UNIVERSITÀ
DEGLI STUDI
FIRENZE

BROAD
INSTITUTE

Machine Learning techniques for Genomics

Lorenzo Buffoni

Machine Learning at GGI

2022

Acknowledgements

Mapping algorithm
- Gabriele Scalia

Brain image latent space
- Lorenzo Buffoni

Semantic segmentation
- Ziqing Lu
- Aman Sanger

Data analysis
- Raghav Avasthi
- Neriman Tokcan

Data generation
- Asa Segerstolpe
- Naeem Nadaf
- Inbal Avraham-Davidi
- Chuck Vanderburg
- Meng Zhang
- Xiaowei Zhuang

Various help
- Mor Nitzan
- Nik Brown
- Duccio Fanelli
- Sanja Vicovic

Graphics and Artwork
- Leslie Gaffney
- Anna Hupalowska

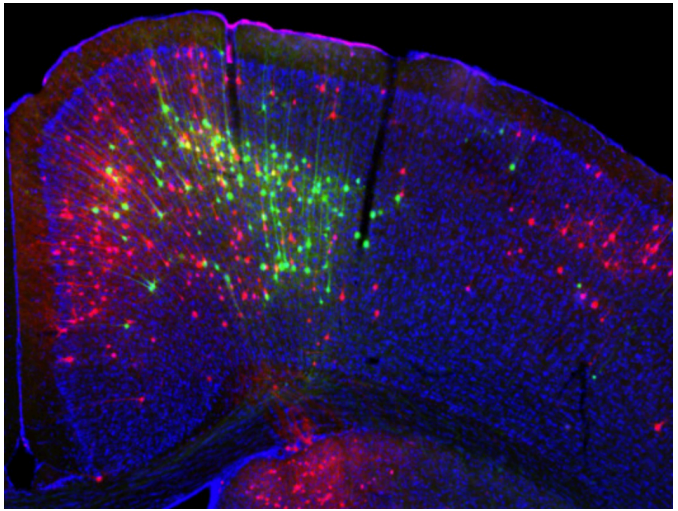
Mouse Melanoma
- Eran Hodis
- Elena Torlai Triglia
- Saurabh Parkar

Lab Leaders
- Aviv Regev
- Evan Macosko

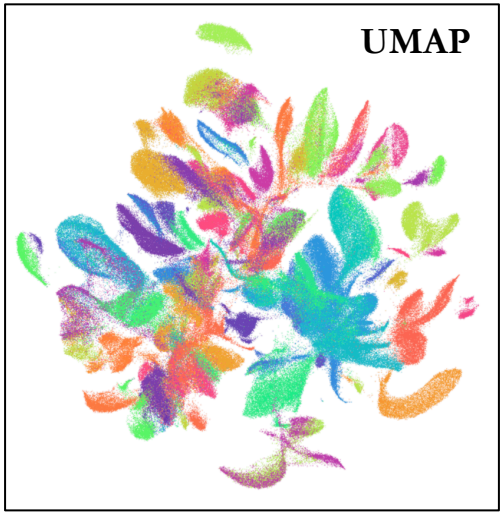
Project Leader
- Tommaso Biancalani

We are building a Human Cell Atlas

Source: *HubMAP*



Source: *BICCN*



Source: *CZI*

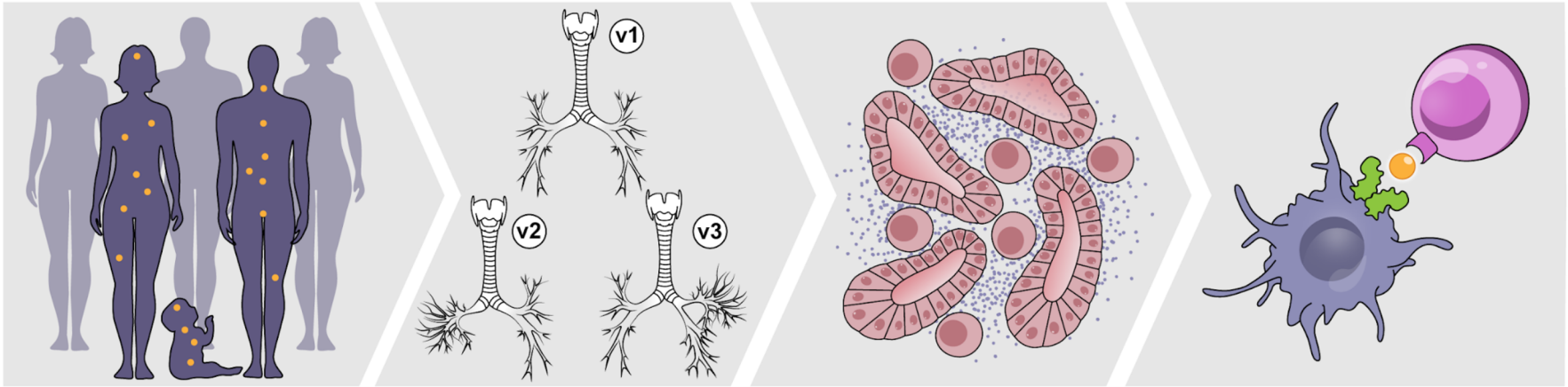
BICCN

CHAN ZUCKERBERG INITIATIVE

HuBMAP
Human BioMolecular Atlas Program

HUMAN CELL ATLAS

Biological atlases require integration of diverse datasets at different scales

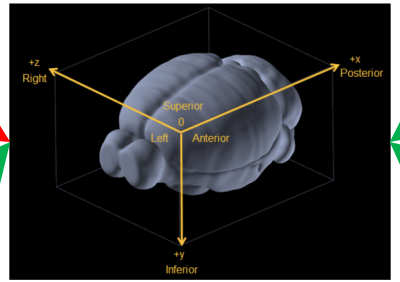


Macro Meso Histology Cellular

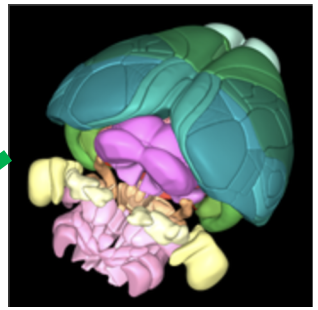
The mouse brain atlas sets the paradigm for biological atlases

Single cell data
(eg *Macosko/Regev*)

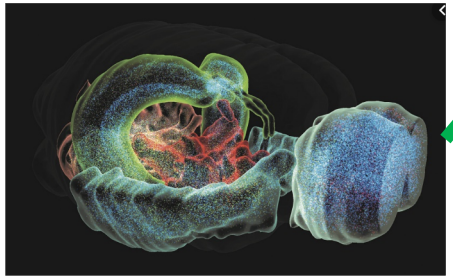
?



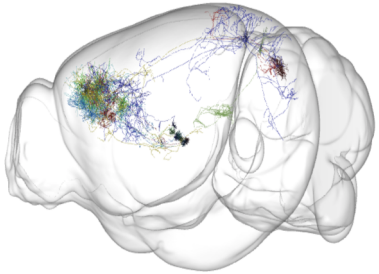
Common Coordinate Framework



Ontology
(eg *Allen*)



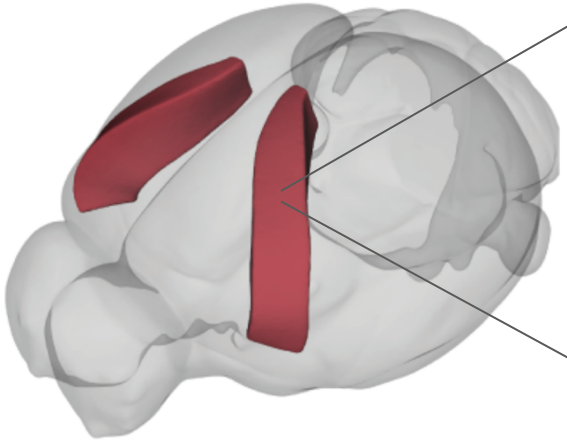
Cell positions (eg *Blue brain)*

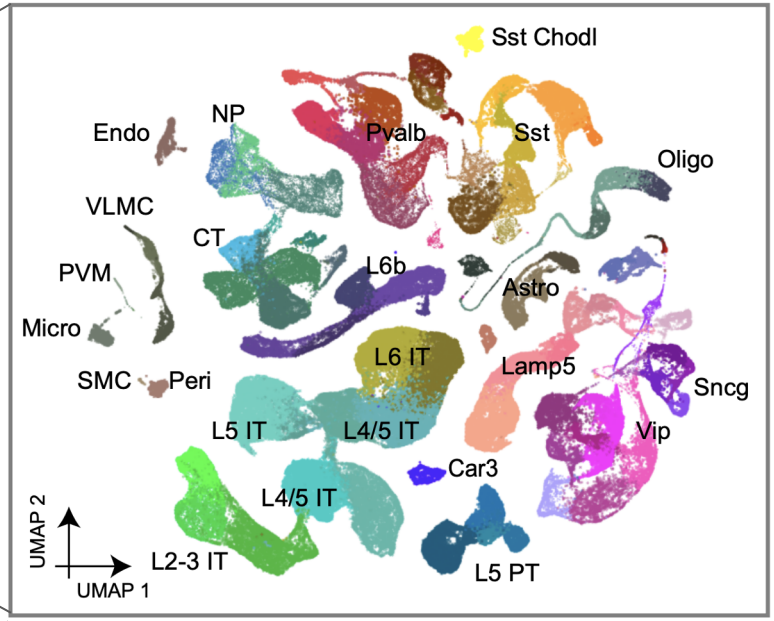


Connectivity
(eg *MouseLight*)

We understand the cell type ontology of the primary motor area

Primary motor area *MOp*

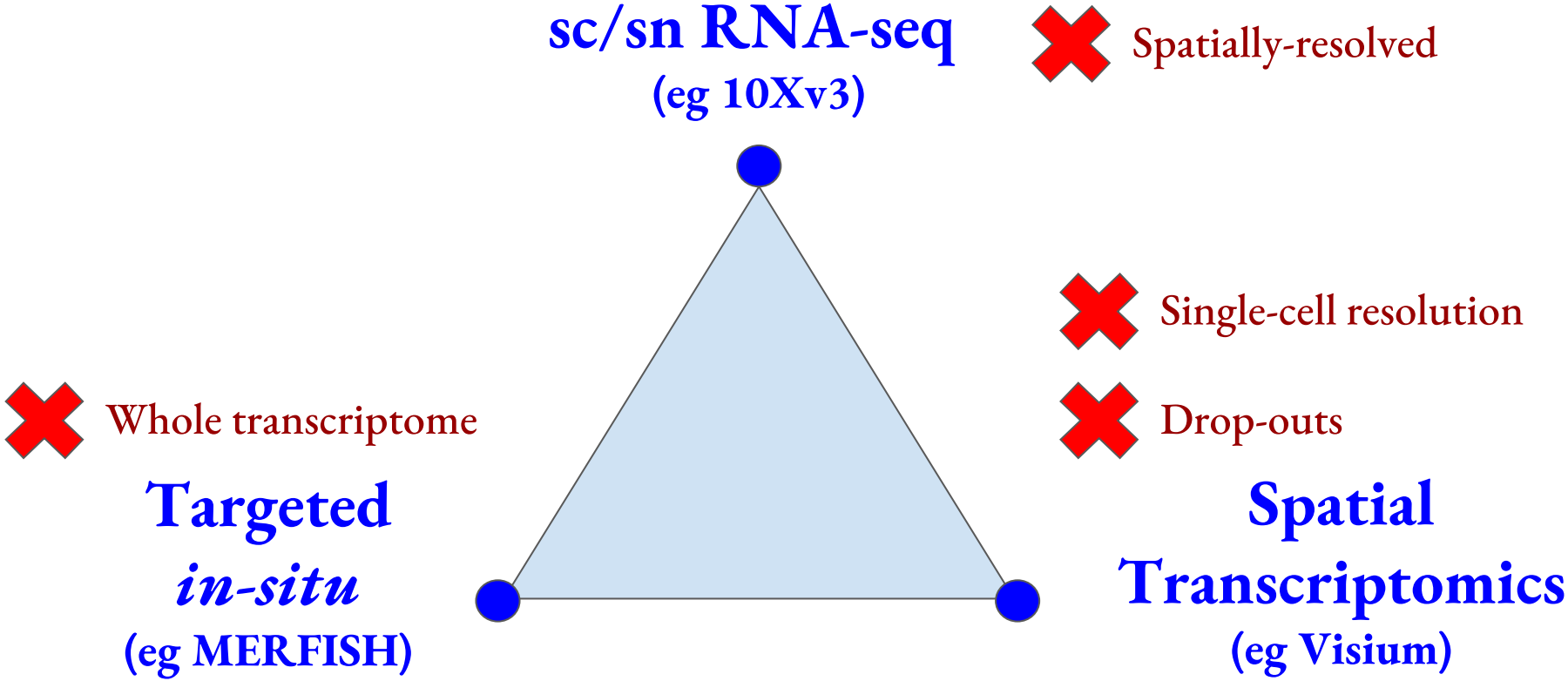




BICCN

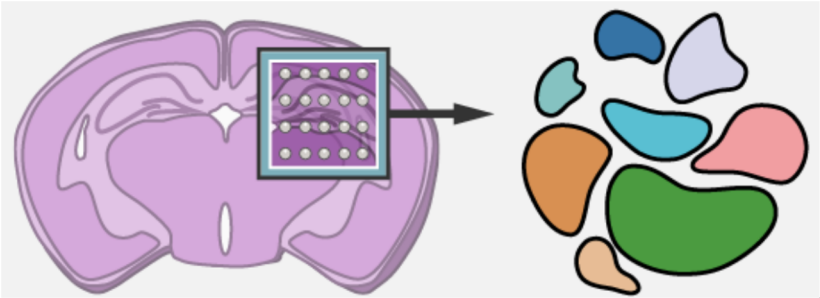
Hao et al. - *An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types* (Nature)

