By: **Nadeen Ali** - 22110361

**Multi-Omics Data Integration for Downstream Analysis & Biomarker Discovery**

Technical Report

Course: Bioinformatics – Spring 2024/2025

Course Instructor: Dr. Rami Al-Ouran

Table of Contents

[**1 – Introduction & Overview** 4](#_Toc200229023)

[**2 – Dataset Description** 4](#_Toc200229024)

[**3 – Preprocessing (First Notebook)** 5](#_Toc200229025)

[Saving Files 5](#_Toc200229026)

[Reading Files 5](#_Toc200229027)

[Removing Missing Values 6](#_Toc200229028)

[Filtering Low Quality Data 6](#_Toc200229029)

[Finding Common Samples (Patients) 7](#_Toc200229030)

[Outlier Detection 8](#_Toc200229031)

[**4 – MOFA Object Creation** 10](#_Toc200229032)

[Entry Point 10](#_Toc200229033)

[Reading Data 11](#_Toc200229034)

[Disabling Scaling Views 11](#_Toc200229035)

[Passing Data 11](#_Toc200229036)

[Model Configuration 12](#_Toc200229037)

[Training Options 12](#_Toc200229038)

[Running and Saving Model 12](#_Toc200229039)

[**5 – Downstream Analysis** 12](#_Toc200229040)

[Mapping Features to Genes 12](#_Toc200229041)

[Plotting Weights, Both Views (Rna and Mutation) 13](#_Toc200229042)

[Weights Ranked 14](#_Toc200229043)

[Variance Explained 15](#_Toc200229044)

[Factors’ Correlation 16](#_Toc200229045)

[Factors and Features Heatmap 17](#_Toc200229046)

[Weights Dot Plot (RNA) 18](#_Toc200229047)

[Weight Dot Plot (Mutation) 19](#_Toc200229048)

[Scatterplot, Scaled Weights 19](#_Toc200229049)

[Revising Heatmap (10 Features) 21](#_Toc200229050)

[Highest Genes in Each Factor 22](#_Toc200229051)

[**6 – Pathways and Ontologies** 22](#_Toc200229052)

[**References** 23](#_Toc200229053)

# **1 – Introduction & Overview**

This project aims to collect and study data regarding the TCGA-STAD project, focusing on **stomach adenocarcinoma**, the *most* common type of stomach cancer causing digestive issues and abdominal pain [1][2], to discover potential biomarkers which are *clear* indicators of the cancer’s presence.

To do this, two types of omics data were used from the *same* patients:

1. **RNA-seq:** Measures gene expressions
2. **Mutation data:** binary values indicating gene mutation status (0 = no mutation exists, 1 = mutation does exist)

Each dataset was cleaned, preprocessed, and visualized *independently* before applying MOFA+, an AI based python library for multi-omics data integration, to combine the views (datasets) for further and deeper biological analysis to identify key pathways and ontologies that may be involved in stomach cancer.

# **2 – Dataset Description**

As mentioned above in the Introduction & Overview section, this data is collected focusing on the *most* common stomach cancer - stomach adenocarcinoma found from the website **LinkedOmics**. During the initial analysis, 3 datasets were used. Here are some of their details before diving into section 3 detailing all the steps:

1. **Rna-seq**

* Shape *before* preprocessing: 20,225 rows x 415 columns
* Shape *after* preprocessing: 205 rows (patients) x 19112 columns (genes)
* After MOFA+ integration: used as **view0**

1. **Mutation data**

* Shape *before* preprocessing: 13724 rows x 393 columns
* Shape *after* preprocessing: 205 rows (patients) x 1068 columns (genes)
* After MOFA+ integration: used as **view1**

1. **Clinical data**

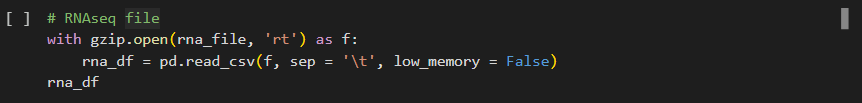
* Shape *before* preprocessing: 15 rows x 443 columns
* Shape *after* preprocessing: 205 rows (patients) x 14 columns (clinical features)
* Only used for high level analysis and *not* integrated into MOFA+

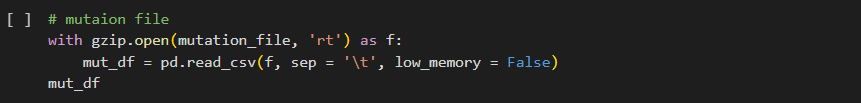
*Note* that all datasets **have the same number of samples (patients)** after preprocessing. This is to ensure that we are comparing all data across the *same* patients for consistent, accurate, and interpretable results.

