Lab Meeting - Data

Diya

Project Overview

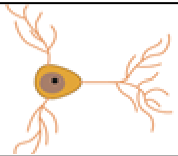
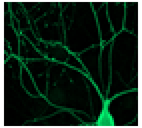
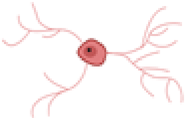
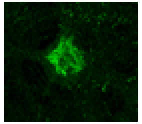
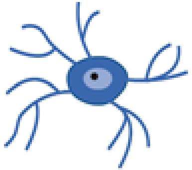
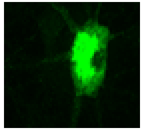
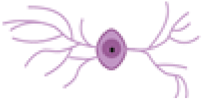
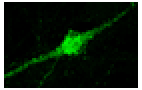
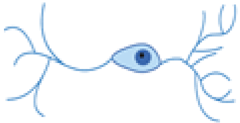
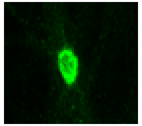
Previous project: Compare the organization of Munc13 and PSD-95 at Ex→Ex synapses vs Ex→PV (parvalbumin-expressing) synapses

Overall goal: Investigate the organization of excitatory synapses onto different inhibitory neuron subtypes (in addition to PV)

1. Develop method(s) to identify inhibitory subtypes
    a. Identify markers for inhibitory subtypes
    b. Determine the populations of various subtypes, ensuring they are significant to investigate
    c. Generate an inhibitory specific cell fill

Inhibitory Neuron Subtype Markers

Cell type	Percent CA1 neurons
Axo-axonic (AAC)	0.04%
Parvalbumin-expressing basket cells (PVBCs)	1.5%
Bistratified Cells (BiCs)	0.7%
Cholecystokinin-expressing basket cells (CCKBCs)	1%
Dendrite targeting CCK expressing Interneurons	0.4%
Oriens lacunosum-moleculare interneurons (O-LMs)	0.5%
Neurogliaform (NGFCs)	1%
Ivy cells (IvCs)	2.5%
Interneuron selective interneurons (ISIs)	2.2%

Inhibitory Neuron Type(s)	Antibody/ Neuropeptide Marker	Probable Morphology	Staining
Axo-axonic and PV-expressing basket cells	Anti-Parvalbumin (PV)		
CCK-expressing basket cells/ CCK interneurons	Anti-Cholecystokinin (CCK)		
Oriens lacunosum-moleculare interneurons	Anti-Somatostatin (SST)		
Neurogliaform and ivy cells	Anti-Neuropeptide Y (NPY)		
Interneuron selective interneurons	Anti-Vasoactive Intestinal Peptide (VIP)		

Results: Larger and Denser Munc13 Clusters in Ex→Inh Synapses, but Variable Munc13 to PSD-95 Ratios Among Inhibitory Neuron Subtypes

Table 1. Cell Type Percentages

Subtype Marker	Average Percentage Per Coverslip
GFP	10.74%
CCK	7.55%
PV	4.95%
SST	2.85%
NPY	3.00%

Table 2. Sample Sizes for Synaptic Analysis

Subtype Marker	Total Number of Cells Analyzed	Total Number of Synapses Analyzed
GFP	30	11519
CCK	14	1112
PV	16	2769
SST	47	6251
NPY	26	4015

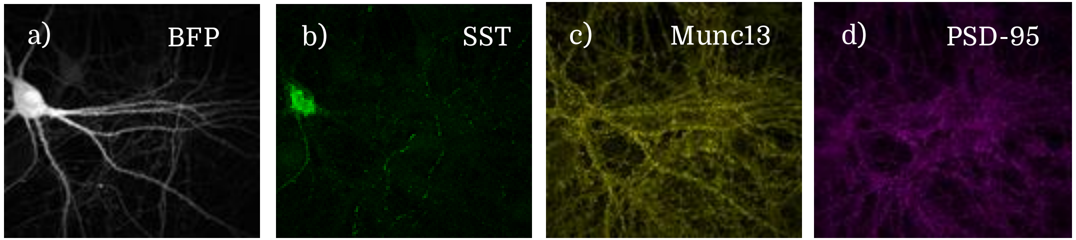
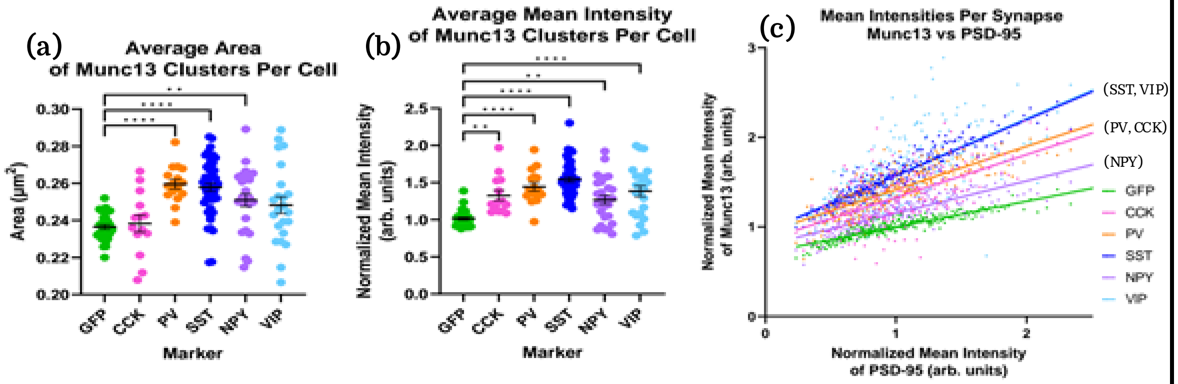


Figure 3. a) The inhibitory specific BFP cell fill visualizes the neuron's dendrites. b) The anti-SST antibody confirms the presence of the SST neuropeptide, identifying the inhibitory subtype (fig 2). c) Synaptic staining shows Munc13 clusters. d) Synaptic staining shows PSD-95 clusters. Taken at 60x.

