Consulting Project

Faculty Supervisor: Masanao Yajima
Teaching Fellow: Shiwen Yang
Group Member: Huaijin Xin, Chenghao Xia, Bolong Xian

2023-12-15

Introduction

The topic of this consulting project is gene distribution in Entorhinal Cortex. Entorhinal Cortex is anatomically positioned between the neocortex and the hippocampus, and its major role is to bridge information exchange between the two regions. Our client, Ana Morello who is a graduate student at department of Anatomy & Neurobiology in School of Medicine in Boston University, recently is doing research about the gene distribution of cells in EC region. They took a monkey's brain and cut it into slices to have several layers of EC section.

She utilized a technique called in-situ hybridization to dye different genes into different colors. In-situ hybridization is a powerful technique used in molecular biology to detect and localize specific DNA or RNA sequences within a tissue section or cell sample. This method involves hybridizing a labeled complementary DNA or RNA probe to the target nucleic acid sequence within the tissue or cells. The probe's label, which can be radioactive or fluorescent, allows for the visualization of the hybridization location, thereby indicating where the specific sequences of interest are expressed within the sample. The specific technique she utilizes is called the RNAscope Multiplex Fluorescent Assay v2 which is a more advanced version of in-situ hybridization designed specifically for the simultaneous detection of multiple RNA targets within a single sample. This technique employs fluorescent labeling, enabling researchers to visualize and quantify the expression of several different RNA molecules at once. The "multiplex" nature of the assay allows for the co-localization of different RNA species within the same sample, providing a comprehensive understanding of gene expression patterns and interactions. Different fluorescent dyes for multiplex fluorescence imaging: Opal 520, 570, 620, 690. Number represents the wavelength in nanometer of light and those also represent different genes in the datasets. The measurement she got is fluorescent intensity which is A measure of the amount of fluorescence emitted by a sample. Fluorescence is a phenomenon where certain molecules absorb light (photons) at one wavelength and then re-emit light at a longer wavelength. Higher the Fluorescent Intensity means higher the concentration of certain gene in the selected cell.

The datasets we get are 3 layers of different fluorescent intensity measures from the reflection of different wavelengths (520, 570, 620, 690) in different cells and the datasets also consists of the horizontal distance between the cell and the edge of the slice of the EC region. And it also has a column which represents which of the 4 genes is positive for this cell. There are still lots of variables in the raw data that we did not use in this project such as the x axis and y axis of the cell.

The goal of the project is firstly count the number of positive cells for different genes, secondly show the correlation between different genes, thirdly show the distribution of four type of genes, and lastly find the relationship of cells between each layers.

Data Cleaning

We have divided the three layers into two datasets: one comprises cells containing Opal_520, while the other includes all cells, whether or not they contain Opal_520. Typically, we utilize the dataset where all cells contain Opal_520. Here is an example showing the first five rows of this dataset:

Class	Opal_520	Opal_570	Opal_620	Opal_690	Distance
520:570:690	0.3483	0.1596	0.0225	0.1164	2871.8301
520:570:690	0.2152	0.1041	0.0196	0.1136	2866.8936
520:570:690	0.5518	0.0258	0.016	0.2296	2861.261
520:690	0.4816	0.0202	0.02	0.3153	2868.6372
520:570	0.2459	0.1088	0.0211	0.0229	2918.8682

In a separate dataset, we assess the presence of genes in each cell, incorporating four additional columns with boolean outputs. This dataset is exclusively employed for generating 3D plots. Below is an illustration featuring the first five rows of this particular dataset:

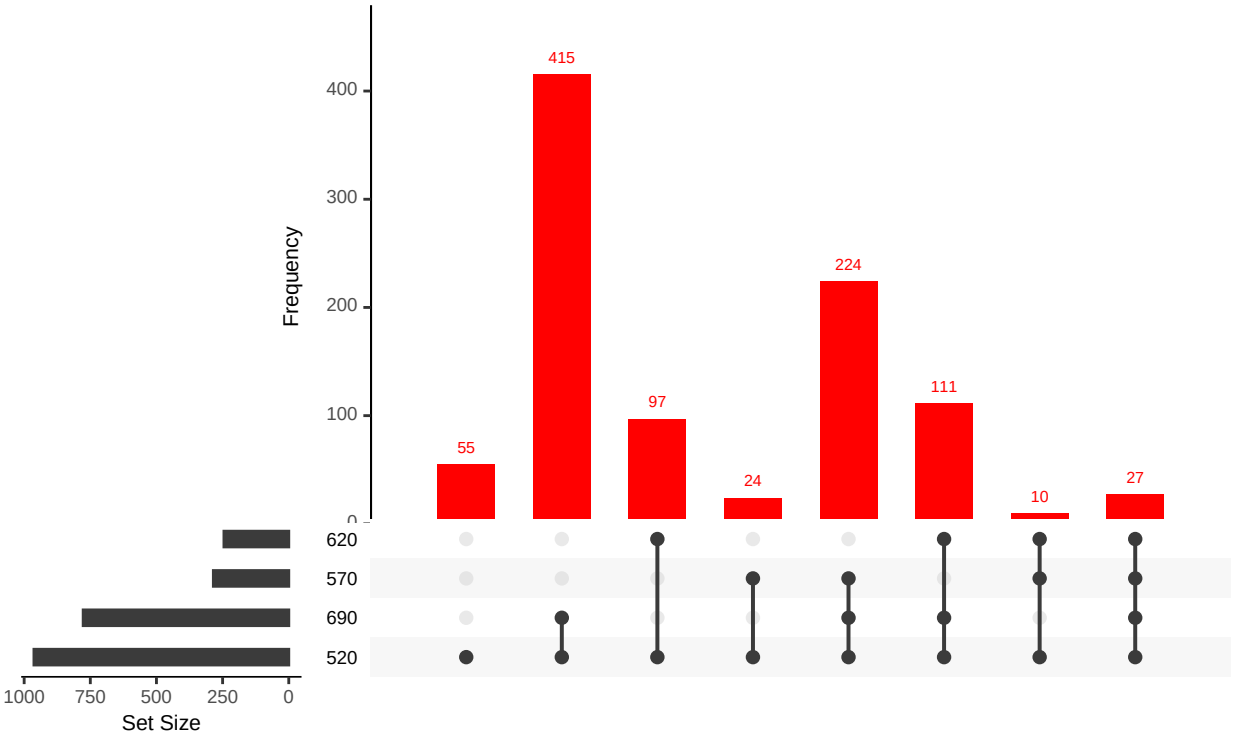
MFI520	MFI570	MFI620	MFI690	dist	IND520	IND570	IND620	IND690
0.3483	0.1596	0.0225	0.1164	2871.8301	TRUE	TRUE	FALSE	TRUE
0.2152	0.1041	0.0196	0.1136	2866.8936	TRUE	TRUE	FALSE	TRUE
0.5518	0.0258	0.016	0.2296	2861.261	TRUE	TRUE	FALSE	TRUE
0.4816	0.0202	0.02	0.3153	2868.6372	TRUE	FALSE	FALSE	TRUE
0.2459	0.1088	0.0211	0.0229	2918.8682	TRUE	TRUE	FALSE	FALSE

Both datasets undergo a cleaning process wherein values associated with fluorescent intensity equal to 0 are eliminated. Additionally, we have renamed certain column names for better clarity and understanding. When examining the fluorescent intensity for each gene, we employ the logarithm to enhance our ability to visualize the distribution.

