of an overfit model can simply be optimized away from. Thus, the standard solution is to generate uniformly random masks

$1_{M(p)} \in \mathbb{R}^{S \times T}$ where

$$1_{M(p)}(s,t) \sim \text{Bernoulli(p)}.$$

Our cross-validation error is then

$$\epsilon_q = \frac{1}{R}\sum_{r=1}^{R}(\|1_{M(p)_r^C} \odot X - \widehat{d}_q(\widehat{e}_q(1_{M(p)_r^C} \odot X))\|_2^2$$

where

$$\widehat{d}_q, \widehat{e}_q = \widehat{\text{NMF}}(1_{M(p)_r} \odot X, q)).$$

The optimum number of components is then

$$\widehat{q} = \arg\min_q \epsilon_q.$$

*Stabilizing NMF* The NMF program is non-convex, and, empirically, individual replicates will not converge to the same optima. One solution therefore is to run multiple replicates of the NMF algorithm, cluster the resulting vectors. This approach raises the questions of how many clusters to use, and how to deal with stochasticity in the clustering algorithm itself. We address this issue through the notion of clustering stability (**?**).

The clustering stability approach is to generate $L$ replicas of k-cluster partitions $\{C_{kl} : l \in 1 \dots L\}$ and then compute the average dissimilarity between clusterings

$$\xi_k = \frac{2}{L(L-1)}\sum_{l=1}^{L}\sum_{l'=1}^{l} d(C_{kl}, C_{kl'}).$$

Then, the optimum number of clusters is

$$\hat{k} = \arg\min_k \xi_k.$$

A review of this approach is found in **?**. Intuitively, archetype vectors that cluster together frequently over clustering replicates indicate the presence of a stable clustering. For $d$, we utilize the adjusted Rand Index - a simple dissimilarity measure between clusterings. Note that we expect to select slightly more than the $q$ components suggested by cross-validation, since archetype vectors which appear in one NMF replicate generally should appear in others. We then select the $q$ clusters with the most

archetype vectors - the most stable NMF results - and take the median of each cluster to create a sparse representative archetype.

# SUPPLEMENTAL EXPERIMENTS

## *Establishing a lower limit of detection*

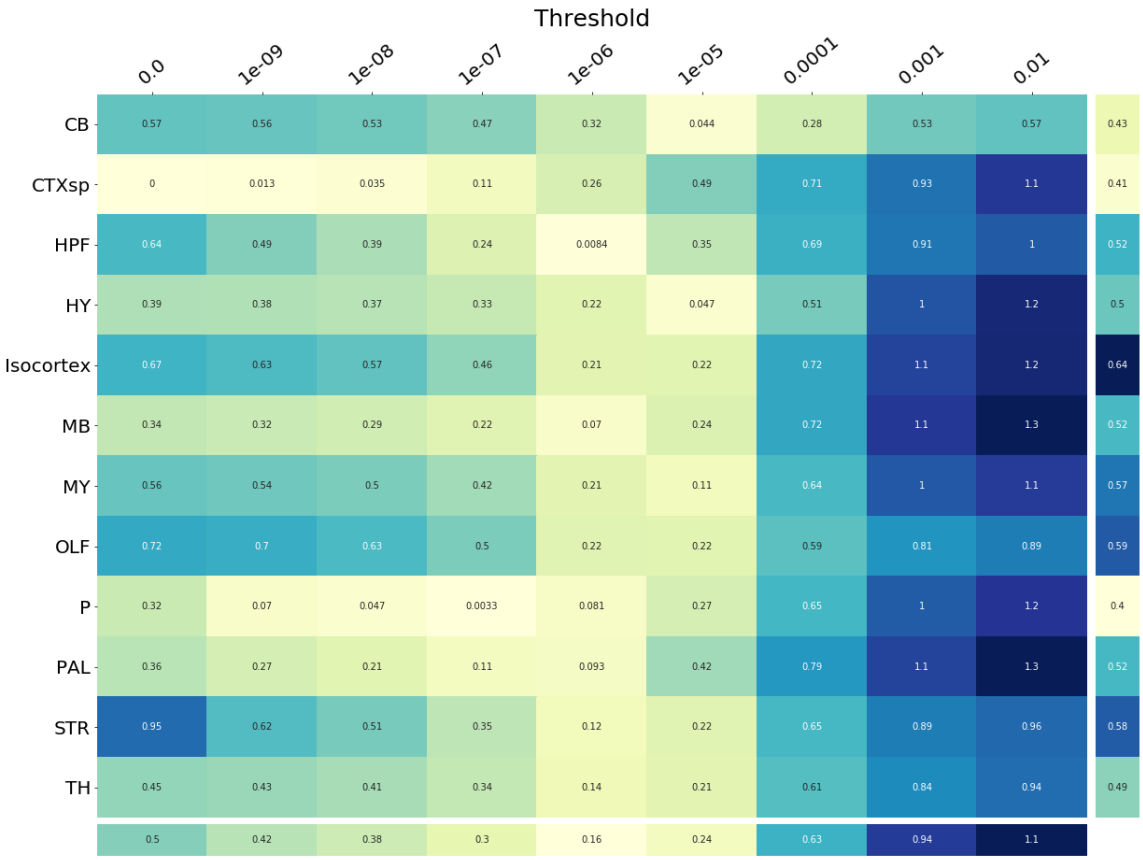



Figure 6: $\tau$ at different limits of detection.

