# Discovering Hidden Disease Subtypes from Proteomic Data Using Unsupervised Clustering

Shri Chada, Vinaika Maruvada
Department of Biomedical Engineering, University of Houston, Houston, TX

## Introduction

- Clinical diseases are commonly classified and treated based on symptoms, clinical measurements, and diagnostic tests. While this framework is effective for identifying broad disease categories, it often fails to explain why **patients with the same diagnosis** can experience **very different disease progression and treatment responses**. In chronic and immune-mediated diseases, therapies that work well for some patients may be ineffective for others, delaying effective care and allowing disease to worsen.
- This discrepancy arises because patients who share the same clinical diagnosis may differ substantially at the molecular level. Traditional disease labels group patients by **observable features** but **do not fully capture the underlying biological processes driving disease**. As a result, biologically distinct patients are often treated as if they are the same.
- Advances in **high-dimensional proteomics** now allow **thousands of proteins** to be measured simultaneously from patient samples, providing insight into immune activity, cellular stress, signaling pathways, and tissue remodeling. However, **identifying meaningful biological structure within such complex datasets remains challenging**. Most existing approaches rely on predefined disease labels, which can obscure molecular heterogeneity.
- To address this limitation, this project uses an **unsupervised, network-based computational approach** to identify **biologically meaningful patient subgroups** directly **from proteomic data, without relying on clinical diagnoses.**

## Project Significance

### Societal Impact:

- Understanding molecular heterogeneity within diseases has direct implications for patient care and treatment effectiveness. When patients are grouped solely by **clinical diagnosis**, therapies are often applied using a **one-size-fits-all approach**, despite large **biological differences between individuals**. This can result in **ineffective treatments, unnecessary side effects, and delayed disease management**.
- By identifying **molecularly defined patient subgroups**, this research supports a shift toward **biology-driven patient stratification.** Such an approach can help clinicians better **understand why certain patients respond differently to the same therapy** and guide the development of **more targeted and personalized treatment strategies**. Ultimately, this may **improve patient outcomes, reduce trial-and-error prescribing, and lower healthcare costs.**

### Scientific Importance:

- From a scientific perspective, this work addresses a key **challenge** in modern biomedical research: **extracting biologically meaningful patterns from high-dimensional proteomic data.** Traditional analysis methods often focus on individual proteins or predefined disease categories, which can miss coordinated biological processes acting across multiple proteins.
- This research introduces **a network-based framework** that **groups proteins into functional modules** and uses these modules to **characterize patient biology**. By capturing coordinated biological activity rather than isolated signals, the approach provides a systems-level view of disease mechanisms. Applying this framework across multiple diseases and sample types also allows assessment of generalizability.

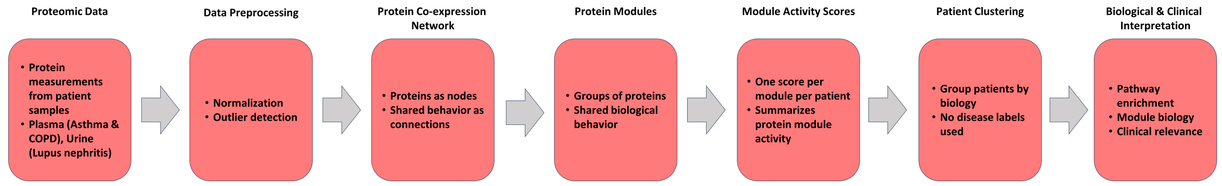
## Research Objectives

The primary objectives of this research are threefold:

- First, to develop an **unsupervised computational pipeline to identify biologically meaningful protein modules** based on shared expression patterns, **without relying on predefined disease labels**.
- Second, to use module-level activity patterns to **cluster patients into molecular subgroups** and assess whether these subgroups **reveal heterogeneity within and across clinically defined diseases**.
- Third, to **evaluate the generalizability of this approach** by applying **the same pipeline to** independent proteomic datasets from **different diseases and biological sample types**.

## Procedure and Methodology

**Figure 1. Overview of the computational pipeline for identifying molecular patient subgroups**




This figure summarizes the data-driven pipeline used to identify molecular patient subgroups. Proteins with similar behavior are grouped into modules, patients are clustered based on module activity, and the resulting groups are interpreted using biological pathway analysis.