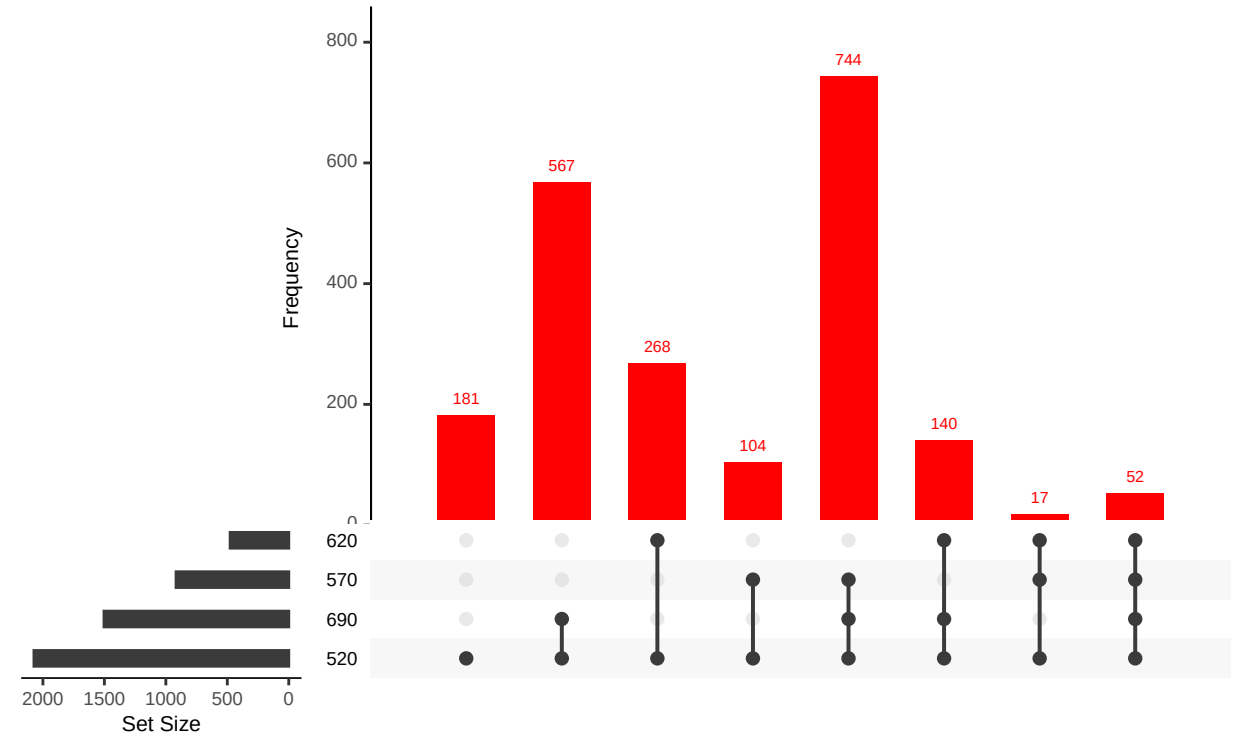
Visualization

Upset Plot

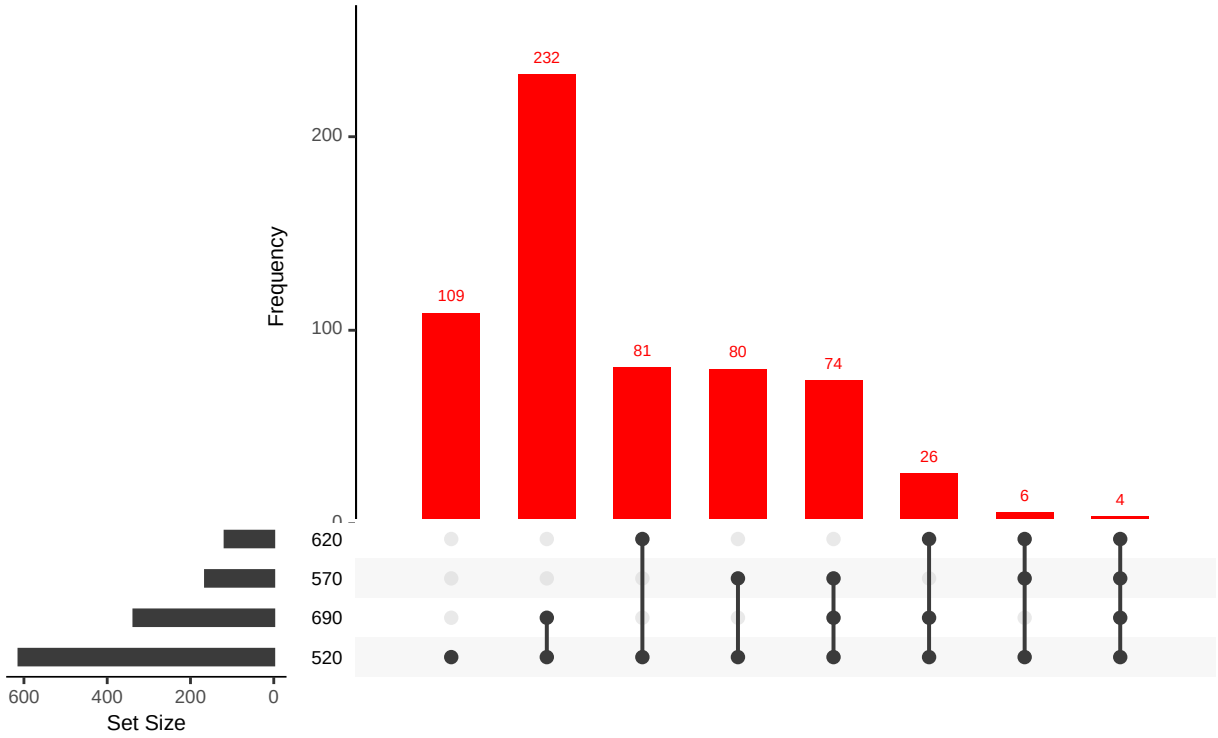
We employ the UpSet plot, a powerful visualization tool, to gain insights into the distribution of elements across three distinct layers in our dataset. The UpSet plot provides a comprehensive overview of intersecting sets, showcasing the frequency and relationships among them. Specifically, it allows us to explore how various combinations of elements from the three layers contribute to the overall distribution, enabling a nuanced understanding of patterns and overlaps in the data.



The above figure illustrates the distribution for layer12_4. By employing data that includes only cells containing Opal_520, the count of Opal_520 serves as a direct indicator of the number of cells in the dataset, totaling around 1000. Notably, within this layer, the cell combination 520:690 exhibits the highest frequency, with a count of 415.



The figure above depicts the distribution for layer12_5. Utilizing data that exclusively includes cells containing Opal_520, the count of Opal_520 directly represents the number of cells in the dataset, totaling approximately 2000. Notably, within this layer, the cell combination 520:570:690 exhibits the highest frequency, with a count of 744.

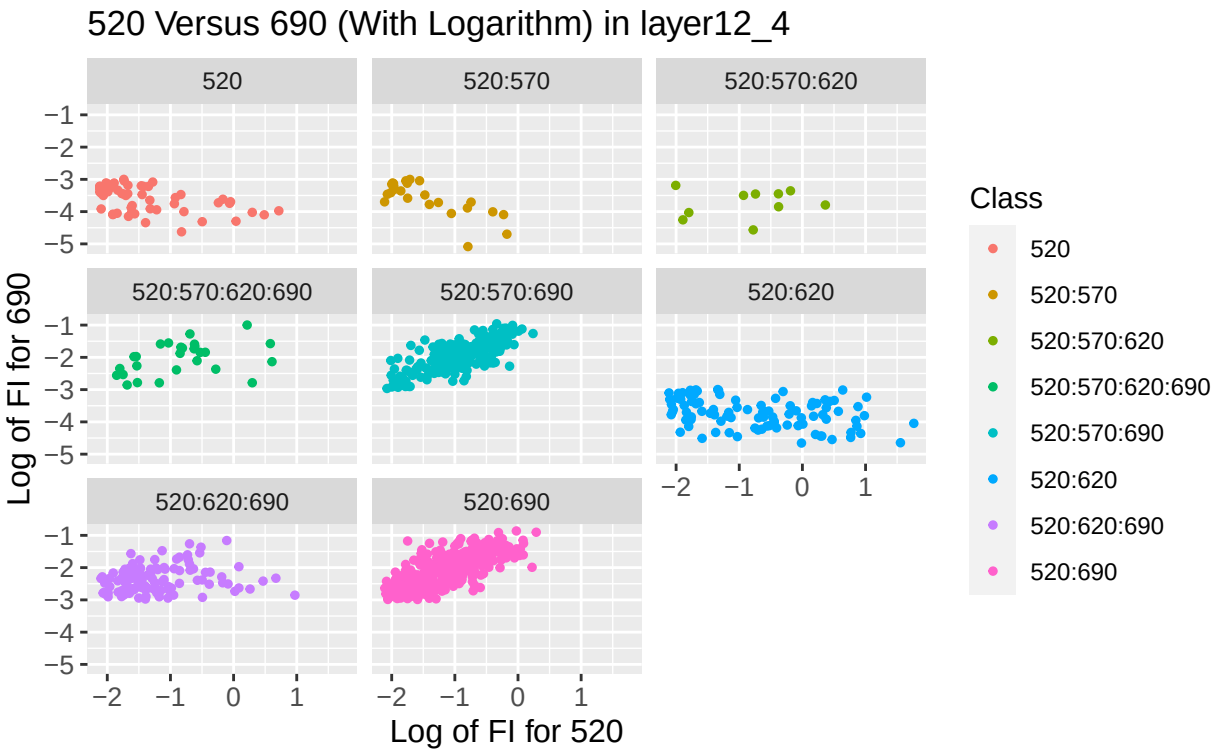


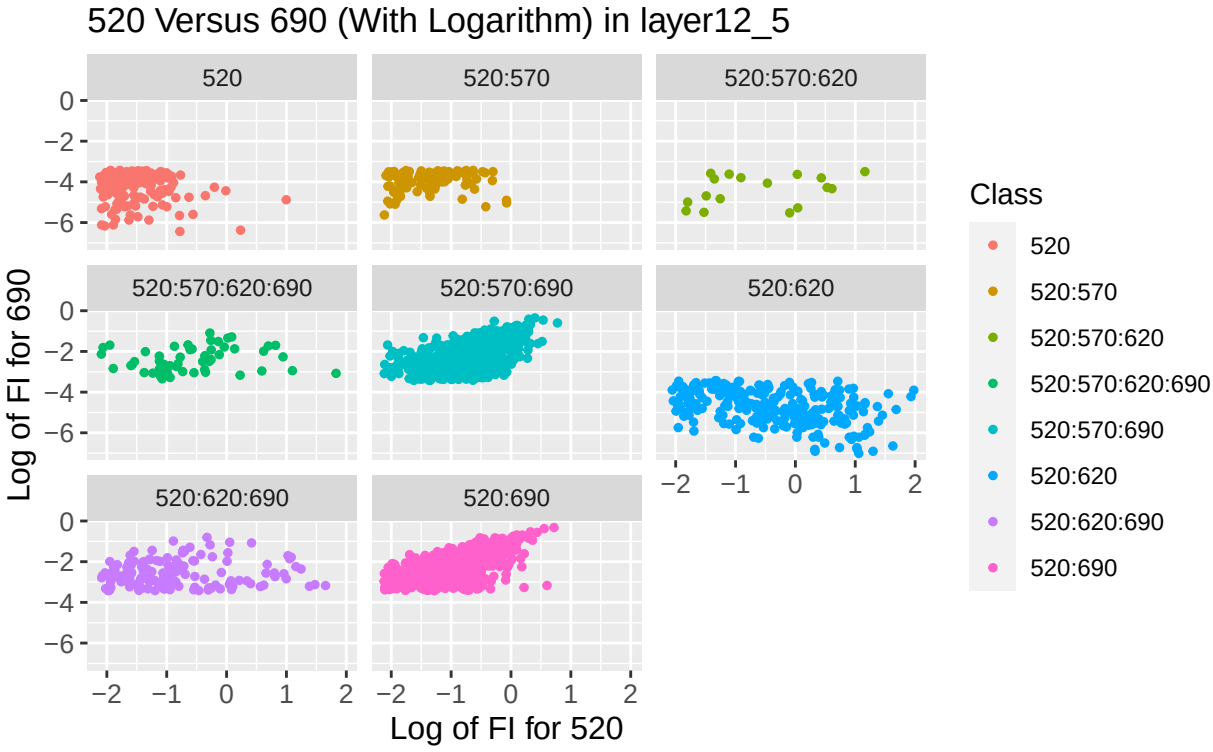
The figure above illustrates the distribution for layer20_5. Utilizing data comprising cells with Opal_520, the count of Opal_520 directly corresponds to the number of cells in the dataset, totaling approximately 600. Notably, the cell combination 520:690 exhibits the highest frequency, with a count of 252 within the layer.