# **3 – Preprocessing (First Notebook)**

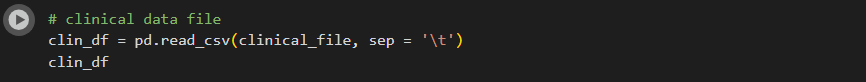
### Saving Files

* Used gzip to open rna-seq and mutation data





* While clinical data did not need any extra functions to open



### Reading Files

* Problem 1: First column contained gene and clinical data was named ‘attrib\_name’
* Solution 1: Renamed column to ‘gene’ and ‘clinical\_data’ as fit based on df
* Problem 2: Index was not set correctly
* Solution 2: Set ‘gene’ and ‘clinical data’ as indices
* Problem 3: Samples were columns and features were rows
* Solution 3: Transposed so samples became rows and features were columns
* Optional: Set new index name (rows, patients) to ‘sample\_id’
* Printed head to make sure all changes are as needed

A computer screen shot of text

AI-generated content may be incorrect.

A black screen with colorful text

AI-generated content may be incorrect.

A computer screen with text

AI-generated content may be incorrect.

### Removing Missing Values

* Checking no null values in datasets
* Problem: Founds 505 null values in clinical data
* Solution: Dropped them
* Check none left

A black and grey striped background

AI-generated content may be incorrect.

### Filtering Low Quality Data

* Problem: Not *all* data in the datasets are crucial: some genes for examples are rarely expressed or rarely mutated
* Solution (rna-seq): Genes *must* be expressed in at least 20% of samples
* Result: Small change, but regardless is a change

A screenshot of a computer program

AI-generated content may be incorrect.

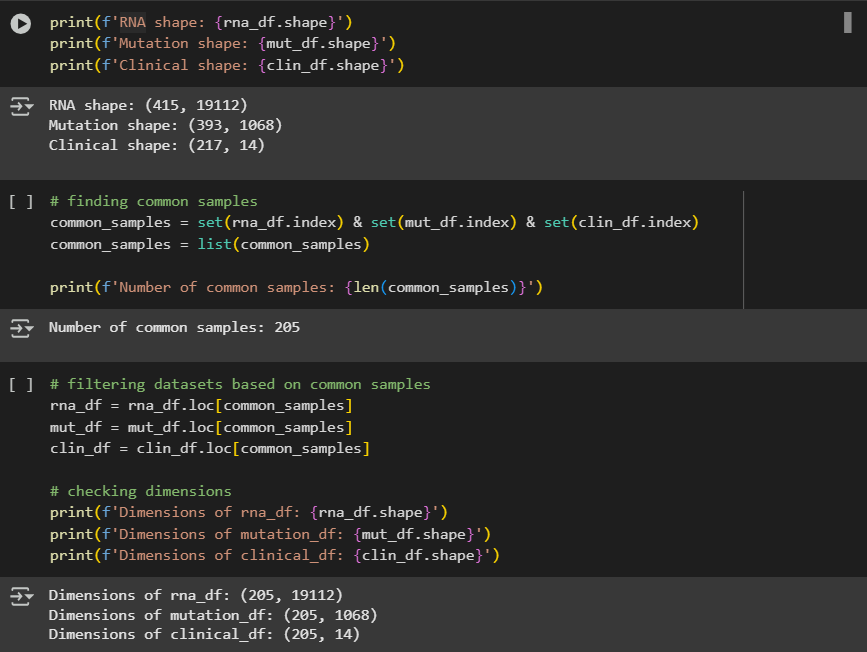
* Solution (mutation data): Genes *must* be mutated in a at least 5% of samples
* Result: Huge change noticed (thus why only 5% taken and not more)

A computer screen with text

AI-generated content may be incorrect.

