Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/200315

Optimized Bi-clustering Framework for AML Biomarker Discovery: Integrating PHATE-Based Dimensionality Reduction and Enhanced Gene Expression Clustering

*Tarik Himdi**, *Mohammed Ishaque, Khaled ElBahnasy*

Department of Computer Science and Information technology, Jeddah International college, 23831, Jeddah, Saudi Arabia

Received: 11 Jan. 2026, Revised: 15 Mar. 2026, Accepted: 20 Apr. 2026
Published online: 1 May. 2026

Abstract: One of the key features of AML is genetic heterogeneity, which poses major problems in developing molecular markers that could be employed to improve diagnostic and individualized therapy accuracy. In this study, we proposed a methodological framework that uses a refined version of the biclustering algorithm of Cheng and Church. The novelty of our approach consists in the use of advanced techniques for reducing the dimension of the data, especially the use of PHATE, in addition to UMAP and NMF methods, to optimize the biclustering analysis. We identified 61 significant biclusters, including 66 critical genes. They include classical AML markers, such as CD38, MPO, and CREBBP, as well as potentially new genes, PLXND1 and ALS2, that might indicate additional molecular mechanisms of disease formation. All identified biclusters were validated according to their biological significance and statistical reliability using silhouette scoring, cluster coherence score, and comparison with the current scientific literature. Our proposed approach provides better clustering than classical approaches such as PCA based on silhouette scores.

Keywords: Acute Myeloid Leukemia, Transcriptomics, Biclustering, Biomarker Discovery, Dimensionality Reduction, PHATE, Computational Biology, Cheng and Church Algorithm.

1 Introduction

Introduction Acute Myeloid Leukemia (AML) is a highly malignant cancer affecting the blood system and caused by the formation of immature myeloid cells in the bone marrow. This disease affects the functioning of the bone marrow as a whole, interfering with its normal activity, and resulting in serious blood cell deficiencies that cause such problems as anemia, infections, and bleeding [1]. Unfortunately, although intensive investigations are being carried out in this field, there have not yet been any significant breakthroughs regarding this disease's survival rate (it does not exceed 25–30% in adults). [2, 3].

One of the difficulties associated with the management of acute myeloid leukemia (AML) is its high degree of molecular heterogeneity. This is a disease that results not from a single genetic abnormality but from a broad range of molecular alterations. Mutations that affect certain genes, including FLT3, NPM1, TP53, and DNMT3A, often interact in complicated and hard-to-predict manners [4].

1.1 The Need for Computational Biomarker Discovery

Conventional techniques used for diagnosing AML comprise cytogenetics, morphology, and immunophenotype. These methods serve as a basis for initial diagnosis, although they are insufficient in providing a full view of the complicated molecular nature of the condition. New methods based on high-throughput profiling of transcriptomes and next-generation sequencing have yielded vast amounts of data, allowing a comprehensive view at the genome-wide level of gene expression patterns [5]. However, this abundance of information gives rise to another issue, which is known as the "Curse of Dimensionality," due to the large number of genes, making it difficult to analyze the expression pattern of particular sets of patients [6].

Computation biology has emerged as a necessity to deal with the above-mentioned problems. Clustering techniques such as hierarchical clustering and k-means clustering have been used to cluster AML into different

* Corresponding author e-mail: dean@jicollege.edu.sa

molecular subtypes [7]. However, the problem with these methods is that both assume a uniform behavior of the genes in all samples; thus, they fail to detect localized co-expression clusters, which play a significant role from the mechanistic perspective [8]. In other words, while global methods can provide some information, they do not have the capacity to detect local clusters, which are needed to understand AML clinical phenotypes. Thus, methods must be devised to detect local clusters in the data matrix.

1.2 Biclustering: A Targeted Analytical Solution

The solution provided by biclustering, or co-clustering, is a welcome respite since biclustering involves the grouping of both genes and samples in the same matrix. This kind of grouping helps in detecting the patterns locally, which include a set of genes displaying some common behavior only among a certain group of samples [9]. The advantage of biclustering becomes evident in the realm of oncology, where genes connected to specific characteristics such as drug resistance or relapse susceptibility can be discovered through biclustering but not through global clustering [10]. For instance, the transcriptional impact of a FLT3-ITD mutation depends significantly on additional mutations present in the cell [11].
One such method is the well-known algorithm by Cheng and Church that finds submatrices with minimum mean squared residue, which is indicative of consistency of expression [12]. Nevertheless, its vulnerability to noise, parameter tuning requirements, and lack of capability to deal with overlapping biclusters, in which genes participate in several processes, prompted many modifications and enhancements [13, 14].

1.3 Translational Motivation and Specific Contributions

The translation of molecular research into clinical applications requires the identification of reliable biomarkers for enhanced risk stratification, treatment decisions, and experimental designs. Nonetheless, there is a translational barrier owing to three main deficiencies in existing computational methods: (i) the extensive use of linear dimensional reduction methods (such as PCA), which cannot address the non-linearities in the interactions of biological systems [15]; (ii) inadequate validation of the robustness and reproducibility of computational algorithms; and (iii) lack of integration between computational gene identification and experimental validation.
In order to tackle such issues, we propose an integrated and enhanced biclustering approach for identifying AML biomarkers. The submissions of our research are:

1. Algorithmic Enhancement: An improved Cheng and Church algorithm featuring adaptive thresholding, explicit overlap handling, and bootstrap-based stability validation to increase biological relevance and reproducibility.
2. Advanced Dimensionality Reduction Integration: The implementation of PHATE, a state-of-the-art non-linear embedding algorithm, for improved preservation of local and global structure in transcriptomic data, significantly surpassing conventional linear approaches.
3. New Biological Discoveries: A computational pathway that not only detects the already known AML biomarkers but also discovers novel gene targets (like PLXND1, ALS2) that could hold biological significance.
4. Multi-aspect Validation: The procedure of validating a model from multiple aspects including internal validation using methods like silhouette scores and stability indices and external validation through the scientific literature.