We cannot spatially resolve whole transcriptomes at single cell resolution

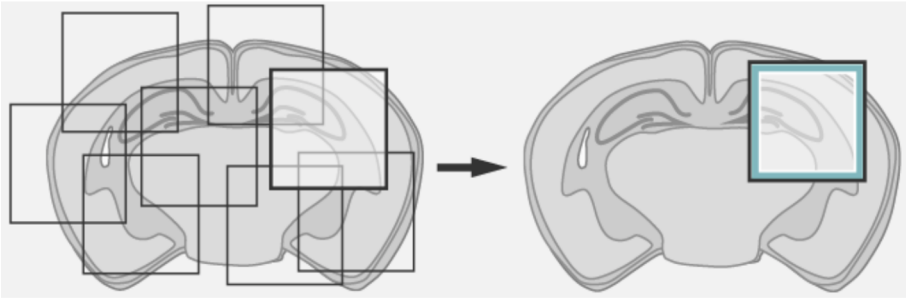


Our contribution: Integrated spatially resolved whole transcriptomes of single cells

Integration of scRNA-seq data with spatial data



Integration of spatial data with histology

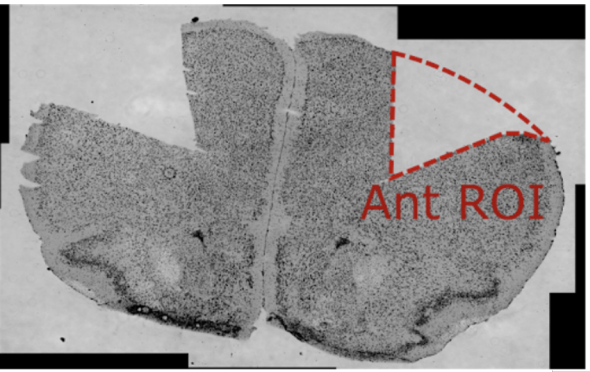


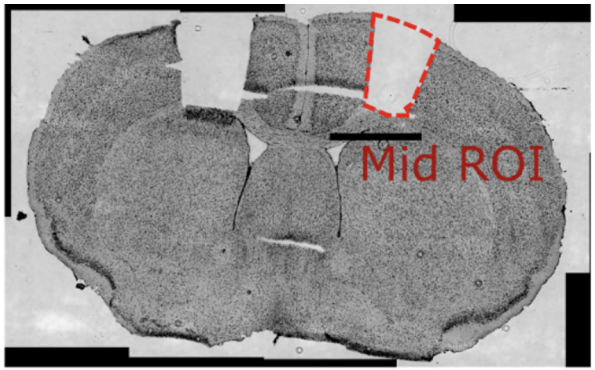
T. Biancalani*, G. Scalia*, L. Buffoni et al.
Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram (Nature Methods 2021)

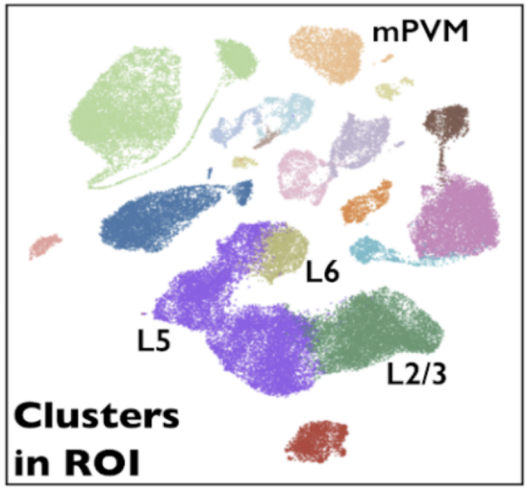
Part I: Mapping

We collect snRNA-seq data from healthy adult mouse brains

- Data collected from three ROIs from the primary motor area, *MOp*.
- 160k cells annotated into 22 cell types (from Macosko lab).





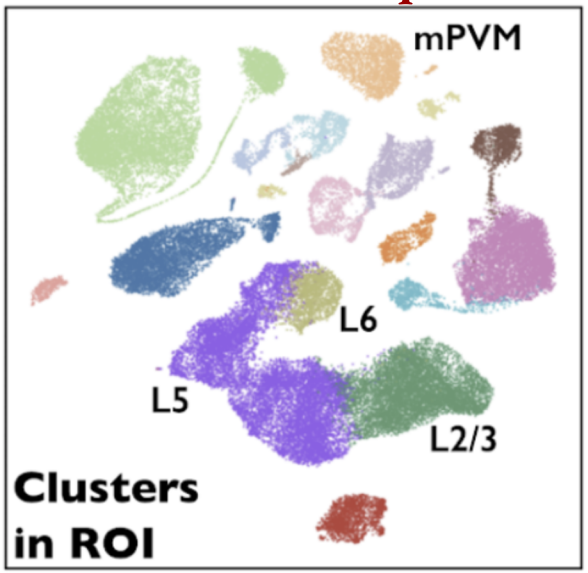


BICCN

Images by Charles Vanderburg

We map single-nuclei data onto a *MERFISH* dataset using Tangram

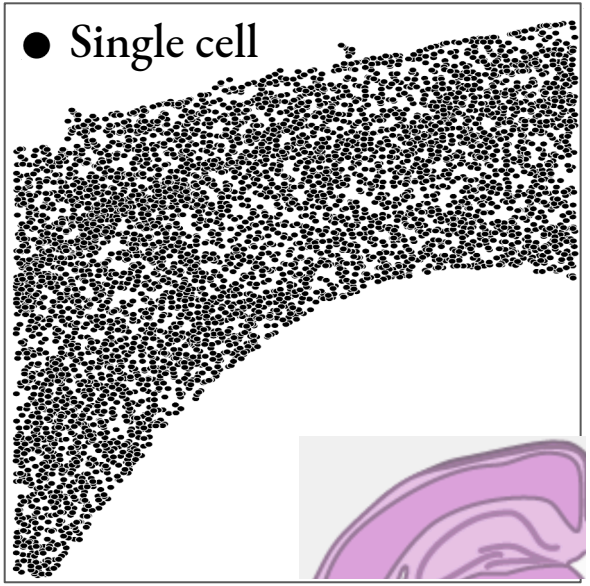
snRNA-seq



- 160k cells
- ~ 30k genes / cell

Mapping

MERFISH



- ~4k cells
- 254 genes / cell

How Tangram works (in a nutshell)

Notation:

Index i is for cells (snRNA-seq data) and has dimension n_{cells}

Index j is for spatial voxels and has dimension n_{voxels}

Index k is for genes and has dimension n_{genes}

We voxelize at the finest possible resolution for the technology used (MERFISH, Visium, ...)

We have two matrices:

- One from snRNA-seq that has dimensions $n_{cells} \times n_{genes}$ and we'll call it S.
- One from the spatial technology that has dimensions $n_{voxels} \times n_{genes}$ and we'll call it G.

Plus a n_{voxels} vector of cell densities $\vec{d}$

We aim at finding a mapping matrix M that tells us the probablity of the cell i being in voxel j.

How Tangram works (in a nutshell)

We minimize the following cost function using PyTorch

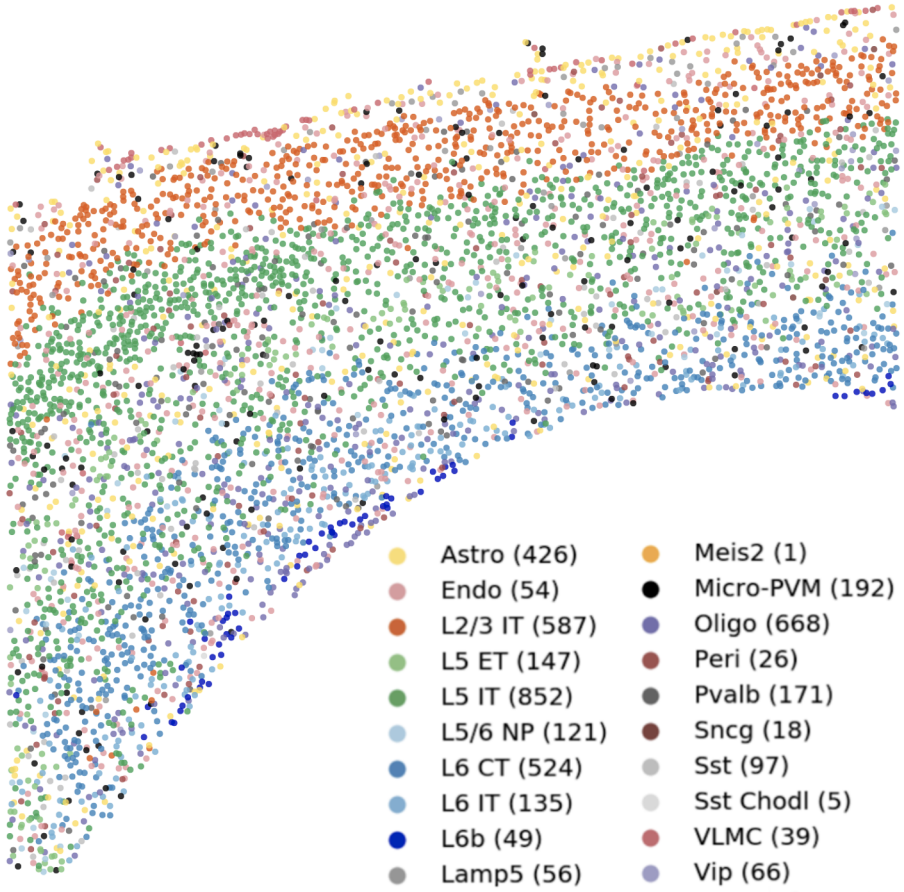
$$\Phi(\tilde{M}) = KL(\vec{m}, \vec{d}) - \sum_{k=1}^{n_{genes}} \cos_{sim}\left((M^T S)_{*,k}, G_{*,k}\right) - \sum_{j=1}^{n_{voxels}} \cos_{sim}\left((M^T S)_{j,*}, G_{j,*}\right)$$

Using $M = softmax(\tilde{M})$ ensures that $0 \leq M_{i,j} \leq 1$ and $\sum_j M_{i,j} = 1$

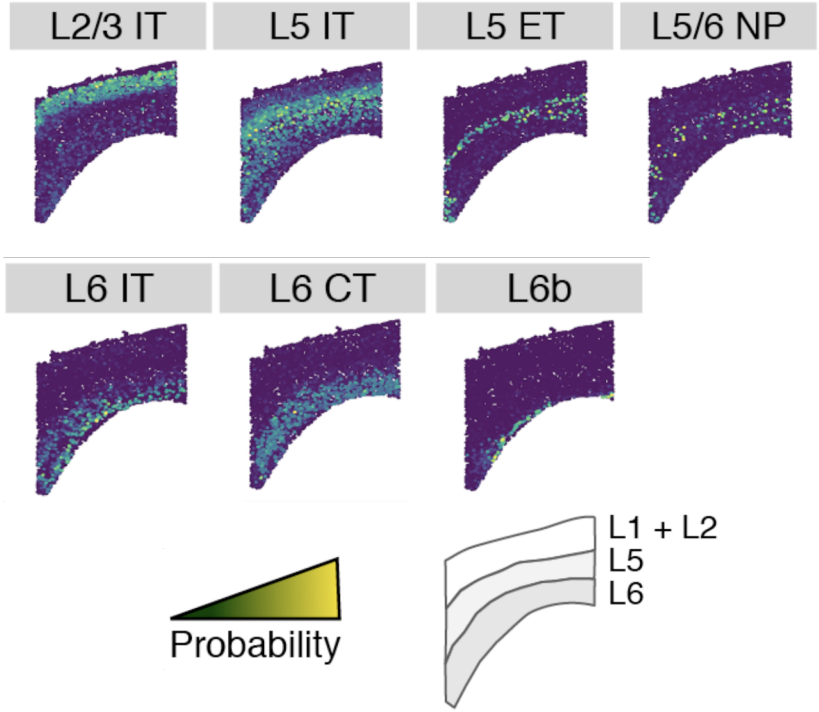
$\cos_{sim}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$ is the cosine similarity function

$\vec{m}$ is the cell density vector for the mapping $m_j = \sum_i \frac{M_{ij}}{n_{cells}}$

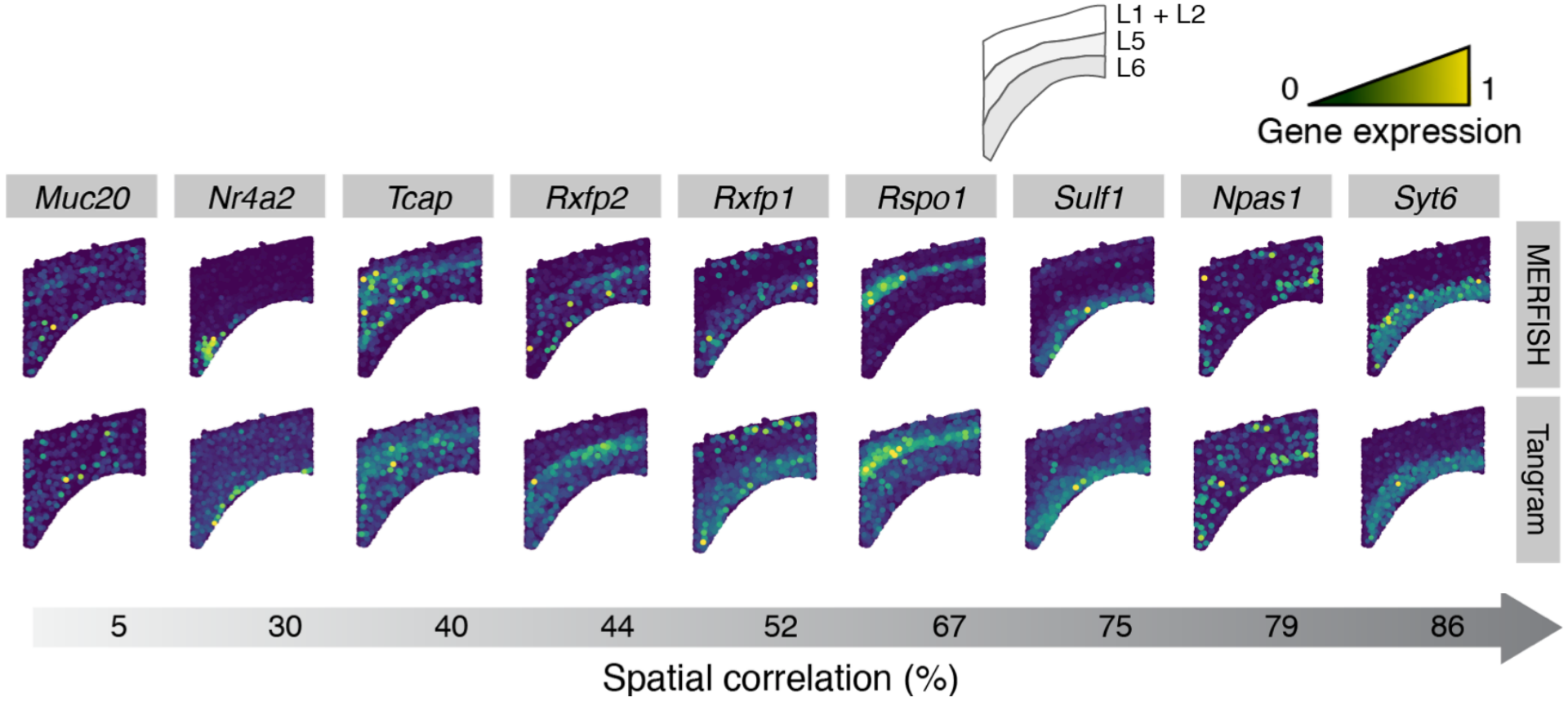
Mapping allows for spatial localization of cell types