Figure 4. Average area (a) and intensity (b) of Munc13 clusters collected from confocal images is compared by synapse subtypes shown in figure 2. Asterisk (*) identifies a significant increase compared to Ex→Ex (GFP). c) Munc13 intensity graphed against PSD-95 intensity indicates SST and VIP, PV and CCK, and NPY have statistically different protein ratios.



Further Analysis - Summer Project

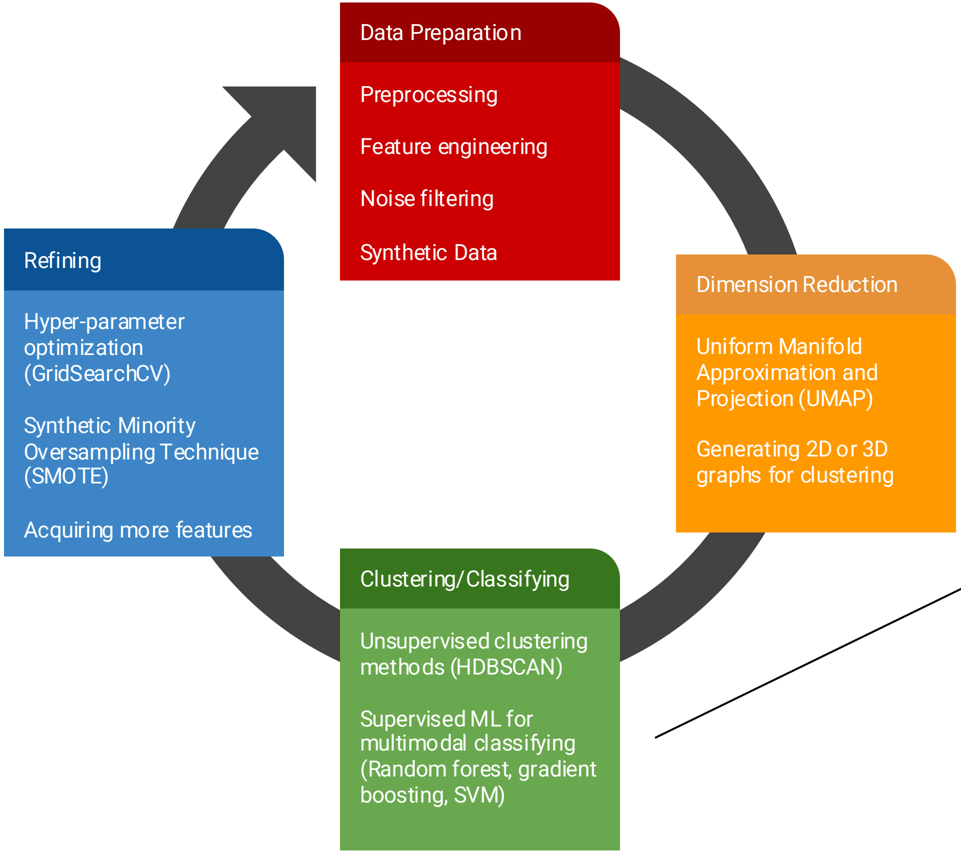
Overview: Categorize Ex->Inh Synapses and Identify Key Features Using Machine Learning Methods

Goals:

1. Determine synapse types/subtypes based on the organization of Munc13 and PSD95
   a. Determine how closely those types correlate to postsynaptic cell type
2. Identify synaptic protein organization features we can leverage to distinguish between synapses with different postsynaptic cell types

A. Combine multiple methods to create data analysis workflow and test/evaluate already existing tools on our data set
B. Compare several algorithms and perform parameter sweeps at each step to define an optimal approach

Analysis Process



Algorithms to Compare

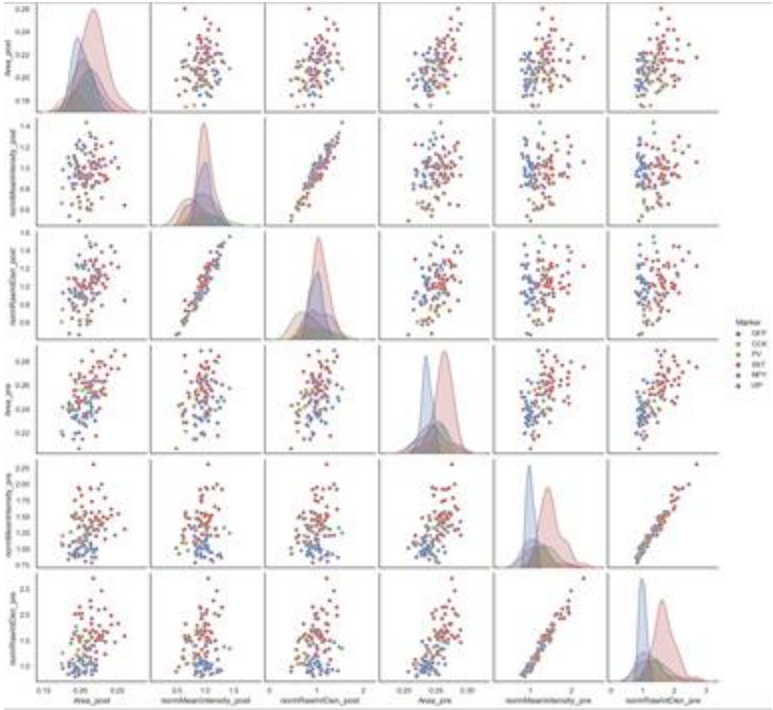
Support Vector Machine: best for high-dimensional data, small to medium-sized datasets, using strategies like one-vs-one or one-vs-all, can handle noisy data with the use of the kernel trick and soft margin approach