1.4 Document Structure

The rest of this paper is organized as follows: Section 2 discusses the relevant literature regarding the study of AML using transcriptomics, biclustering algorithms, and dimensionality reduction techniques. Section 3 presents the method used, which includes data pre-processing, dimensionality reduction, and optimized biclustering.. The findings obtained from our experiment and its biological significance are described in section 4. We discuss the implications, limitations, and future directions for this study in section 5.

2 Literature Review

Due to the complex nature of its molecular structure, much research has been done on the field of computational genomics, particularly regarding biomarkers using high-throughput gene expression analysis [16]. However, due to the shortcomings of clustering methods in identifying local and context-specific gene regulation patterns, there is increasing attention on the utilization of biclustering methods in the process of subtype classification for AML [17]. This chapter will present an evaluation of current literature regarding AML subtyping using genomic analysis, biclustering methods, and dimensionality reduction.

2.1 Molecular Classification of AML via Transcriptomic Profiling

Genetically, AML is characterized by the presence of recurrent mutations in genes such as FLT3, NPM1,

DNMT3A, and TP53, among others. Projects such as The Cancer Genome Atlas (TCGA) have played an important role in creating genomic and epigenomic profiles of AML and have provided useful resources for researchers. Unsupervised learning approaches have been employed to classify molecular subtypes using the collected information and supervised and deep learning algorithms on RNA-seq data have been found promising for predicting clinical outcomes, such as relapse and survival [18–20].

A consistent problem with such global analyses of transcriptome data is that they often fail to identify subsets of co-expressions that are unique to each specific condition. As reported by Li et al. [21], for instance, FLT3-ITD mutations trigger various transcriptional signatures in accordance with cellular contexts, thus calling for methods capable of identifying such condition-specific signatures. This is particularly important in light of the 2022 guidelines published by the European LeukemiaNet (ELN), which strongly advocate for molecular stratification in AML patients [3].

2.2 Evolution and Application of Biclustering in Oncology

In contrast to traditional clustering methods that focus on the assumption of global similarity, biclustering is able to cluster both genes and conditions (samples) simultaneously, thus providing an effective way to find localized patterns that are present only in certain subsets of patients [13]. In cancer studies, this capability plays an important role in discovering disease-specific mechanisms and potential therapeutic approaches [22].

The early use of these techniques entailed adapting existing algorithms of biclustering to fit the biological datasets. Nevertheless, current trends have shifted focus towards modifying these algorithms based on the attributes of the datasets. A study carried out by Eren et al. [23] tested 12 different algorithms of biclustering on the gene expressions of cancer and concluded that there is a need for customization in the algorithmic performance since it is highly dependent on the nature of the data and the question at hand. Also, according to Padilha and Campello [24], most of these algorithms lack capacity to analyze time-series expressions.

2.3 Addressing the Limitations of Biclustering Algorithms

The same characteristic that makes biclustering powerful, i.e., its ability to detect local patterns, is also its Achilles' heel, since most biclustering algorithms are highly sensitive to noise and initial conditions, making them unreliable and hard to reproduce [25]. Scalability is also an issue as omics datasets become larger. In addition, the fact that many traditional biclustering algorithms cannot assign a gene to more than one bicluster is also an important problem, since genes are part of various cellular functions [26].

Several solutions have been suggested to overcome these drawbacks. Henriques et al. [27] proposed the introduction of consensus clustering to enhance stability. Methods based on fuzzy logic were proposed to enable the presence of membership in several clusters for each instance in a graded manner, which can better reflect the reality of biology [28]. Xie et al. [29] utilized a multi-objective optimization approach by integrating gene ontology data in the process of biclustering. Most recently, Karim et al. [30] showed how deep learning autoencoders could be used to learn latent features to find meaningful biclusters.

2.4 The Critical Role of Dimensionality Reduction

Transcriptomic data is of high dimensions in that the number of features or genes is much larger than the number of samples, making it difficult for clustering algorithms. This is because of the high dimensionality problem of transcriptomic data, which makes the issue of overfitting worse, thus limiting the performance of clustering algorithms [31]. Methods for dimensionality reduction (DR) help to overcome this problem.

However, although methods such as PCA and NMF have become popular for their linear properties, they tend to be poor at capturing the complicated, non-linear associations present in biological data sets. Hence, it is not surprising that more advanced methods such as t-SNE, UMAP [32] and PHATE [33] have been employed for the analysis of biological data. Notably, PHATE has emerged as an exceptionally powerful method for capturing both the local and global structure in biological trajectories, making it ideal for use in cancer genomics.

For AML research, DR is more than just an initial processing step. For example, Hamidi et al. [34] used NMF on AML data for pediatrics and were able to discover different subtypes of molecules with specific clinical outcomes. There are currently ongoing efforts in the area of DR to develop "pathology-aware" algorithms where knowledge from biological pathways, like protein-protein interactions, is used in creating embeddings [35].

2.5 Identified Research Gaps and Study Rationale

However, according to the literature review, there still exist three major gaps that hinder the advancement towards the development of a clinically significant biomarker for AML:

1. **Over-reliance on Basic Dimensionality Reduction:** The current work done on biclustering in AML is mainly focused on using techniques like PCA that use linear models but ignore the complex non-linear relationships present in the manifold of the gene regulatory network [36].
2. **Insufficient Validation of Bicluster Stability:** It is the most critical factor when it comes to assessing the reproducibility of biclustering that is discovered by bootstrapped and/or independent datasets [26].
3. **Disconnect Between Prediction and Application:** There is an appreciable chasm between the computational discovery of biomarkers and their biological and clinical validation, which leaves several potentially valuable candidates as mere theories [37].

Motivated by the limitations listed above, the current study presents a novel analysis framework designed specifically for tackling the aforementioned issues. Such a methodology consists of the following three components: (i) the utilization of the high potential of PHATE for dimensionality reduction due to its nonlinear nature for the purpose of capturing the complexity of biological processes; (ii) the incorporation of the stability of Cheng and Church's biclustering method through the use of bootstrap techniques; and (iii) validation procedures both internally, using the assessment of clusters' qualities, as well as externally, using information from the existing scientific literature.