### Finding Common Samples (Patients)

* Problem: Data has been cleaned, but the rows (samples or patients) are *still* not in common across all 3 datasets
* Solution: Must filter all 3 datasets based on found common samples by using & in a set (to avoid duplicated sample ids)
* Results: Samples *shrunk* to 205 across all 3 datasets



### Outlier Detection

* Problem: Outliers are extreme values that deviate from the rest of the data and can throw off analysis and results especially when integrating datasets.
* Solution: Applying dimensionality reduction and outlier detection using PCA, TSNE (with PCA), *and* UMAP (also with PCA)
* **PCA** applied first to reduce noise and capture *most* important variation in rna-seq
* **TSNE** applied to preserve *local* structure
* **UMAP** applied to preserve a balance between *both* local and global structures
* Result: Not much variance found, with highest (PC1) at below 14% **only**.

A graph of a graph with blue bars

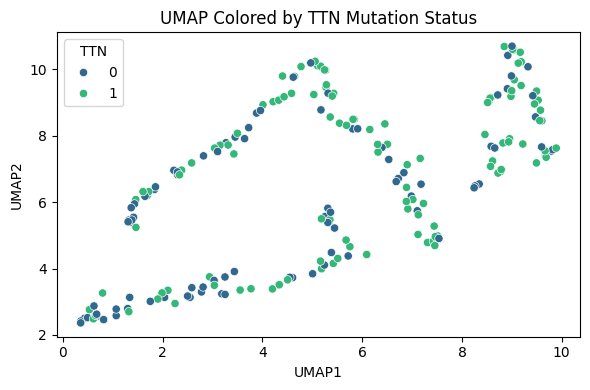
AI-generated content may be incorrect.

* As expected due to low explaining variance, no clear outliers were shown across all 3 dimensionality reduction methods.

A chart of blue and green dots

AI-generated content may be incorrect.A diagram of a number of dots

AI-generated content may be incorrect.

 A diagram of a diagram

AI-generated content may be incorrect.

A diagram of a number of blue dots

AI-generated content may be incorrect. A map of the united states

AI-generated content may be incorrect.

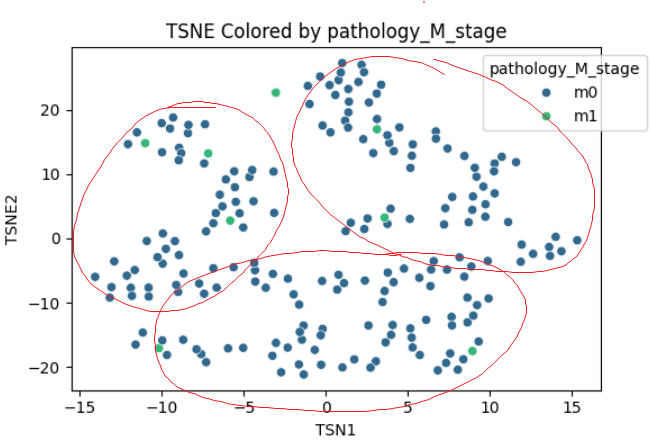
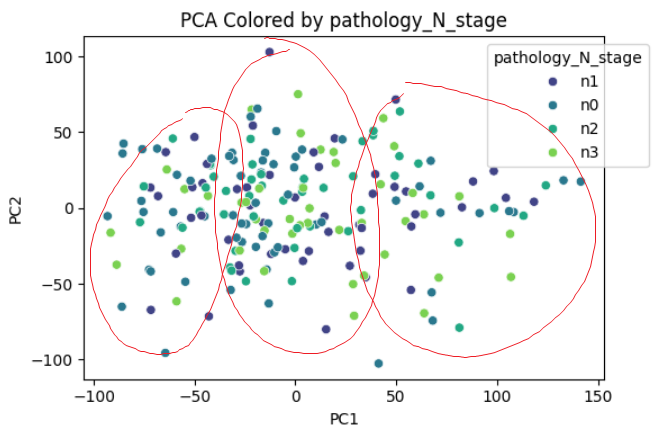
* Double Checking: Using IsolationForest to fit on PCA to ensure *no* aggressive outliers exist indeed.
* Result: 5% show not many outliers, 10% begins to show seemingly normal points as *outliers*, thus being too aggressive.
* **Decision: No outliers, thus no removal of any data**