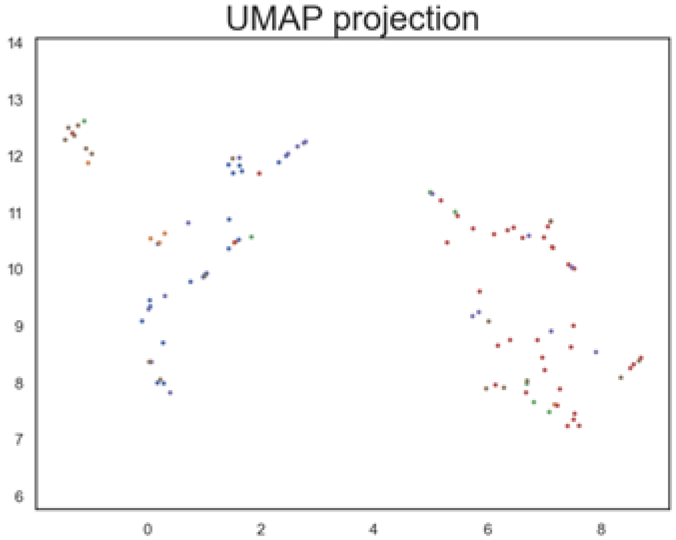
Gradient Boosting: handles complex relationships between features, supports multiclass classification, can handle noisy data well by focusing on difficult-to-classify instances

Random Forest Classifier: handles datasets with a large number of features, inherently supports multiclass classification, robust to noise and can handle overlapping data well by averaging predictions from multiple trees

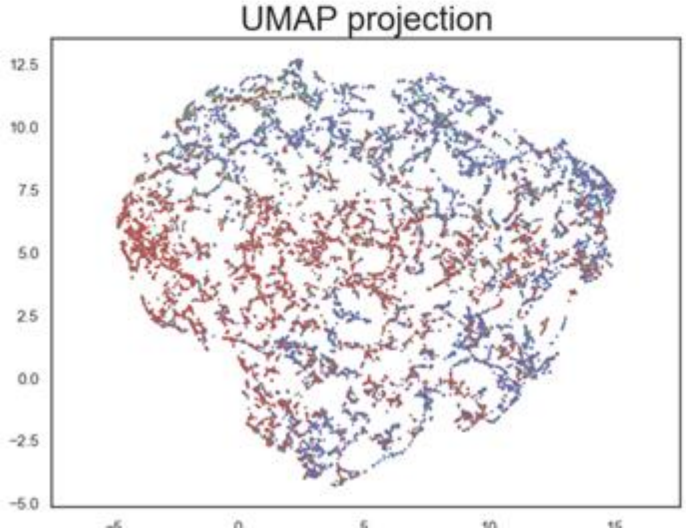
Uniform Manifold Approximation and Projection

6 dimensions to 2 dimensions





Means per Cell

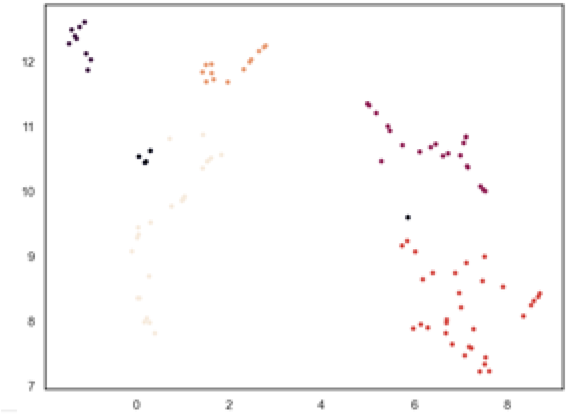


Synaptic

Clustering Algorithm - HDBSCAN

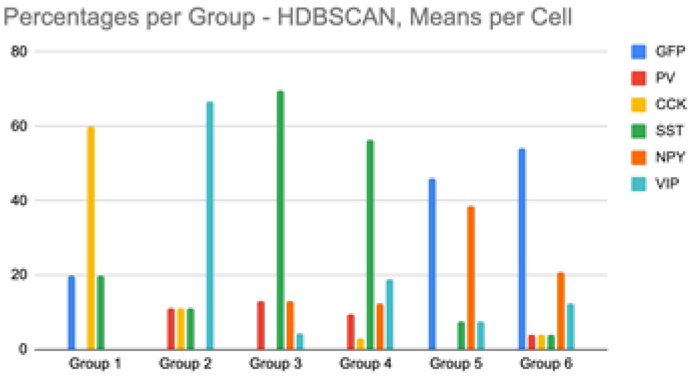
Min_cluster_size = 5

Mean Per Cell

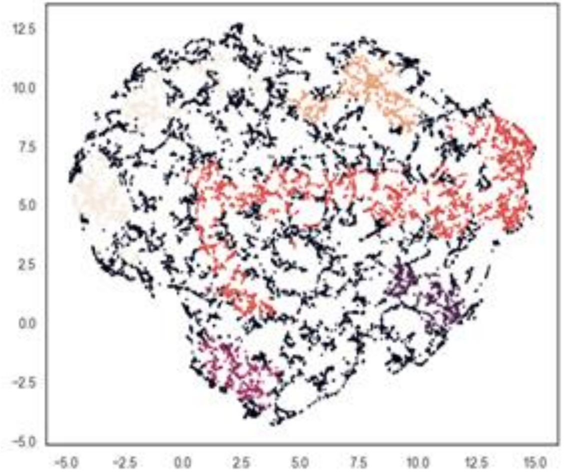


Percentage of each subtype per group

Group	%GFP	%PV	%CCK	%SST	%NPY	%VIP
Group 0	20.0	0.0	60.0	20.0	0.0	0.0
Group 1	0.0	11.11	11.11	11.11	0.0	66.67
Group 2	0.0	13.04	0.0	69.57	13.04	4.35
Group 3	0.0	9.38	3.12	56.25	12.5	18.75
Group 4	46.15	0.0	0.0	7.69	38.46	7.69
Group 5	54.17	4.17	4.17	4.17	20.83	12.5

