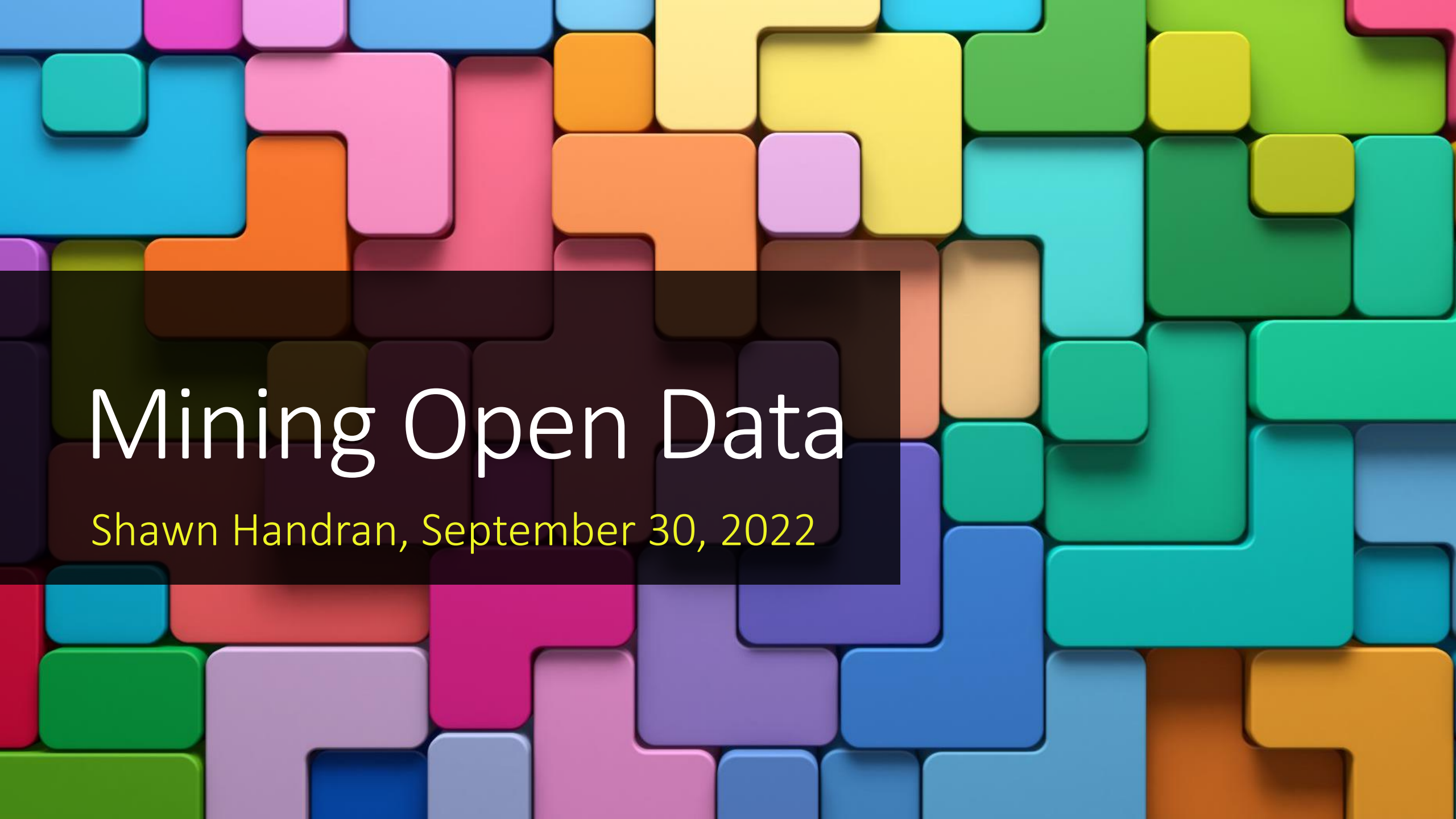

Mining Open Data
Shawn Handran, September 30, 2022

Can you mine open data, completely *in silico*, to make new discoveries?

If yes, who is capable of success?

- Only specialized data scientists and programmers?
- Grad students, post-docs, and PIs? (SMEs)
- High school and undergrad students?
- Citizen scientists?

If yes, what level of success?

Publication-worthy original research manuscripts?

Senior- or masters-level thesis project?

Classroom projects for HS and undergrad students?

The answer so far (IMHO)…

Data scientists and programmers?	Yes
SMEs (Grad, Post-docs & PIs)	Probably yes
Undergrad and HS (typical students)	Probably no
Citizen scientists (non-neuroscientists)	Probably no
Publication-quality original research	Uncertain
Senior- or masters-level thesis projects	Probably yes
Classroom projects (HS & undergrad)	No

Updates since last time

- Completed machine learning Udemy course
- Started another Udemy course on advanced pandas
- Fixed python environment issues (thanks Qiao!)
- Reported patch-seq error to Allen Inst. & they fixed it!
- Reported human patch-seq file problem to Allen Inst.
- Discovered that raw patch-scRNA data is not available
- Got in touch with Nick Turner about his Cell paper

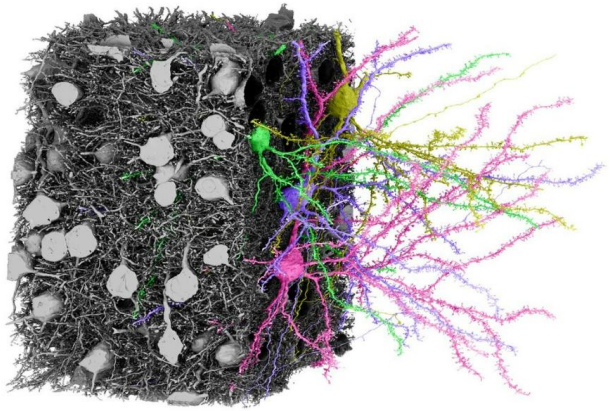
What about encoder/NN analysis?

- I’m not there yet… not sure if I ever will be!!
- I’m still fumbling around with handling data frames in pandas, looking for data on github, trying to understand archived scripts and notebooks, etc.
- I see a huge potential for encoder-based analysis of morphological data from the serial EM volumes

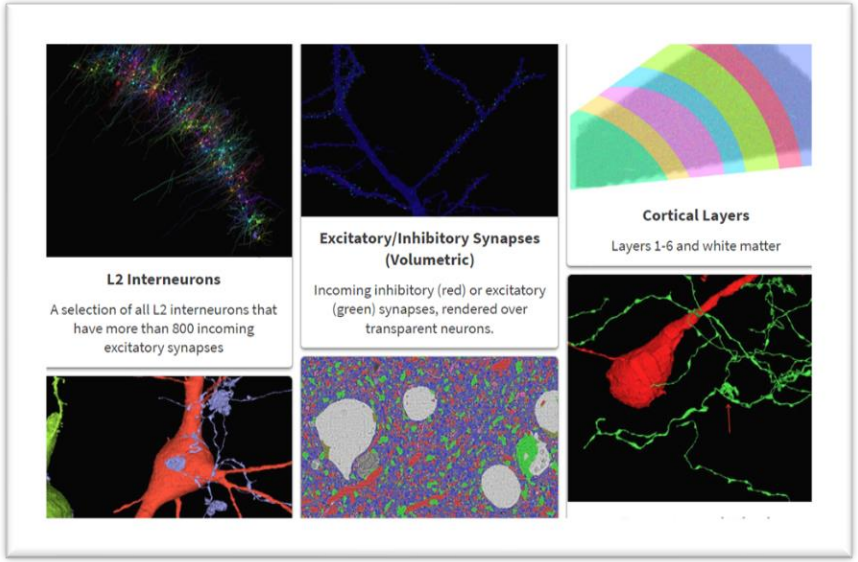
Show and Tell 展示和讲述
（学校针对幼童的一种教学活动）

- Neuroglancer/MicronsExplorer of Layer2/3 EM volume and new mitochondria visualization (Nick Turner)
- New RNAseq tool at Allen Brain Atlas