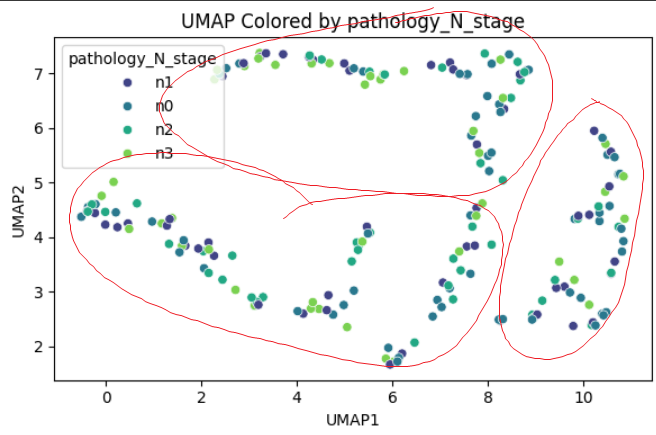
A diagram of a number of dots

AI-generated content may be incorrect.A diagram of red and blue dots

AI-generated content may be incorrect.

* Observation: When colouring PCA, TSNE, *and* UMAP plots by different variables (mutations, clinical data), the structure and pattern of plot remains the same
* **Is it a problem? No.** Could mean mutation and clinical data are *not* primary drivers of expression data (later proven in MOFA integration)
* Suggests that latent factors drive data.
* Any groups or clusters noticed? Yes, ***slightly***. Very similar across all plots
* Very rough clusters, but still existing.
* Meaning: Samples (patients) that are like each other, showing similar behaviour
* How: Similar gene expressions, mutations, biological responses, etc.
* What Next: Using MOFA data integration to uncover latent variables explaining separation between groups.





## **4 – MOFA Object Creation**

* Purpose: Building the MOFA+ models that’ll learn latent, shared patterns across multi-omics data (rna-seq *and* mutation data)

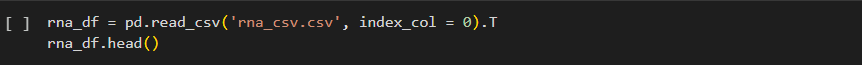
### Entry Point

* Creation of the object to set up MOFA+



### Reading Data

* Loading data to pass onto MOFA object later
* Transposing so rows = patients and columns = genes





### Disabling Scaling Views

* Views (rna-seq and mutation data) are *not* scaled
* Why? Rna-seq is *already* normalized, and mutation data is binary



### Passing Data

* Converting pandas df to matrices for MOFA compatibility
* Transposing *inside* data setup: columns *must* be equal
* Rna-seq data assigned ‘gaussian’ likelihood due to its continuous data
* Mutation data assigned ‘bernoulli’ likelihood due to its binary values (0s and 1s)
* View Names: view0 = rna data, view1 = mutation data

A screenshot of a computer program

AI-generated content may be incorrect.

### Model Configuration

* factors = 10: number of latent (hidden) factors model learns to explain variation across omics data (rna and mutation)

A black rectangle with white text

AI-generated content may be incorrect.

### Training Options

A black rectangle with white text

AI-generated content may be incorrect.

### Running and Saving Model

* .hdf5 file format for downstream analysis (covered in next section)