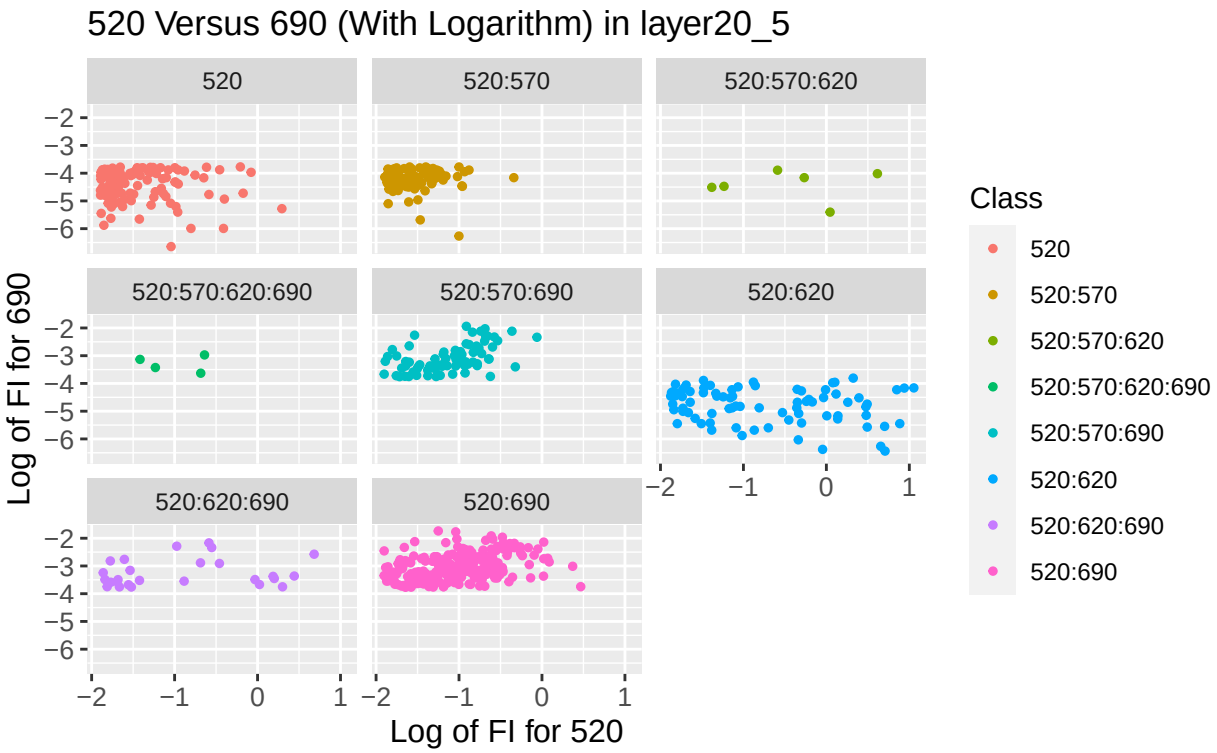
In summarizing the insights from these three figures, a notable observation is that Opal_690 appears more frequently than Opal_570 and Opal_620 across layer20_5, layer12_4, and layer12_5. Both layer20_5 and layer12_4 exhibit 520:690 with the highest frequency in their respective layers. However, in layer12_5, while 520:690 boasts a substantial frequency, 520:570:690 holds the highest frequency value. Interestingly, Opal_520 and Opal_690 frequently co-occur. Moving forward, our analysis aims to delve deeper into understanding the relationship between Opal_690 and Opal_520.

Correlation

We employ a scatterplot to visually analyze the distribution of Opal_520 and Opal_690. In this plot, the x-axis represents the logarithmic values of the fluorescent intensity for Opal_520, while the y-axis depicts the logarithmic values of the fluorescent intensity for Opal_690. This visualization allows us to explore the relationship and potential patterns between the two variables in a quantitative manner.


520 Versus 690 (With Logarithm) in layer12_4
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Log of FI for 690
Log of FI for 520
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690

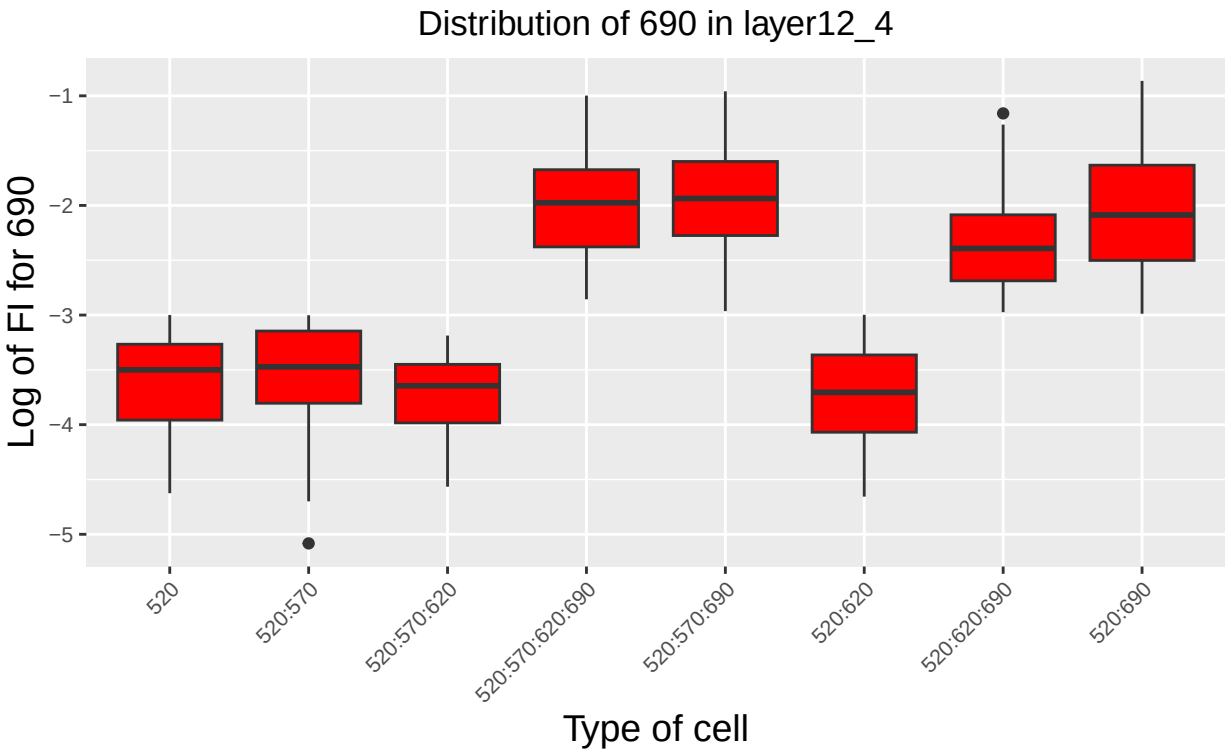

520 Versus 690 (With Logarithm) in layer12_5
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Log of FI for 690
Log of FI for 520
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690