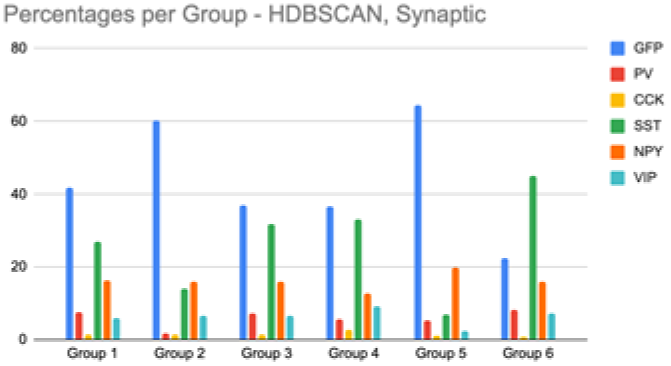


Synaptic



Min_cluster_size = 300

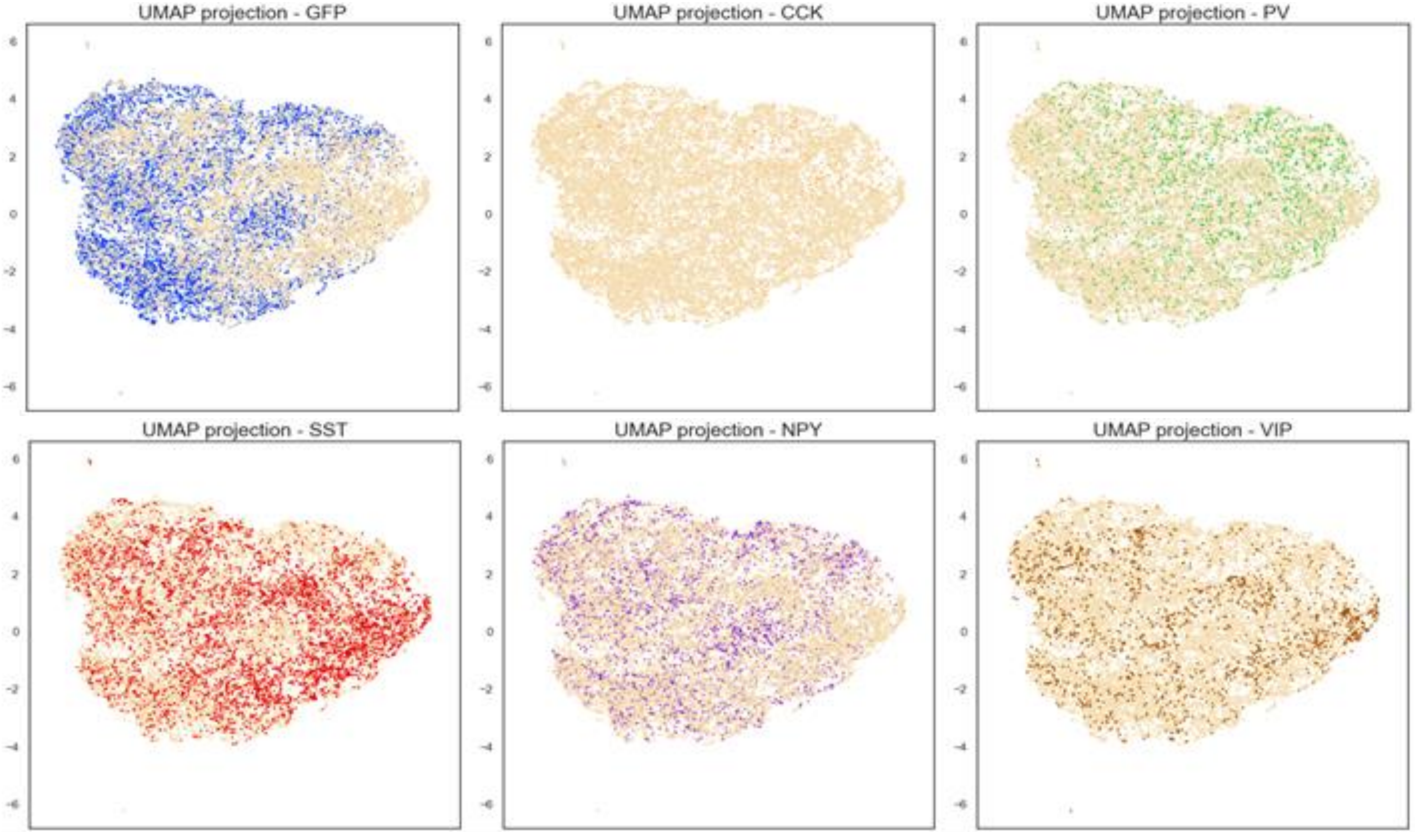
Group	%GFP	%PV	%CCK	%SST	%NPY	%VIP
Group 0	41.85	7.49	1.5	27.08	16.21	5.88
Group 1	60.34	1.72	1.33	14.06	15.92	6.63
Group 2	37.0	7.19	1.53	31.8	15.9	6.57
Group 3	36.78	5.62	2.61	33.15	12.64	9.2
Group 4	64.34	5.43	1.11	6.87	19.93	2.33
Group 5	22.6	8.27	0.81	45.13	15.88	7.31



Acquiring additional features

Reran acquisition in ImageJ to include additional measurements

32 features for Munc13 and PSD-95 (64 total)

