Exploratory Analysis of Multi-Modal Integration Using MOFA

Table of Contents

1. Motivation and Scope

1.1 Project Motivation and Goal

The goal of this analysis is to integrate multiple data modalities from a PDAC dataset using MOFA, identify latent factors, and explore whether these factors capture biological signals associated with clinical outcomes. The analysis is purely exploratory and intended for hypothesis generation.

1.2 Selected Data Modalities

The modalities available in the assignment were: clinical data, circRNA, miRNA, mRNA, phosphoproteome, proteome, and SCNA.

For this analysis, I reduced the number of modalities to allow easier interpre-

tation, selecting clinical data, mRNA, proteome, and SCNA: - Clinical data - main source of outcome labels, essential for analysis - mRNA - gene expression, commonly used in prognostic analyses - Proteome - protein abundance per gene, captures pathway activation and disease state, provides information about which proteins/signaling pathways are upregulated in high-risk patients
- SCNA - Somatic Copy Number Alterations per gene, represent genomic structural changes (e.g., oncogene amplification, tumor suppressor deletion), can drive both mRNA and protein changes

Details about data preprocessing are provided in Section 4.

2. Key Results

2.1 MOFA Model Summary

I used three modalities to train the MOFA model: mRNA, proteome, and SCNA. This approach allows MOFA to learn the underlying biological drivers for each factor, which can then be correlated with survival outcomes from clinical data. The model was initialized with 10 factors and converged to 8 final factors after training. The training code can be found in the `3_train_mofa.ipynb` notebook.

Analysis of the resulting factors was performed using the MOFAx package, following the official tutorial:

Figure 1. Variance explained per factor and per modality.

Key observations: 1. Different factors are driven by different modalities, e.g., Factor 1 is primarily driven by mRNA and Factor 3 is primarily driven by SCNA. 2. Most factors show variance explained by multiple modalities, indicating that biological signals are shared across data types.

2.2 Association of Latent Factors with Survival

2.2.1 Survival Analysis Using MOFA Factors To identify latent factors potentially relevant to clinical outcome, I calculated the association between each MOFA factor and overall survival using Cox proportional hazards models. Each factor was tested individually using its sample-level factor values.

Factor	HR	Coef	P-value
Factor2	0.753896	-0.282500	0.002358
Factor8	1.529151	0.424712	0.016868
Factor4	1.263506	0.233890	0.022660
Factor6	0.831411	-0.184631	0.208350
Factor3	1.021573	0.021344	0.804748
Factor7	1.025310	0.024995	0.858935
Factor1	1.006422	0.006402	0.915671
Factor5	1.006285	0.006265	0.957607

Table 1. Association between MOFA factors and survival.

Factors 2, 8, and 4 show nominally significant associations with survival ($p < 0.05$). Factor 2 is associated with improved survival (negative coefficient), while Factors 8 and 4 are associated with worse prognosis (positive coefficients). Factors 2 and 8 were selected for downstream interpretation, as they show opposite associations with survival and likely represent contrasting biological states.

3. Interpretation of Selected Factors

3.1 Interpretation of Factor 2

3.1.1 Modality Contribution and Loading Structure Factor 2 variance is primarily explained by the mRNA modality, which also has the largest number of non-zero feature loadings (Figure 2). Proteomic loadings are more symmetrically distributed around zero, while SCNA loadings are sparse but display a pronounced negative peak (Figure 3).

This suggests that Factor 2 captures transcriptional variation, with some genomic alterations reflected in the SCNA modality.

Figure 2. Variance explained per modality and number of non-zero loadings for Factor 2.

Figure 3. Distribution of feature loadings for Factor 2 across modalities.

3.1.2 Identification of Genomic Regions Associated with SCNA Signal To investigate whether the negative SCNA loading peak corresponds to localized genomic events, I examined the genomic positions of genes within the lowest 5% of SCNA loadings using the `mygene` annotation package.

Chromosome	Gene count
3	61
17	46

Table 2. Chromosomal distribution of genes with lowest SCNA loadings.

These genes are highly concentrated within relatively small genomic regions on chromosomes 3 and 17, spanning 17 Mb and 1.41 Mb respectively (Figure 4), suggesting the presence of recurrent copy number alterations.

Figure 4. Genomic start position distributions of negatively weighted SCNA genes on chromosomes 3 and 17.

To further assess the relationship between the identified SCNA regions and Factor 2, I computed pairwise Spearman correlations between mean SCNA values in the chromosome 3 and chromosome 17 regions, Factor 2 values, and survival time (Table 3).