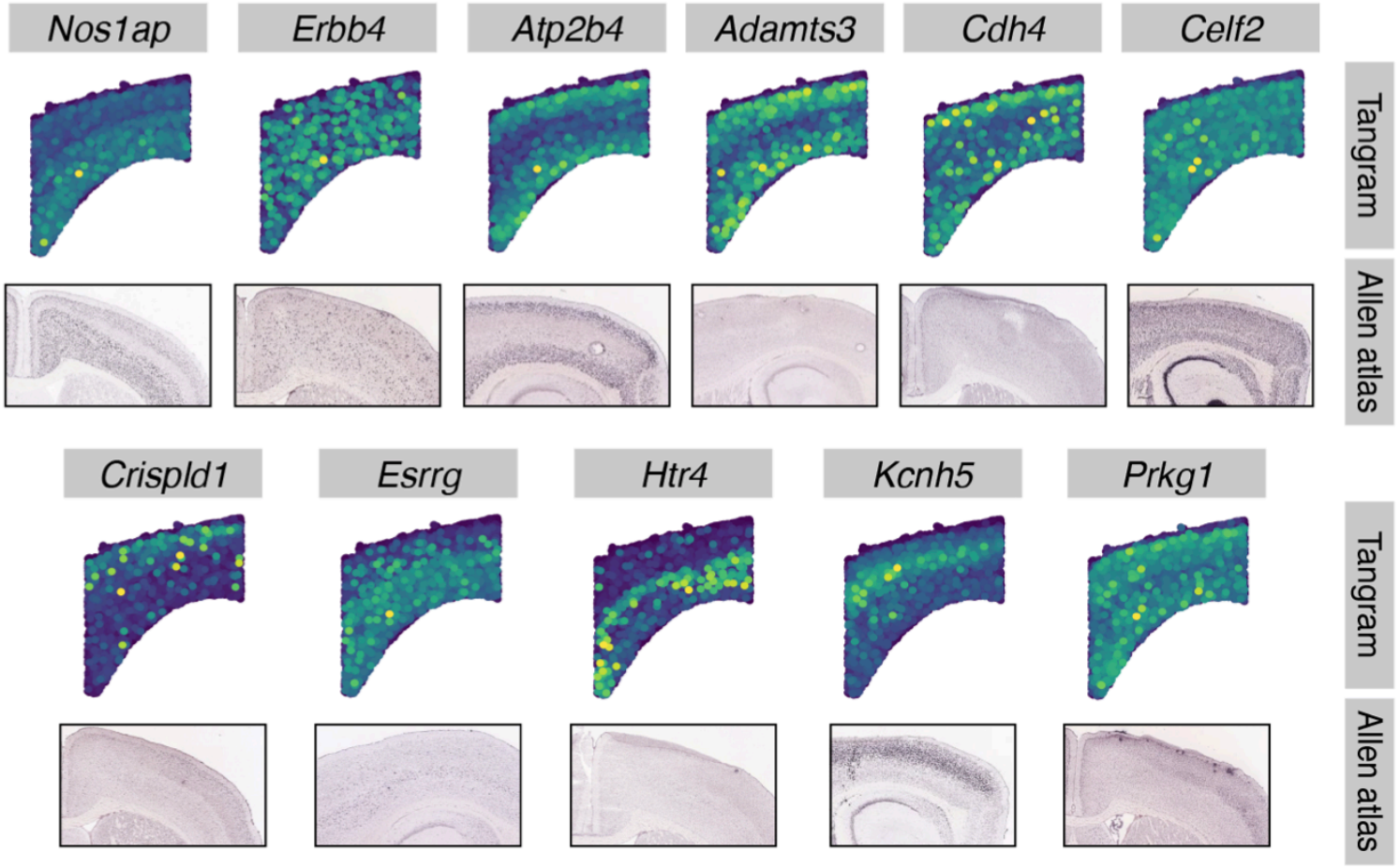
Probability maps for excitatory neurons



Mapping allows us to predict spatial gene expression

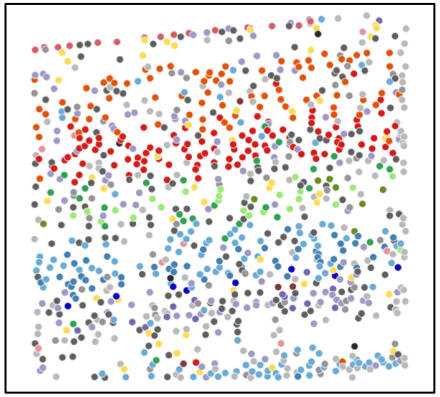


Mapping increases gene throughput to ~30k genes

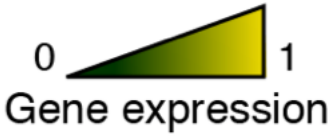


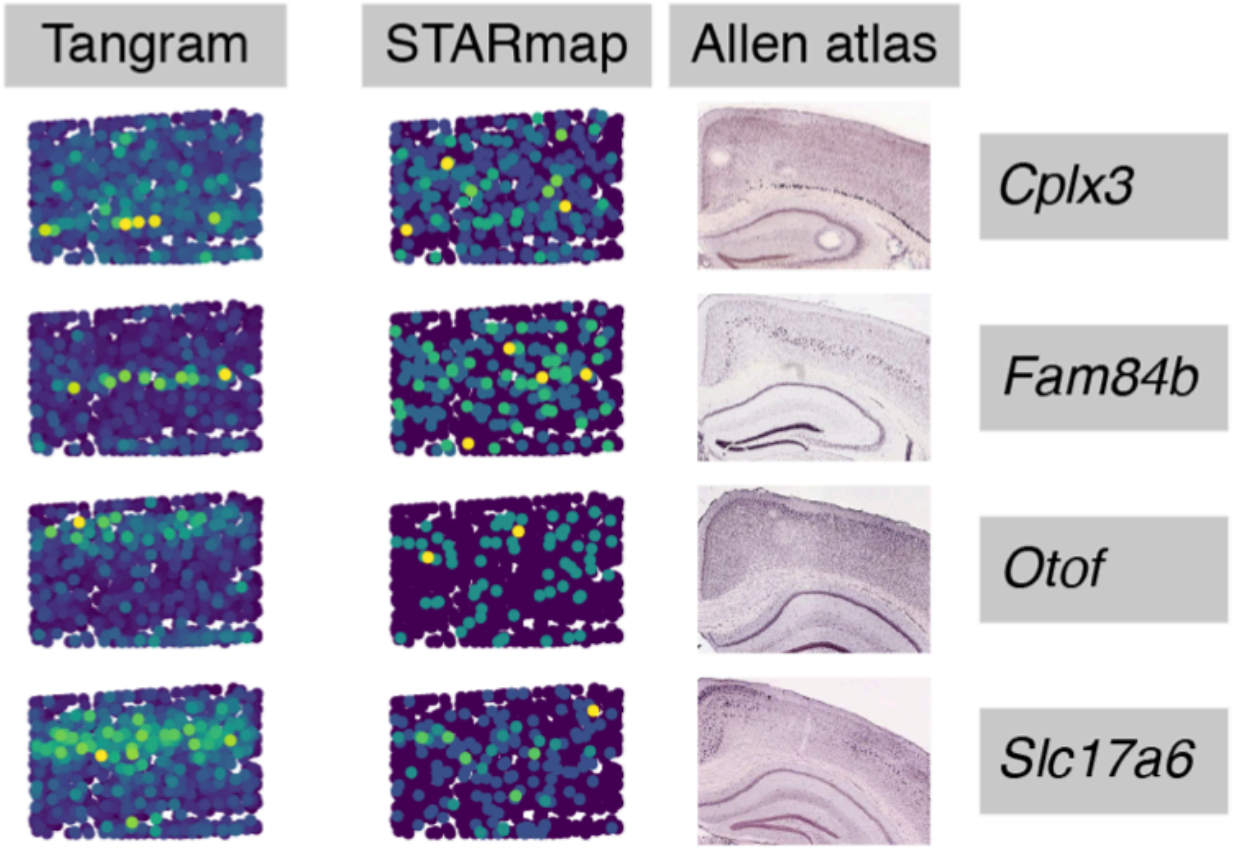
Mapping corrects low-quality genes

STARmap



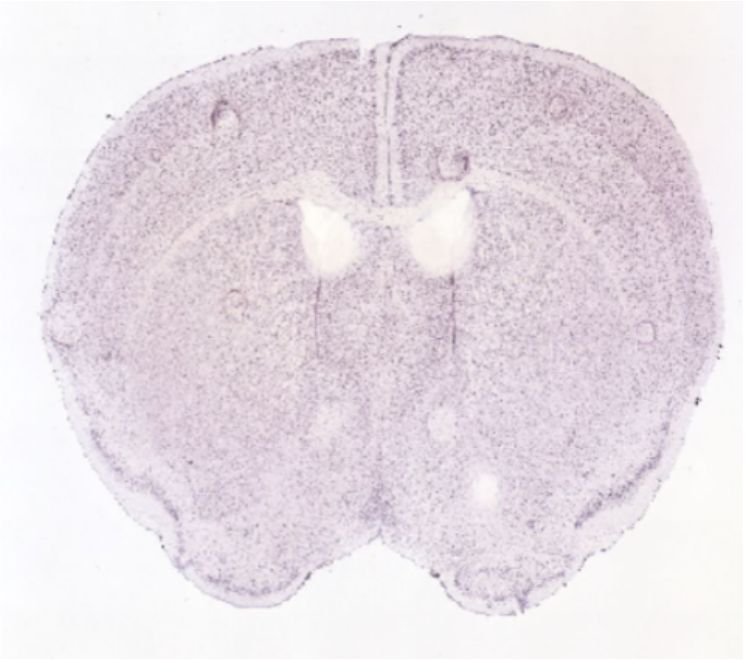
- ~1k cells
- ~1k genes/cell





Part II: Integration with histology

Image registration requires human supervision



Experimental image

?

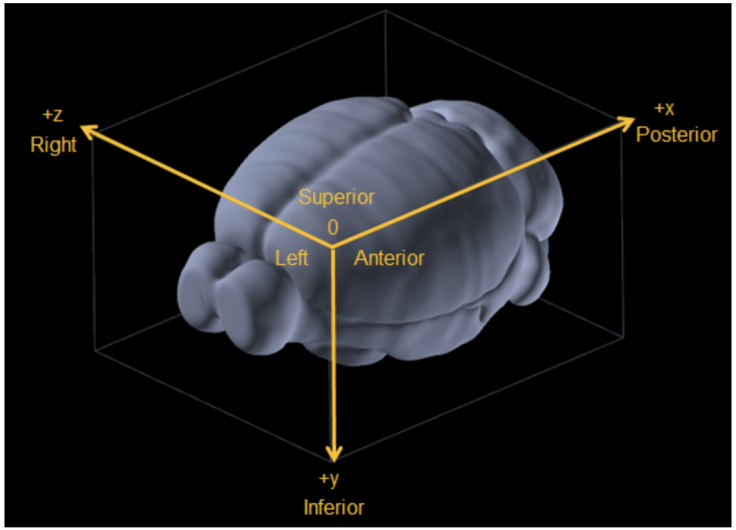
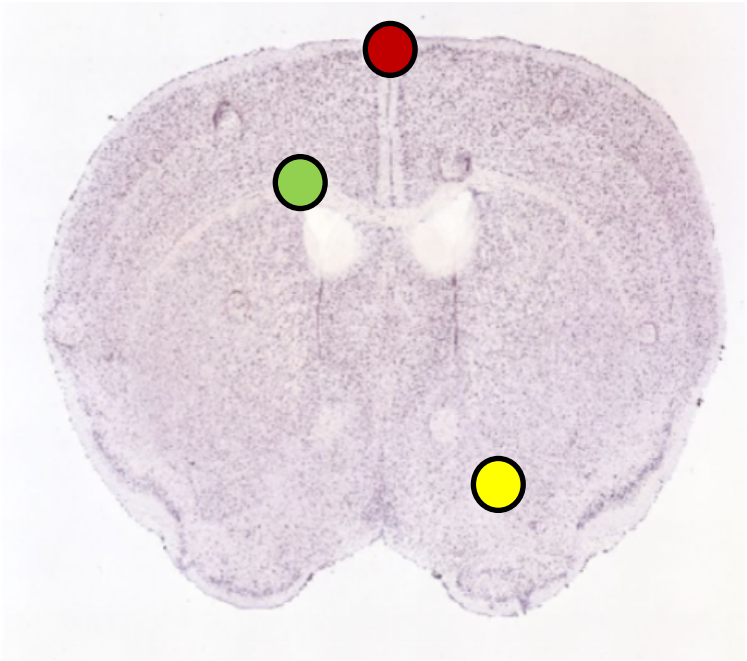
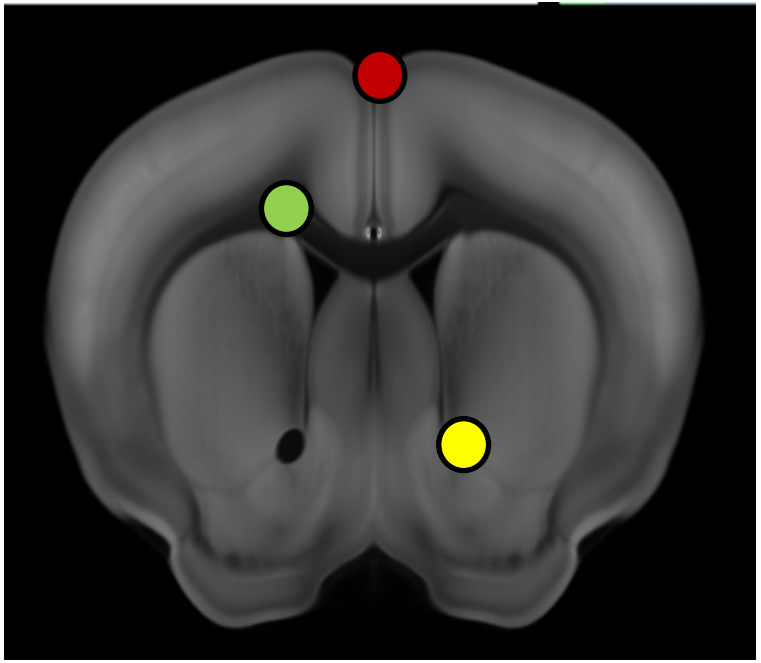