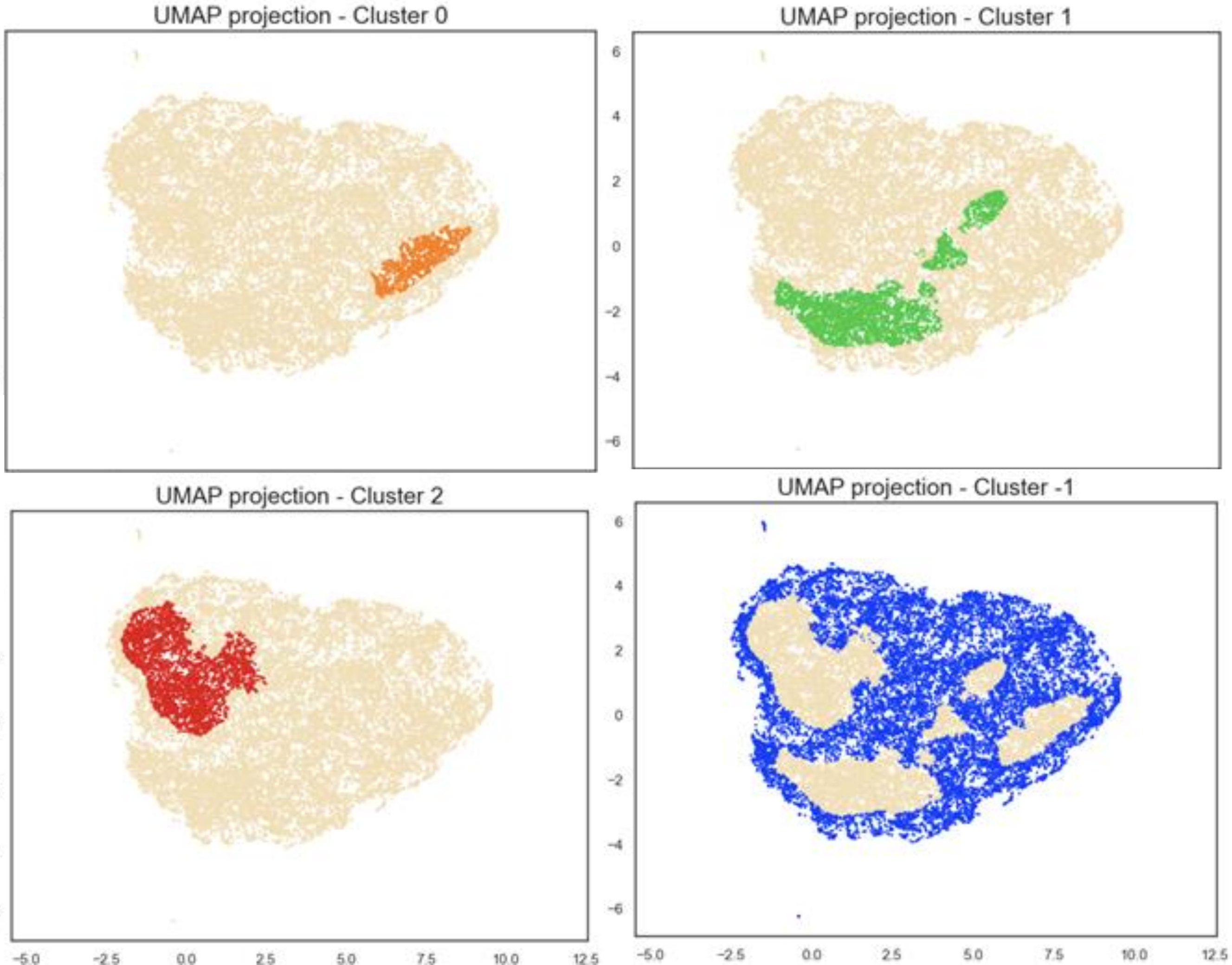


HDBSCAN

Orange - mostly SST
Red - mostly GFP

Group	%GFP	%PV	%CCK	%SST	%NPY	%VIP
Group -1	39.98	11.38	4.26	23.15	15.19	6.04
Group 0	8.73	17.53	3.99	48.08	13.02	8.65
Group 1	51.85	6.64	3.00	20.00	13.84	4.67
Group 2	51.32	6.69	4.43	17.35	14.29	5.91

Min_cluster_size=
1000



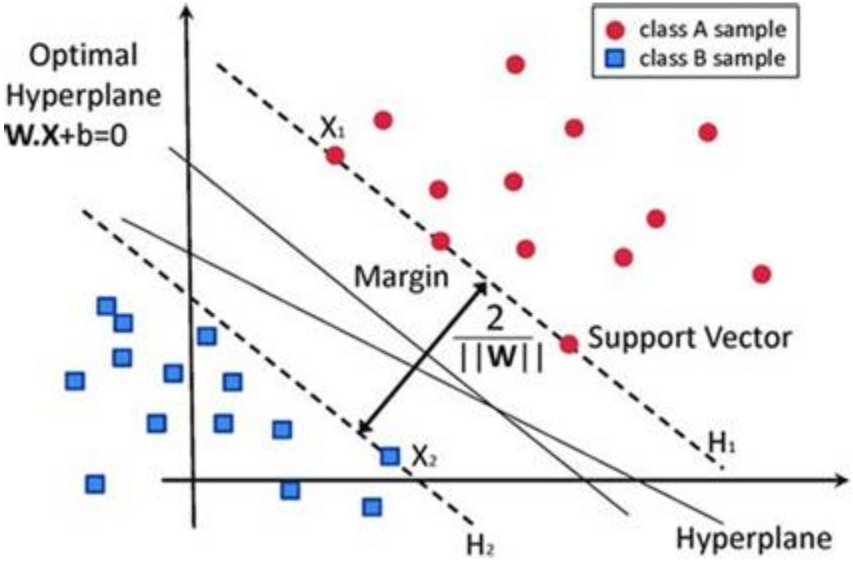
Limitations and Conclusion

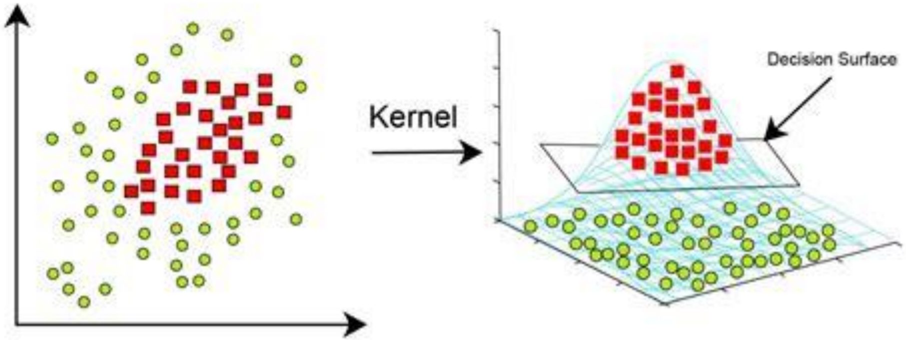
1. Co-expression of markers across multiple inhibitory subtypes
    a. Bistratified cells - SST+ and NPY+
2. Highly variable synaptic staining within certain subtypes
3. Relatively low sample size for ML analysis
4. Features limited to only Munc13 and PSD-95
    a. Unable to create robust protein profile for each synapse type
5. Limited by confocal data rather than super-resolution

