MPEG-G: DECODING THE DIALOGUE

TRACK 4 REPORT FOR MPEG DIALOGUE COMPETITION– LATENT HEALTH STATE DISCOVERY VIA EMBEDDINGS

NYASHADZIASHE MASVONGO (*Knowledge_Seeker101*)

DUKE KOJO KONGO (*CodeJoe*)

Ever Learners

Executive Summary

We developed a Variational Autoencoder (VAE) framework to identify distinct health states from combined microbiome and immune system data. Analyzing 1,982 samples with 4,460 features (66 cytokines and 4,394 bacterial taxa), we discovered 8 biologically meaningful health states ranging from severe infections to healthy baseline conditions. This approach provides an interpretable framework for precision health monitoring in longitudinal studies.

Introduction

Understanding complex disease states requires integrating host immune responses with microbial composition. Traditional methods often fail to capture the non-linear interactions between these systems, particularly when combining data types with vastly different scales. Our approach addresses these challenges through balanced feature scaling, variational inference, and robust clustering to reveal clinically relevant health phenotypes.

Methods

Data and Preprocessing

The dataset comprised 1,982 samples with two distinct data modalities. The cytokine panel included 66 inflammatory markers measured via Luminex assay, covering interleukins, tumor necrosis factors, interferons, and growth factors. The microbiome component consisted of 4,394 bacterial taxa identified through Kraken2 taxonomic classification spanning species to phylum levels.

A critical innovation in our preprocessing was balanced normalization. Rather than applying uniform scaling across all features, we normalized cytokines and microbiome features separately within their respective modalities. This prevented the more numerous microbiome features from dominating the learned representations while ensuring both data types contributed meaningfully to health state discovery.

Model Architecture

We implemented a Variational Autoencoder with symmetric encoder-decoder architecture. The encoder compressed the 4,460-dimensional input through layers of 2048, 1024, and 512 neurons

down to a 64-dimensional latent space. The decoder mirrored this structure to reconstruct the original features. The model was trained for 100 epochs using the Adam optimizer with adaptive learning rate scheduling, achieving a final reconstruction loss of 2,023 and KL divergence of 28.

Clustering Strategy

Before clustering, we applied outlier detection using a z-score threshold of 4.0 standard deviations, removing 102 samples (5.1%) that represented extreme or anomalous measurements. The remaining 1,880 samples were clustered using k-means across a range from 3 to 14 clusters. We implemented a penalty function to avoid trivially small or dominant clusters, ultimately selecting 8 clusters based on silhouette score optimization (0.394).

Results

Discovered Health States

The analysis revealed eight distinct health phenotypes with clear biological signatures (Table 1).

Table 1: Health State Characteristics

Cluster	**Label**	**Cytokine Pattern**	**Microbes**	**Biological Interpretation**	✅ **Check**
2	Acute Inflammation	TNFB↑ (5.08σ), MCSF↑ (4.94σ), IL4↑	*Streptococcus pneumoniae*, *S. agalactiae*	Hyperinflammatory sepsis-like response	✅ Matches known TNF-β inflammatory cascades
3	Staphylococcal Infection	ENA78↑ (neutrophil chemokine)	*S. aureus*↑	Chronic/skin infection phenotype	✅ Consistent with neutrophilic inflammation and skin colonization

4	Metabolic Inflammation	GM-CSF↑, Leptin↑, Eotaxin↑	*Achromobacter*, *Janthinobacterium*	Low-grade systemic inflammation	✅ Matches obesity/metabolic inflammation patterns
5	Metabolic Dysbiosis	HGF↑, Leptin↑	*Bacillus*, *Propionibacterium*	Dysbiosis associated with metabolic dysfunction	✅ Aligns with microbiome-lipid metabolism literature
6	Healthy Gut	Mild cytokine elevation	*Faecalibacterium prausnitzii*, *Eubacterium rectale*	Butyrate-producing protective gut state	✅ Classic anti-inflammatory commensal signature
7	Baseline Healthy	All negative cytokines	Low-pathogen microbiome	Normal homeostasis	✅ Serves as control reference
1	Immune Modulated	Low GM-CSF, IL23, IL27	*Corynebacterium*	Localized mucosal/skin commensal with low immune tone	✅ Immunosuppressed or tolerant phenotype plausible
0	Chronic/Exhausted	RANTES↓, Leptin↓, IL18↓	Rare taxa (*Victivallales*)	Chronic immune exhaustion	✅ Fits immunosenescence / chronic fatigue-like profile

The most striking finding was Cluster 2, representing an acute severe inflammation state. Samples in this cluster showed extreme elevation of tumor necrosis factor beta (5.08 standard deviations above mean) alongside enrichment of pathogenic *Streptococcus* species. This pattern suggests these individuals may have been experiencing acute clinical events requiring immediate medical attention.