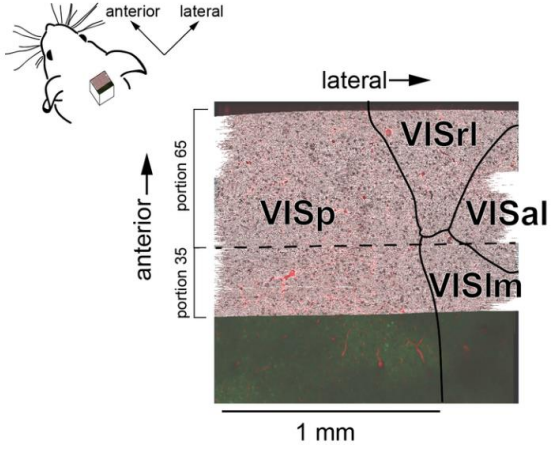
Serial EM volumes available



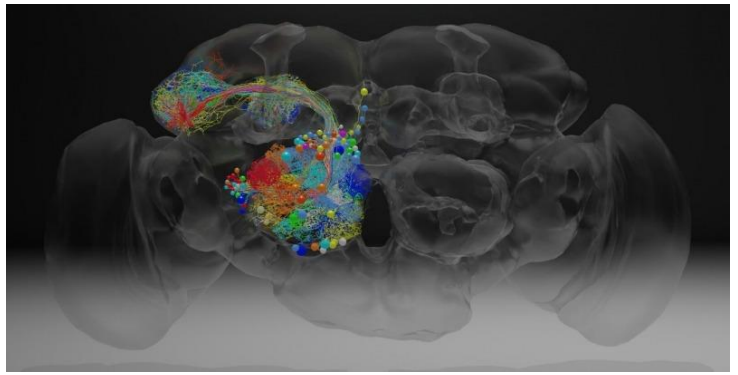
Allen Institute
Mouse
L2/3 volume
0.04 mm^3



Harvard
Human
1 mm^3 volume



Allen Institute
Mouse
1 mm^3 volume



HHMI/Janelia
Drosophila
0.015 mm^3

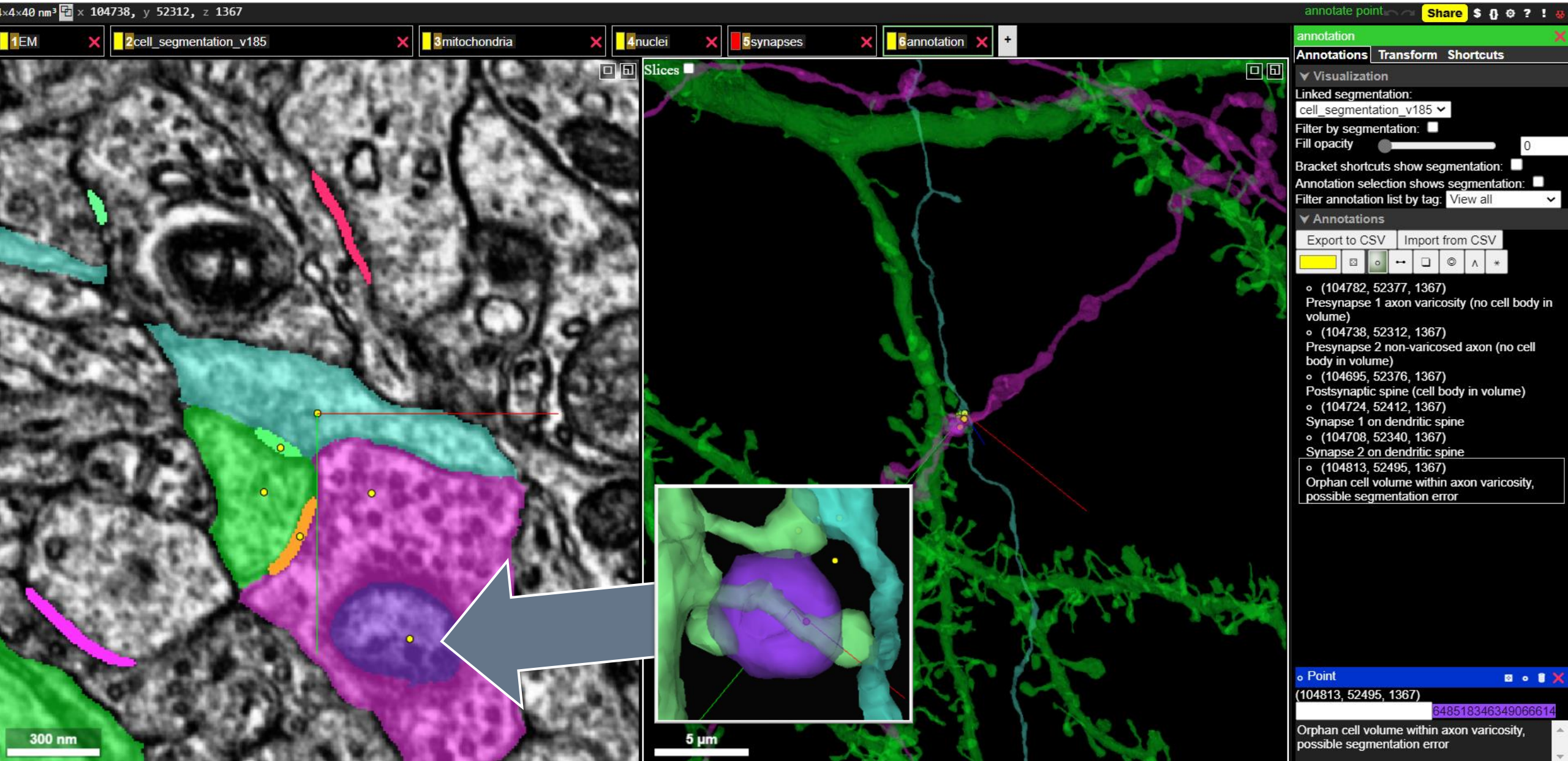
Neuroglancer

- Developed by Seung lab, Princeton
- 3D visualization and segmentation of serial EM reconstructions from brain volumes
- Current technology limits volume size to ~1 mm^3
- Segmentation for nucleus, mitochondria, & synapses with ~90% accuracy (~10% missed or misidentified)
- No segmentation for other organelles or synaptic vesicles