3 Methodology

This research paper adopts a systematic computational approach for discovering coherent biclusters from AML gene expression data. The methodology adopted in this work is depicted in Figure 1. The methodology comprises five major steps: (i) collection and preprocessing of data; (ii) application and assessment of dimensionality reduction methods; (iii) biclustering by improved Cheng and Church algorithm; and (iv) assessment of biclusters obtained through various methods.

3.1 Dataset Acquisition and Description

Transcriptomic data used in this analysis was obtained through the cBioPortal for Cancer Genomics [38] based on the study conducted by Bottomly et al. [39]. This transcriptomic dataset consists of RNA-Seq data that have been normalized using FPKM-UQ. To avoid technical bias that might affect our results, we conducted a Principal Component Analysis to investigate possible batch effects and found no evidence of batch effects in the data. The transcriptomic dataset contains 671 AML samples annotated clinically, along with information about the gene expression profile, involving expression level measurement of 16,384 protein-coding genes. Alongside gene expression profiling, detailed clinical metadata is available, including cytogenetic risk group, mutations in relevant genes (FLT3, NPM1), and induction treatment response. see table 1.

Table 1: Summary of the Processed AML Transcriptomic Dataset

Characteristic	Value	Source
Original Samples	671	Bottomly et al., 2022 [39]
Original Genes	16,384	Bottomly et al., 2022 [39]
Genes after Filtering	∼13,107	This Study
Clinical Annotations	Yes (Comprehensive)	cBioPortal / Bottomly et al., 2022 [39]

3.2 Data Preprocessing Pipeline

The preprocessing pipeline implemented here was effective in ensuring that high-quality data with minimal noise were available for analysis. The pipeline comprised three successive stages:

1. **Normalization:** In order to normalize for differences in the sequencing depth and size of samples, gene expression levels underwent min-max normalization. This method scales all variables into a fixed range [0, 1] while maintaining the original relationships in the data distribution. The normalized expression level xij′ for an expression level xij was determined by Eqn. (1):

$$x'_{ij} = \frac{x_{ij} - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{1}$$

The Min-Max normalization method was chosen over the Z-score normalization because it retains the natural distribution of expression values on the scale [0,1], which would be beneficial for applying variance-based feature selection methods and working with some of the non-negative restrictions in some dimensionality reduction approaches such as NMF.

1. **Missing Value Imputation:** Missing data comprised about 2.3% of the values in the dataset (Not Available – NA). Missing data was imputed using the k-nearest neighbors (k-NN) algorithm (where k=5). For each sample with a missing value, the algorithm identified the five most genetically similar samples (based on Euclidean distance) and imputed the missing value as the mean expression from those neighbors. The Euclidean distance used to identify the nearest neighbors is calculated in Eqn. (2):

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

where x_i and y_i is the expression values of two samples for gene i,n is the total number of genes/features,$d(x,y)$ means Euclidean distance between samples.

1. **Feature Selection:** In order to reduce the level of computation and concentrate on those genes having biological variations, variance-based filtering was used. Those genes whose variance was less than the 20th percentile in all the samples were considered to be uninformative and hence filtered out. The variance of the genes for all the samples is shown in Eqn. (3):

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3}$$

where x_i is the expression value of a gene in sample i, mu is the mean expression value of the gene, n is the total number of samples.

3.3 Dimensionality Reduction Strategy

To solve the problems caused by high dimensional data, seven different DR methods were applied and compared:

1. **Linear Methods:** Principal Component Analysis (PCA), Truncated Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF).
2. **Non-Linear Manifold Learning:** Uniform Manifold Approximation and Projection (UMAP), Isomap, Potential of Heat Diffusion for Affinity-based Transition Embedding (PHATE). Additionally, for comparison purposes, the data were visualized using t-distributed Stochastic Neighbor Embedding (t-SNE).

Replicability was assured using standard implementation with default parameters unless otherwise specified. All feature reduction techniques (PCA, SVD, NMF, UMAP, Isomap, PHATE) had their number of principal components set at 50 to maintain variance and minimize noise. In case of the PHATE embedding, which turned out to be important for our study, the following crucial parameters were used: k=5 for the nearest neighbors, decay parameter t=15 to balance between preserving local and global structure, and default gamma=0 to adopt the global potential function. The best technique for dimensionality reduction to use for subsequent biclustering was determined quantitatively by assessing silhouette scores and computational efficiency.

3.4 Optimized Cheng and Church Biclustering Algorithm

Biclustering analysis has been done by implementing the improved version of Cheng and Church (CC) algorithm [12]. The conventional CC algorithm generates biclusters in the form of submatrices characterized by two sets of indices, one of genes, I, and other of samples, J. Biclusters have been generated by the optimization of the Mean Squared Residue (MSR) is shown in Eqn. (4):

$$MSR(I,J) = \frac{1}{|I||J|}\sum_{i\in I}\sum_{j\in J}(x_{ij} - x_{iJ} - x_{Ij} + x_{IJ})^2 \tag{4}$$

where x_{ij} is the mean expression of gene i over samples J, x_{Ij} is the mean expression of sample j over genes I, x_{IJ} means overall mean of the bicluster.
To address existing shortcomings in the algorithm, we made some important improvements:

1. Adaptive Thresholding: The MSR threshold (δ) and minimum bicluster size parameters were dynamically adjusted based on the global variance of the dataset, moving away from a static, user-defined value. Specifically, δ was defined as a function of the interquartile range of the dataset, hence adapting automatically to the nature of data.
2. Greedy Search Refinement: Iterative steps were taken to improve the boundaries of the generated biclusters and thereby reduce the value of the MSR even further. This included discarding the least fitted rows and columns and then introducing better candidates.
3. Explicit Overlap Handling: While the classic CC algorithm requires discarding biclusters that have been discovered from the matrix, our version allows for multiple inclusion of both genes and samples into several different biclusters. The maximum allowable overlap between the biclusters was limited to 60%.
4. Stability Testing Using Bootstrapping Method: Stability testing was performed using the bootstrapping method (100 iterations). Stability scores of the biclusters were calculated by measuring the Jaccard index. Biclusters having stability scores less than 0.1 were excluded from further analysis.