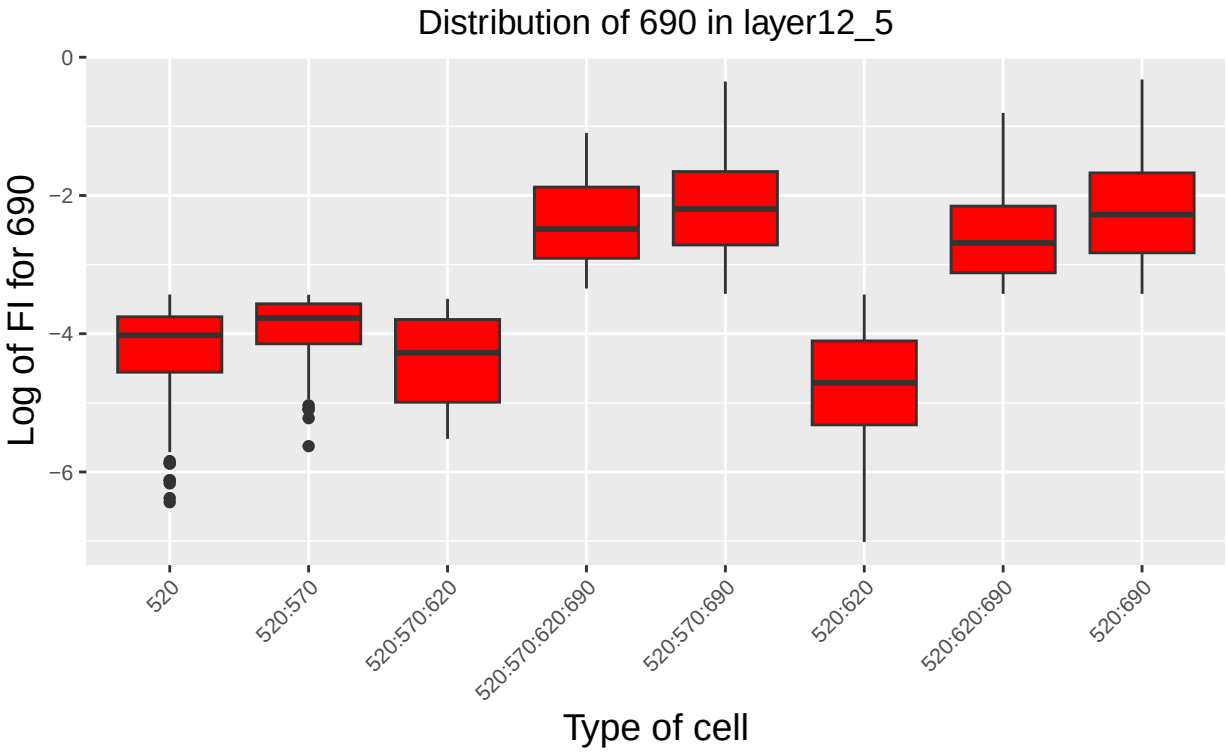


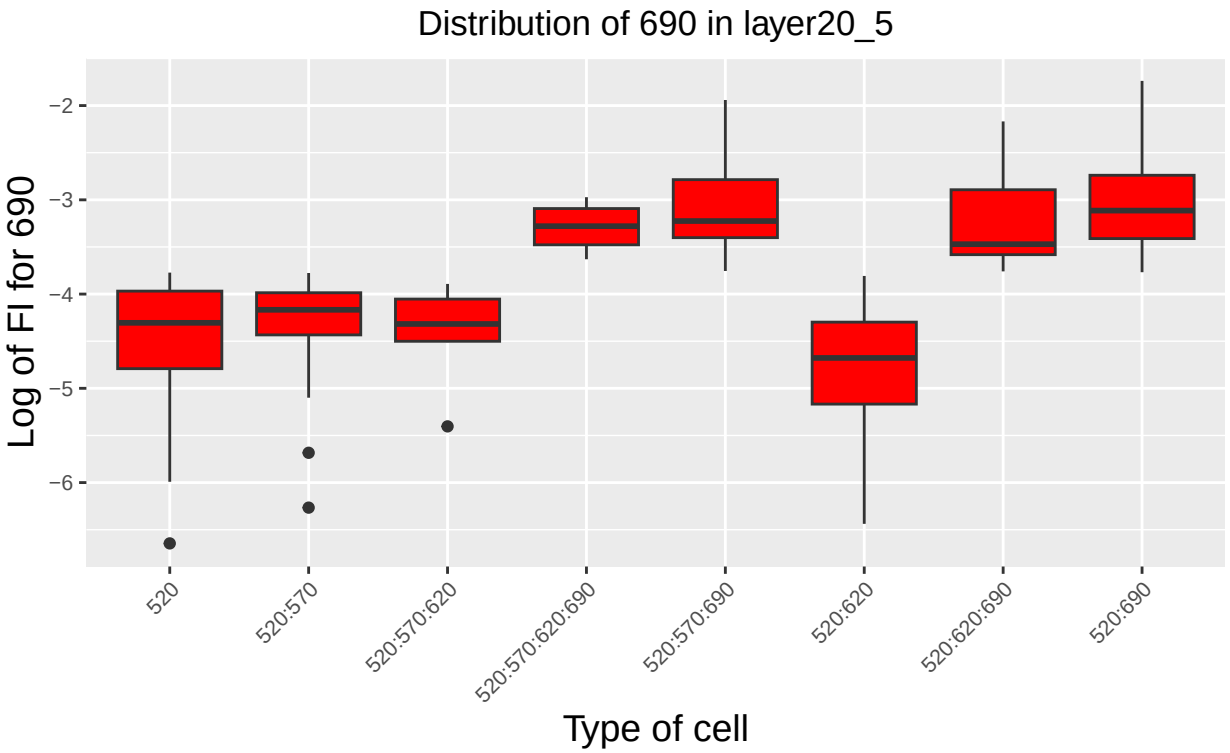
Based on the figures above, it is evident that cells containing Opal_690 tend to exhibit higher values for Opal_690, which aligns with expectations. In both layer12_4 and layer12_5, the cells with combinations 520:570:690 and 520:690 demonstrate a positive correlation between Opal_520 and Opal_690. As the value of Opal_520 increases, there is a corresponding increase in the value of Opal_690. However, in layer20_5, a notable correlation between Opal_520 and Opal_690 is not readily apparent. The presence of a subtle positive correlation in layer20_5 raises the question of whether this correlation is a common feature across all layers or if it is specific to layer12_4 and layer12_5. Further investigation is required to ascertain whether the observed correlation is consistent across all layers or if it exhibits variation, with layer12_4 and layer12_5 demonstrating a more pronounced association compared to layer20_5.

Boxplot

The utilization of boxplots offers a comprehensive and insightful analysis of the distribution of Opal_690 across eight distinct cell types. Boxplots provide a visual summary of the central tendency, spread, and potential outliers within each cell type, allowing for the efficient comparison of distributional characteristics. By employing boxplots, we can discern not only the variations in Opal_690 expression across different cell types but also gain a nuanced understanding of the overall distributional patterns and potential heterogeneity within each cell type.


Distribution of 690 in layer12_4
Log of FI for 690
−1
−2
−3
−4
−5
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Type of cell

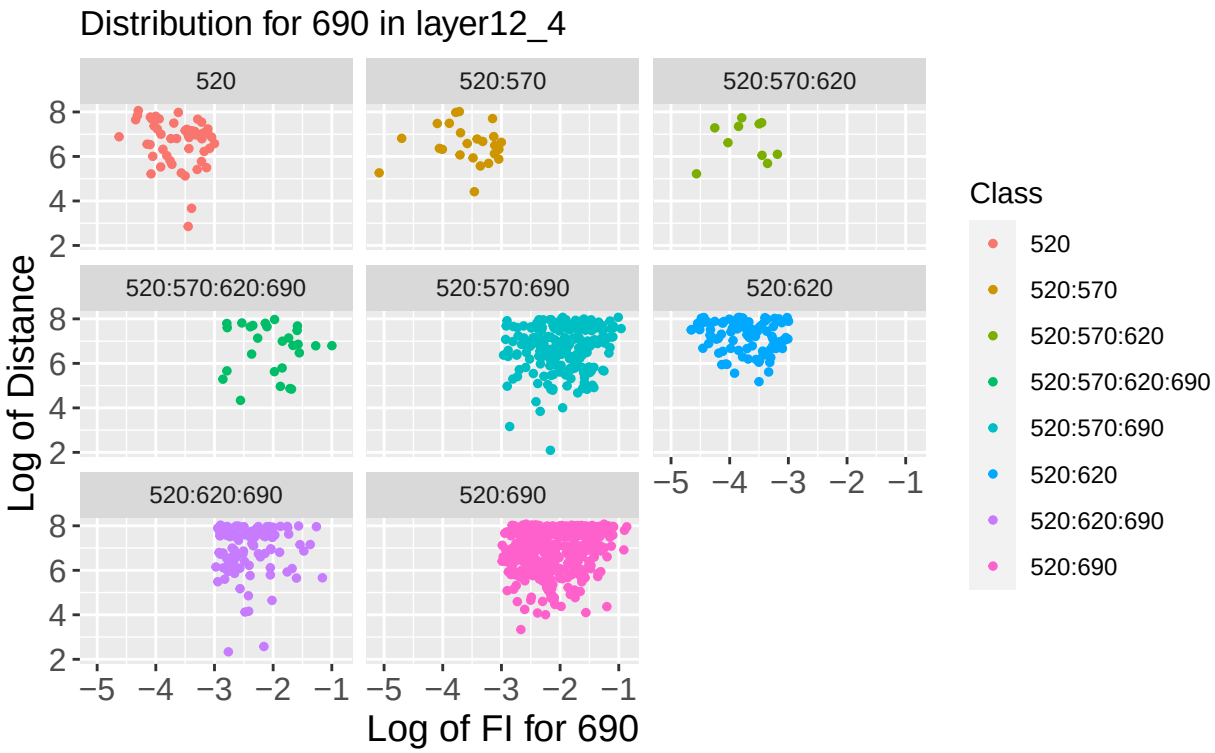

Distribution of 690 in layer12_5
Log of FI for 690
0
−2
−4
−6
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Type of cell