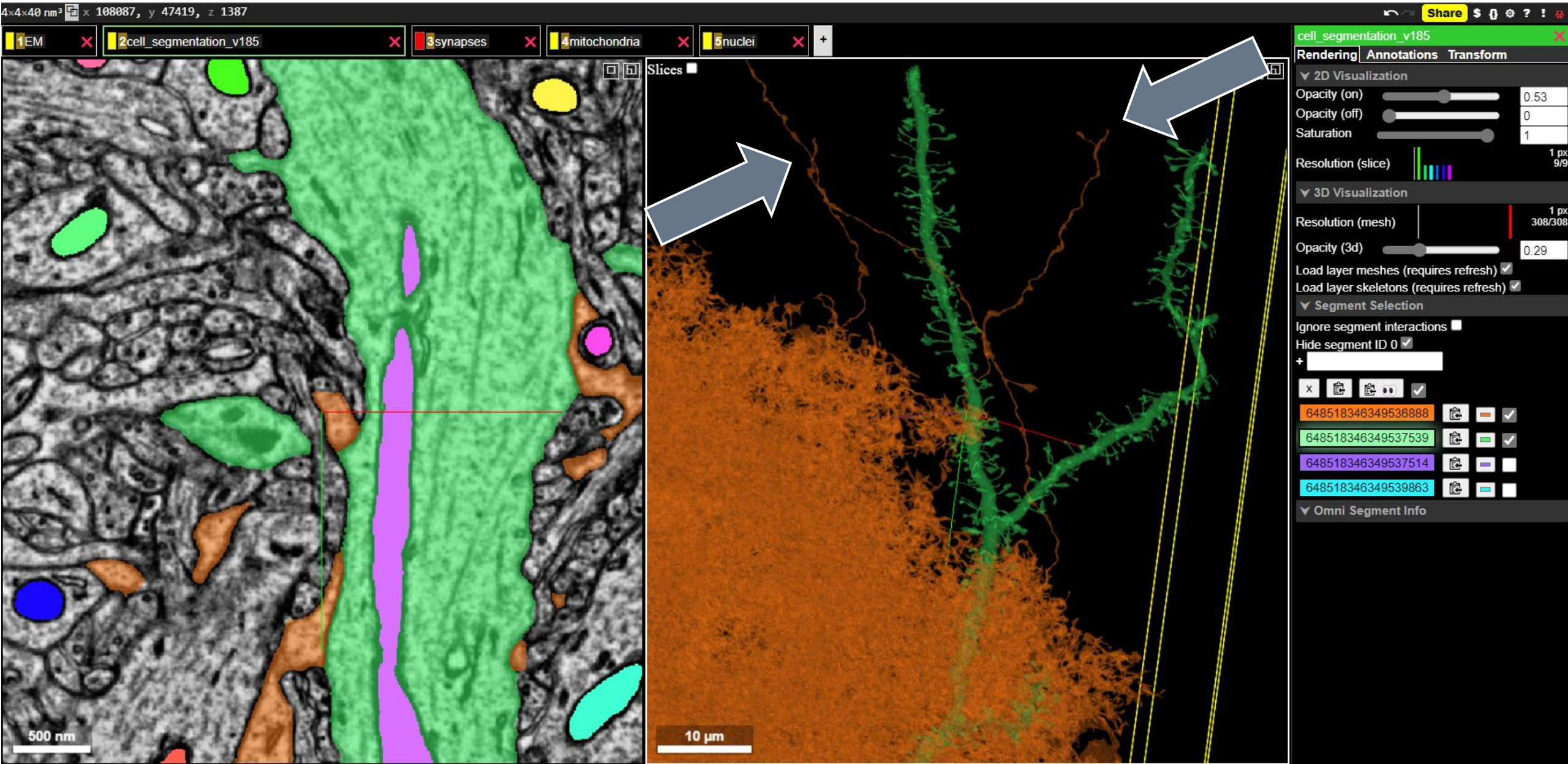
My workflow

- Search for cool looking features in Neuroglancer ☺
- Copy state into a Google Sheets tracking spreadsheet
- Add notes, cell IDs, synapse ID, other feature IDs, etc.
- Some interesting features I've found:
  - Unidentified cell volume inside a compound synapse
  - An orphan axon mis-segmented to an astrocyte
  - Interesting mitochondria near a synapse