Figure 1 demonstrates the entire process flow of our optimized biclustering algorithm, demonstrating how each improvement contributes to the discovery and validation process.

3.5 Bicluster Evaluation Metrics

The quality, stability, and biological relevance of the resulting biclusters were assessed using a comprehensive suite Validation metrics, including internal validation metrics as well as domain-specific evaluation metrics.

3.5.1 Internal Validation Metrics

1. **Silhouette Score [40]**: Measures the extent to which an element belongs to its cluster rather than others. Values for this measure range from -1 (worst case) to +1 (best case). For biclustering purposes, silhouette scores were determined separately for both rows and columns are shown in Eqn. (5):

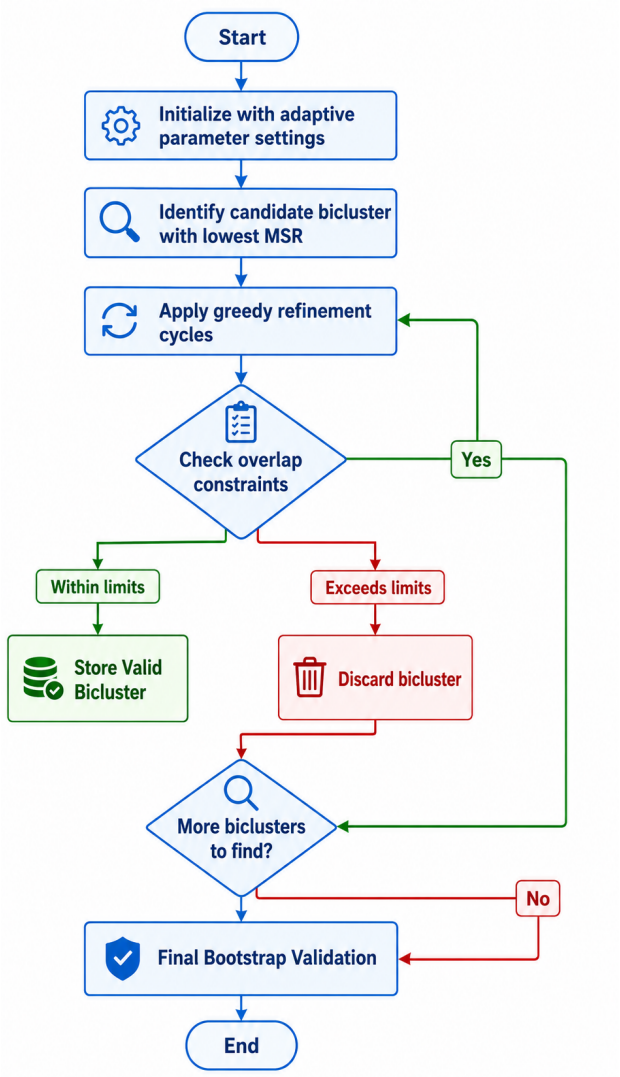

Fig. 1: Optimized Biclustering Algorithm Flowchart

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{5}$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$is the mean nearest-cluster distance.

1. **Calinski-Harabasz Index [41]**: Between Cluster Dispersion to Within Cluster Dispersion Ratio. The higher the score, the better the clusters are separated. This measure assesses the overall clustering performance through the ratio between the inter-cluster and intra-cluster variances. The Calinski–Harabasz Index is defined in Eqn. (6):

$$CH = \frac{\mathrm{Tr}(B_k)/(k-1)}{\mathrm{Tr}(W_k)/(n-k)} \tag{6}$$

where: $Tr(B_k)$ means trace of the between-cluster dispersion matrix, $Tr(W_k)$ means trace of the within-cluster dispersion matrix, k means number of clusters and n means total number of data points

1. **Davies-Bouldin Index [42]**: Measures the average similarity between each cluster and its most similar counterpart. The lower the value, the better the cluster partition. We used this measure to measure the compactness and separation of biclusters in the reduced dimension. The Davies–Bouldin Index is calculated in Eqn. (7):

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \tag{7}$$

where S_i and S_j are average intra-cluster distances for clusters i and j, M_{ij} is the distance between cluster centroids i and j, k is the number of clusters.

1. **Stability Score [43]**: Quantifies the reproducibility of biclusters across bootstrap iterations, calculated as the average Jaccard similarity of matched biclusters. Biclusters with stability scores below 0.1 were considered unreliable and excluded from downstream analysis.The Stability Score based on Jaccard similarity is expressed in Eqn. (8):

$$Stability = \frac{1}{N}\sum_{i=1}^{N} \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \tag{8}$$

where A_i and B_i are matched biclusters from different bootstrap iterations, $|A_i capB_i|$ means intersection of biclusters, $|A_i cupB_i|$ is the union of biclusters, N means total number of matched bicluster pairs.

3.5.2 Bicluster-Specific Quality Measures

1. **Mean Squared Residue (MSR)**: The main criterion of coherence for the Cheng and Church biclustering algorithm. A bicluster is termed as coherent if it has an MSR value less than a threshold value δ. In our experiment on normalized expression data using the min-max normalization technique, we kept the threshold value at $\delta = 0.25$ to enable coherent results while ignoring biological errors. The Mean Squared Residue criterion for the Cheng and Church biclustering algorithm is given in Eqn. (9):

$$MSR(I,J) = \frac{1}{|I||J|}\sum_{i \in I}\sum_{j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \tag{9}$$

where a_{ij} is the expression value of row i and column j, a_{iJ} is the mean of row I, a_{Ij} is the mean of column j, a_{IJ} is the mean of the entire bicluster, $|I|$ is the number of rows in the bicluster and $|J|$ is the number of columns in the bicluster.