**Methods and Data Overview**

**Datasets**

- **Asthma–COPD plasma dataset:** Plasma protein measurements from patients diagnosed with asthma or chronic obstructive pulmonary disease (COPD), comprising **85 patients** and **1233 measured proteins** per sample.
- **Lupus nephritis urine dataset:** Urine protein measurements from patients with lupus nephritis (LN), Diabetic Chronic Kidney Disease (DCKD), Bladder Cancer (BC), Urology Clinic controls (UC, presenting with hematuria) and Healthy Controls (HC), comprising **85 patients** and **7596 measured proteins** per sample.

**Data Preprocessing**

- Protein measurements log-transformed to reduce skew and enable fair comparison across patients.
- Small number of extreme outlier samples removed.
- Low-variance and highly redundant proteins filtered prior to network construction to focus the analysis on informative molecular signals.

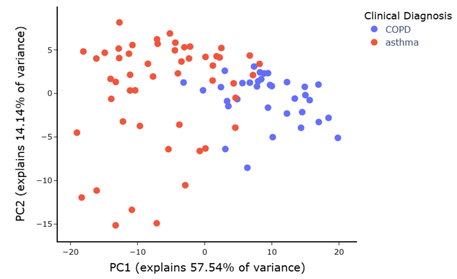
**Module and Patient Analysis**

- Proteins organized into co-expression modules using a network-based community detection approach (Walktrap algorithm).
- For each patient, a **module activity score** was calculated for every protein module.
- Patients were then clustered into **molecular patient groups** based on module activity profiles, without using clinical disease labels.

**Biological Interpretation and Generalization**

- Protein modules showing strong differences across molecular patient groups were interpreted using pathway and gene ontology enrichment analysis (Reactome, Gene Ontology, KEGG).
- To assess generalizability, the same pipeline was applied to an independent urine proteomic dataset containing lupus nephritis and other disease and control groups.

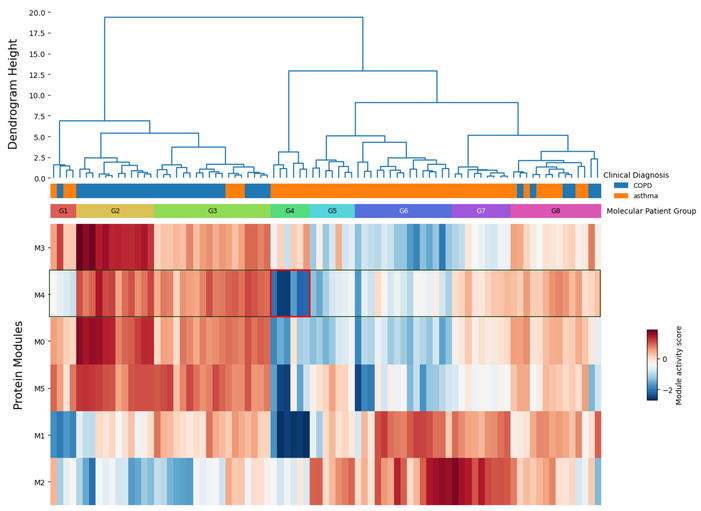
**Figure 2. Exploratory PCA of patient protein profiles by clinical diagnosis**




Principal component analysis (PCA) of patient proteomic profiles from the COPD–asthma dataset. Each point represents a patient and is colored by clinical diagnosis. Substantial overlap between asthma and COPD patients indicates that diagnosis alone does not fully capture underlying molecular differences, motivating the need for unsupervised, biology-driven patient stratification.

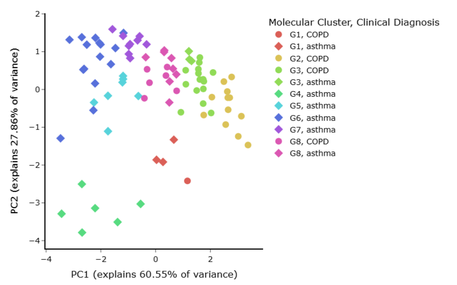
## Results

**Figure 3. Hierarchical clustering based on protein module activity reveals distinct molecular patient groups**



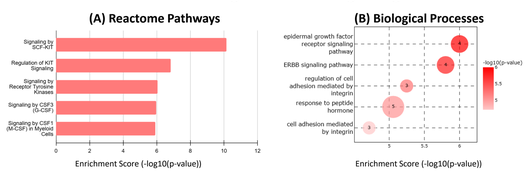

Hierarchical clustering of patients based on protein module activity reveals distinct molecular subgroups. Annotation bars indicate clinical diagnosis and molecular cluster assignment, while the heatmap shows the module activity scores that drive patient similarity.