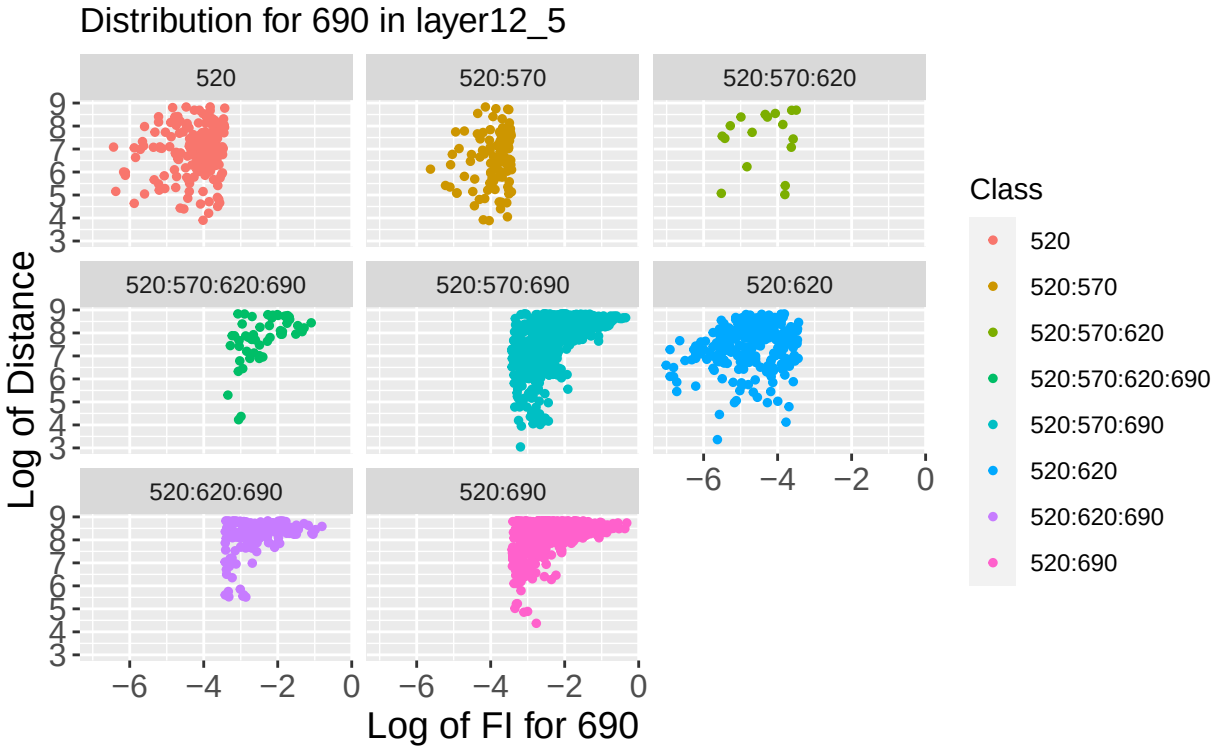


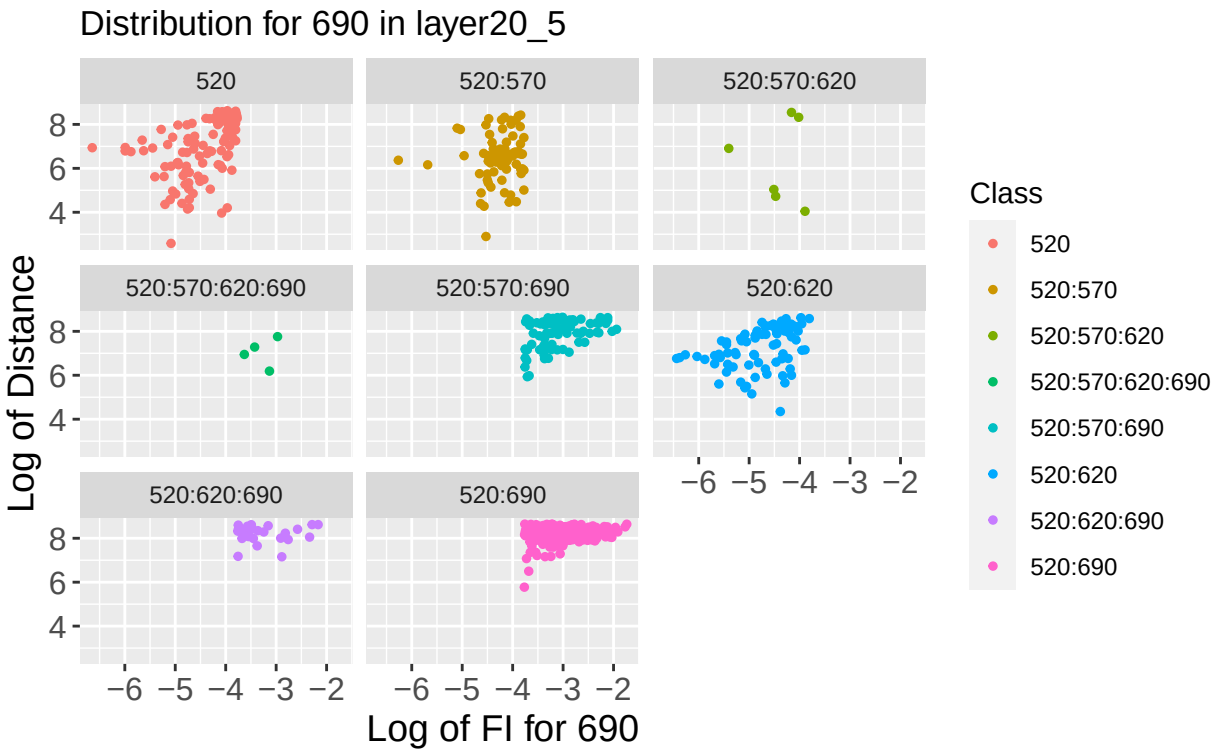
The observed figures clearly indicate a substantial distinction between cells containing Opal_690 and those without it. Specifically, cells that include Opal_690 exhibit significantly higher values compared to their counterparts lacking Opal_690. This stark contrast underscores the impact of Opal_690 on the measured values, suggesting a discernible influence of Opal_690 presence on the observed fluorescent intensities. The visual representation of this difference in values between the two groups provides a clear and immediate understanding of the impact of Opal_690 on the distribution within the dataset.

Distribution for Distance and Opal_690

Given that Opal_690 exhibits some correlation with Opal_520, we aim to explore any specific correlation between Opal_690 and Distance. In the plot, the x-axis represents the logarithmic values of the fluorescent intensity for Opal_690, while the y-axis illustrates the logarithmic values of the Distance.


Distribution for 690 in layer12_4
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Log of Distance
Log of FI for 690
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690

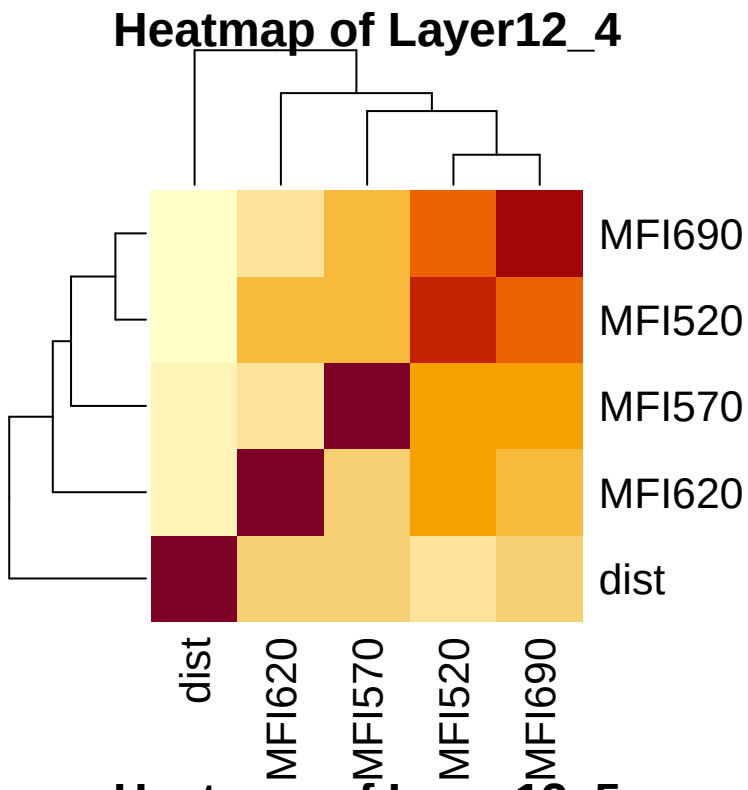

Distribution for 690 in layer12_5
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Log of Distance
Log of FI for 690
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690