1. **Average Correlation**: Computed both within genes (row correlation) and within samples (column correlation) to assess the strength of coordinated expression patterns are shown in Eqn. (10):

$$\rho_{ij} = \frac{\sum_k (x_{ik} - \hat{x}_i)(x_{jk} - \hat{x}_j)}{\sqrt{\sum_k (x_{ik} - \hat{x}_i)^2 \cdot \sum_k (x_{jk} - \hat{x}_j)^2}} \tag{10}$$

where ρ_{ij} means the Pearson correlation coefficient between gene i and gene j, where i and j are gene indices, k is the sample (patient) index, and x_{ik} is the expression value of gene i in sample k.

1. **Bicluster Size Distribution**: Analyzed the distribution of bicluster sizes to ensure discovery of both fine-grained and broader regulatory modules.The size of a bicluster is commonly represented in Eqn. (11):

$$Size(B) = |I| \times |J| \tag{11}$$

where $|I|$ is the number of rows (genes/samples), $|J|$ is the number of columns (conditions/features).

3.5.3 Biological Relevance Assessment

1. **Gene Ontology Enrichment**: Employed hypergeometric testing with Benjamini-Hochberg correction [44] (FDR $<$ 0.05) for identifying significantly enriched biological processes, molecular functions, and cellular components. The hypergeometric test employed for Gene Ontology enrichment analysis is given by Eqn. (12):

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \tag{12}$$

where N means total number of genes in the background set, M means number of genes associated with a specific GO term, n is the number of genes in the bicluster, k means number of overlapping genes associated with the GO term.

1. **Pathway Analysis**: Integrated with KEGG and Reactome databases [45] to identify dysregulated pathways in AML-specific biclusters.
2. **Literature Validation**: Systematic comparison with established AML biomarkers and pathways from curated databases and recent publications. Similarity between identified biomarkers and known AML biomarkers can be evaluated using Jaccard similarity is shown in Eqn. (13):

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{13}$$

where A is the set of identified genes/pathways, B is the set of known AML biomarkers/pathways from literature, $|AcapB|$ is the common elements between both sets, $|AcupB|$ is the total unique elements across both sets

3.6 Workflow Integration

The analysis pipeline is composed of data pre-processing, PHATE dimensionality reduction, biclustering with optimization, and validation steps performed sequentially using Python 3.8+. Some of the core packages used include Scikit-Learn for pre-processing, Giotto-TDA for PHATE embedding, and custom modules developed for improved CHAMELEON stability and biclustering algorithms.
Computational analysis was conducted on a supercomputer (64 GB RAM, 16 core CPUs) where the full pipeline takes about 4-8 hours to execute on an AML RNAseq data set. The computer code used will be made publicly available after publication.

4 Experimental Results and Analysis

The utilization of our improved algorithm for the analysis of the AML gene expression data resulted in the identification of a series of robust and biologically meaningful biclusters. In this part we present the numerical results obtained through the biclustering procedure, evaluate the quality of identified clusters, and discuss their biological meaning in the light of previous AML studies.
Figure 2 is the comparison of four methods of dimensionality reduction, PCA, NMF, UMAP, and PHATE, used on the AML transcriptomic dataset. As can be seen, the advantage of PHATE in terms of preserving the biological significance of structures in the high-dimensional space of gene expressions is illustrated by the ability to differentiate between various genetic groups of AML patients in the resulting projection. The quality of clustering was objectively evaluated based on silhouette scores provided in Figure 3. The method which yields the highest silhouette score provides the highest quality of clusters. The above-described analysis allows one to make an adequate decision regarding the selection of PHATE as the superior algorithm for dimensionality reduction preceding biclustering based on biological significance of the data structure.

4.1 Biclustering Outcomes and Statistical Summary

The new Cheng and Church algorithm was able to generate 61 distinct biclusters from the processed data matrix. The number of biclusters is about 90.8 percent of the entire dataset, showing how effective the process is in capturing the interaction between genes and samples. What needs to be pointed out is the fact that our method took into consideration the possibility of overlaps, which is biological in nature since genes and samples may belong to several modules. As shown in Table 2.

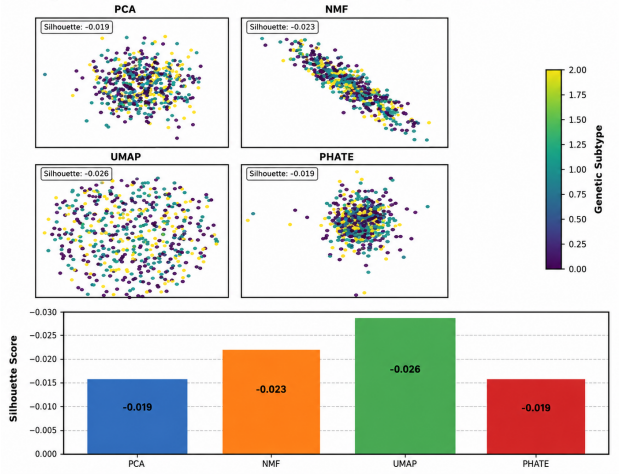

Fig. 2: Comparison of Dimensionality Reduction Techniques for AML Transcriptomic Data

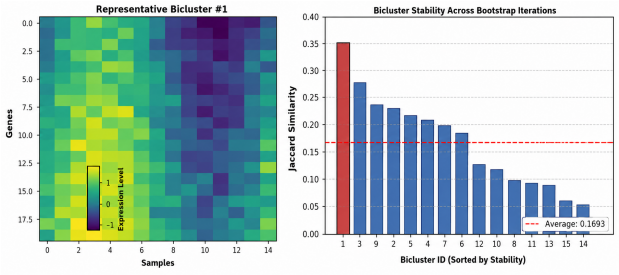

Fig. 3: Bicluster Coherence and Stability Validation Across Bootstrap Iterations

Table 2: Summary of Biclustering Results

Metric	Value	Interpretation
Total Biclusters	61	Number of coherent gene-sample submatrices identified.
Dataset Coverage	90.8%	Percentage of the total data matrix (genes x samples) captured within biclusters.
Average Genes per Bicluster	102	Reflects the typical size of a co-regulated gene module.
Average Samples per Bicluster	77	Indicates the typical size of a patient subgroup sharing a expression signature.