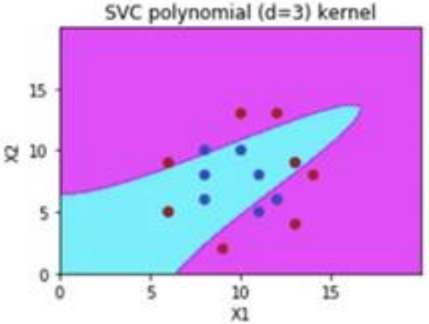
Mainly due to a low sample size and limited features, unsupervised clustering is inconclusive, so we must rely on the classes we identified via staining for further analysis using supervised methods.

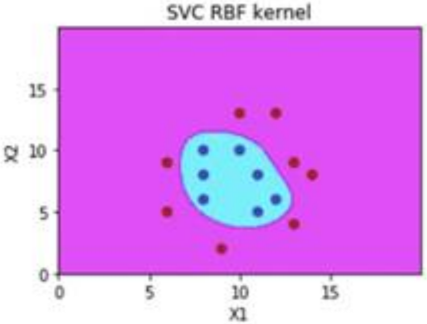
Support Vec

- Establishes hyperplane (decision bound
- Finds optimal equation for the line, with
- Kernels trick - projects data into higher c
  polynomial and Radial Basis Function (F