Image registration requires human supervision



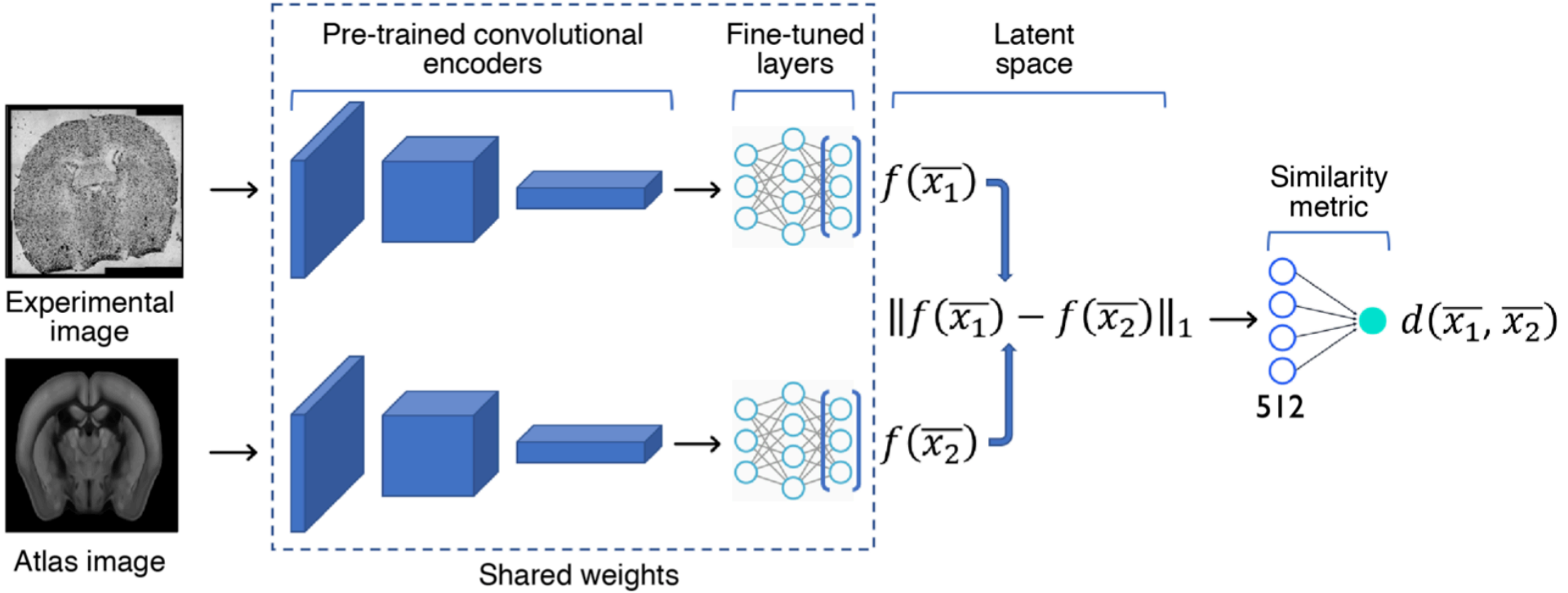
ANTs



Experimental image

From reference atlas

We trained a “face recognition” model on histological images of mouse brains



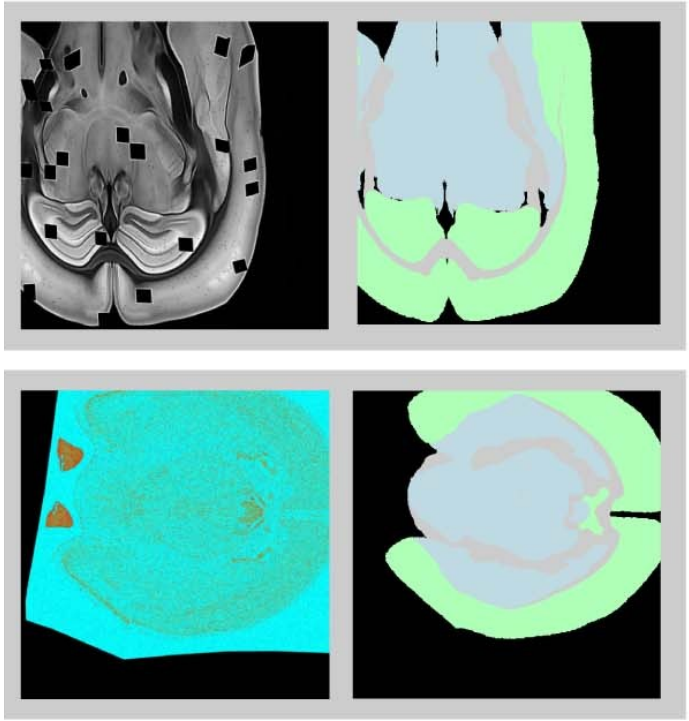
We trained a “face recognition” model on histological images of mouse brains

Model details:
- Pretrained encoder DenseNet169
- Pretrained on ImageNet
- Fine-tuned the last convolutional layer + 2 fully connected layers.

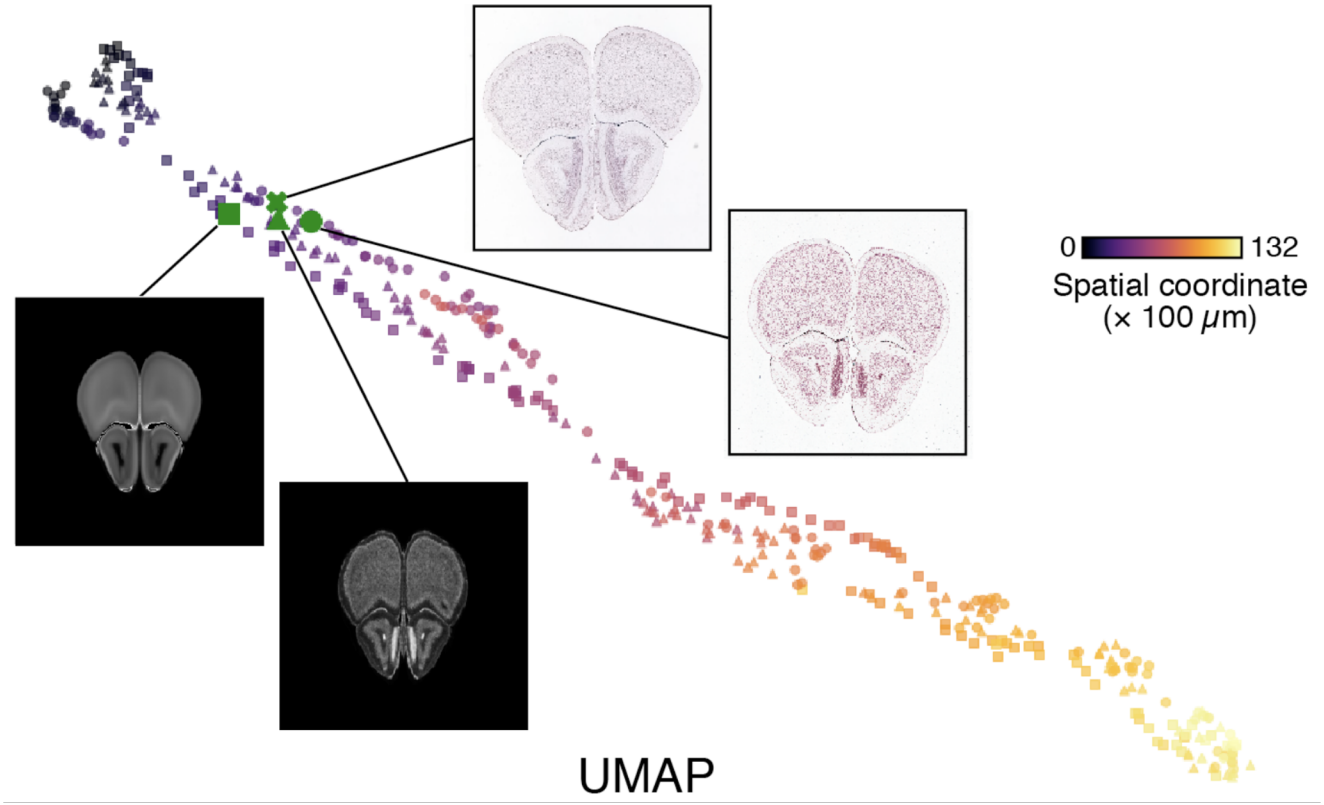
Trained for 50 epochs using 18000 image pairs per epoch in batches of 16.

$$MSE(\hat{d}, d) = \frac{1}{N}\sum_{i=1}^{N}(d_i - \hat{d}_i)^2$$

Needed heavy augmentation to for training.



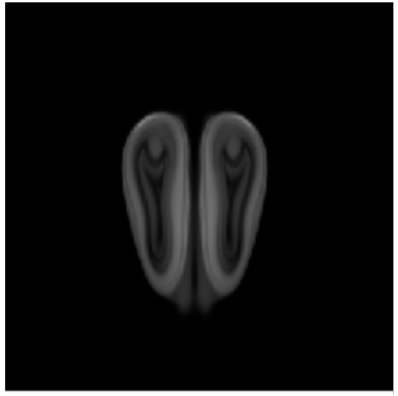
In the latent space, the geometrical distance represents the anatomical distance

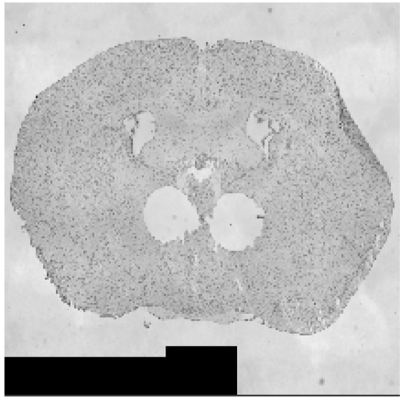


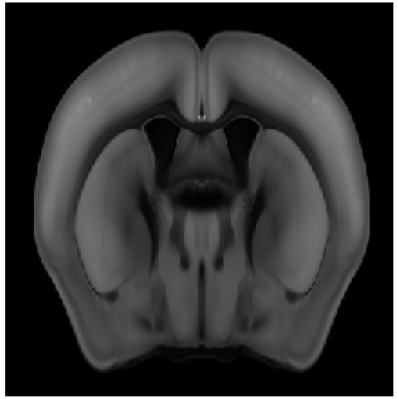
Model predictions are used for “depth calling”