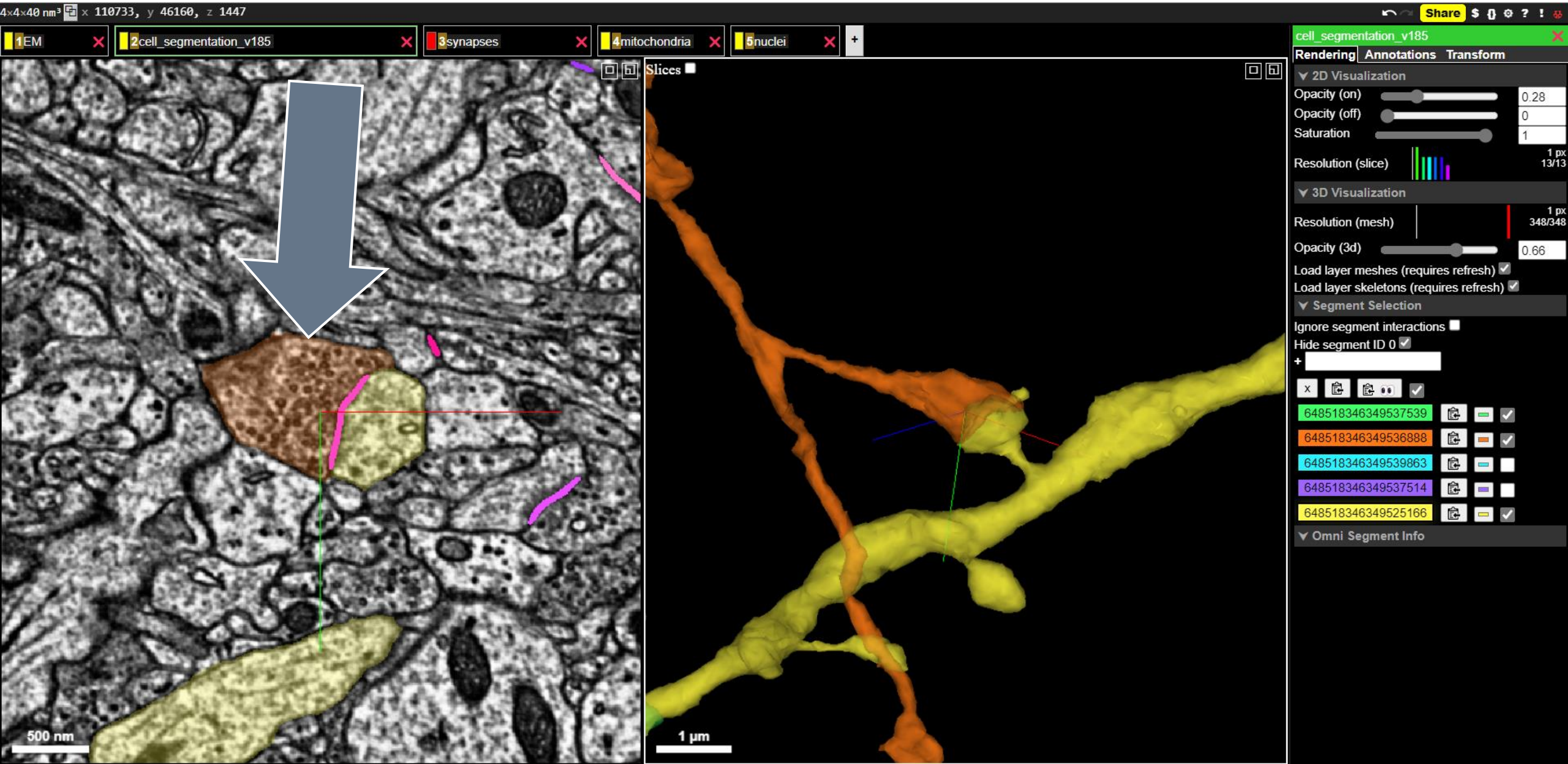
Unidentified cell volume



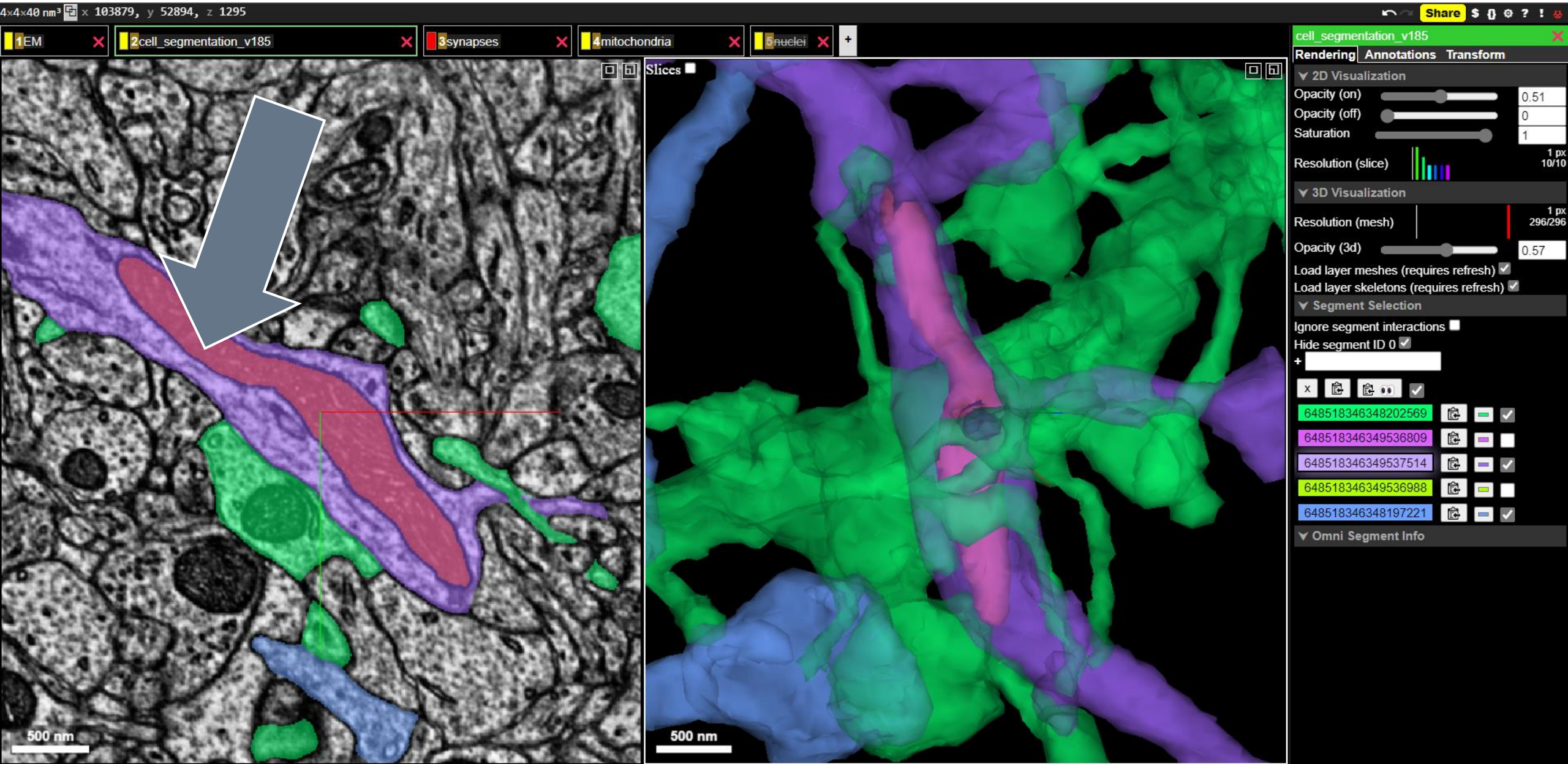
Process mis-identified as an astrocyte



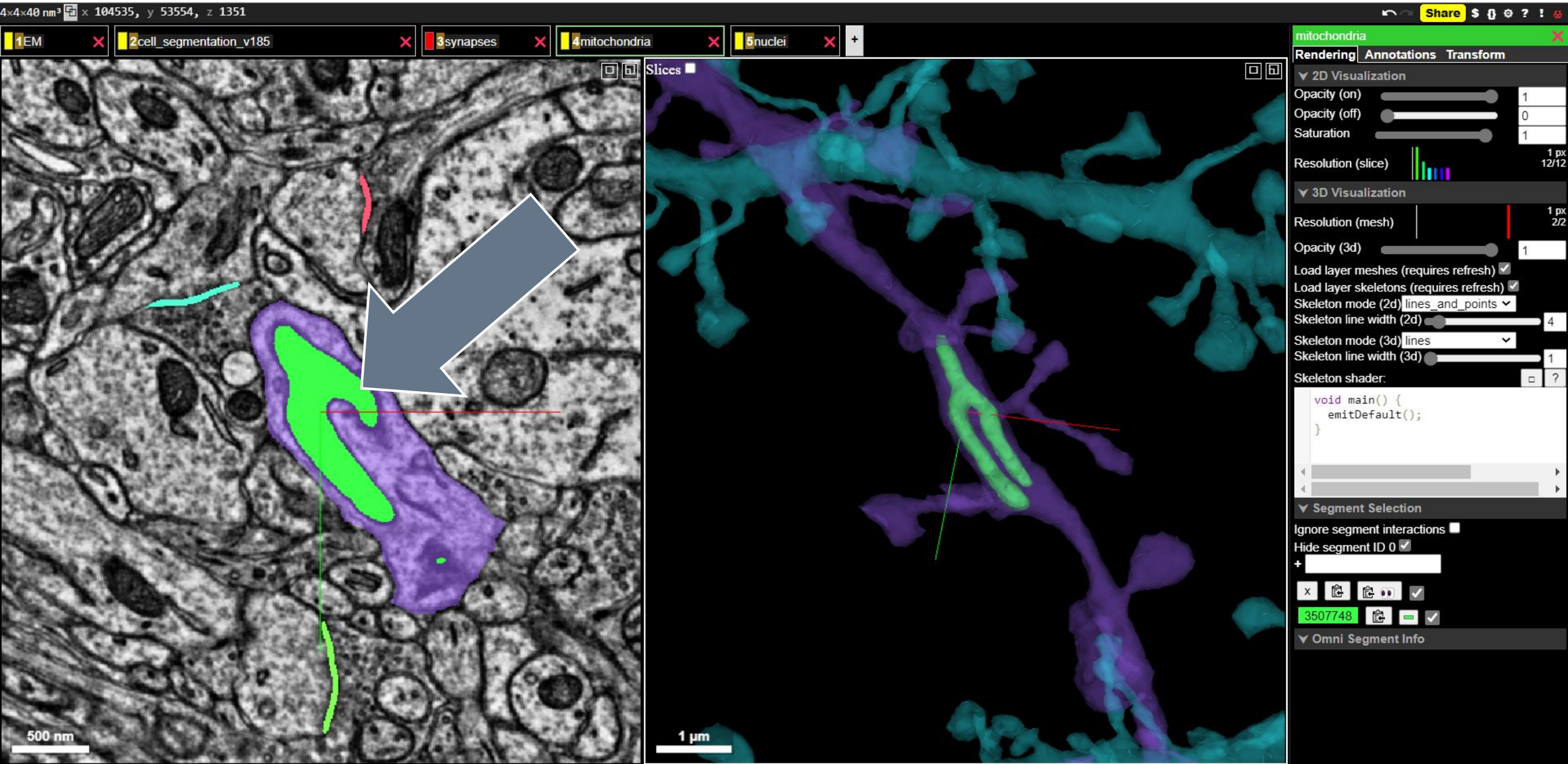
“Astrocyte” process is clearly an axon!