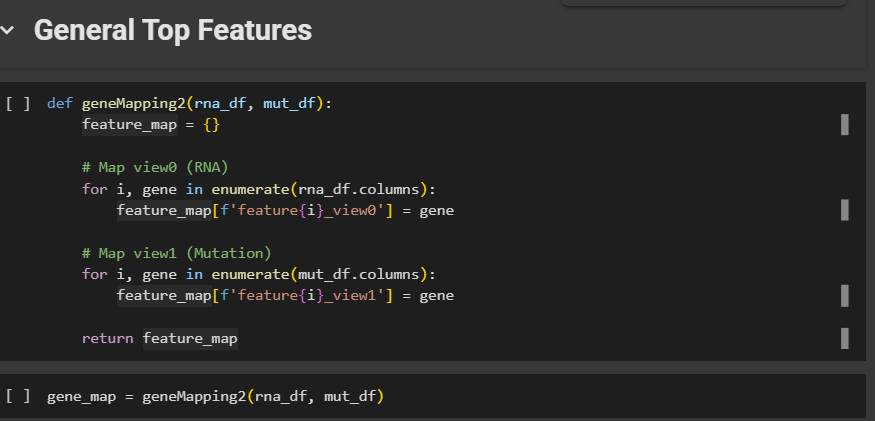
A black rectangle with white text

AI-generated content may be incorrect.

## **5 – Downstream Analysis**

### Mapping Features to Genes

* Problem: MOFA did not pick up gene names during training
* Solution: Created a function that maps feature names to gene names from original dfs



### Plotting Weights, Both Views (Rna and Mutation)

A screenshot of a graph

AI-generated content may be incorrect.

* Rna view (view0) sees strong features from factors 1 – 4 *mainly* (our focus it seems)
* Mutation view (view1) is quite weak without strong features on any factors

### Weights Ranked

A graph with a line

AI-generated content may be incorrect.

* This plot shows a factors’ strongest features
* X axis = feature rank
* Y axis = feature’s weight
* For example: Factor 1 shows *very strong* features. So do factors 2, 3, *and* 4 (again, our focus lining up as expected)

### Variance Explained

A group chart with text

AI-generated content may be incorrect.

* As mentioned, view1 (mutation) is not indicative of any driving factors.
* *However*, factor 1 is driving view0 (rna) *incredibly strong*.
* Factors 2 and 3 very *lightly* follow on driving the rna view.