Our image

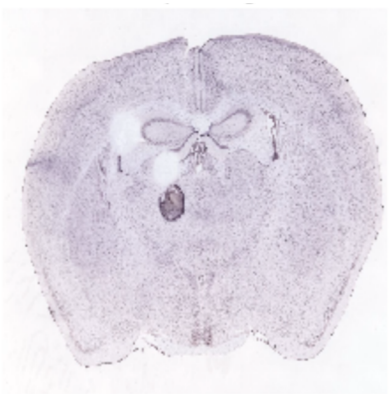
Reference atlas



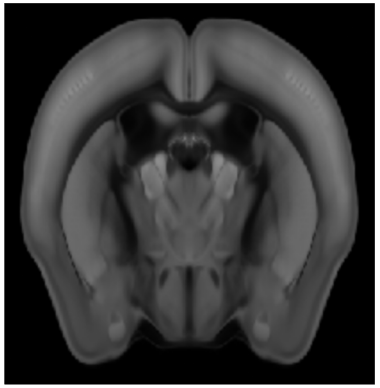


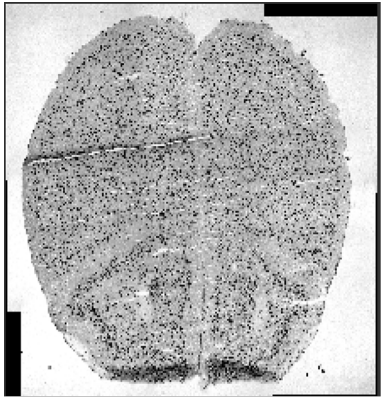


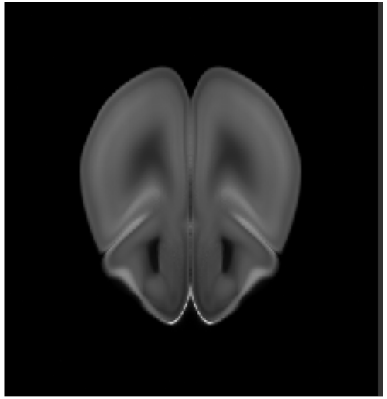
Our image



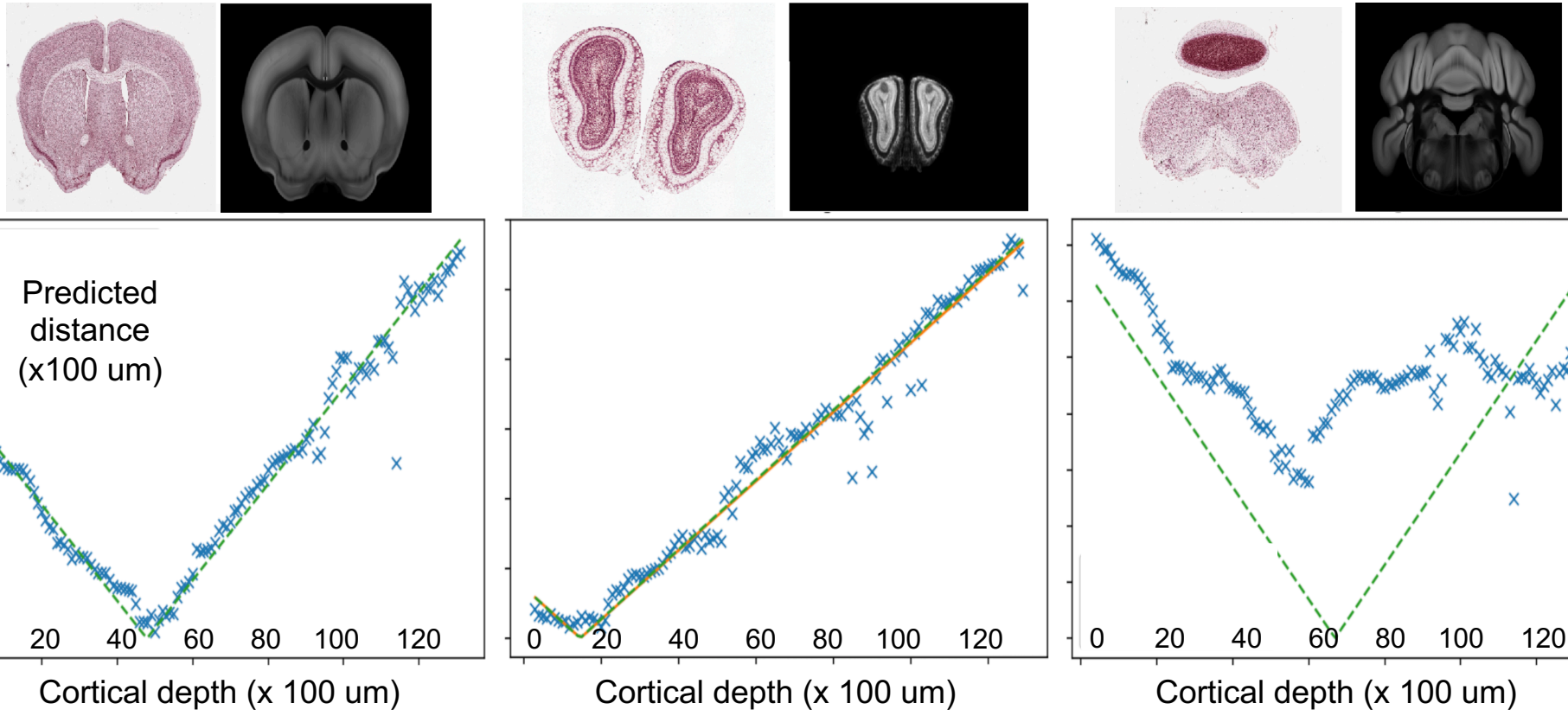
Reference atlas



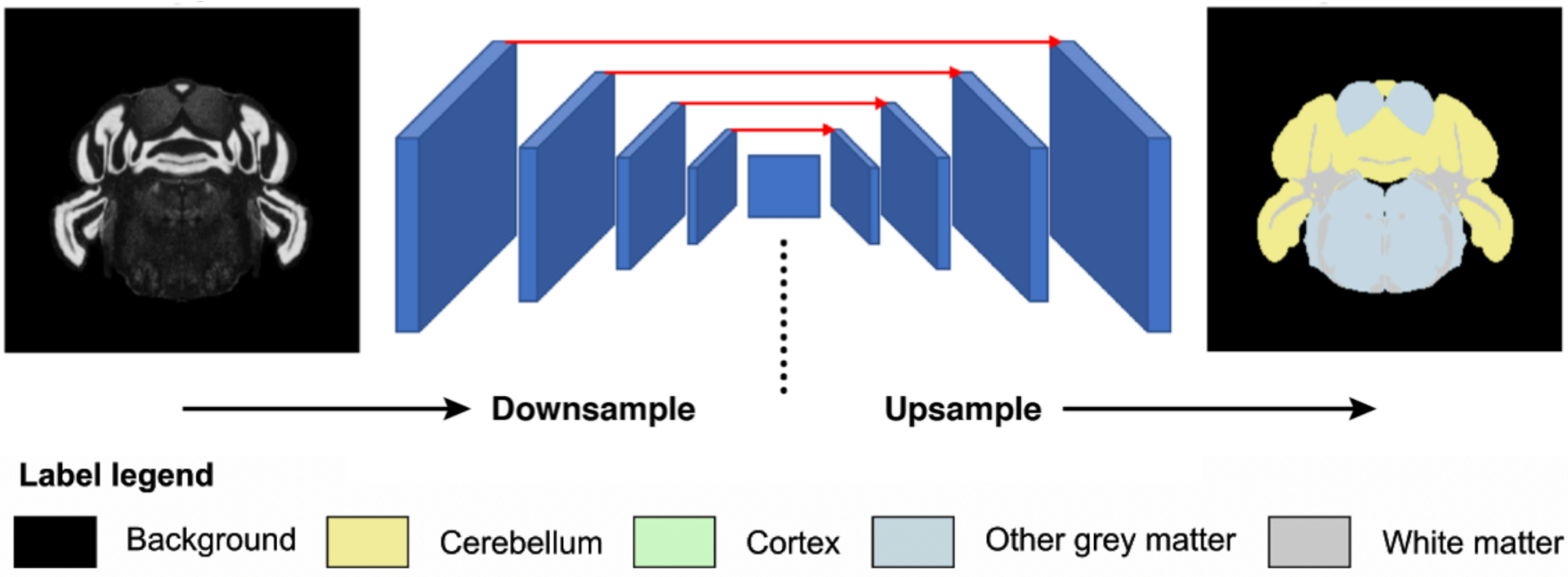




Predictions are accompanied by uncertainty estimation

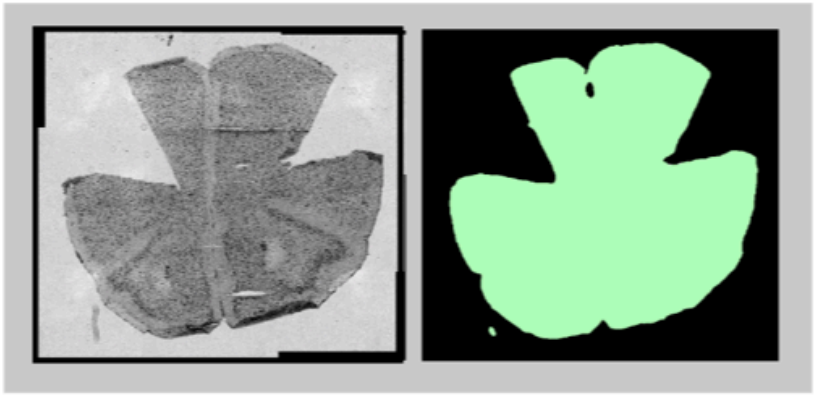


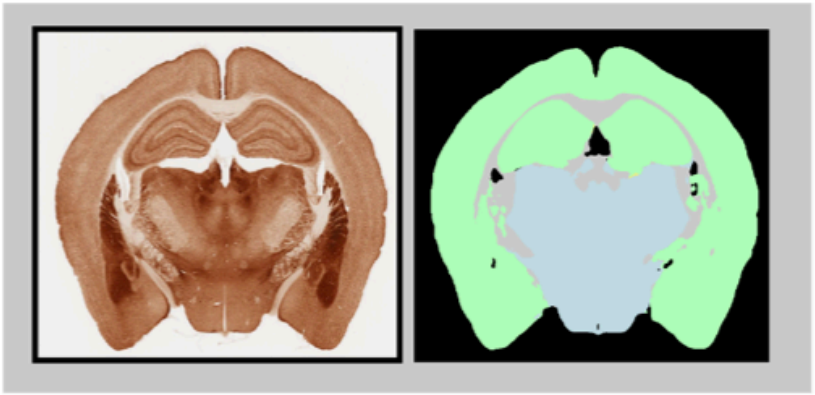
We perform anatomical region calling via semantic segmentation

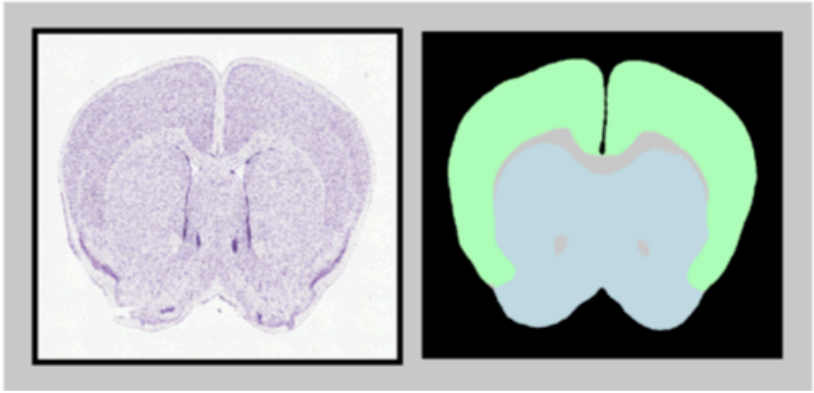


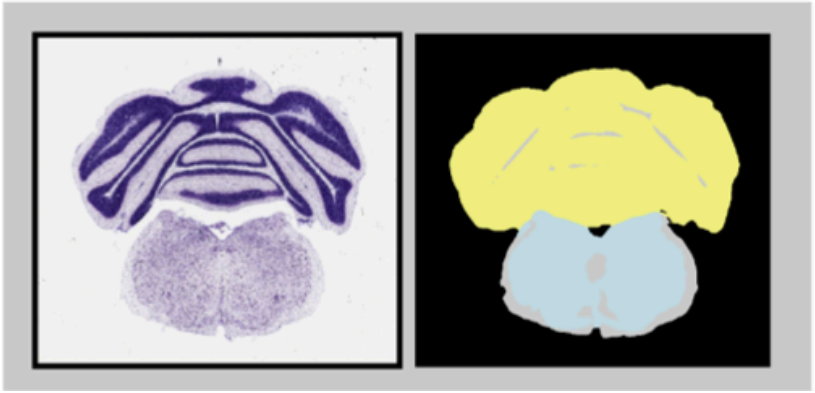
Contribution of Ziqing Lu and Aman Sanger

We produce consistent masks for each image

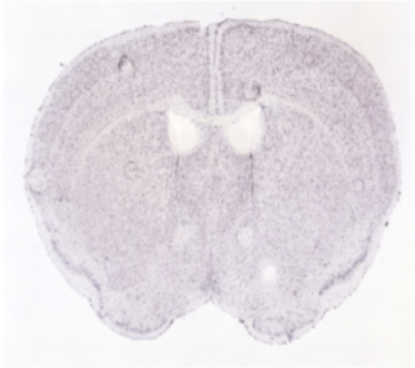




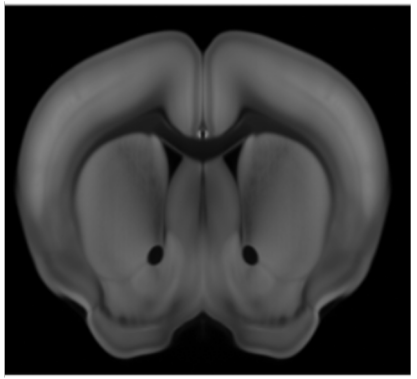




The two models combined provide a fully automated registration pipeline

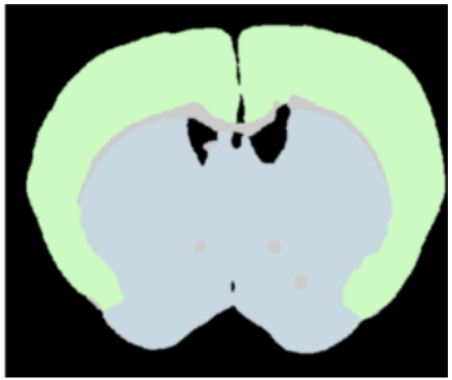


(I)
Depth calling



(III)
Produce mask

(II)
Extract mask

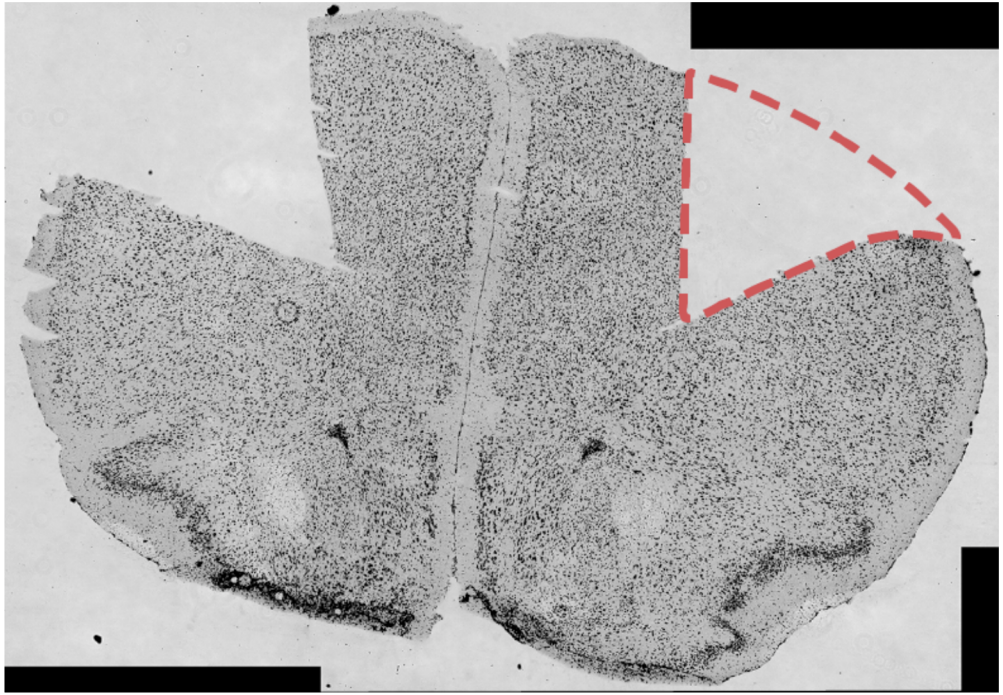


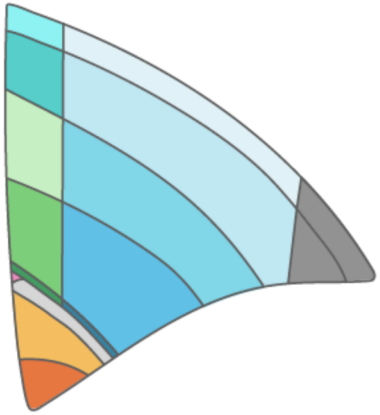
(IV) Registration

ANTs

Contribution of Neriman Tokcan

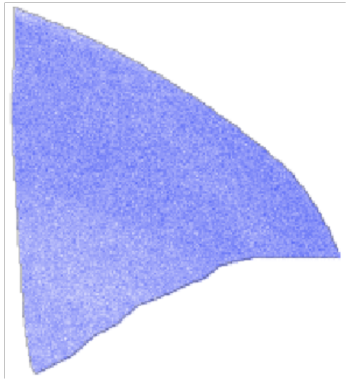
Using the pipeline, we identify the anatomical/cell maps for each ROI