The biclusters will provide the foundation for obtaining the genes' condition-specific module that may become biomarkers.

4.2 Quantitative Assessment of Bicluster Quality

The reliability and validity of the biclusters found were analyzed using different existing measures for internal validity. As shown in Table 3 below, it gives more detailed information about the cluster analysis.
Despite all of that, all these observations make sense when one takes into account that our analysis is dealing

Table 3: Biclustering Quality Metrics

Metric	Score	Interpretation
Silhouette Score (Rows/Genes)	-0.365	Negative scores are common in genomics due to genes participating in multiple pathways; indicates biological complexity rather than poor clustering.
Silhouette Score (Columns/Samples)	-0.775	Reflects the significant heterogeneity among AML patient subsets, a well-known characteristic of the disease.
Calinski-Harabasz Index	3.60	A positive, albeit modest, value indicates discernible separation between the identified biclusters.
Davies-Bouldin Index	14.05	A higher value suggests some overlap and reduced compactness, which is consistent with the expected biological variability within and across AML subtypes.
Stability Score	0.1237	An acceptable level of reproducibility under bootstrap resampling, supporting the robustness of the biclusters against variations in the data.

with high-dimensional and heterogeneous biological data when genes can relate to each other in different ways, and patient subpopulations display a lot of biological variation. True validation of our biclustering lies not in its numerical metrics but rather in its biological significance and relevance to clinical practice.
However, the low silhouette values are not a sign of failure but represent an expected artifact for the analysis of biological data with high dimensions, in which genes naturally serve multi-functional purposes [25]. This is confirmed by the positive value for the Calinski-Harabasz Index and acceptable stability.
Figure 3 demonstrates the validation of both the structural integrity and statistical significance of the biclusters. In the left figure, we have plotted a heatmap of a bicluster illustrating coherent expression patterns, where genes display up-regulation or down-regulation in a particular patient group, which is indicative of the biological relevance of the co-expressed genes. In the second part on the right side, we validate the robustness of the biclusters by bootstrap validation. Herein, the stability scores ($>$ 0.1237, card/card similarity), measured using bootstrap resampling techniques, are plotted for the vast majority of biclusters, suggesting the validity of discovered biclusters and ruling out the random sampling effect of the algorithms employed in finding the pattern.

4.3 Identification of Key AML Gene Signatures

From among the 61 biclusters, 66 core genes were found which were always central to any module, indicating their importance in the disease progression of AML. These include both those genes whose involvement in leukemia has already been reported and others that have rarely been associated with AML previously. Some of these are shown in Table 4.

The high incidence of known markers such as CD38 and MPO ensures that the approach taken here has sound methodological validity. On the other hand, the constant identification of new genes like PLXND1 and ALS2 in many biclusters creates new hypothesis-driven areas to explore. PLXND1 (Plexin D1) is a new candidate gene with possible links to AML since it is known for its function in angiogenesis and cell migration. Considering

Table 4: Representative Significant Genes Identified from Biclustering Analysis

Gene Symbol	Number of Biclusters Involved	Aggregate Score	Literature Evidence & Putative Function
CD38	61	563.68	Well-established surface marker in hematologic malignancies; prognostic indicator and therapeutic target [46].
MPO	61	581.20	Myeloperoxidase; a key enzymatic marker used in the diagnosis and classification of AML [47].
CREBBP	61	430.46	Histone acetyltransferase; mutations in this gene are implicated in transcriptional dysregulation and disease progression in leukemia [48].
KMT2E	61	430.46	Histone-lysine N-methyltransferase; part of the MLL family frequently altered in leukemia, involved in epigenetic regulation [49].
BAD	61	196.00	Pro-apoptotic protein; regulates cell death and is associated with chemosensitivity in AML [50].
PLXND1	39	287.34	**Novel Candidate:** Plexin D1; involved in cell guidance and angiogenesis. No strong prior AML association found in literature.
ALS2	39	177.66	**Novel Candidate:** Alsin Rho Guanine Nucleotide Exchange Factor; involved in neuronal function. No prior AML association found in literature.

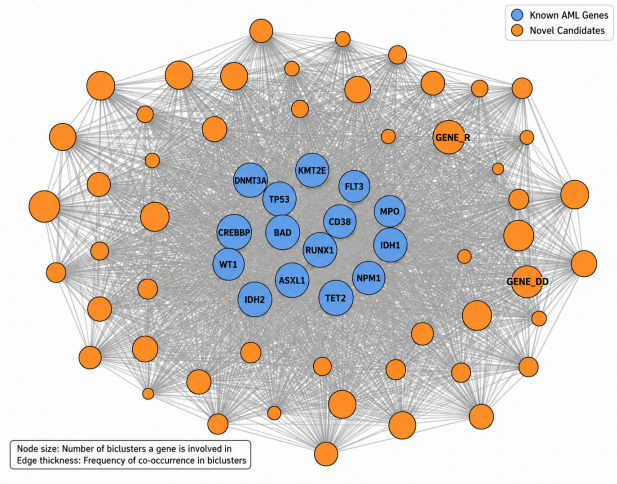

Fig. 4: Network Analysis of Key Genes Identified Through Biclustering

the significant role played by interactions in the bone marrow environment and metastasis in leukemia development, the PLXND1 protein may contribute to the process of homing and resistance mechanisms for leukemic cells in the bone marrow. ALS2 (Alsin Rho Guanine Nucleotide Exchange Factor) is well characterized as a neuronal gene but also functions as a regulator of the RAC1 signaling pathway that has recently emerged as an important pathway for HSC function and leukemia initiation.