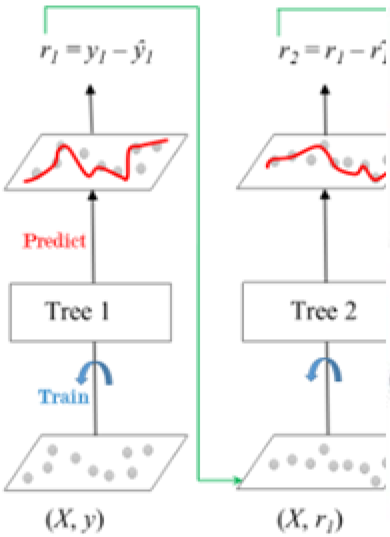


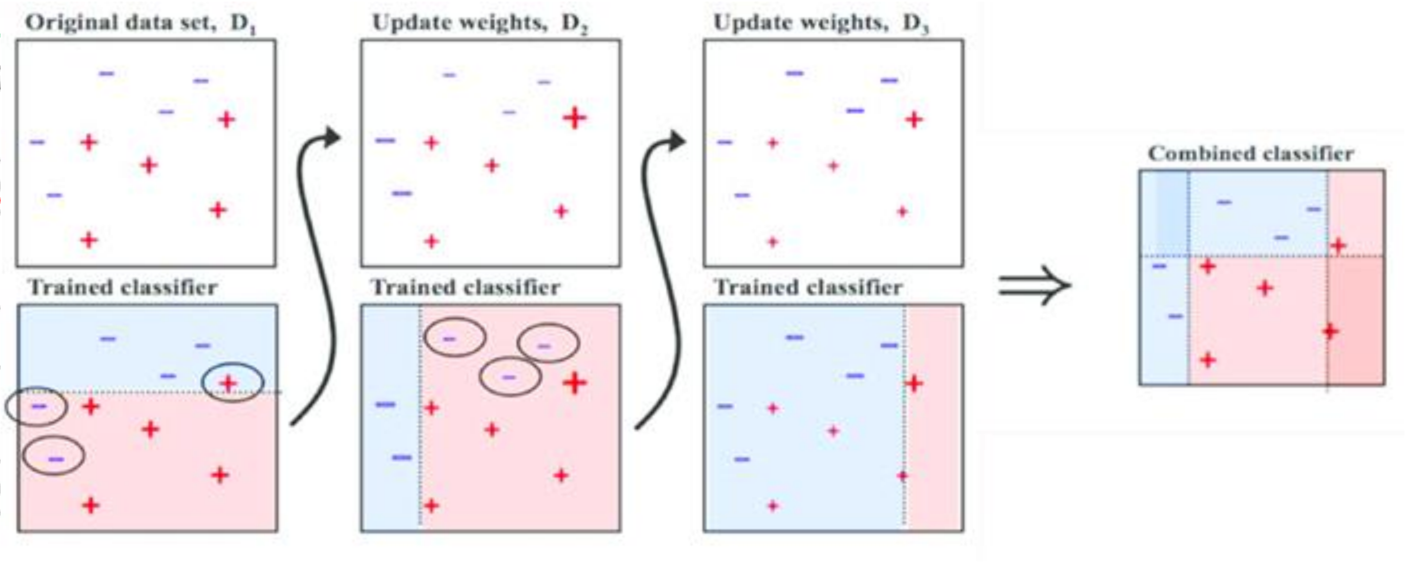




Gradient Boosting

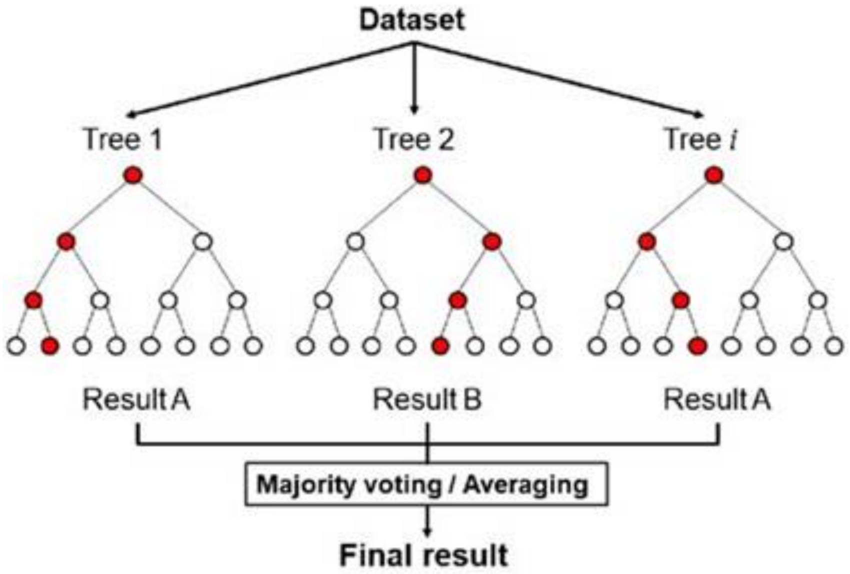
- Ensemble Learning method which trains the model sequentially and each new model tries to correct the previous mode
- Computes the gradient of the loss function, trains a new weak model to minimize this gradient, and adds new predictions to ensemble, repeating until a stopping criterion is met

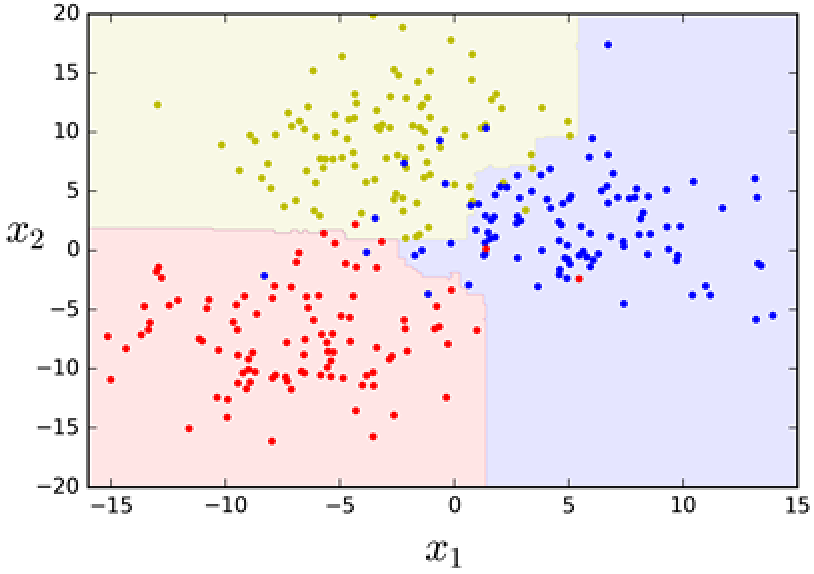




Random Forest Classification

- Also an ensemble method so it combines predictions from other models.
- Each of the smaller models in the random forest ensemble is a decision tree
- Multiple decision trees created using different random subsets of the data and features and most popular prediction among trees chosen



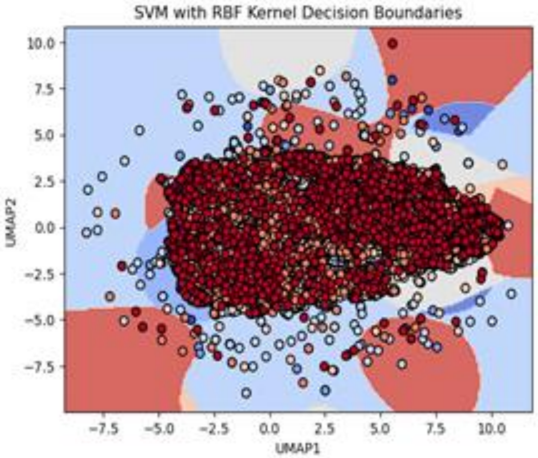


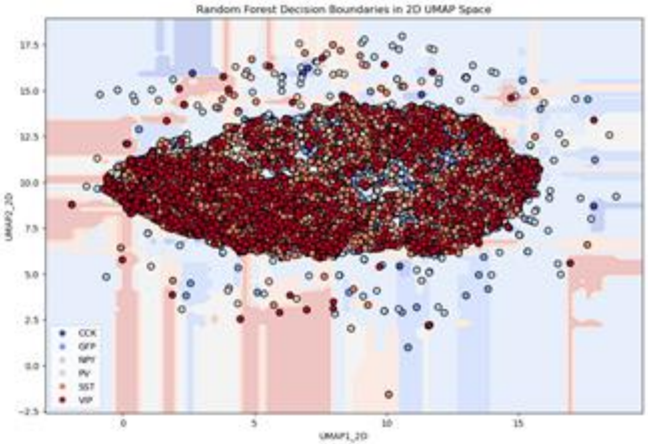
Supervised Learning Algorithms

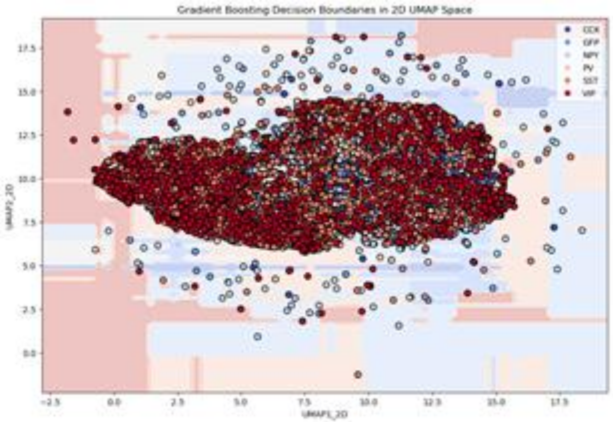
Support Vector Machine: best for high-dimensional data, small to medium-sized datasets, using strategies like one-vs-one or one-vs-all, can handle noisy data with the use of the kernel trick and soft margin approach