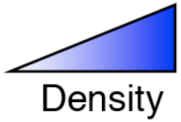


MOp1
MOp2/3
MOp5
MOp6a
MOp6b
MOs1
MOs2/3
MOs5
MOs6a
MOs6b
cing
CP
fa
scwm
SSp-m

Anatomical region map
(from *Allen CCFv3*)



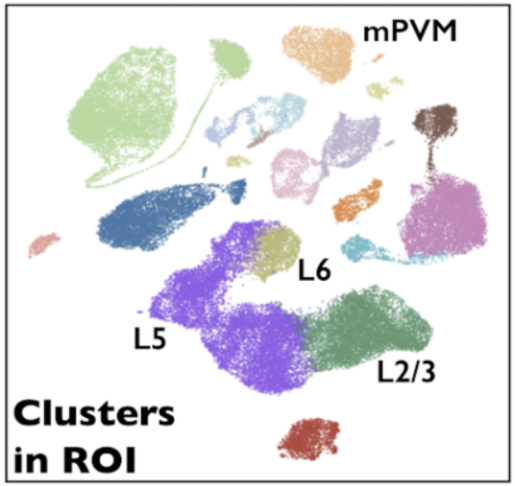
Cell density map
(from *Blue Brain Cell Atlas*)



Part III: Mapping on Spatial Transcriptomics

We map snRNA-seq data data onto a Visium dataset

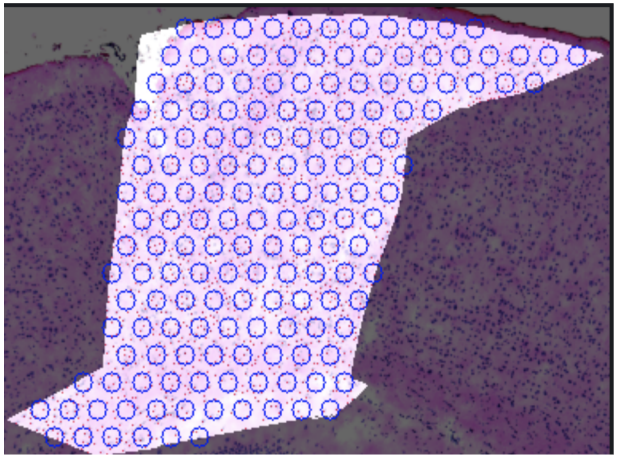
snRNA-seq data

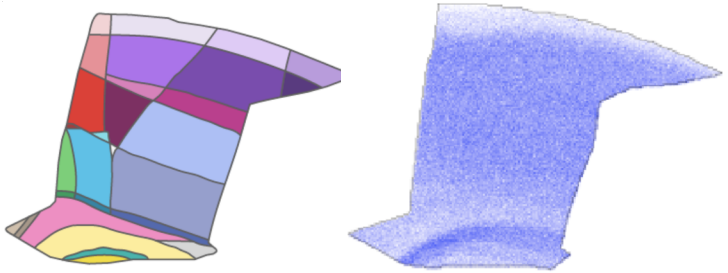


Mapping

Fitting gene expression on ~1k marker genes

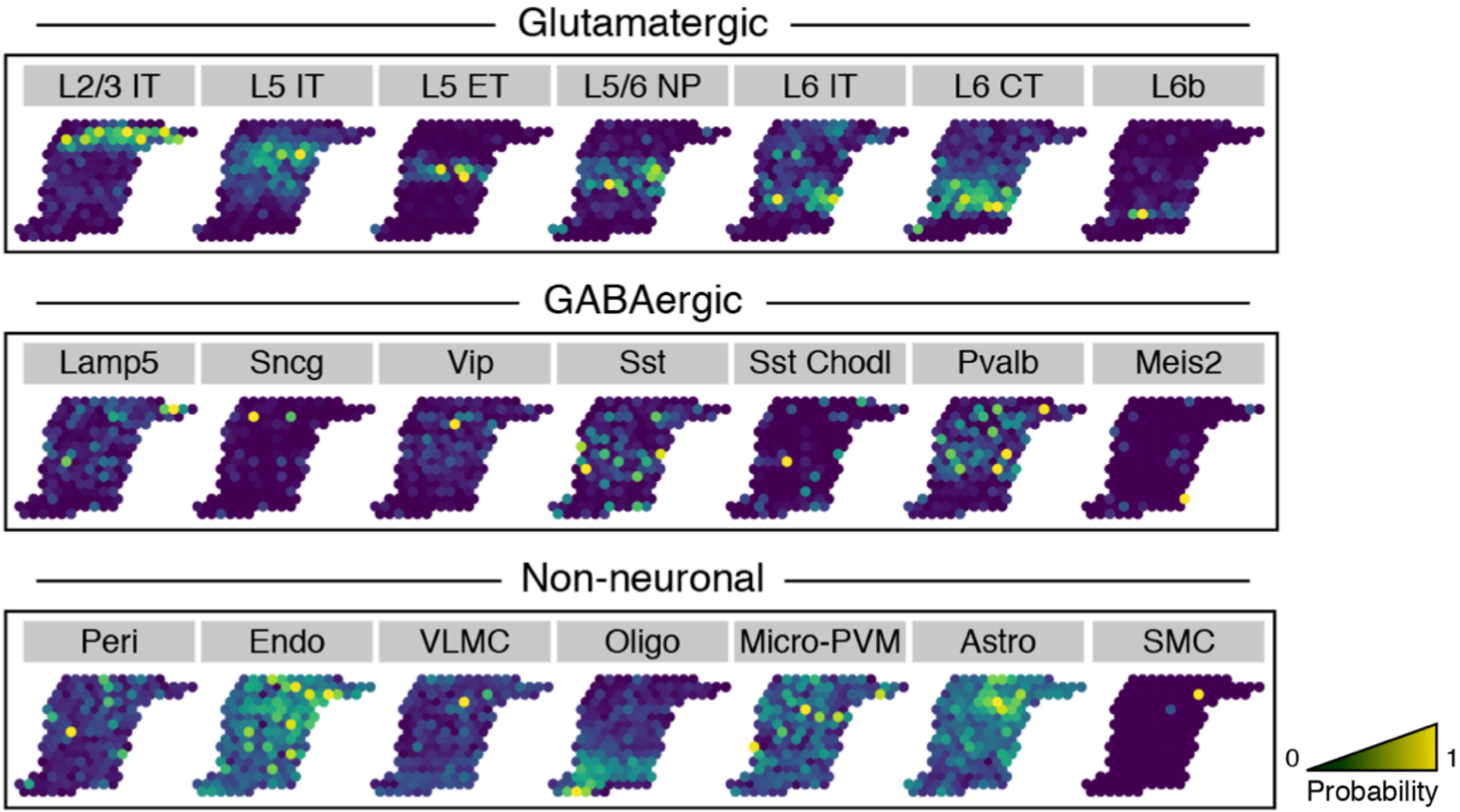
Visium



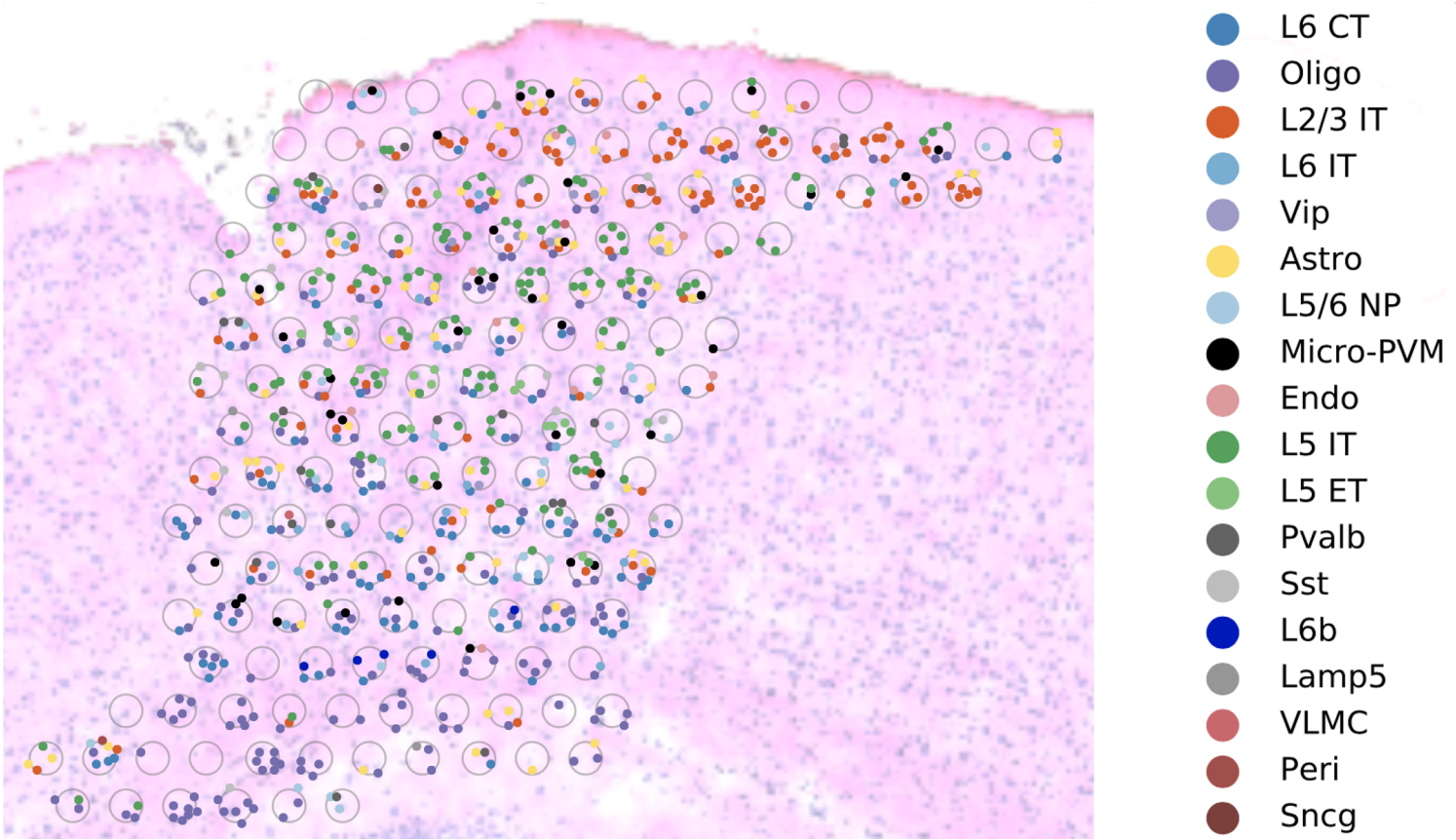


Contribution of Raghav Avasthi

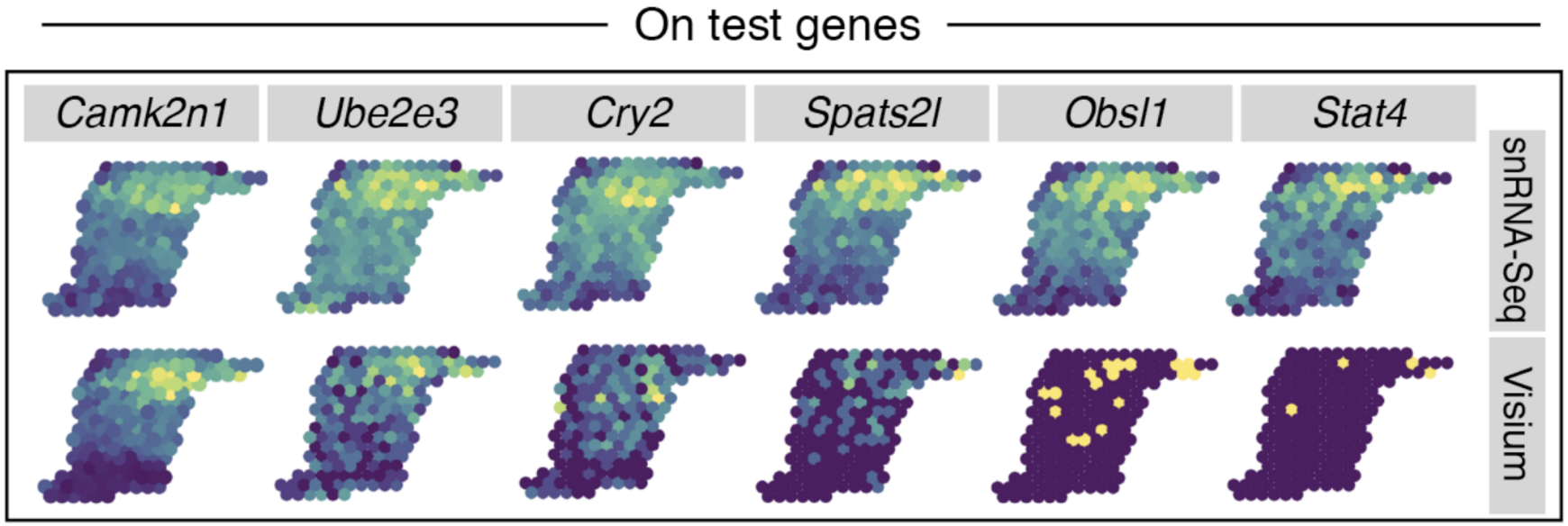
We predict cell type localization on the *Visium* ROI