Mitochondrion near synapse



Branched mitochondrion in dendrite



Mitochondria segmentation (New!)

neuroglancer (neuromancer-seung-import.appspot.com)

See: Nicolas Turner, Cell 2022

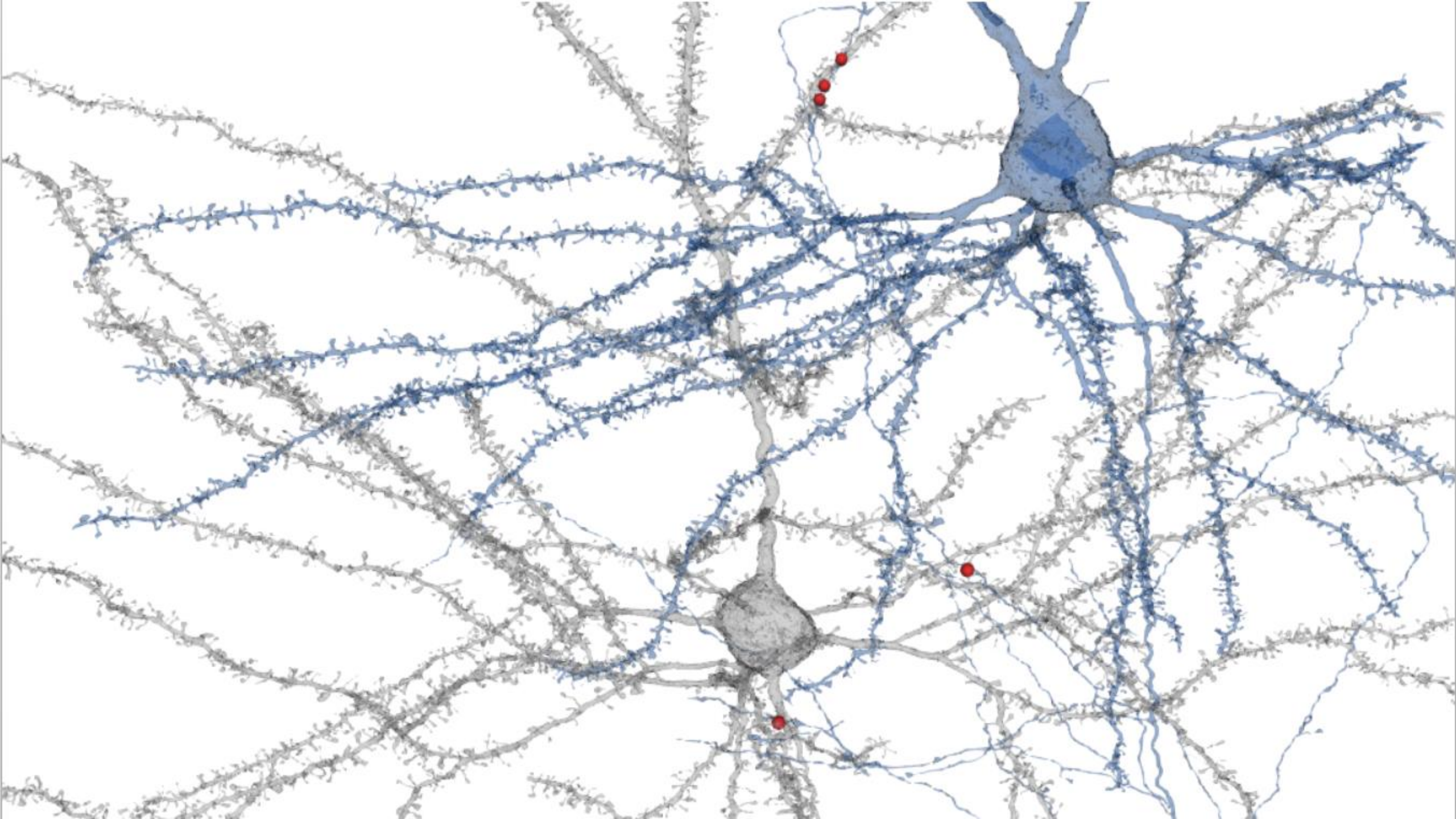
3D Visualizer in Python

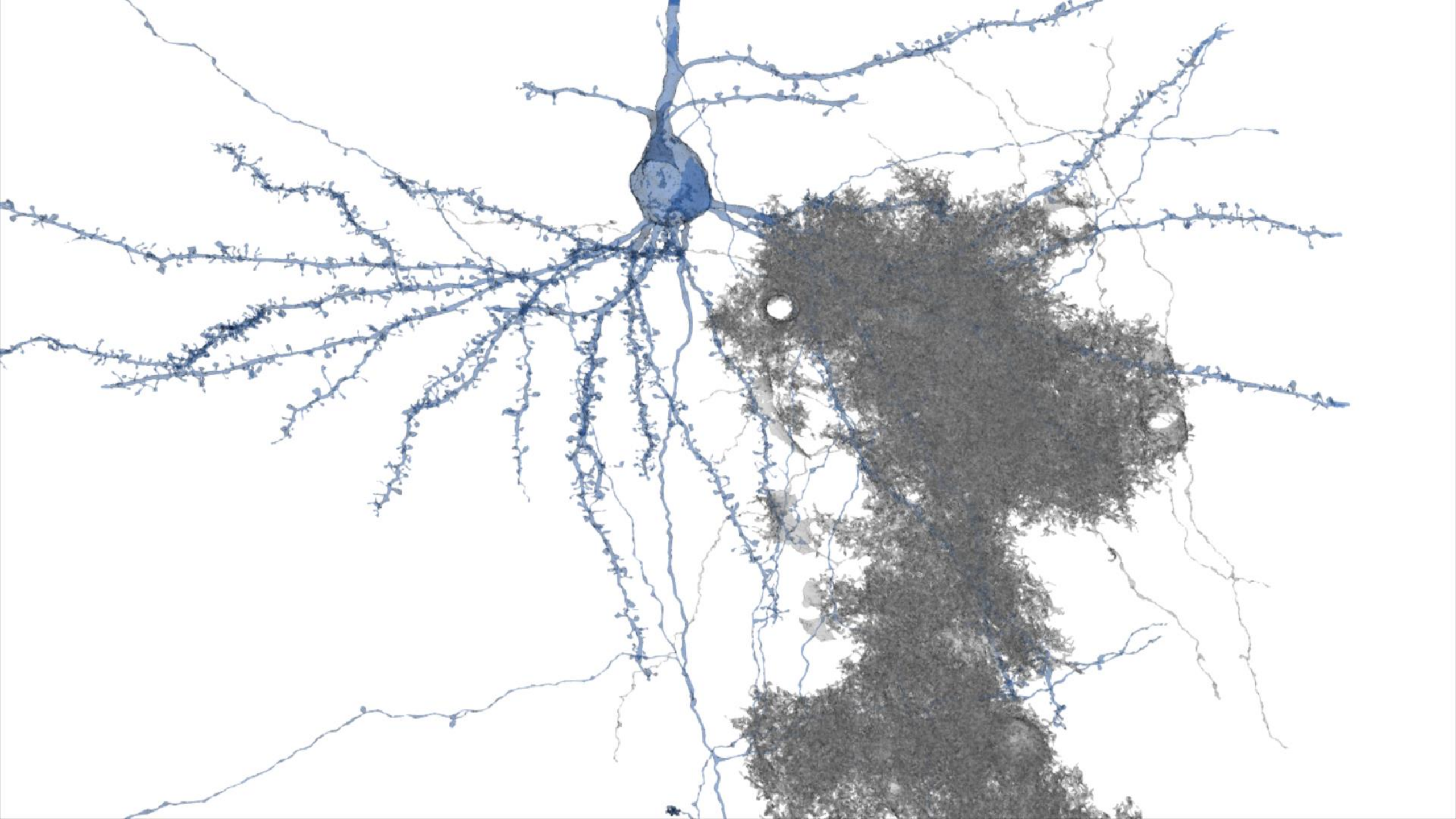
Local > Allen Institute > Microns Explorer > AnalyzingAndVisualizingMeshes.ipynb