Correlation Pair	Spearman r	p-value
Chr3 mean vs Chr17 mean	0.3230	0.0002007
Chr3 mean vs survival time	-0.0679	0.4463
Chr17 mean vs survival time	-0.1810	0.04086
Factor 2 vs Chr3 mean	-0.2304	0.008877
Factor 2 vs Chr17 mean	-0.2960	0.000693

Table 3. Spearman correlations between regional SCNA values, Factor 2, and survival time.

Mean SCNA values in the chromosome 3 and chromosome 17 regions are correlated with each other. Importantly, Factor 2 shows significant negative correlations with SCNA means in both regions, indicating that this latent factor captures variation in these genomic alterations. While only chromosome 17 shows a direct association with survival time, Factor 2 is strongly correlated with both regions. This suggests that the factor integrates multiple coordinated SCNA signals.

3.1.3 Functional Characterization of mRNA and Proteome Signals with gProfiler Inspection of MOFA weights within the identified chromosomal regions shows negative loadings for chromosome 17 (Figure 5), consistent with the direction of association observed in the survival analysis.

Figure 5. Distribution of MOFA weights within the chromosome 3 and 17 regions.

To functionally characterize this signal, I performed pathway enrichment analysis using g:Profiler on the 100 most negatively weighted mRNA features associated with Factor 2.

The enriched pathways are predominantly related to cytokine signaling, inflammatory response, and immune cell migration (Table 4), suggesting that lower expression of inflammatory and immune-related genes is associated with improved survival.

Source

Name

p-value

Intersection Size

GO:BP

cytokine-mediated signaling pathway

0.000128

16

KEGG

Cytokine-cytokine receptor interaction

0.000237

14

KEGG

Viral protein interaction with cytokine and cytokine receptor

0.003449

10

GO:BP

cellular response to cytokine stimulus

0.008103

16

GO:BP

granulocyte chemotaxis

0.010858

10

GO:BP

inflammatory response

0.015952

18

GO:BP

leukocyte chemotaxis

0.024014

11

GO:BP

myeloid leukocyte migration

0.024014

11

GO:BP

response to cytokine

		0.028645	16
GO:BP	response to peptide	0.045664	16
GO:BP	granulocyte migration	0.046819	10

Table 4. Significantly enriched pathways for negatively weighted mRNA features in Factor 2.

A similar analysis of the 50 most negatively weighted proteomic features revealed enrichment for DNA replication, cell cycle regulation, and DNA damage response pathways (Table 5).

Source	Name	p-value	Intersection Size
GO:BP	double-strand break repair via break-induced replication	9.823730e-07	5
REAC	Unwinding of DNA	9.096544e-06	5
KEGG	DNA replication	4.744610e-05	5
GO:BP			

regulation of DNA-templated DNA replication initiation

1.266312e-04

5

GO:BP

DNA replication initiation

1.656172e-04

6

GO:BP

defense response to fungus

2.688609e-04

6

GO:BP

response to fungus

1.346474e-03

6

REAC

Activation of the pre-replicative complex

4.047180e-03

5

REAC

Activation of ATR in response to replication stress

8.069051e-03

5

KEGG

Neutrophil extracellular trap formation

8.234258e-03

5

GO:BP

regulation of DNA-templated DNA replication

9.950821e-03

5

REAC	DNA strand elongation	1.216417e-02	5
KEGG	Cell cycle	1.488358e-02	5
KEGG	Leishmaniasis	1.625210e-02	4
REAC	Orc1 removal from chromatin	2.563355e-02	6

Table 5. Enriched pathways for negatively weighted proteomic features in Factor 2.

Together, these results suggest that Factor 2 captures a biological state characterized by reduced inflammatory signaling and lower proliferative activity, which is associated with improved patient survival.

An important caveat is that enrichment results depend strongly on the selected feature cutoff and should be interpreted as hypotheses for further validation.

3.2 Interpretation of Factor 8

To contrast with Factor 2, I performed the same exploratory analysis for Factor 8, which showed a significant association with worse survival in the Cox regression analysis.

3.2.1 Modality Contribution and SCNA Structure Factor 8 shows substantial contribution from the proteome modality. SCNA loadings (Figure 7) have several peaks. I analyzed one peak that corresponds t most positively weighted features, and the other corresponds to most negatively weighted features.

Figure 6. Variance explained per modality and number of non-zero loadings for Factor 8.

Figure 7. Distribution of feature loadings for Factor 8 across modalities.