Mapping allows cell type localization at single-cell resolution



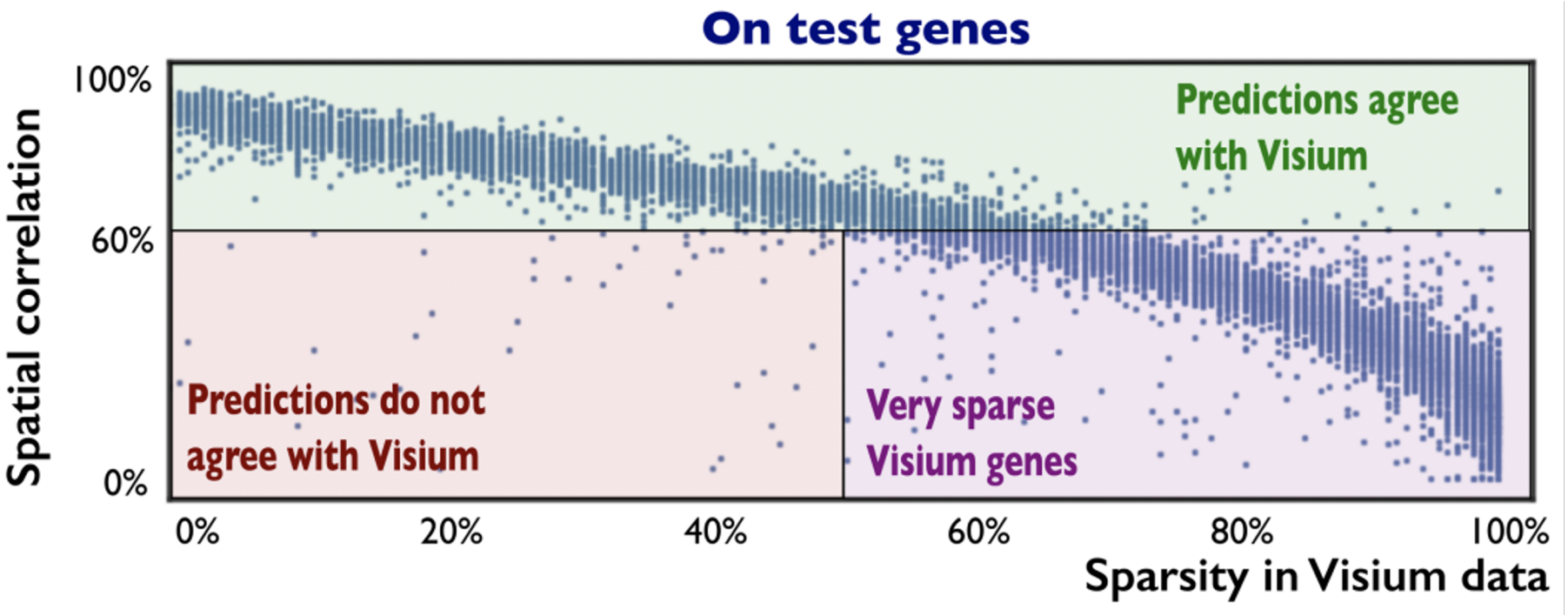
Agreement on test genes decreases as data becomes sparser



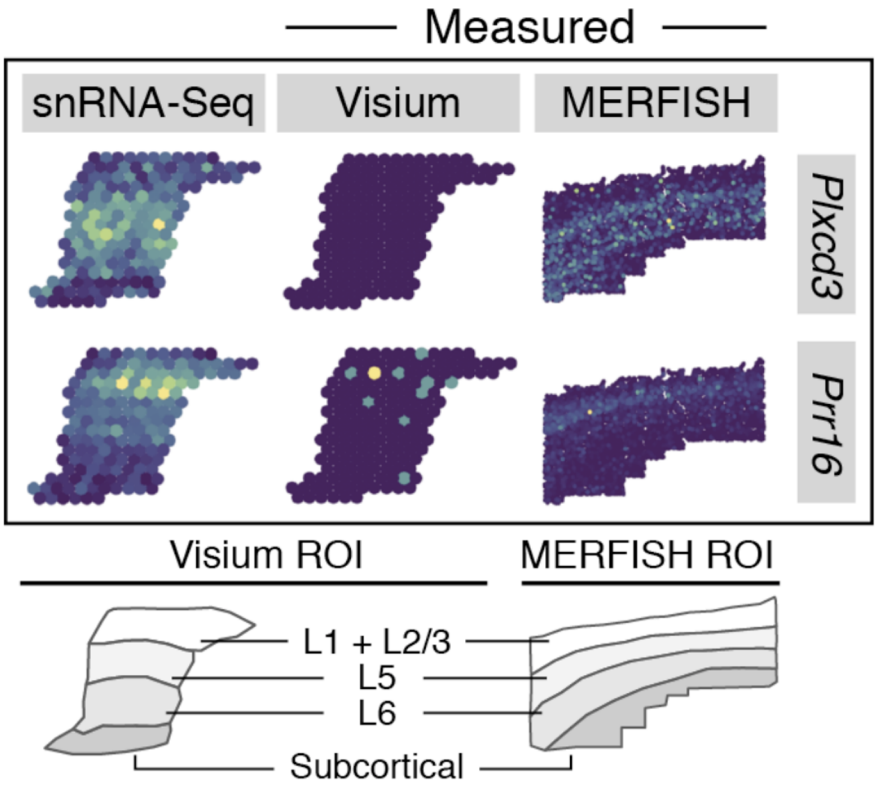
Agreement against predictions

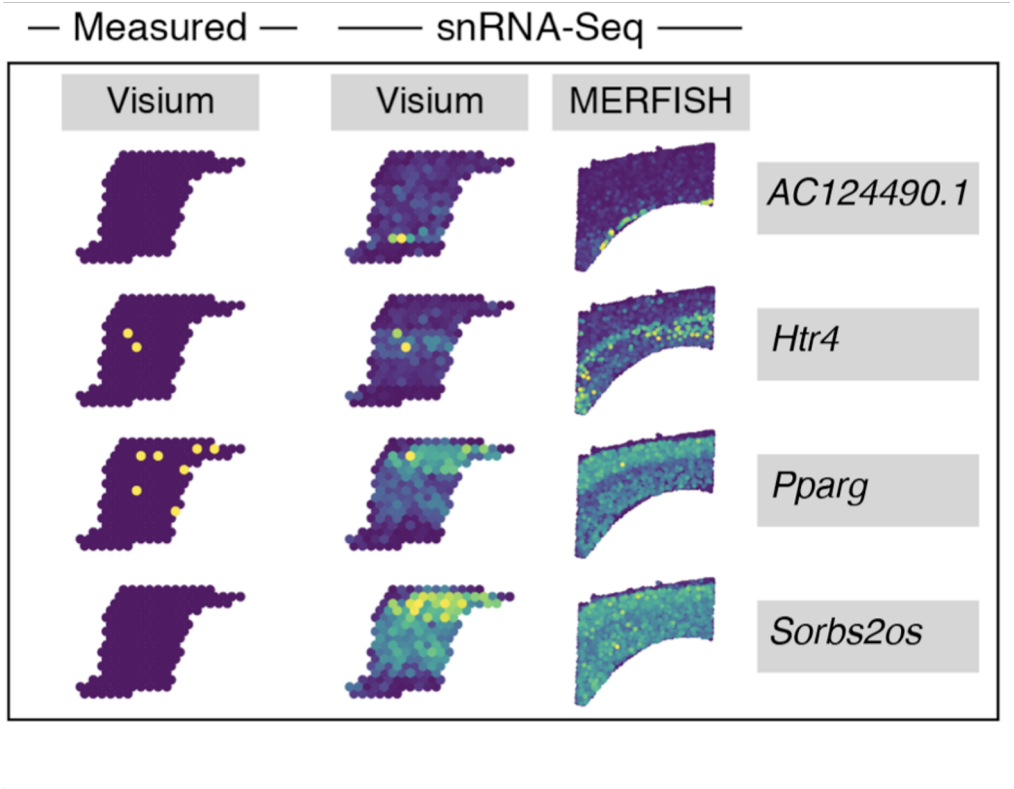
Sparsity of Visium data

We partition the transcriptome according to method performance



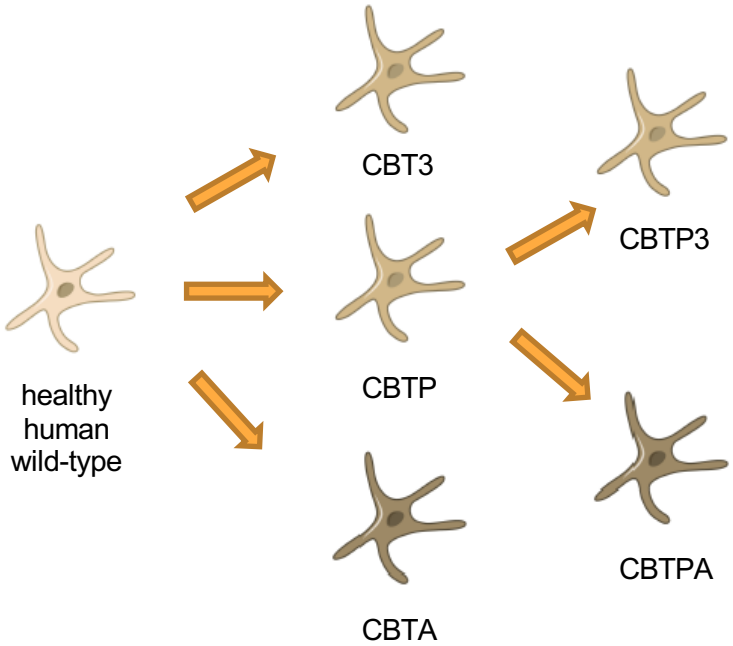
Predictions are validated against *MERFISH* data



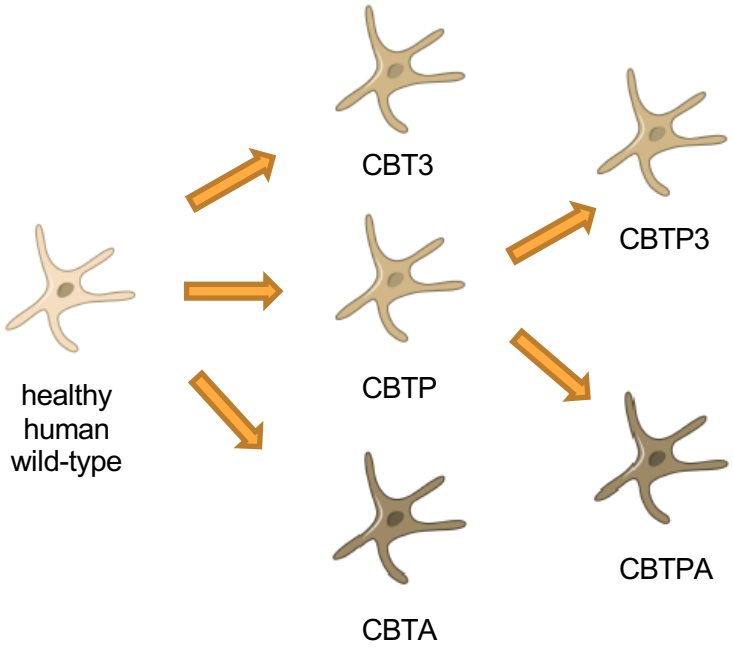


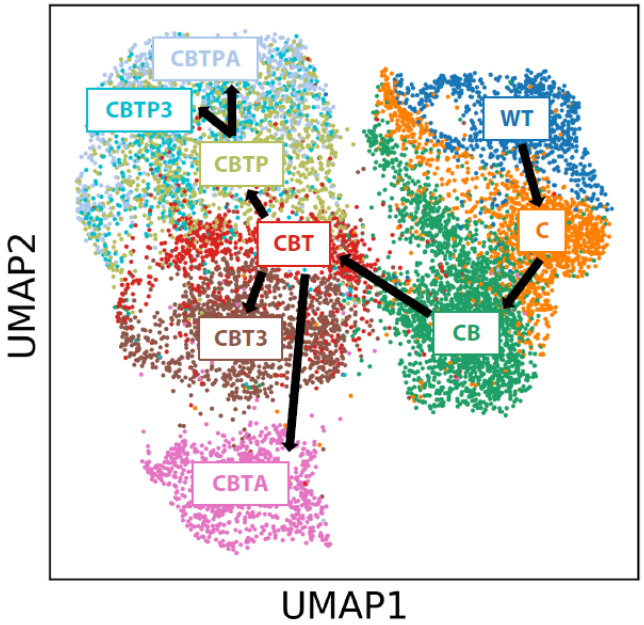
Part IV: The inverse problem

Starting from a «blank» cell we can engineer single mutations in the genome that will lead to a melanoma



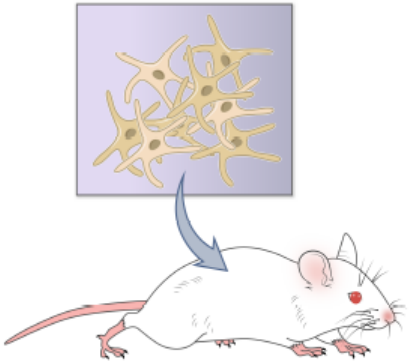
Starting from a «blank» cell we can engineer single mutations in the genome that will lead to a melanoma



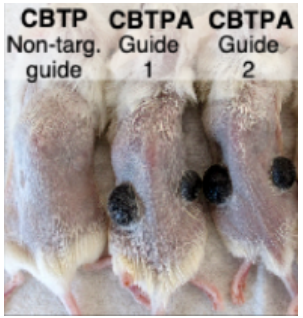


Melanoma is grown on mice and histology is collected

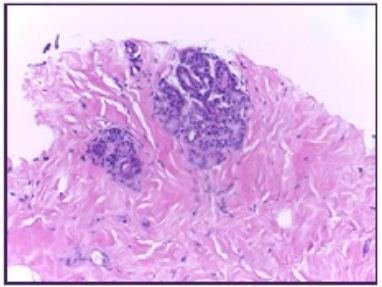
Injection of cells
into **immunodeficient** (NSG) mice

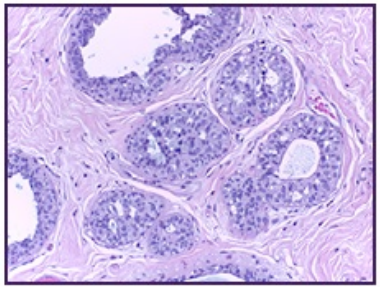


Observe phenotypical differences

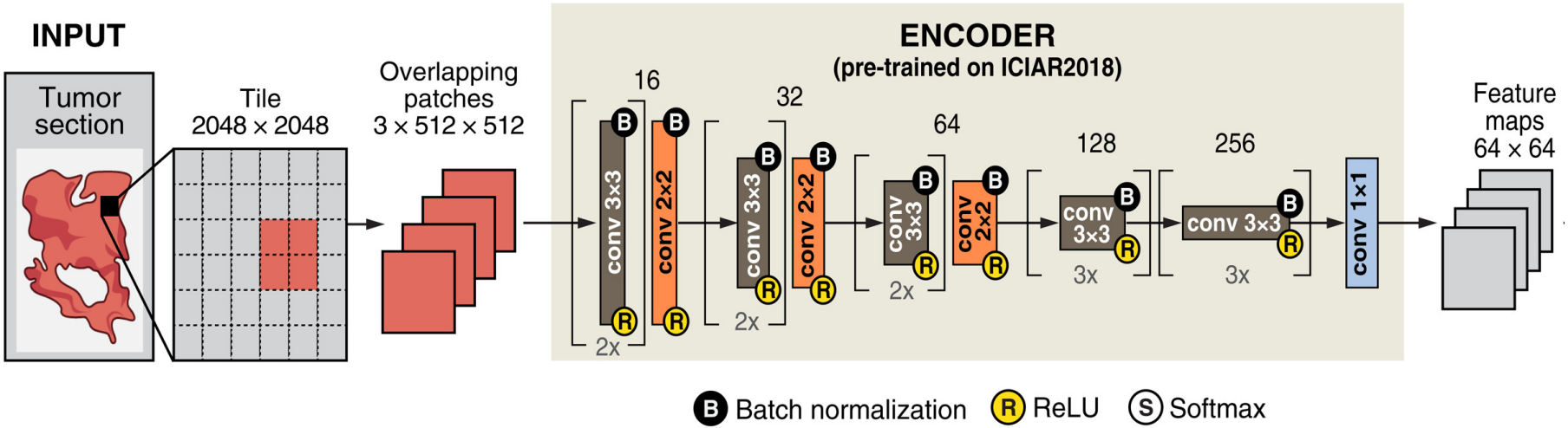


Can we train a classifier to recognize
the genotype of a tumor from the histology?