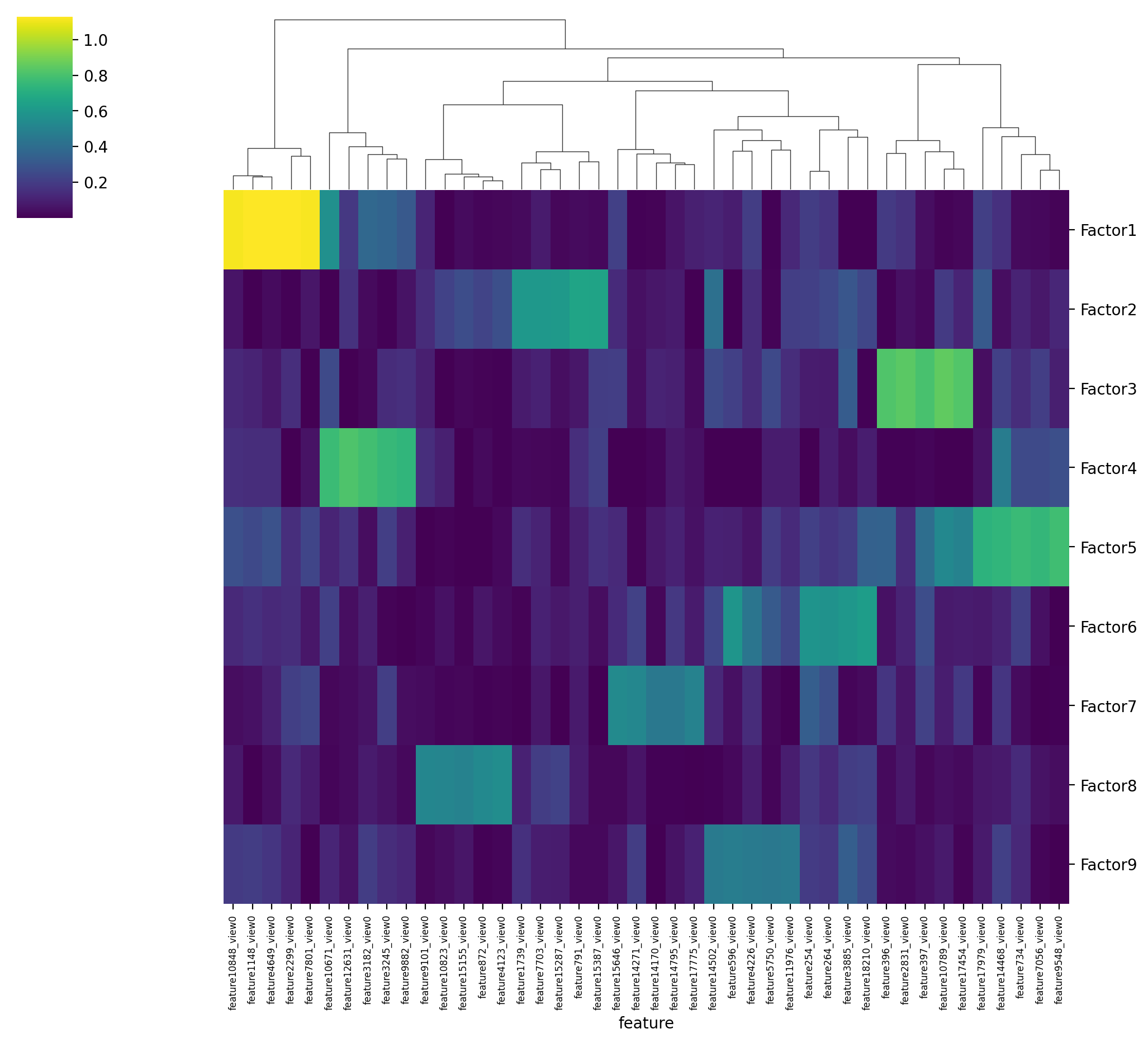
### Factors’ Correlation

A screenshot of a graph

AI-generated content may be incorrect.

* This plot shows the relationship between factors with one another.
* Shows *weak* correlation, thus meaning no overlap between factors.
* Can be a good thing as it means factors show distinct biological information (no redundancy in data or genes)

### Factors and Features Heatmap



* Once again, another heatmap but this time showing the relationship between factors and features.
* Order of seemingly strongest relationships based on brightness:
  + Factor 1 🡪 Factor 3 🡪 Factor 4 🡪 Factor 5 🡪 Factor 2 🡪 remaining are weak.
  + Factor 5 a new focus added? Potentially

### Weights Dot Plot (RNA)

A graph of different colored dots

AI-generated content may be incorrect.

* Plots above further proves that factor 5 *may indeed* be a new added focus

### Weight Dot Plot (Mutation)

A graph of a number of dots

AI-generated content may be incorrect.

* *Only* factor 1 has strong coloured dots: but *still*, values are insignificant with highest being slightly above 0.08 so we ignore it.

### Scatterplot, Scaled Weights

* This compares the density and direction of 2 factors’ features
* We have *many* factors: which do we compare? **Based on variance explained!**
* Ranking variance -
  + # 1. factor 1
  + # 2. factor 2
  + # 3. factor 3
  + # 4. factor 4
  + # 5. factor 5
  + # 6. factor 6
  + # 7. factor 7
  + # 8. factor 8
  + ignoring factor 9 because it is below threshold of > 2% (this threshold I came up with for balanced inclusion while also ignoring low values)
* How will we go about the plot?
  + we will compare factor 1 with ALL others bc of its strong R2
  + if good results, same will be strat applied with factors 2 and 3
  + avoiding comparing mid factors with other mid factors
  + including comparisons of 2, 3, and 4 together

A group of graphs on a white background

AI-generated content may be incorrect.

* Results:
  + Quite strong clusters in factors 2 – 5 vs. factor 1
  + No clusters, greater sparsity in factors 6 - 9 (expected)
  + So, we focus on factors 1 - 5 (as previously thought)
  + Factor 1 vs. 2 - 5 very orthogonal (unique bio info, we will see later)
* Now we compare F2 with F3 – 5, F3 with F4 – 5, and F4 with F5