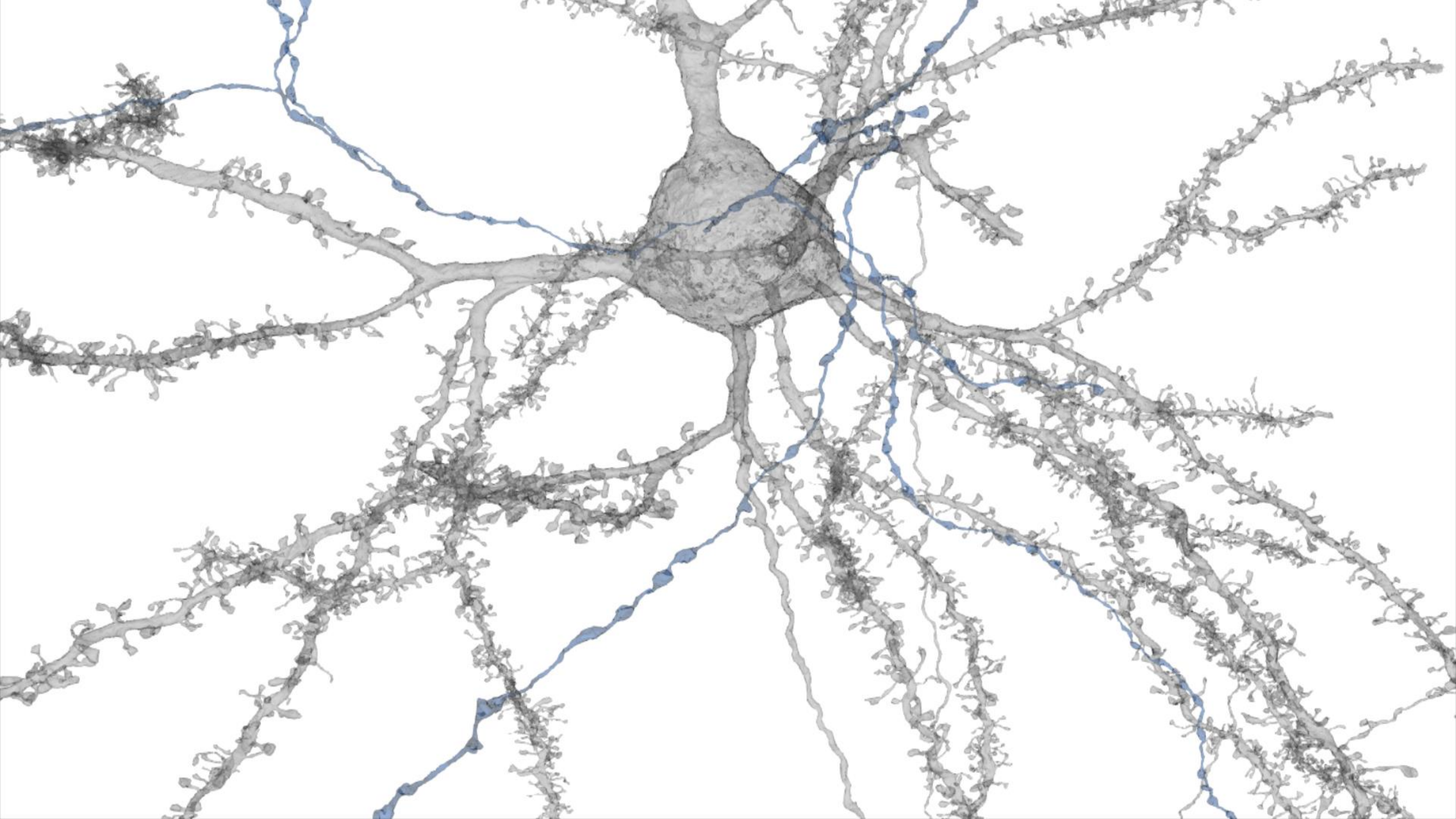
Local > Allen Institute > Microns Explorer > AnalyzingAndVisualizingMeshes_Copy2 Neuron and astrocyte.ipynb

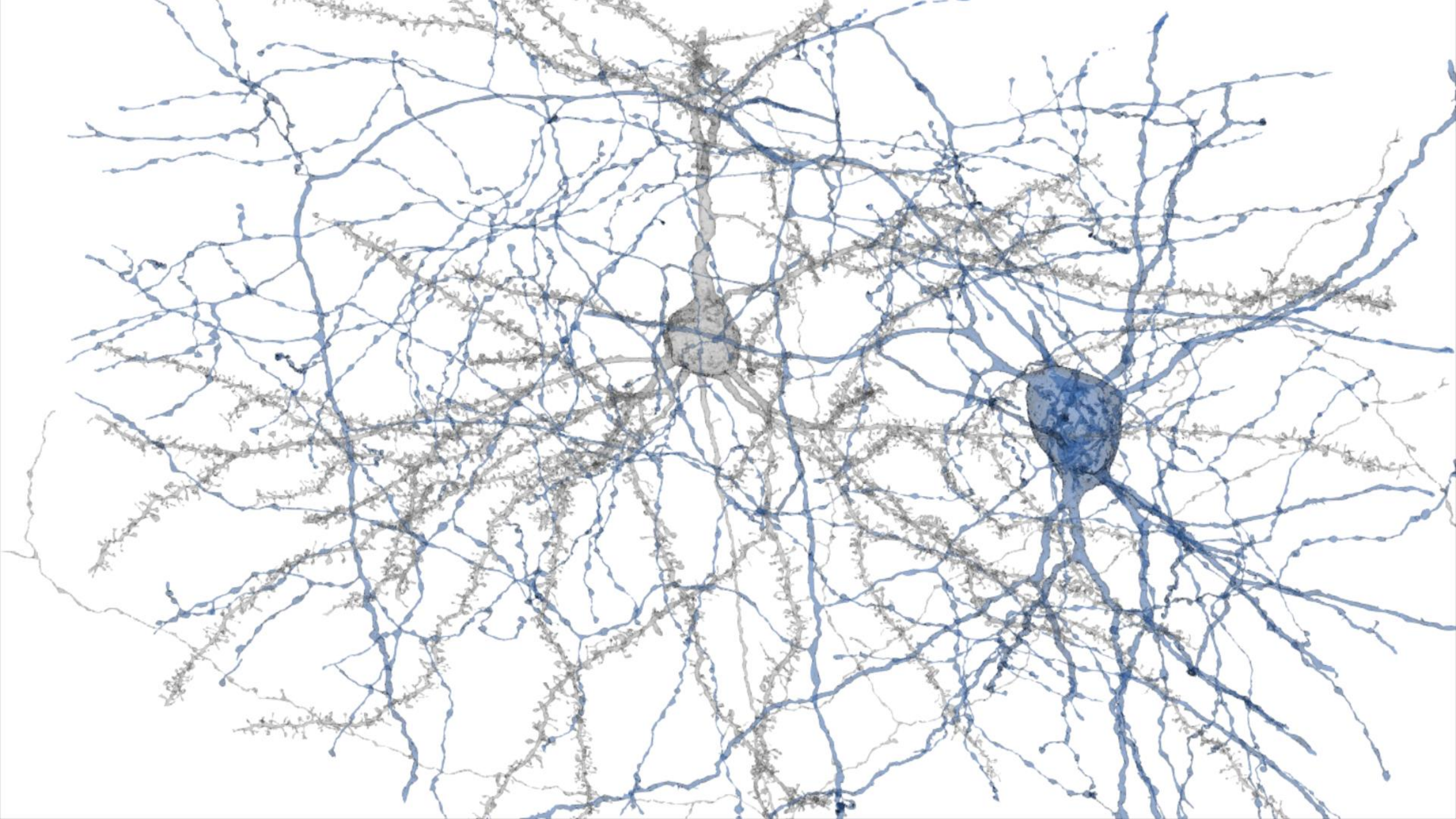
Local > Allen Institute > Microns Explorer > AnalyzingAndVisualizingMeshes_Copy3 Pre axon outside volume.ipynb