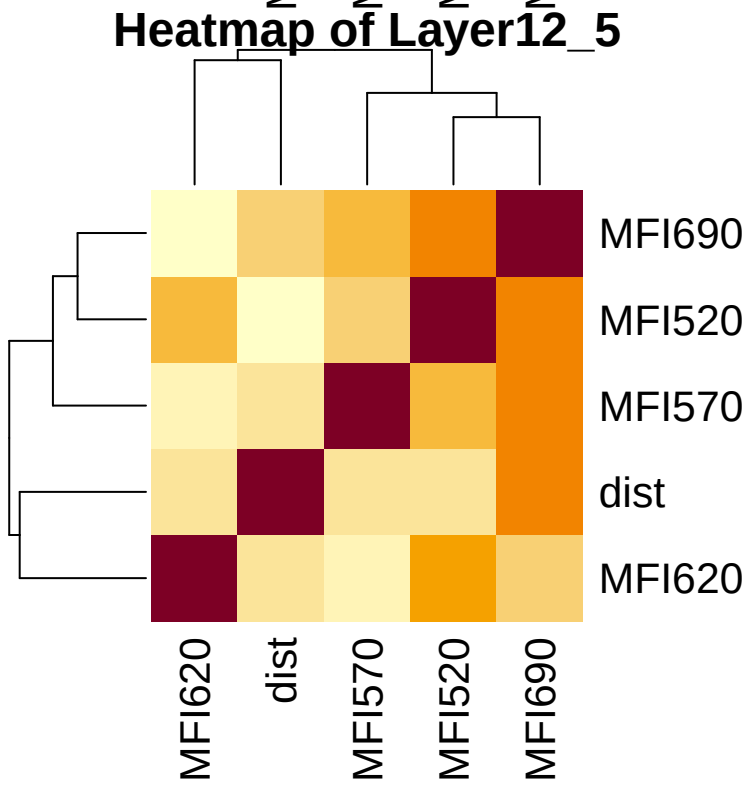


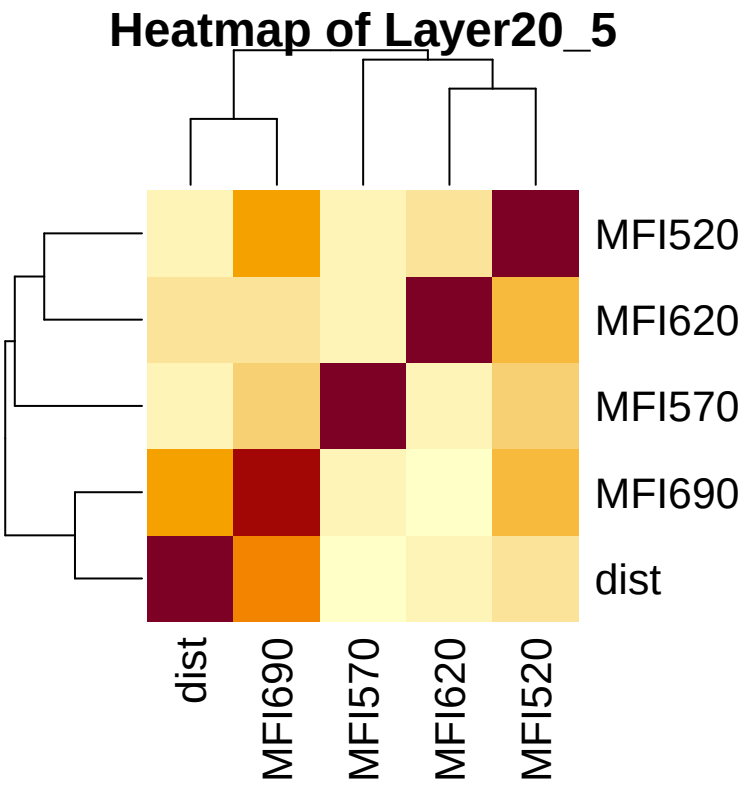
In the above figures, our focus centers on cells with combinations 520:690, 520:570:690, 520:620:690, and 520:57:620:690. Notably, in layer12_5, a noteworthy pattern emerges, indicating that when the value of Opal_690 is small, the distance exhibits a considerable range from low to high. However, as the value of Opal_690 increases, the distance tends to contract into a narrower range characterized by larger values. This trend suggests a distinctive relationship between Opal_690 and distance in layer12_5. On the other hand, for layer20_5, a distinct pattern is observed, wherein a small range with substantial distance values prevails regardless of whether the value of Opal_690 is high or low. This disparity in patterns underscores the importance of considering layer-specific nuances in interpreting the relationship between Opal_690 and distance.

Heatmap

We also came up with a heat map to indicate the relationship of different genes. And we can see that there is some difference between the correlations among 3 layers.


Heatmap of Layer12_4
MFI690
MFI520
MFI570
MFI620
dist
dist
MFI620
MFI570
MFI520
MFI690

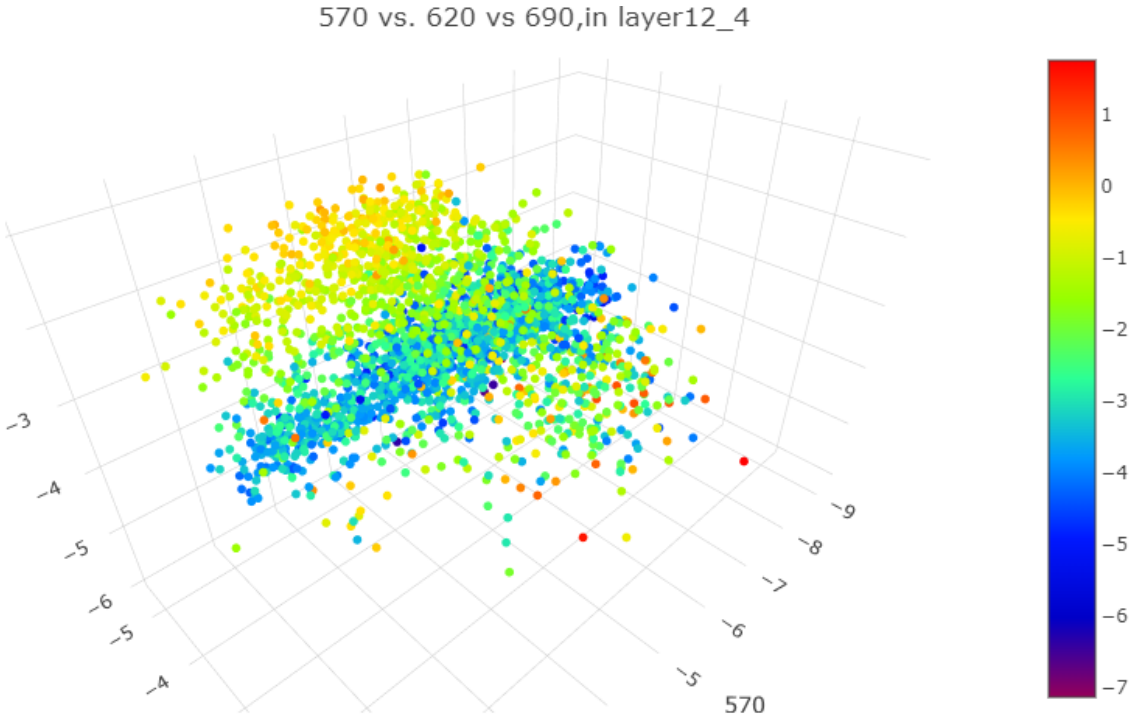

Heatmap of Layer12_5
MFI690
MFI520
MFI570
dist
MFI620
MFI620
dist
MFI570
MFI520
MFI690

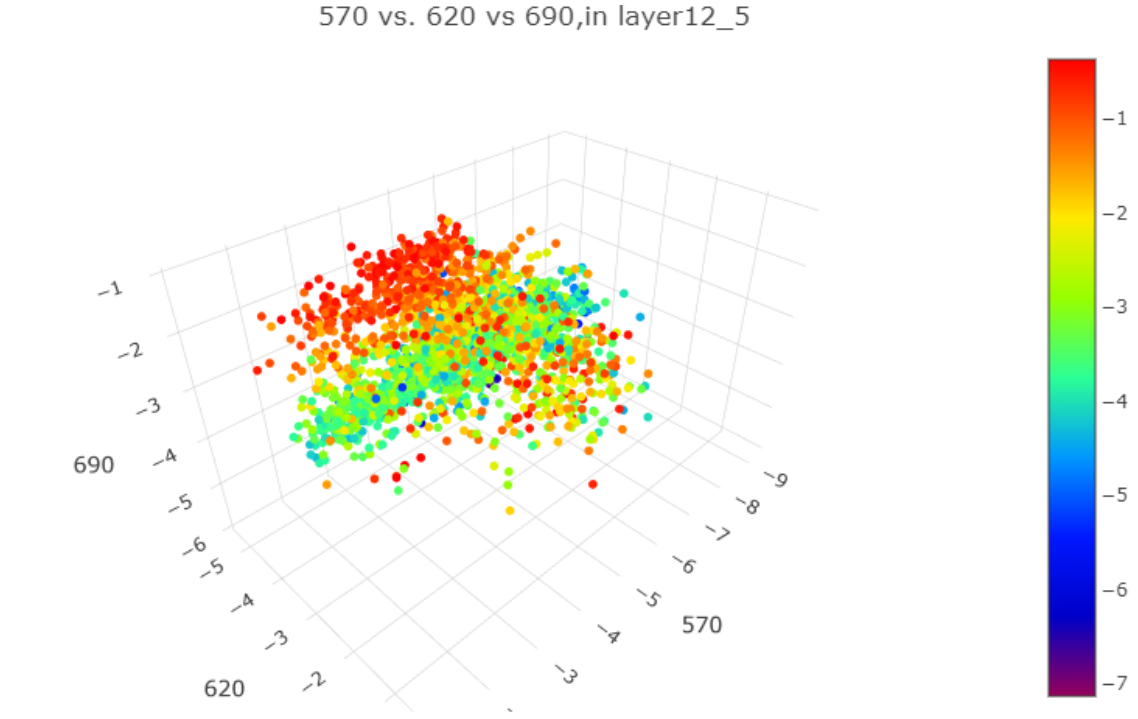


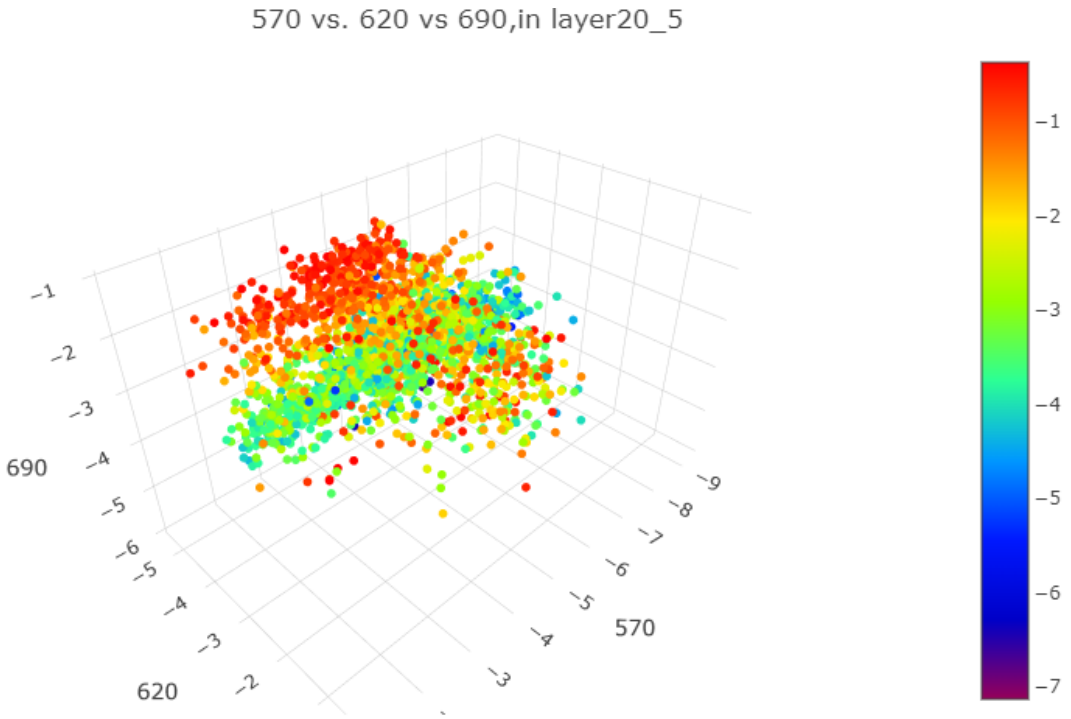
For the figures above, the intensity of color serves as an indicator of the strength of correlation, with darker shades representing greater correlation. Notably, if a line connects two variables, it signifies that these two variables exhibit the highest correlation compared to any others. Examining both layer12_4 and layer12_5, a notable line connects Opal_690 and Opal_520, providing empirical support for our hypothesis regarding their high correlation. Contrary to this, as previously mentioned, in layer20_5, the correlation between Opal_690 and Opal_520 is not particularly significant; however, it is intriguing to observe a connecting line between Opal_690 and Distance, indicating a noteworthy correlation between these variables. This unexpected correlation in layer20_5 adds a layer of complexity to our understanding, highlighting the importance of considering layer-specific dynamics in interpreting correlation patterns.