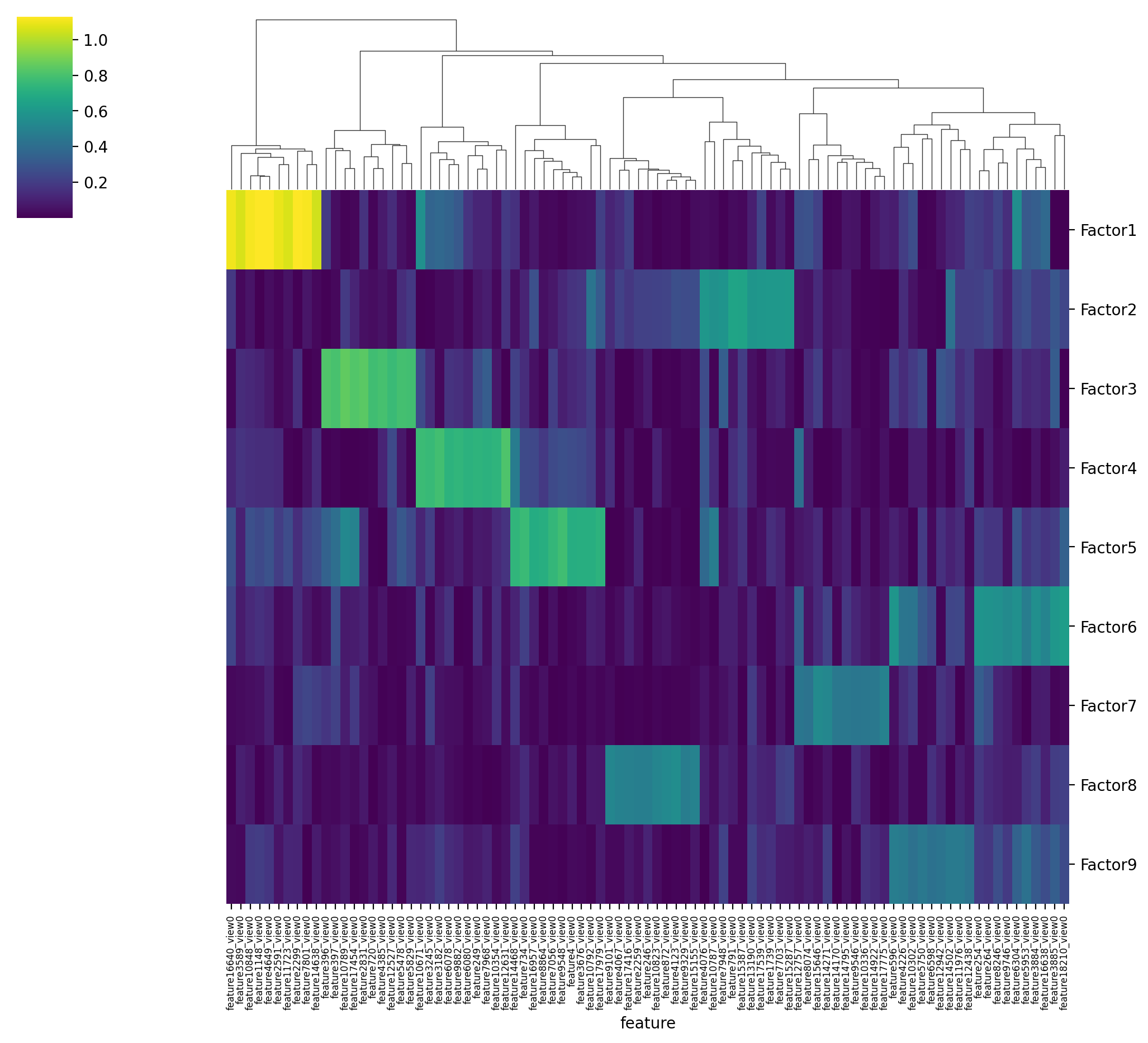
A collage of graphs

AI-generated content may be incorrect.

