**[BIOINFO 2025]**

# Patient Phenotype Prediction from Single-cell RNA-seq Data through Annotation-free Multiple Instance Learning

Janghyun Noh[1], Min Kim[2], Yoobin Shin[2], and Minsik Oh[1,*]

[1] *Department of Data Technology, Myongji University*
[2] *Department of Convergence Software, Myongji University*
* *Corresponding author: msoh@mju.ac.kr*

Patient phenotypes can be determined by specific cell populations with abnormal functions, making their prediction a key challenge in precision medicine. In this context, single-cell RNA sequencing (scRNA-seq) provides a powerful resource for identifying disease-associated cell populations but variability in cell numbers, high dimensionality, and technical noise hinder predictive performance. Multiple Instance Learning (MIL)-based approaches address these issues by representing each patient as a bag of cells, using attention mechanisms to identify informative subpopulations. Recent studies have extended this idea with hierarchical attention, leveraging predefined annotations to improve interpretability. However, reliance on annotation quality can compromise robustness and limit predictive performance.

To overcome this limitation, we propose AMIL (Annotation-free Multiple Instance Learning), a framework that predicts patient phenotypes without requiring predefined annotations. AMIL generates cell groups directly from gene expression and progressively refines them through adaptive merging in the scGPT embedding space pretrained on large-scale single-cell data. This refinement enforces intra-group compactness and inter-group separation, preserving local gene expression variation while incorporating global patterns from scGPT embeddings. By applying multiple merging thresholds, AMIL produces a set of candidate groupings, from which the optimal is selected using a composite clustering score. These groups are then integrated into a MIL framework with dual-level attention, capturing informative signals at both the cell and group levels.

Experiments on benchmark datasets demonstrate that AMIL outperforms existing methods, achieving higher AUROC scores while discovering biologically meaningful subpopulations. These results establish AMIL as a robust, annotation-free framework for phenotype prediction, with strong implications for precision medicine and disease research.