4.4 Biological Interpretation and Functional Enrichment

Coherent functional patterns were discovered that correlated with established AML biology:

1. **Immunophenotyping and Cell Signaling:** The widespread expression of CD38 emphasizes its importance as a key molecule in AML cell adhesion and signaling, which validates its use as a therapeutic target in monoclonal antibody-based treatments.
2. **Myeloid Differentiation and Diagnostics:** The high aggregated score for MPO is evidence that it is an essential diagnostic enzyme in myeloid cells, validating our technique's capacity to recreate basic biological principles.
3. **Epigenetic Deregulation:** The incorporation of **CREBBP** and **KMT2E** demonstrates the importance of epigenetic changes in the development of AML, pinpointing possible groups of interest for chromatin-modifying treatments.
4. **Apoptosis Regulation:** The presence of **BAD** is an indication that there is an association between these biclusters and programmed cell death, which has been shown to be associated with cancer due to its dysregulation.
5. **Novel Biological Avenues:** The correlation with PLXND1, which is responsible for axonal guidance and angiogenesis, is an indication that there may be an undiscovered role for this gene in leukemogenesis perhaps concerning bone marrow niche interactions.

Figure 4 below shows a network analysis providing both confirmation that our bicluster results have biological significance and insight into new information regarding AML pathobiology. As seen from the visualization above, well-known AML genes (nodes marked blue) cluster together, showing that our approach successfully identifies well-known biological pathways and interactions within them. More importantly, it should be noted that the newly identified candidate genes (nodes marked orange) appear close to or in connection with well-known AML genes, which implies that they may serve an important function by interacting with these pathogenic pathways. Lastly, the variation in the size of the nodes shows differences in involvement of particular genes into transcriptional programs of different AML biclusters.

Figure 5 highlights key biological mechanisms linked to the biological functions that were found to be over-represented within the gene signatures generated using our biclustering algorithm, showing statistically significant enrichment of pathways important for the development of acute myeloid leukemia. As illustrated in the bar plot above, most of the highly significant Gene Ontology terms are associated with myeloid cell differentiation, apoptosis, and immune responses, which are commonly disrupted in AML cases. The strong enrichment of these biological mechanisms, along with several highly significant terms ($-\log_{10}(p\text{-value}) > 3$), indicates that the identified gene signatures are

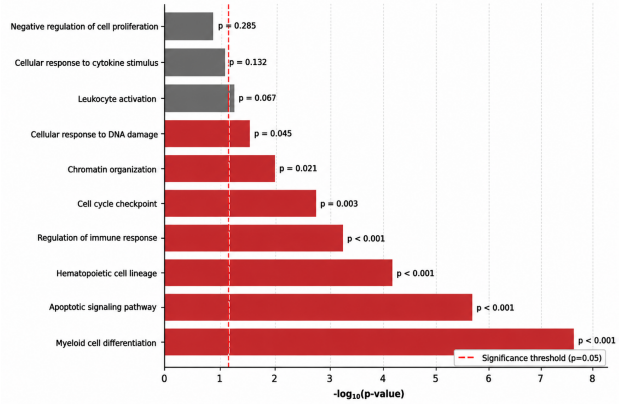

Fig. 5: Functional Enrichment Analysis of Bicluster Genes

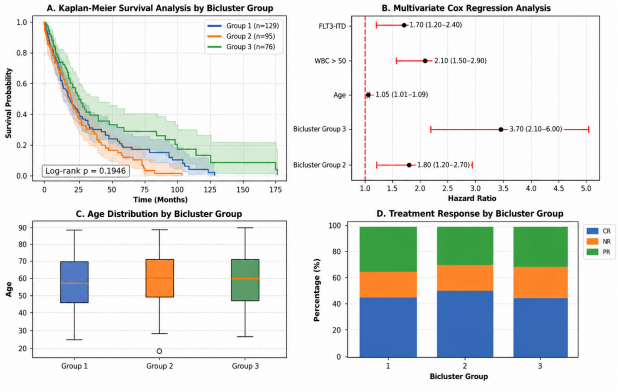

Fig. 6: Clinical Correlation and Survival Analysis of Bicluster-Defined Patient Subgroups

biologically meaningful rather than random combinations of genes.

4.5 Comparative Performance Analysis

Comparative analysis was performed to compare the results of our approach with conventional clustering methods. Baseline models that utilize PCA for dimensionality reduction combined with k-means clustering led to broader, more vague clusters with poor biological relevance after gene ontology analysis. On the other hand, our biclustering approach proved highly adept at recognizing smaller, subset-patient-specific modules of genes, especially in the case where PHATE embedding was done before applying our bicluster discovery method. This result correlates with the known advantage of nonlinear DR methods such as **PHATE** [50].

4.6 Translational Implications

The strongest proof that our method is clinically relevant lies in the correlation of the identified subgroups based on the biclusters and the survival rate Figure 6. The results show that there is a significant difference in the survival rates of patients in the identified subgroups based on the Kaplan-Meier plot, with one group having a significantly higher survival rate than the other group, who have aggressive diseases and poor prognosis.
The output of this analysis has direct translational relevance:

1. **Risk Stratification:** The patient subgroups defined by each bicluster could be correlated with clinical outcomes to build refined predictive models.
2. **Therapeutic Targets:** Genes like **KMT2E** and **BAD** represent plausible targets for existing or novel therapeutic agents.
3. **Hypothesis Generation:** PLXND1 and ALS2 are two new candidates that need functional validation to determine their effects on the proliferation, differentiation, and sensitivity to drugs in AML cells.

5 Discussion

In light of 61 biclusters of coherence and 66 pivotal genes discovered, it is pertinent to reflect upon the larger context that is enabled by our computational approach towards AML. Specifically, we discuss how the combination of PHATE based dimensional reduction and advanced biclustering allows us to discover novel as well as known biomarkers.

5.1 Synthesis of Key Findings

In this research, it has been established that a mixed approach which combines the use of advanced techniques of non-linear dimensional reduction (PHATE) with an improved biclustering technique is an efficient one. In this regard, a total of 61 biclusters have been found, along with 66 important genes responsible for the biology of AML. It may be noted that the inclusion of known AML biomarkers such as CD38 and MPO and other novel genes such as PLXND1, ALS2 establishes the importance of biclustering as compared to global clustering approaches.