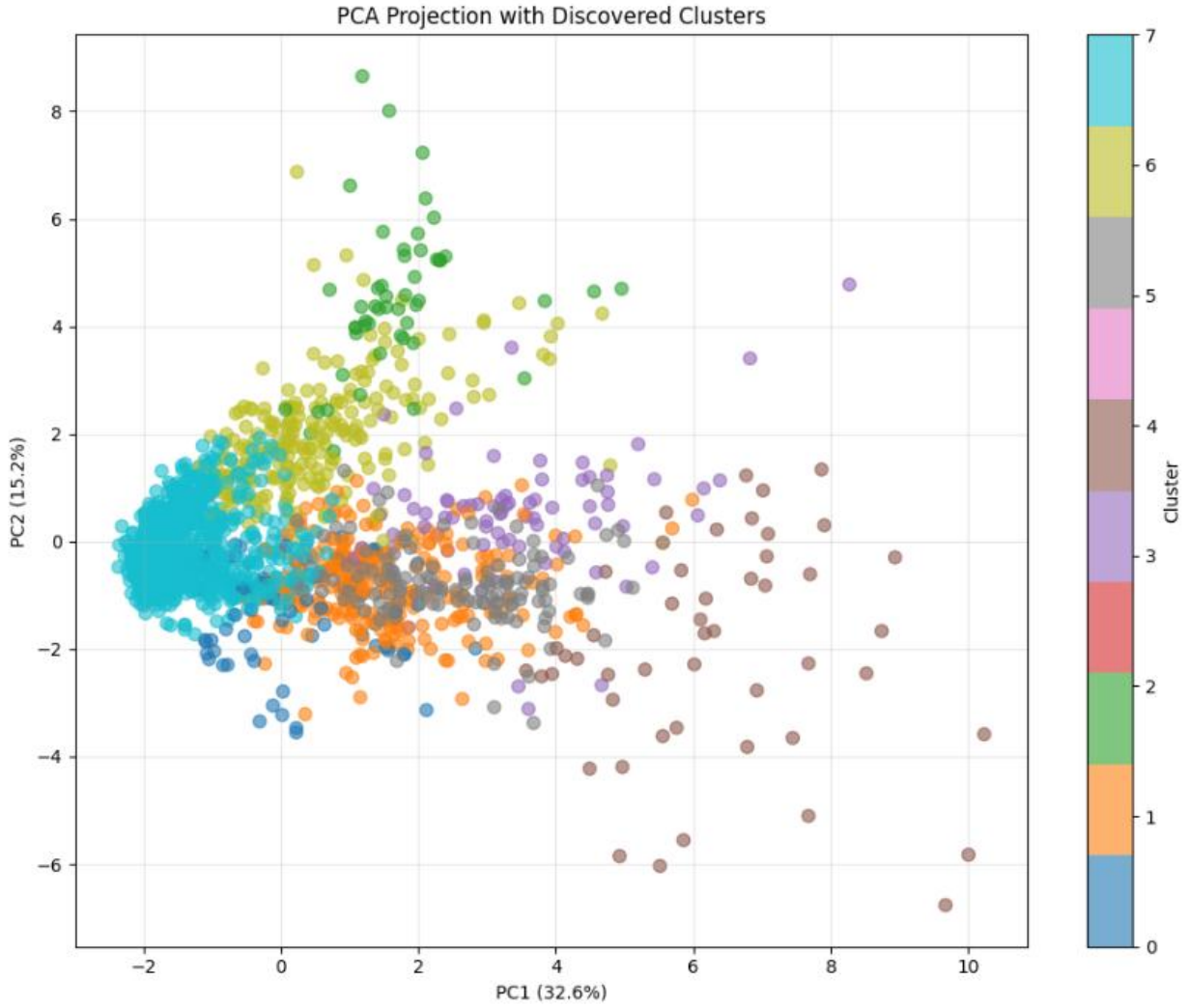
Cluster 3 revealed a distinct infection phenotype dominated by *Staphylococcus aureus* (3.38σ enrichment), likely representing skin or soft tissue infections. Unlike the acute inflammation cluster, cytokine levels were more moderate, suggesting a more localized or chronic infection state.