## *Loss subsets*

The

## *Matrix Factorization*

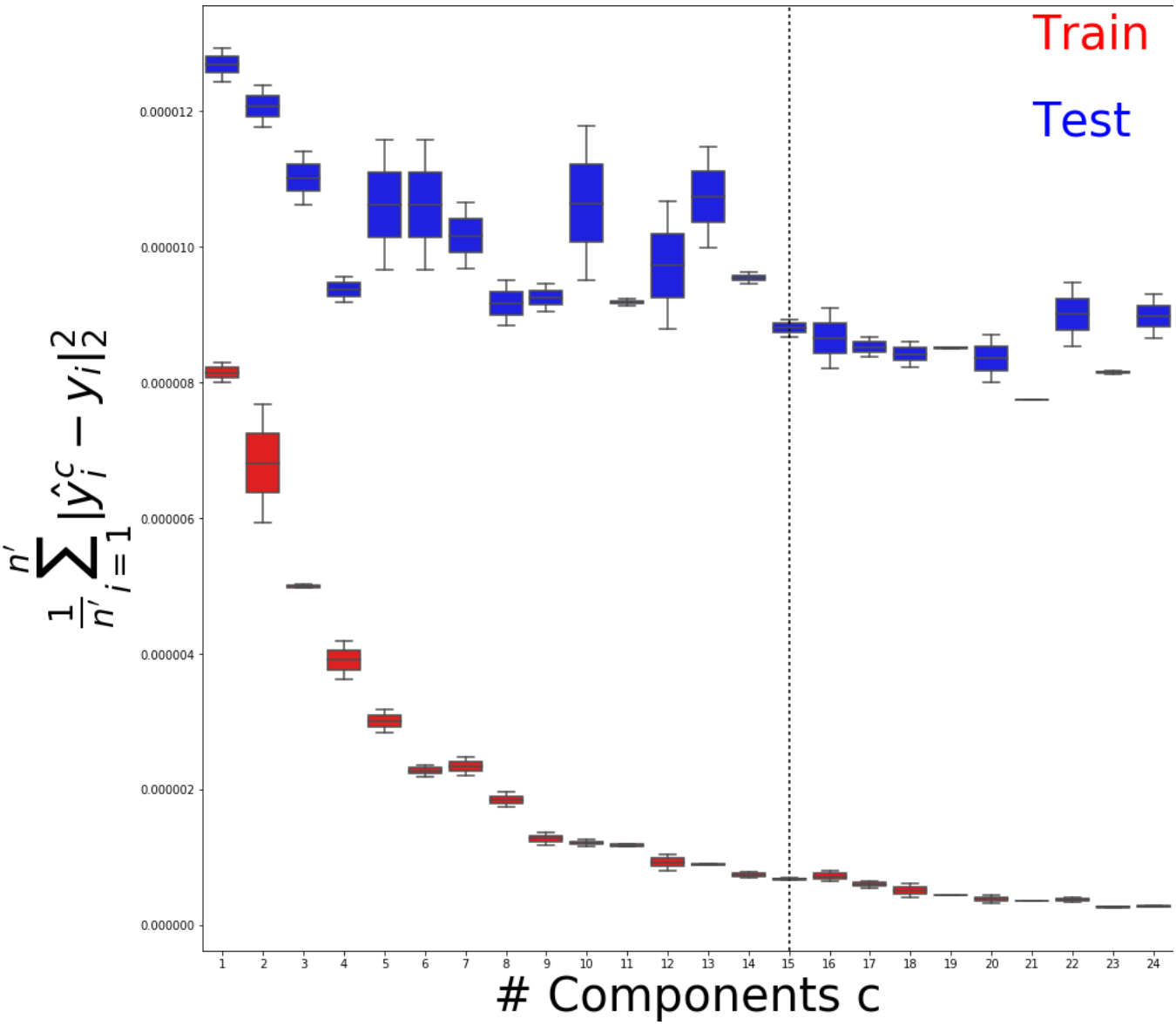



Figure 7: Train and test error across 2 (SK's comment:**increase**) replicates using NMF decomposition.