3D Plot

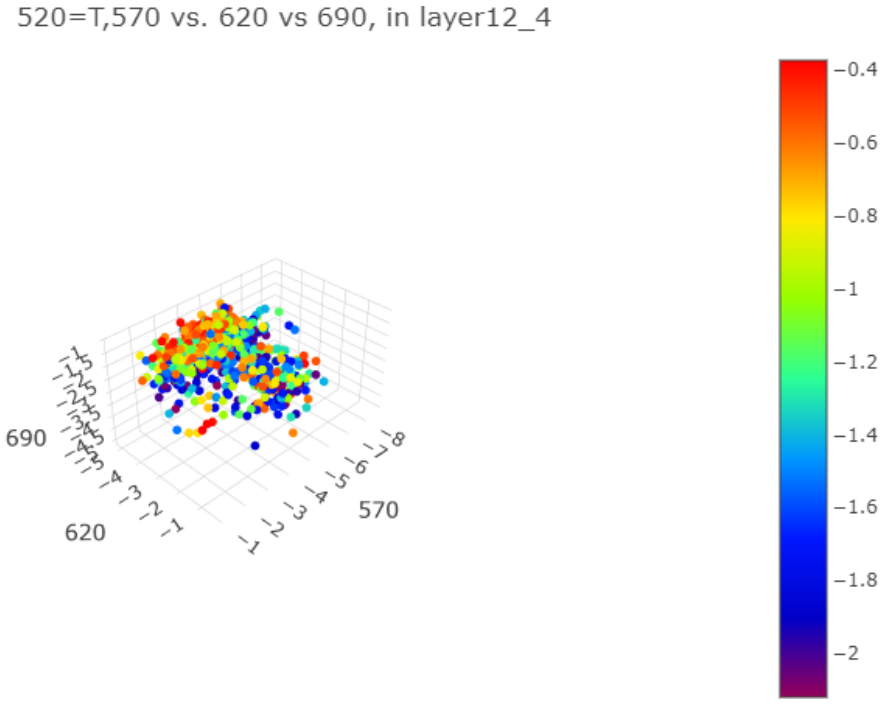
Since there is more than two variables we're dealing with, an idea of visualization in 3-D version is inspired by our advisor. But knitting the moving pictures of 3D plots is not available, besides the screenshots provided with only one or two perspectives, we will give you our original codes so that you can reproduce what we got so far. And please be aware that the scales labeled below are all after log-transformation aimed for a more clear view.

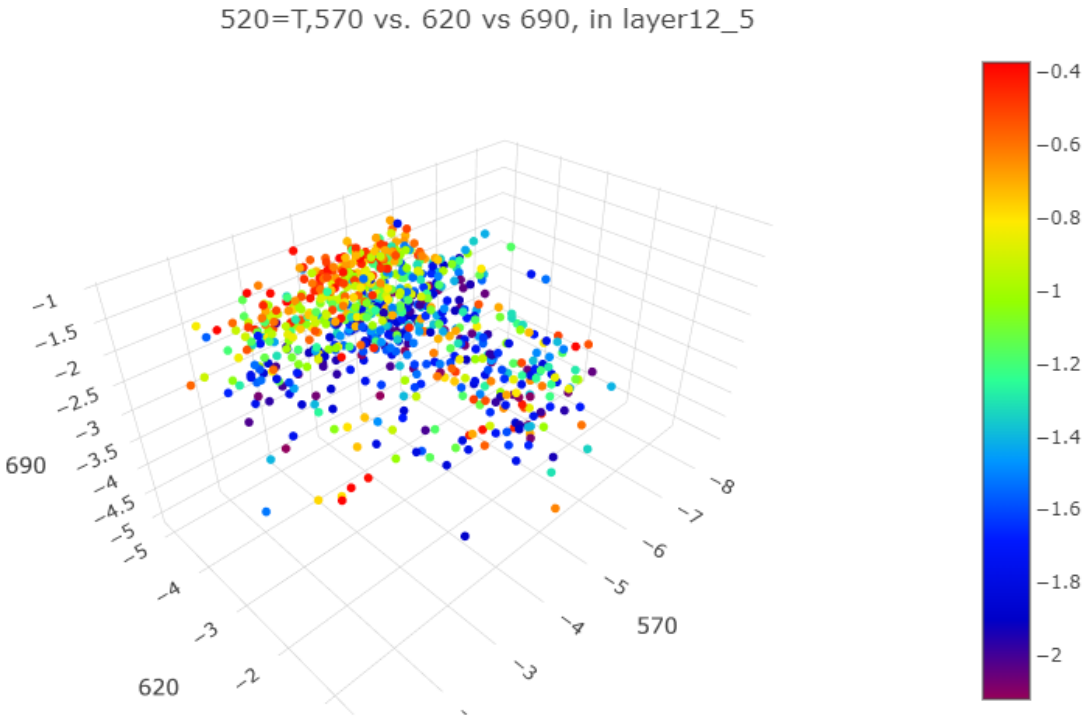

570 vs. 620 vs 690,in layer12_4
−3
−4
−5
−6
−5
−4
570
−5
−6
−7
−8
−9
1
0
−1
−2
−3
−4
−5
−6
−7

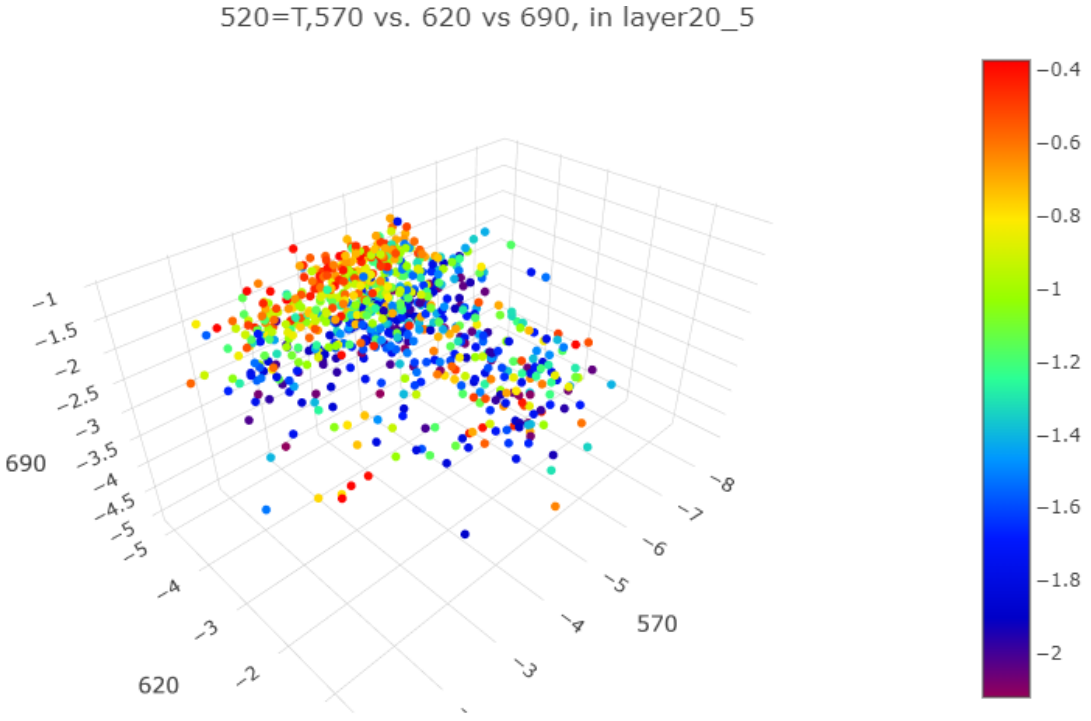

570 vs. 620 vs 690,in layer12_5
−1
−2
−3
690
−4
−5
−6
−5
−4
−3
620
−2
−3
−4
−5
−6
−7
−8
−9
570
−1
−2
−3
−4
−5
−6
−7