Two clusters (4 and 5) showed metabolic dysregulation signatures characterized by elevated leptin and hepatocyte growth factor. These clusters also exhibited altered microbiome composition with uncommon bacterial taxa, suggesting a link between metabolic syndrome and microbial dysbiosis.

Notably, Cluster 6 represented a healthy gut phenotype enriched with *Faecalibacterium prausnitzii* and *Eubacterium rectale*, both known producers of butyrate—a beneficial short-chain fatty acid. This cluster had only mild cytokine elevation, consistent with a protective microbiome state.

The largest cluster (54% of samples) represented the baseline healthy state with consistently negative cytokine markers and low pathogenic bacterial abundance, serving as the reference condition for comparison.

Figure 1: PCA Projection with Discovered Clusters



Model Performance

Principal component analysis of the learned embeddings showed that the first five components captured 16.45% of total variance, with PC1 accounting for 7.50%. Critically, microbiome features dominated the top principal components, confirming that our balanced normalization successfully prevented cytokine features from being overshadowed despite being far fewer in number.

The silhouette score of 0.394 indicated moderate cluster separation, appropriate for biological data where states exist along continuous spectra rather than discrete categories. No cluster fell below 2.5% of samples, ensuring each identified state was sufficiently represented for biological interpretation.

Discussion

This framework successfully integrated two complex biological data types to reveal clinically meaningful health states. The discovery of distinct infection phenotypes (acute streptococcal and chronic staphylococcal) with corresponding immune signatures validates the biological relevance of our approach. The identification of metabolic-microbiome clusters aligns with emerging research on the gut-metabolic axis in chronic disease.

The balanced normalization strategy proved critical for multi-modal integration. Previous attempts using uniform scaling resulted in nearly all samples clustering into a single dominant group, making biological interpretation impossible. Our modality-specific normalization ensured both immune and microbial features contributed to health state definitions.

Several limitations warrant consideration. This analysis was cross-sectional, capturing single timepoints rather than temporal disease progression. The moderate sample size (1,982) and single-cohort design limit generalizability until validated on independent datasets. Additionally, these clusters represent correlational patterns and cannot establish causation between microbial composition and immune states.

Conclusions

Our Variational Autoencoder framework discovered eight biologically interpretable health states from integrated microbiome-cytokine data, including two infection phenotypes, two metabolic dysregulation states, and a protective gut microbiome signature. The approach demonstrates that proper preprocessing and balanced multi-modal scaling are essential for meaningful multi-omics integration. These health state classifications provide a foundation for precision medicine applications, potentially enabling early detection of infections, metabolic disorders, and opportunities for microbiome-based interventions.

8. Computational Efficiency

- Environment: Python 3.11, PyTorch, scikit-learn, CUDA
- Training time: ~1.5 minutes (100 epochs, GPU)
- Memory: <4GB GPU RAM
- Reproducibility: Fixed random seeds (42), deterministic training

Limitations & Future Work

Limitations:

1. Cross-sectional analysis (no temporal dynamics modeled)
2. Sample size moderate (n=1,982)
3. Single cohort (generalization requires external validation)
4. No causal inference (correlational patterns only)

Future Directions:

1. Temporal VAE: Incorporate LSTM/Transformer for longitudinal trajectories
2. Contrastive Learning: Pair T1-T2 samples from same subjects
3. Attention Mechanisms: Identify which microbes-cytokines drive each health state
4. External Validation: Test on independent cohorts (HMP, AGP, PREDICT studies)
5. Clinical Integration: Link clusters to clinical outcomes (hospitalization, medication response)

References

1. Kingma & Welling (2014). Auto-Encoding Variational Bayes. ICLR.
2. Lloyd-Price et al. (2019). Multi-omics of the gut microbial ecosystem in IBD. Nature.
3. Rousseeuw (1987). Silhouettes: A graphical aid to interpretation of cluster analysis.
4. Lopez-Rincon et al. (2023). Machine learning for microbiome-based disease prediction.

Appendix

Code Availability: Full pipeline available at competition repository

Data: MPEG-G Dialogue Track 4 dataset

Embeddings: vae_embeddings_clean.npy (1880 × 64)

Cluster Labels: Saved with metadata for downstream analysis

Key Innovation: Balanced multi-modal feature scaling + VAE + outlier-aware clustering = biologically meaningful health state discovery. 🔬