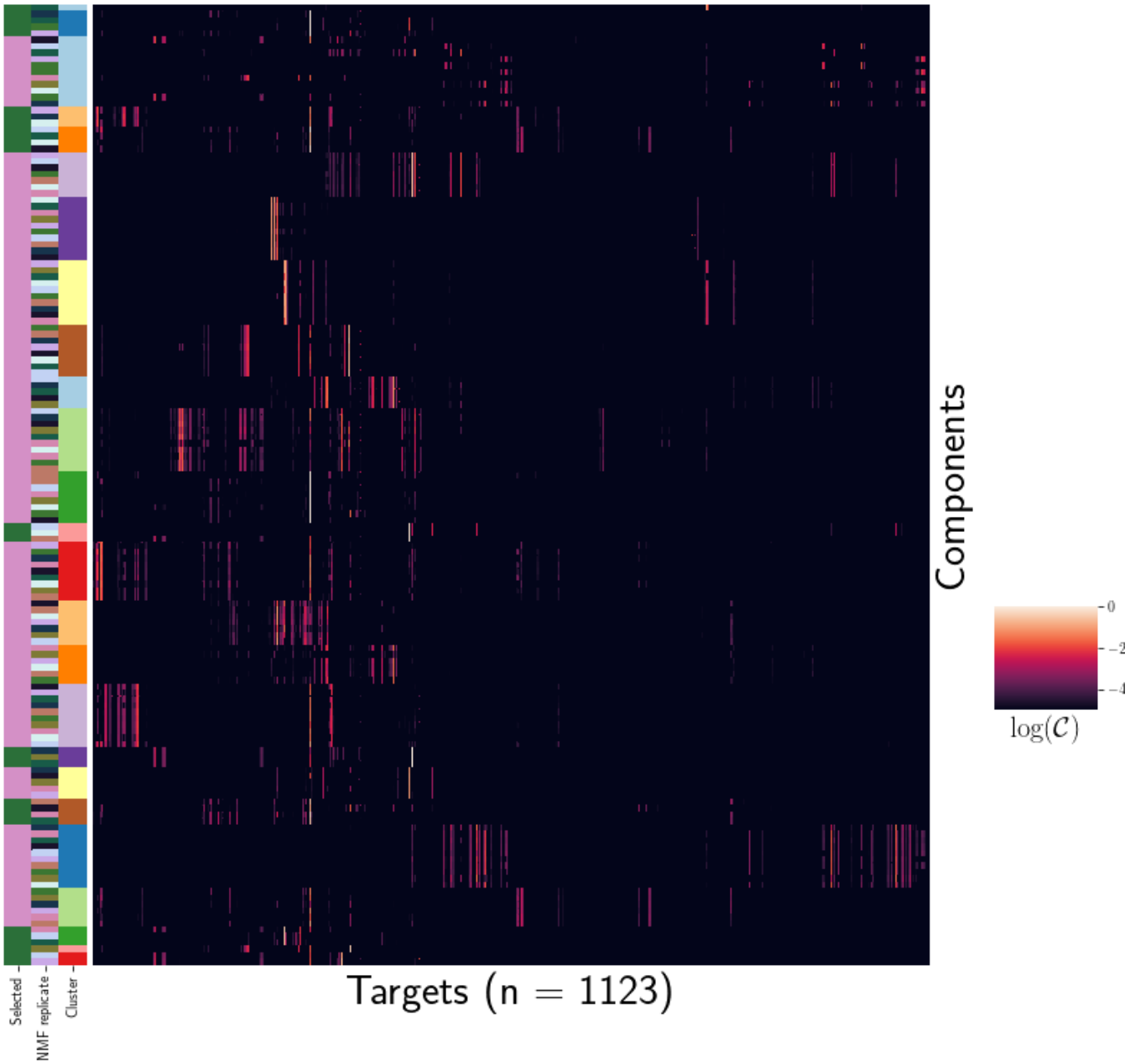



Figure 8: Stability of NMF results across replicates.

## COMPETING INTERESTS

This is an optional section. If you declared a conflict of interest when you submitted your manuscript, please use this space to provide details about this conflict.

## TECHNICAL TERMS

All NETN article types require Technical Terms.

Identify approximately 10 key terms that are mentioned in your article and whose usage and definition may not be familiar across the broad readership of the journal. Provide brief (20-word or less) definitions for each term, avoiding in these definitions the use of jargon, or highly technical or specialized language. When the article is typeset, the Technical Terms will appear in the margins at or near their first mention in the text.

In your manuscript, bold the first occurrence of each **Technical Term** and then provide a list of the terms and their definitions at the end of the manuscript after the references.

**Technical Term** a key term that is mentioned in an NETN article and whose usage and definition may not be familiar across the broad readership of the journal.