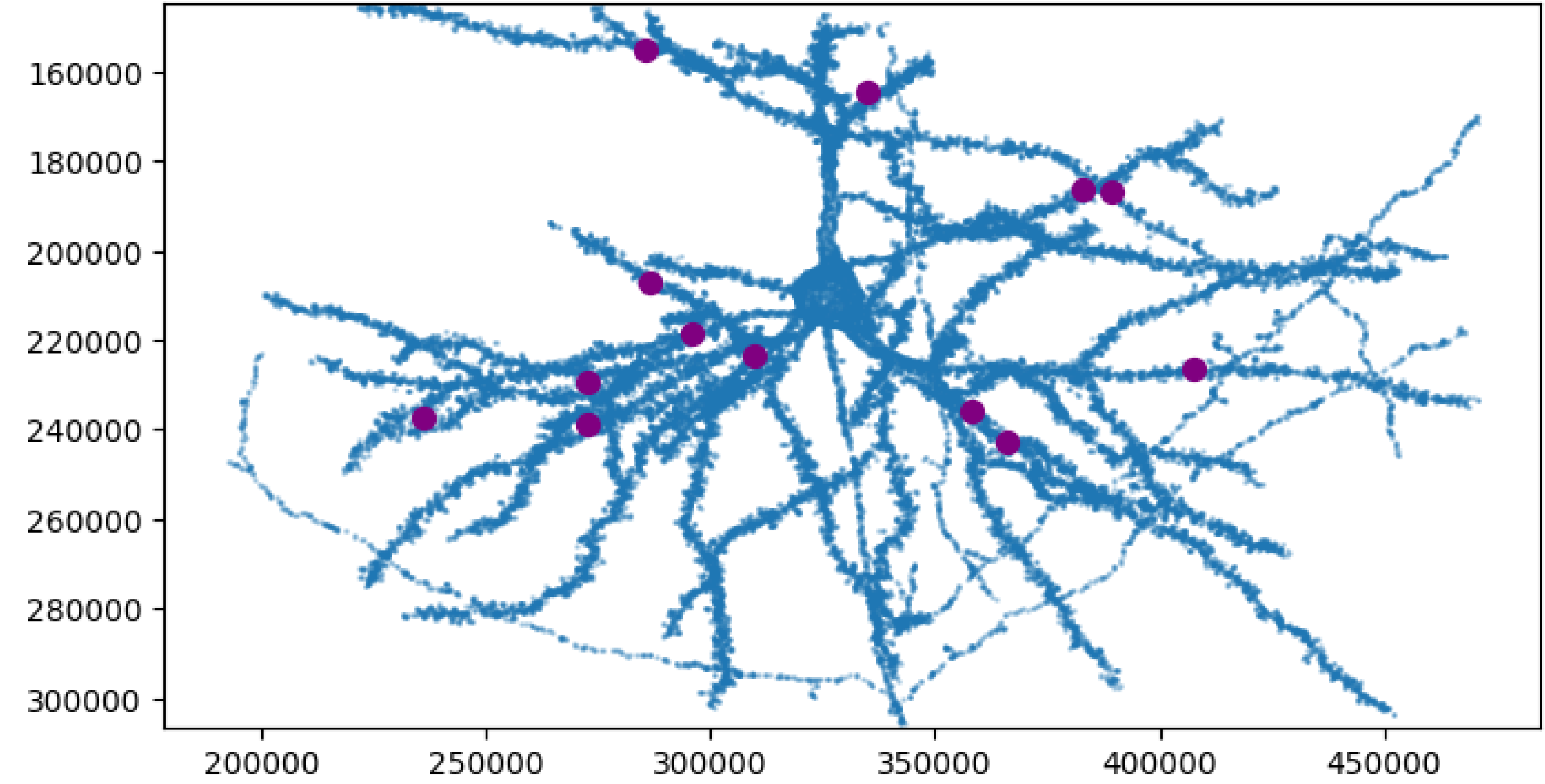
Local > Allen Institute > Microns Explorer > AnalyzingAndVisualizingMeshes_Copy4 presyn mode and postsyn mode number of synapses.ipynb





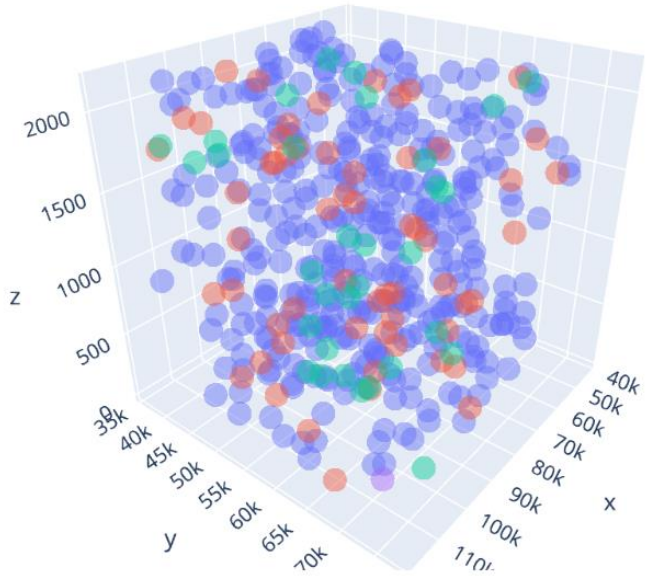


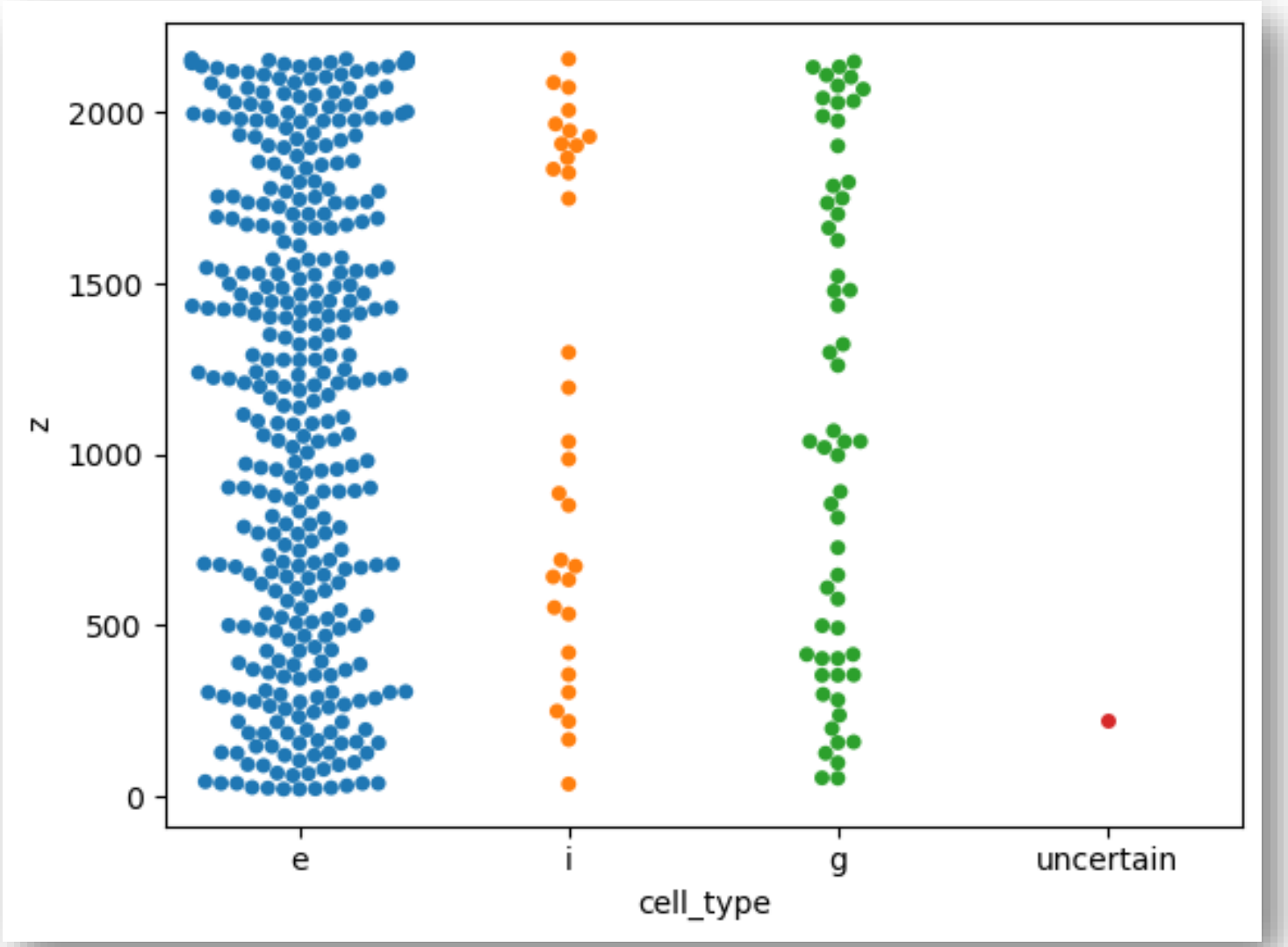



160000
180000
200000
220000
240000
260000
280000
300000
200000
250000
300000
350000
400000
450000

So what? Where's the data?