* No strong overlap between factors: great sign of no redundancy
* But we do see some overlap in factor 5 comparisons...
  + So, we ignore factor 5 as factors 1 - 4 have strong signals already
  + Thus, going back to our original focus of factors 1 – 4
  + Conclusion: high correlation in heatmap is not *always* good and enough

### Revising Heatmap (10 Features)

* After finding our focus factors, we reuse the heatmap but with highest 10 features for each factor now



### Highest Genes in Each Factor

* Factor 1: MYH11, ATP1A2, DES, C2orf40, HSPB6, THBS4, CHRDL1, C7, OGN, SCRG1
* Factor 2: ATP5G2, HPN, SLC34A2, APOA1, SLC5A5, CREB3L3, MUC15, IGF2BP1, PROC, TTR, C12orf56
* Factor 3: AGR2, CAPN8, AGR3, MUC17, TSPAN8, ANXA10, CYP2C18, PIGR, ERN2, FAM3D
* Factor 4: MS4A1, PLA2G2D, CD19, CD79A, LY9, FCRL3, FCRL5, ADAM6, IGJ, MGC29506, S100A2

## **6 – Ontologies**

A group of genes that heavily influence latent factors that MOFA found as important patterns across RNA-seq and mutation data.

### Factor 1

* **GO Biological Process 2025**:
  + Genes are heavily related to heart and contractions and even regulation of blood circulation, which all relate to muscular function of the stomach for digestion for example.
* **GO Cellular Components:**
  + Genes are *also* important for cell communication and movement which heavily impact how stomach cancer grows and spreads.
* **GO Molecular Functions:**
  + These genes are also enriched in sodium/potassium transporter activity and ion binding, which help control cell signal and balance – they’re often altered in cancer to support the growth and survival of the tumour.

### Factor 2

* **GO Biological Process 2025**:
  + Enriched in negative regulation processes, suggesting how tumours suppress immune response or alter tissue structure to spread and survive.
* **GO Cellular Components:**
  + The genes are found in places like vesicles and Golgi which are areas that help the cell make and move proteins; so, these parts in cancer are often changed to help the tumour grow.
* **GO Molecular Functions:**
  + The genes are involved in sodium transport, kinase regulation, and protein interactions, which are important for cell signalling and communication. They play a *crucial* role in how cancer cells grow and respond to signals if weakened.

### Factor 3

* **GO Biological Process 2025**:
  + Genes are crucial in protein unfolding, rna processes, and immune regulation, thus suggesting very strong linkage to tumour behaviour.
* **GO Cellular Components:**
  + Similarly to factor 2 genes, they’re *also* found in vesicle and Golgi, thus pointing to their important in cancer cell signalling.
* **GO Molecular Functions:**
  + Genes involved in immune receptor activities, affecting how the tumour interacts with the immune system or can alter gene expressions.

### Factor 4

* **GO Biological Process 2025**:
  + Genes are linked to B cell activation and immune signalling, suggesting a strong immunity related role.
* **GO Cellular Components:**
  + Genes are once again pointing to involvement in immunity in addition to cell signalling zones.
* **GO Molecular Functions:**
  + Includes ion and metal binding, once again heavily related to immune responses and inflammation.

### Conclusion

1. Factor 1: Plays a role in cell mobility, ion transport, and muscle contraction; processes linked to tumour growth and stomach muscle activity which raises the possibility that stomach cancer influences tissue dynamics and preserves cell viability.
2. Factor 2: Protein transport mechanisms are altered, and immune responses are suppressed in enrichment. These characteristics might aid in tumour growth.
3. Factor 3: Immune control, RNA handling, and unfolded protein response are important mechanisms in this factor which suggests that tumours frequently exhibit stress inducing activities.
4. Factor 4: Strongly linked to immunological function like B cell activation, probably presenting immune-related processes that affect inflammation or tumour presence.

## **7 – Improvements**

* To capture more biological variation, include additional omics layers like proteomics or methylation especially as the mutation omics was weak.
* Instead of keeping clinical data apart, use it as a third view so that MOFA+ may directly learn from signals related to phenotypes.
* To investigate patient subtypes or predict clinical outcomes, use clustering or classification on the latent components.

# **References**

# Rees, M. (2024). What to know about adenocarcinoma stomach cancer. [online] Medicalnewstoday.com. Available at: https://www.medicalnewstoday.com/articles/adenocarcinoma-stomach-cancer#causes [Accessed 5 Jun. 2025].

1. Linkedomics.org. (2025). LinkedOmics :: Data Download. [online] Available at: https://linkedomics.org/data\_download/TCGA-STAD/ [Accessed May 27. 2025].