**Figure 4. PCA of module activity reveals molecular subgroups within and across clinical diagnoses**




Principal component analysis (PCA) of the COPD–asthma proteomic dataset, with patients colored by **molecular patient cluster** assignment. Molecular clustering resolves distinct subgroups within each clinical diagnosis while also grouping some patients with different diagnoses who share similar module activity profiles. This demonstrates how module-based clustering complements, rather than replaces, clinical classification.

**Figure 5. Functional enrichment of Module 4 highlights receptor and growth factor signaling**




**(A)** Reactome pathway enrichment highlights coordinated receptor tyrosine kinase–centered signaling, including EGFR/ERBB, KIT, and CSF-linked pathways. **(B)** Gene Ontology Biological Process terms further support enrichment for growth factor signaling and integrin-mediated cell communication. Dot size reflects the number of module proteins contributing to each term, and color indicates statistical significance.

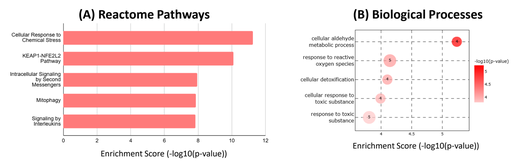
**Molecular Patient Group G4 Aligns with a Distinct Clinical Profile**

Patients in molecular patient group G4 showed clinical characteristics that were not obvious from diagnosis alone. These patients exhibited low blood and sputum eosinophil levels, low fractional exhaled nitric oxide (FeNO), and high Asthma Quality of Life Questionnaire (AQLQ) scores. G4 also showed markedly low Module 4 activity. This alignment between independently measured clinical characteristics and the molecular profile supports the biological relevance of molecular patient grouping.

### Key Takeaways

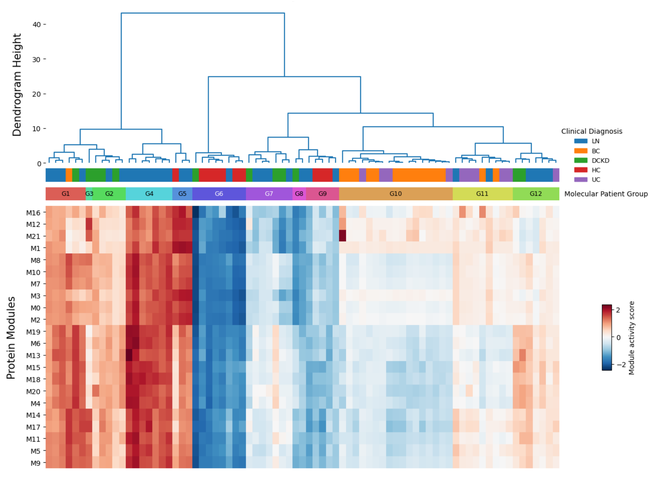
- Molecular clustering reveals patient subgroups that are not fully explained by clinical diagnosis alone.
- Patient subgroups show distinct patterns of protein module activity, including differences in signaling and stress-response programs.
- G4's distinct clinical profile provides evidence for the biological relevance of molecular patient grouping.

**Figure 6. Functional enrichment of Module 3 highlights oxidative stress and detoxification**




**(A)** Reactome pathway enrichment emphasizes **cellular stress and detoxification responses**, including KEAP1–NFE2L2 signaling, mitophagy, and interleukin signaling. **(B)** Gene Ontology Biological Process terms highlight **oxidative stress responses, detoxification, and metabolic processing**.

**Figure 7. Molecular clustering reveals heterogeneity within lupus nephritis in an independent urine dataset**




Applying the same clustering pipeline to an independent urine proteomic dataset shows that patients with the same diagnosis can belong to different molecular groups, demonstrating the generalizability of the approach.

## Conclusions

- This research demonstrates that **unsupervised, data-driven proteomic network analysis** can **reveal biologically meaningful patient subgroups** that are **not captured by traditional clinical disease classifications**. By grouping proteins into functional modules and clustering patients based on module activity, the pipeline uncovered substantial molecular heterogeneity within and across clinically defined diseases.
- In a **plasma proteomic dataset of asthma and COPD patients**, the approach largely separated patients by diagnosis while identifying **multiple molecular subtypes within each disease.** Protein modules differentiating patient subgroups were enriched for biological processes involving immune signaling and oxidative stress responses. These findings provide biological insight into why patients with the **same diagnosis may experience different disease trajectories and treatment responses.**
- Applying the same pipeline to an independent urine proteomic dataset containing lupus nephritis and other diseases **demonstrated the generalizability of the approach**. While patients with some diagnoses clustered tightly, lupus nephritis patients were distributed across multiple molecular groups, indicating substantial biological heterogeneity. Together, these results show that this **framework can uncover hidden molecular structure across diseases and sample types**, supporting a shift toward **biology-driven patient stratification** and personalized medicine.