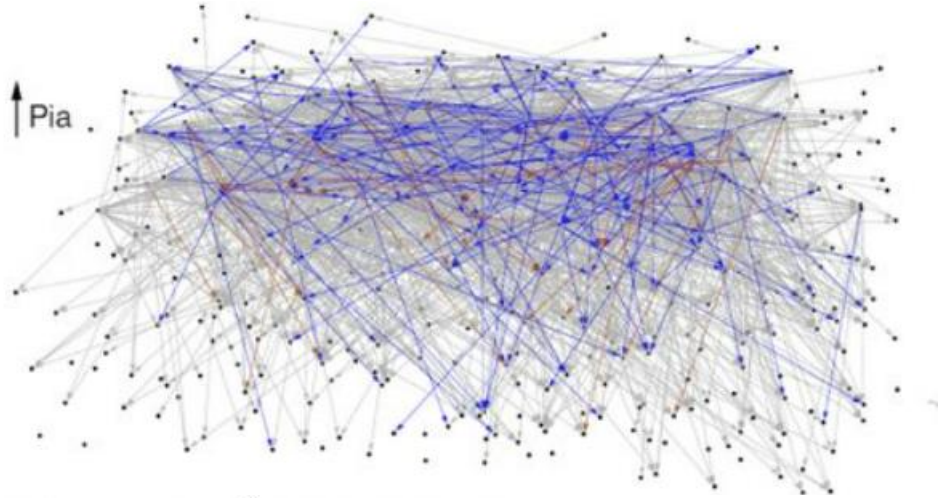
Every cell with a soma in the volume has been identified as excitatory (364), inhibitory (32), glial (59) or uncertain (1)



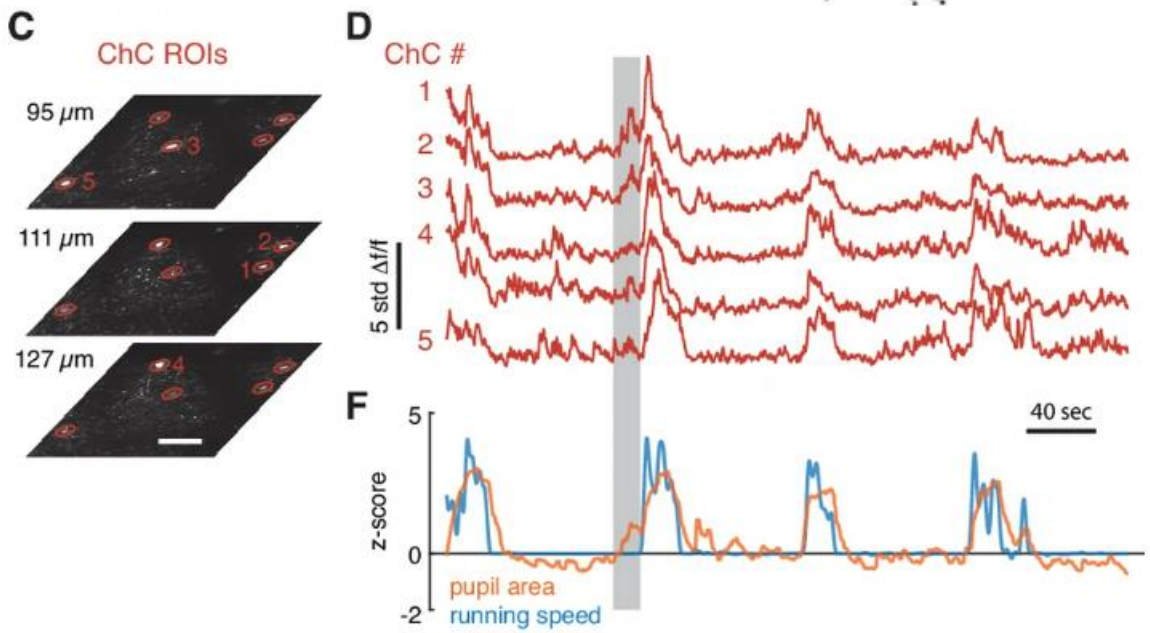


So what? Where's the data?

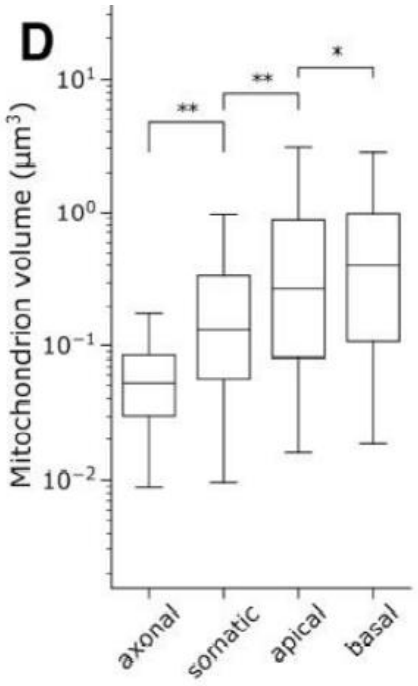
Synapse map?
Dorkenwald biorxiv 2019



Functional correlates?
Schneider-Mizell eLife 2021



Morphological analysis?
Turner Cell 2022



Possible analyses using encoder/NN

- Compartment (4): apical & basal dendrite, soma, axon
- Category (4): excitatory, inhibitory, glial, endothelial
- Neuronal subtype (6): pyramidal, bipolar, basket, chandelier, Martinotti, neurogliaform
- Glial subtype (4): astrocyte, oligodendrocyte, microglia, OPC
- Non-neuronal (3): endothelial, pericyte, OPC-pericyte type
- Morphological (numerous!): distance, volume, branch, etc. of soma, processes, synapses, and mitochondria

Are there still new discoveries waiting to be found in the data?

or did the published data already report most of the potential discoveries?

Allen RNAseq Tool

Transcriptomics Explorer :: Allen Brain Atlas: Cell Types (brain-map.org)

And here:

RNA-Seq Human Data Navigator :: Allen Brain Atlas: Cell Types (brain-map.org)

Unfortunately, full transcriptomic dataset of patch-seq is not released yet (currently limited to cell subtype identification seen on the patch-seq explorer website)

Next steps

- Finish advanced pandas course on Udemy
- Ask Allen Institute about human patch-seq data and other datasets that are still pending/unpublished
- Talk with Nick Turner about how to fix notebook for visualizing mitochondria segmentation in Layer 2/3 volume
- Look at 1 mm^3 EM volumes from Allen and Harvard? (I'm not sure I'm up to this level of complexity)
- Explore HDinHD transcriptomic datasets (Huntington's disease)

Can you mine open data, completely *in silico*, to make new discoveries?

Can it be done (by me)?

- I’m not sure!
- The data handling and programming required to do anything more than a cursory analysis of these datasets is staggering 惊人的
- The burning question I still have is whether original discovery is possible in these large datasets or whether they are simply repositories and archives of already published data with limited exploratory scope