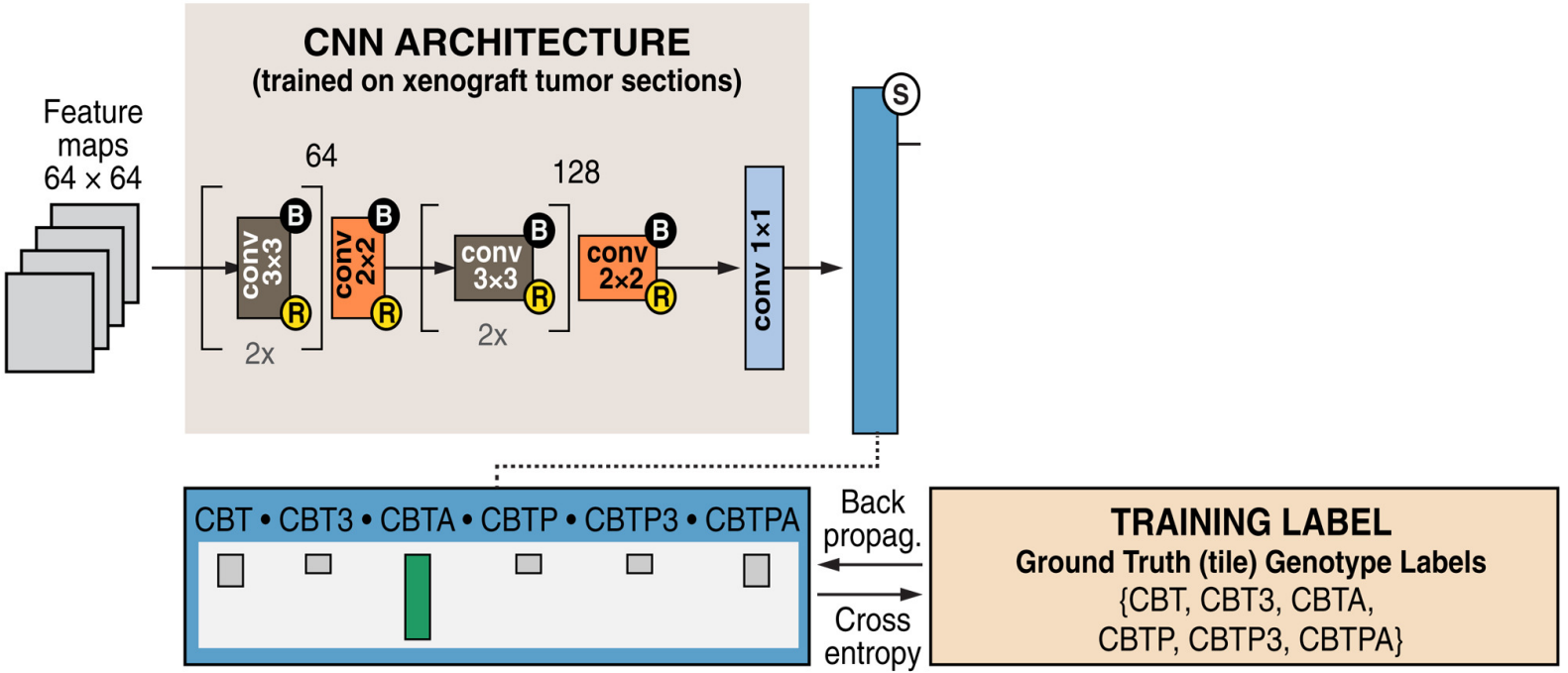


We train a model to predict genotype from histology

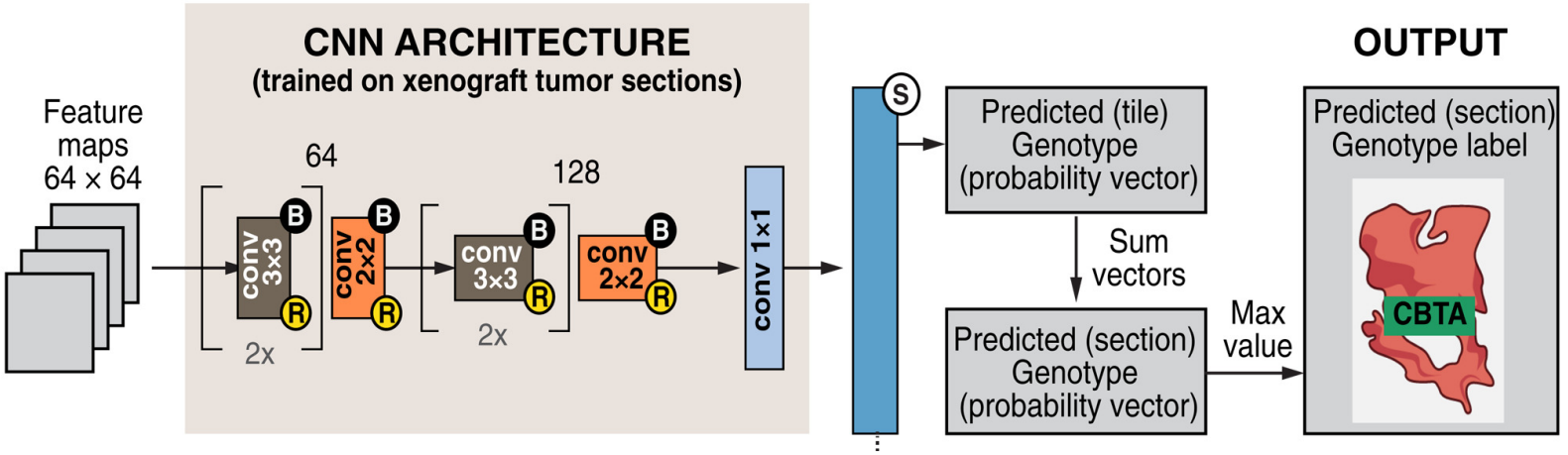


Hodis et al. - *Stepwise-edited, human melanoma models reveal mutations' effect on tumor and microenvironment* (Science)

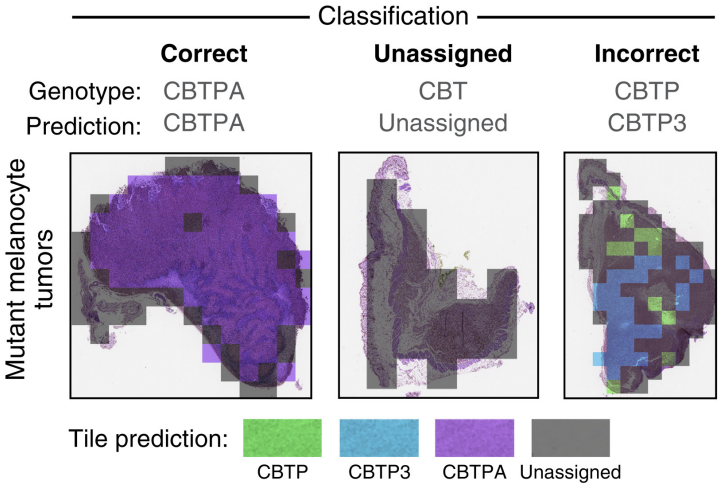
We train a model to predict genotype from histology

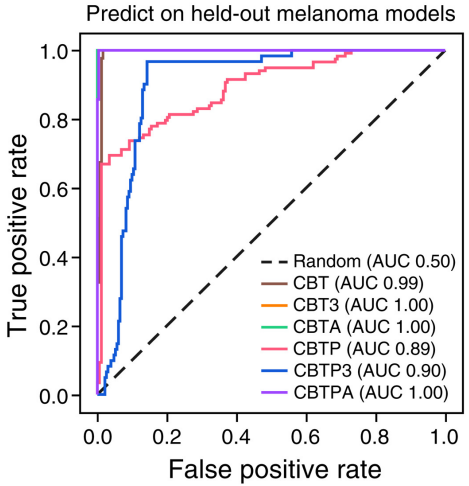


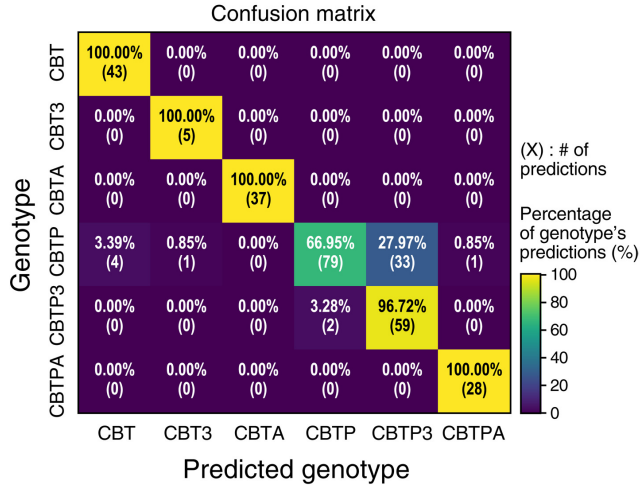
The model is then used to make inference on an entire histology image



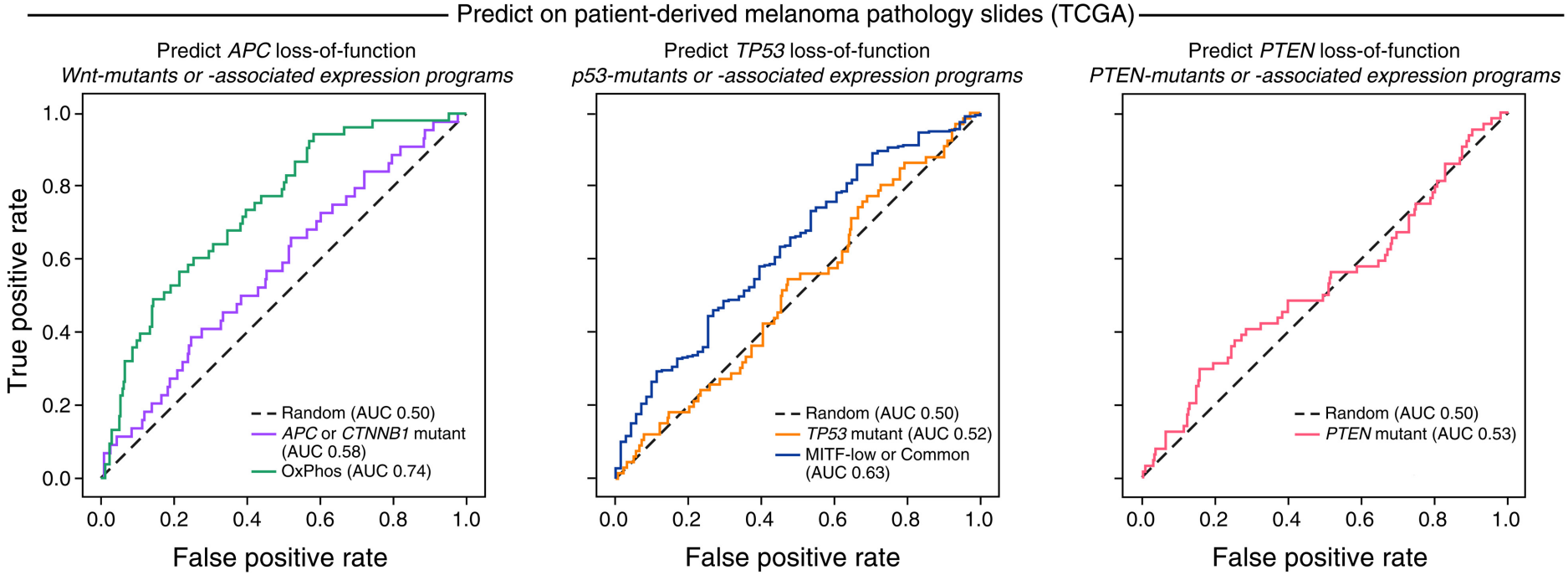
The model consistently predicts genotypes in mice-grown tumors





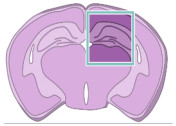


Transfer of the model (mouse data) on real patient's data (TCGA) shows some correlation

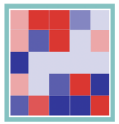


Conclusions.
Using ML we can:

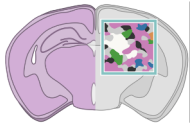
- Relate genes to anatomy.



- Extend gene throughput.



- Increase spatial resolution.



- Identify mutations.