Mapping these SCNA loadings to genomic coordinates revealed two distinct regions: positively weighted genes cluster on chromosome 3 (spanning approximately 28.9 Mb), while negatively weighted genes cluster on chromosome 12 (spanning approximately 13.3 Mb).

3.2.2 Relationship Between SCNA Regions, Factor 8, and Survival

To assess whether these regions are directly associated with survival or primarily reflect latent variation captured by Factor 8, I computed Spearman correlations between regional SCNA means, Factor 8 values, and survival time.

Correlation Pair	Spearman r	p-value
Chr3 mean vs Chr12 mean	0.3000	0.0005821
Chr3 mean vs survival time	-0.0674	0.4495
Chr12 mean vs survival time	-0.0277	0.7564
Factor 8 vs Chr3 mean	0.2734	0.001796
Factor 8 vs Chr12 mean	-0.0556	0.5331

Table 6. Spearman correlations between regional SCNA values, Factor 8, and survival time.

The two SCNA regions show moderate correlation with each other. Neither region is directly associated with survival time. Only the chromosome 3 region shows a significant correlation with Factor 8.

3.2.3 Functional Characterization of mRNA and Proteomic Signals

Given the lack of direct association between regional SCNA values and survival, downstream functional interpretation focused on transcriptional and proteomic signals positively associated with Factor 8.

Pathway enrichment analysis of the top 500 positively weighted mRNA features revealed enrichment for epidermis and skin development-related processes (Table 7), suggesting involvement of epithelial differentiation programs.

Source

Name

p-value

Intersection Size

GO:BP

epidermis development

0.021918

			34
GO:BP	skin development	0.023084	28

Table 7. Enriched biological processes for positively weighted mRNA features in Factor 8.

Enrichment analysis of the top 300 positively weighted proteomic features showed strong enrichment for extracellular matrix organization and collagen-related pathways (Table 8), including collagen biosynthesis and fibril organization.

Source	Name	p-value	Intersection Size
GO:BP	collagen fibril organization	2.26e-08	17
REAC	Collagen biosynthesis and modifying enzymes	5.74e-06	13
REAC	Collagen formation	9.09e-06	15
GO:BP	extracellular matrix organization	1.29e-03	24
GO:BP	extracellular structure organization		

		1.29e-03	24
GO:BP	external encapsulating structure organization	1.29e-03	24
REAC	Extracellular matrix organization	4.07e-02	22

Table 8. Enriched pathways for positively weighted proteomic features in Factor 8.

Factor 8 appears to capture a biological program characterized by extracellular matrix remodeling and epithelial-associated processes, which is associated with poorer survival outcomes.

4. Methods: Data Preprocessing and Model Training

4.1 Clinical Data Preprocessing

Clinical data were used only for downstream association analyses and were not included in MOFA training. Patients with missing values in follow-up time, vital status, or pathological tumor stage were excluded. All remaining patients were retained for survival analysis.

4.2 Omics Data Preprocessing

All omics preprocessing steps were performed before MOFA training (see `2_omics_data_preprocessing.ipynb`).

Only samples present across all selected modalities were retained, resulting in a matched multi-modal dataset.

1. mRNA

The mRNA expression data appeared to be pre-normalized, with no missing values and expression values ranging between 0 and 22. Preprocessing steps included: - Selection of the top 1,500 genes by variance - Z-score normalization across samples for each gene

2. Proteome

Proteomic data contained substantial missingness, with observed values ranging between 12 and 33. The following preprocessing steps were applied: - Removal of proteins detected in fewer than 30% of samples - Imputation of missing values using a protein-wise minimum-based approach (imputed values set below the observed minimum for each protein) - Z-score normalization across samples for each protein

3. SCNA

SCNA data contained some missing values and ranged from -2 to 2. Preprocessing steps included: - Imputation of missing values with zeros - Selection of the top 3,000 genes by variance - Z-score normalization across samples for each gene

4.3 MOFA Model Training

MOFA was trained using the mRNA, proteome, and SCNA modalities. The prepared data matrix consisted of 137 matched samples across all views: - View 0 (mRNA): 137 × 1,500 - View 1 (Proteome): 137 × 9,644 - View 2 (SCNA): 137 × 3,000

The model was initialized with 10 latent factors and sparse feature weights. Training parameters are summarized below (see `3_train_mofa.ipynb` for full details): - Number of factors: 10 - Sparse spike-and-slab feature weights enabled - Automatic relevance determination (ARD) enabled - Maximum iterations: 1,000 - Convergence mode: medium - Factor pruning threshold: variance explained < 0.1% - Random seed: 42

Model training converged and resulted in 8 latent factors retained for downstream analysis.