Gradient Boosting: handles complex relationships between features, supports multiclass classification, can handle noisy data well by focusing on difficult-to-classify instances

Random Forest Classifier: handles datasets with a large number of features, inherently supports multiclass classification, robust to noise and can handle overlapping data well by averaging predictions from multiple trees

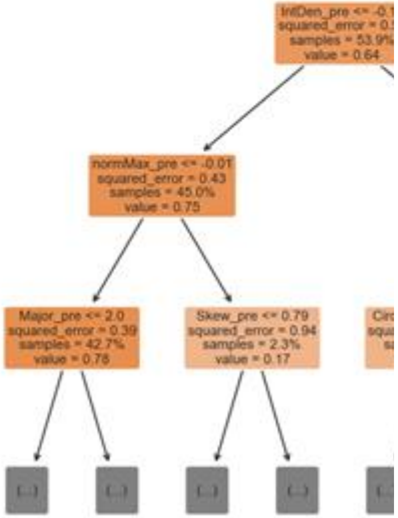


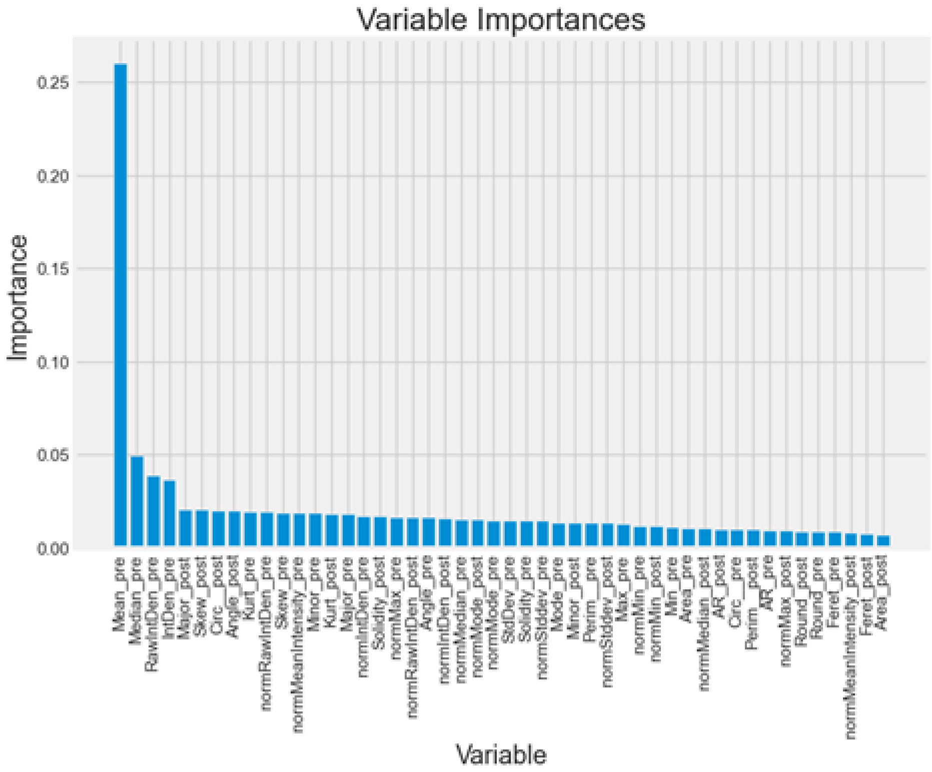




After discussing with Evelina, a research associate from the Ament lab, I learned some techniques that might be a better approach, focusing specifically on data preprocessing and feature selection.

Next Steps - Feature rank/selection





	Feature	Importance
0	Mean_pre	0.260369
1	Median_pre	0.049365
2	RawIntDen_pre	0.038572
3	IntDen_pre	0.036571
4	Major_post	0.020282
5	Skew_post	0.020262
6	Circ__post	0.019980
7	Angle_post	0.019900
8	Kurt_pre	0.019384
9	normRawIntDen_pre	0.019035
10	Skew_pre	0.018881
11	normMeanIntensity_pre	0.018680
12	Minor_pre	0.018482
13	Kurt_post	0.018387
14	Major_pre	0.017992
15	normIntDen_pre	0.017164
16	Solidity_post	0.016866
17	normMax_pre	0.016198
18	normRawIntDen_post	0.016111
19	Angle_pre	0.016065
20	normIntDen_post	0.015956
21	normMedian_pre	0.015218
22	normMode_post	0.014920
23	normMode_pre	0.014812
24	StdDev_pre	0.014803
25	Solidity_pre	0.014778
26	normStddev_pre	0.014670
27	Mode_pre	0.013391
28	Minor_post	0.013229
29	Perim__pre	0.013228
30	normStddev_post	0.013196
31	Max_pre	0.012787
32	normMin_pre	0.011366
33	normMin_post	0.011307
34	Min_pre	0.011174
35	Area_pre	0.010355
36	normMedian_post	0.010251
37	AR_post	0.009755
38	Circ__pre	0.009711
39	Perim__post	0.009683
40	AR_pre	0.009295
41	normMax_post	0.009145
42	Round_post	0.008864
43	Round_pre	0.008686
44	Feret_pre	0.008502
45	normMeanIntensity_post	0.008265
46	Feret_post	0.007465
47	Area_post	0.006643

Next steps - One vs one

- Rather than inputting all the subtypes into one ML algorithm, run a loop that does pairwise comparisons
- Will likely be able to differentiate at least one of the subtypes more successfully than all at once

```
Accuracy: 0.9482273494653911
Confusion Matrix:
 [[3356   97]
 [ 179 1699]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.97      0.96      3453
           1       0.95      0.90      0.92      1878

    accuracy                           0.95      5331
   macro avg       0.95      0.94      0.94      5331
weighted avg       0.95      0.95      0.95      5331
```