## Limitations

- The interpretation of molecular subtypes depends on the availability of clinical metadata. While supporting metadata strengthened interpretation for some patient groups, additional clinical and longitudinal data would further improve biological and clinical insight.

## Future Work

- Future work will focus on extending and refining this framework to further enhance biological interpretation and clinical relevance. Integrating additional data types, such as transcriptomics or longitudinal proteomic measurements, could improve the stability and resolution of molecular subgroups.
- Expanding the analysis to larger and more diverse patient cohorts will help assess the robustness of identified subtypes and their association with clinical outcomes. In particular, linking molecular subgroups to treatment response and disease progression could directly inform precision medicine strategies.
- Finally, refining module-based biomarkers and exploring their potential use in patient stratification or therapeutic targeting may help translate these findings into clinical decision-making tools.

## Select References

1. Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., Stichweh, D., Blankenship, D., Li, L., Munagala, I., Bennett, L., Allantaz, F., Mejias, A., Ardura, M., Kaizer, E., Monnet, L., Allman, W., Randall, H., Johnson, D., Lanier, A., … Pascual, V. (2008). A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity, 29(1), 150–164. https://doi.org/10.1016/j.immuni.2008.05.012

2. Schiess, R., Wollscheid, B., & Aebersold, R. (2009). Targeted proteomic strategy for clinical biomarker discovery. Molecular oncology, 3(1), 33–44. https://doi.org/10.1016/j.molonc.2008.12.001

3. George, B. J., Reif, D. M., Gallagher, J. E., Williams-DeVane, C. R., Heidenfelder, B. L., Hudgens, E. E., Jones, W., Neas, L., Hubal, E. A., & Edwards, S. W. (2015). Data-driven asthma endotypes defined from blood biomarker and gene expression data. PloS one, 10(2), e0117445. https://doi.org/10.1371/journal.pone.0117445

4. Saelens, W., Cannoodt, R., & Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. Nature communications, 9(1), 1090. https://doi.org/10.1038/s41467-018-03424-4

5. Dagher, R., Fogel, P., Wang, J., Soussan, D., Chiang, C. C., Kearley, J., Muthas, D., Taillé, C., Berger, P., Bourdin, A., Chenivesse, C., Leroy, S., Anderson, G., Humbles, A. A., Aubier, M., Kolbeck, R., Pretolani, M., & COBRA Consortium (2022). Proteomic profiling of serum identifies a molecular signature that correlates with clinical outcomes in COPD. PloS one, 17(12), e0277357. https://doi.org/10.1371/journal.pone.0277357

6. Suzuki, M., Cole, J. J., Konno, S., Makita, H., Kimura, H., Nishimura, M., & Maciewicz, R. A. (2021). Large-scale plasma proteomics can reveal distinct endotypes in chronic obstructive pulmonary disease and severe asthma. Clinical and translational allergy, 11(10), e12091. https://doi.org/10.1002/clt2.12091

7. Vasunilashorn, S. M., Dillon, S. T., Chan, N. Y., Fong, T. G., Joseph, M., Tripp, B., Xie, Z., Ngo, L. H., Lee, C. G., Elias, J. A., Otu, H. H., Inouye, S. K., Marcantonio, E. R., & Libermann, T. A. (2022). Proteome-Wide Analysis Using SOMAscan Identifies and Validates Chitinase-3-Like Protein 1 as a Risk and Disease Marker of Delirium Among Older Adults Undergoing Major Elective Surgery. The journals of gerontology. Series A, Biological sciences and medical sciences, 77(3), 484–493. https://doi.org/10.1093/gerona/glaa326

## Acknowledgements

The authors gratefully acknowledge Dr. Chandra Mohan and the members of the Mohan Lab in the Department of Biomedical Engineering at the University of Houston for their guidance and support. The authors also acknowledge the use of urinary proteomic data provided by the Mohan Lab.

**Contact**
**Shri Chada**
shrila777@gmail.com