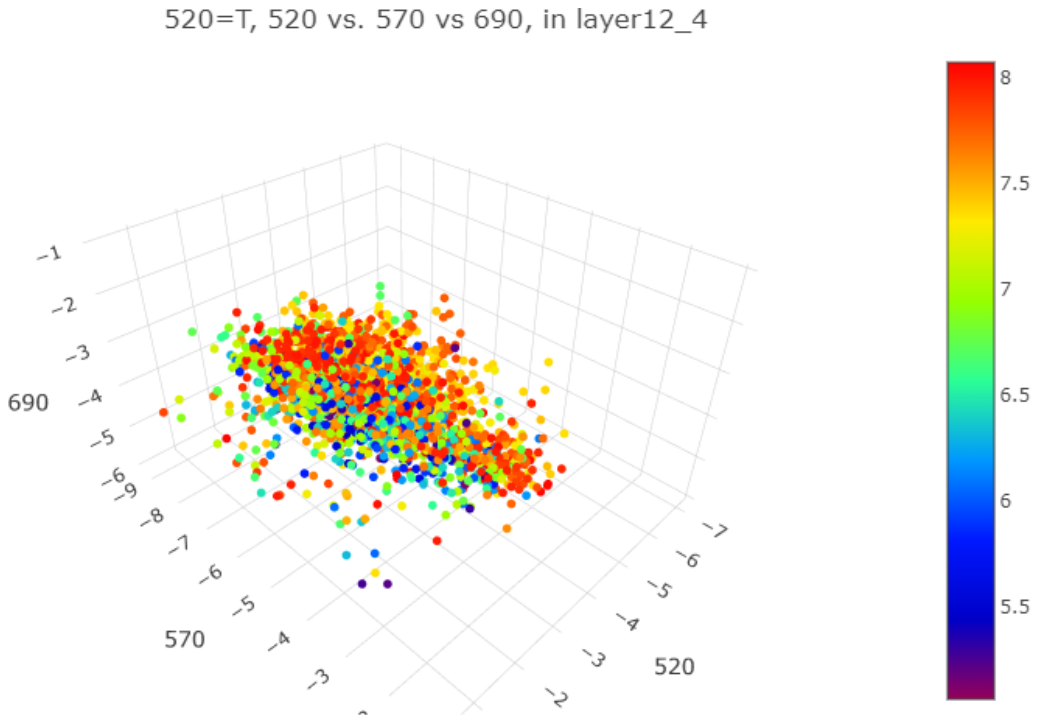
The plots above show the relationship among each gene in terms of intensity level of 520. But these plots accounts all the cells in the sample no matter there is 520 included or not. So it's hard to observe significant relationship among different genes. Next we decided to only include the cells with 520 detected.

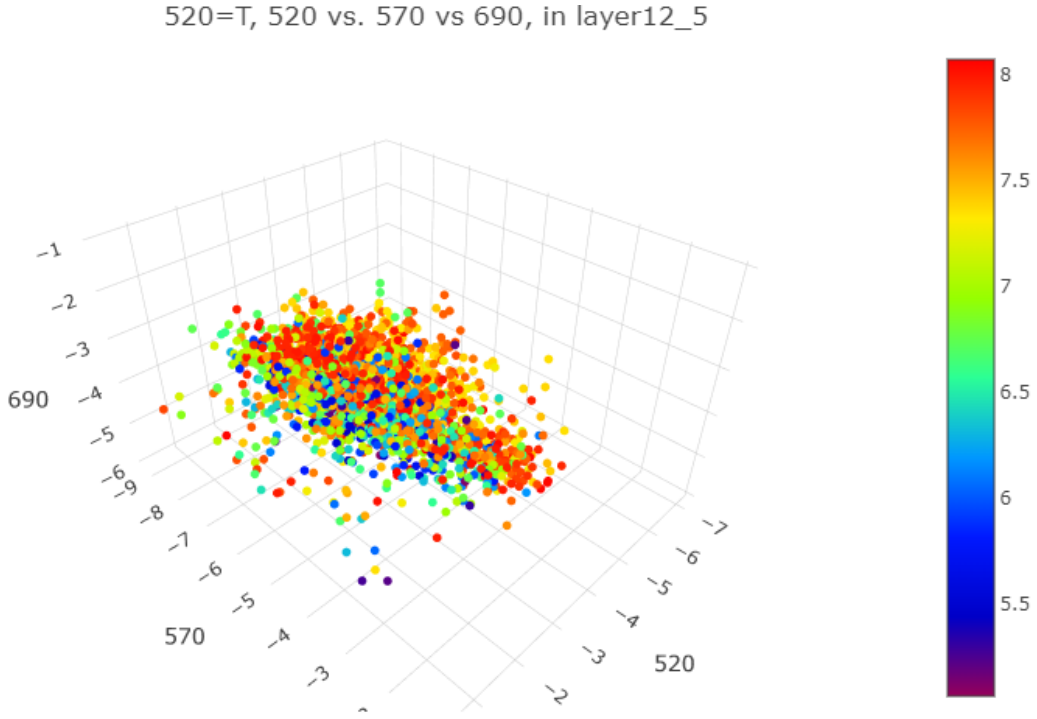


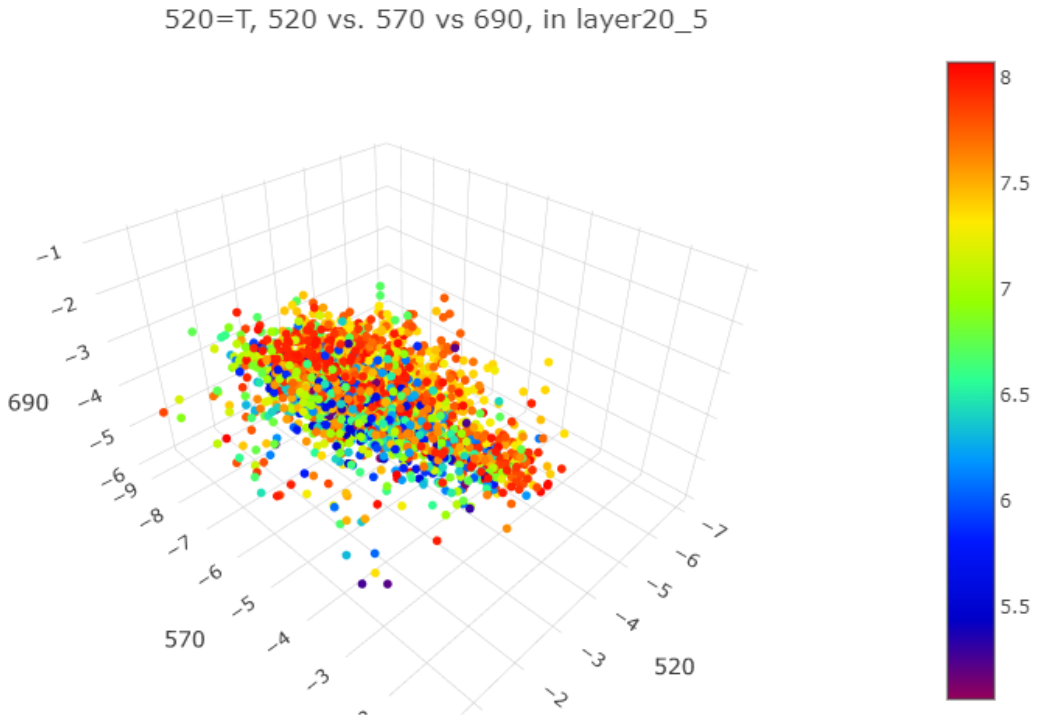




As removing the cells without 520, some relationships could be observed based on the 3D plots. For example, in Layer 2, As the intensity of 520 raise, the intensity of 690 increases while 570 seems to show a reverse relationship with 520 compared to 690. Also, 620 doesn't have obvious relationship with other genes. So for further investigation, we decided to remove 620 which is consider meaningless for showing no any trends with the others in these three plots. Instead 520 is included below.


520=T, 520 vs. 570 vs 690, in layer12_4
690
570
520
8
7.5
7
6.5
6
5.5


520=T, 520 vs. 570 vs 690, in layer12_5
690
570
520
8
7.5
7
6.5
6
5.5



To get 520 involved, the intensity level as colored option is no longer appropriate. Besides intensity level, we thought there might also be some contribution made by distance to the distribution of cells in EC region. So instead of intensity level, we redo the previous process but colored in terms of distance. And we found that In layer 2 as the intensity of 520 decreases, the intensity of 690 also decreases. And as intensity of 520 held constant, the higher the intensity of 690 is the more distant from the edge of EC.

Here are only couple examples from each section of plots. We hope these plots would help our client to seek more essential information for their study.

Conclusion

The project's application of robust statistical methods revealed intricate patterns in gene expression. Notably, a significant correlation between Opal_690 and Opal_520 was found, especially in layers 1 and 2 of the EC. These results highlight the complex interplay of gene expression, which varies distinctly across different layers. Furthermore, the spatial analysis of gene distribution underscored how gene expression intensity is related to proximity to the EC edge, adding a spatial dimension to our understanding of gene distribution.

Throughout the project, we emphasized the clear communication of complex statistical findings to a diverse audience. This involved transforming technical data into comprehensible insights for both scientific and non-expert stakeholders. The visual aids, crafted by our team members, played a crucial role in making intricate patterns more accessible and understandable, enhancing the impact of our research findings.

The integration of statistical rigor with effective communication strategies was key to the project's success. Our approach not only deepened the understanding of specific genes in the EC but also exemplified the importance of multidisciplinary collaboration in research. The team's combined expertise in analysis, visualization, and communication was crucial in delivering a comprehensive and impactful outcome.

In summary, this project exemplifies the power of collaborative research, combining detailed statistical analysis with clear, effective communication. We have not only advanced our understanding of gene distribution in the Entorhinal Cortex but also set a new standard for future interdisciplinary research. Our findings lay

the groundwork for further investigations into the roles of these genes in cognitive processes and, potentially, in the pathology of neurological disorders.