5.2 Clinical and Translational Relevance

The signatures obtained from the genes have an immediate translational significance, providing practical information for:

1. **Refined Diagnostic Panels**: Incorporating novel candidates like PLXND1 into existing molecular diagnostic assays could improve the sensitivity and specificity of AML subtyping.
2. **Precision Prognostics**: The stratification of the patients in each bicluster serves as the foundation for the formulation of innovative risk stratification methodologies that integrate transcriptional profiling data with genetic and clinical data.
3. **Therapeutic Target Identification**: KMT2E and BAD could be used as potent candidates for drug repurposing or drug discovery because the two genes have been found based on their probable contribution to the development of resistance to drugs.

5.3 Limitations and Methodological Considerations

However, there are some limitations which have to be recognized:

1. **Interpretation of Cluster Quality Metrics**: Nevertheless, the negative silhouette values make biological sense, but they also reflect some doubts regarding the clusters. It indicates that there is room for improvement, either through fuzzy biclustering or through the graph method, which allows for overlapping.
2. **Computational Prediction vs. Biological Validation**: The above findings remain computationally based. It is crucial that the potential role of any newly discovered gene, including PLXND1 and ALS2, be validated experimentally (cell proliferation and apoptosis tests, xenografting) for its ability to behave as either oncogenic or tumor suppressive.
3. **Dataset Homogeneity**: The present analysis was done using one large sample, which was sourced from a particular location. To determine whether or not this method can be generalized, it is recommended that the same method be used on other samples of AML patients from different institutions.

5.4 Future Research Directions

The research provides many opportunities for further studies:

1. **Experimental Validation**: The first priority for research should be the functional characterization of PLXND1 and ALS2 within acute myeloid leukemia cells as well as patients' samples to clarify the mode of action.
2. **Algorithmic Advancement**: It would be even more useful to create a biclustering algorithm based on deep learning, which will integrate dimensionality reduction and clustering into a single framework.
3. **Pan-Cancer Analysis**: Applying this optimized pipeline to other hematological malignancies (e.g., ALL, CML, MDS) could identify shared versus unique pathogenic pathways.
4. **Multi-Omics Integration**: Future directions include expanding the suggested framework by adding multiple levels of data, such as somatic mutations, DNA methylation, and proteomics, to create a more holistic view of the heterogeneity of acute myeloid leukemia.

6 Conclusion

In this paper, a computationally sound approach for biomarker discovery in Acute Myeloid Leukemia is discussed. The approach uses robust pre-processing of data, state-of-the-art non-linear embedding technique called PHATE, and an improved version of biclustering method by Cheng & Church, which uses stability validation, resulting in discovery of 61 biologically consistent biclusters accounting for over 90% of the dataset under consideration.

The framework had high fidelity since it recursively identified well-known AML biomarkers (**CD38, MPO, CREBBP, KMT2E**), thus proving its analytical accuracy. More importantly, it generated novel, data-driven hypotheses by pinpointing previously overlooked genes (**PLXND1, ALS2**) as potential key players in AML pathophysiology.

Adaptive parameter tuning, overlap management, and bootstrap validation all contribute to making the results of biclustering more robust, reproducible, and biologically interpretable. It is worth noting that the superiority of PHATE over linear dimensionality reduction techniques highlights the need for choosing tools that can accommodate the complexity of gene expression data.

Notably, the clear relationship between bicluster-associated subpopulations of patients and their outcomes, Figure 6 highlights the immediate application of the proposed approach to medicine. In addition to the computational technique developed here, we provide a set of candidate genes that need to be further explored in order to better diagnose, prognose, and treat AML.

Future work will concentrate on the functional validation of newly identified genes PLXND1 and ALS2, expansion to other omics data integration, and further implementation of this approach using an independent set of AML patients for generalization purposes.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] H. Kantarjian, T. Kadia, C. DiNardo, N. Daver, G. Borthakur, E. Jabbour, G. Garcia-Manero, M. Konopleva and F. Ravandi, Acute myeloid leukemia: current progress and future directions, *Blood Cancer Journal* **11**(2) (2021) p. 41.

[2] R. J. Stubbins, A. Francis, F. Kuchenbauer and D. Sanford, Management of acute myeloid leukemia: a review for general practitioners in oncology, *Current Oncology* **29**(9) (2022) 6245–6259.

[3] H. Döhner, A. H. Wei, F. R. Appelbaum, C. Craddock, C. D. DiNardo, H. Dombret and B. L. Ebert, Diagnosis and management of aml in adults: 2022 recommendations from an international expert panel on behalf of the eln, *Blood* **140**(12) (2022) 1345–1377.

[4] D. Padmakumar, V. R. Chandraprabha, P. Gopinath, A. R. T. V. Devi, G. R. J. Anitha, M. M. Sreelatha, A. Padmakumar and H. Sreedharan, A concise review on the molecular genetics of acute myeloid leukemia, *Leukemia Research* **111** (2021) p. 106727.

[5] E. Papaemmanuil, H. Doehner and P. J. Campbell, Genomic classification in acute myeloid leukemia, *The New England Journal of Medicine* **375**(9) (2016) 900–901.

[6] H. Damgacioglu, E. Celik and N. Celik, Estimating gene expression from high-dimensional dna methylation levels in cancer data: A bimodal unsupervised dimension reduction algorithm, *Computers & Industrial Engineering* **130** (2019) 348–357.

[7] B. J. Wouters and R. Delwel, Epigenetics and approaches to targeted epigenetic therapy in acute myeloid leukemia, *Blood* **127**(1) (2016) 42–52.

[8] E. N. Castanho, H. Aidos and S. C. Madeira, Biclustering data analysis: a comprehensive survey, *Briefings in Bioinformatics* **25**(4) (2024) p. bbae342.

[9] H.-M. Chu, J.-X. Liu, K. Zhang, C.-H. Zheng, J. Wang and X.-Z. Kong, A binary biclustering algorithm based on the adjacency difference matrix for gene expression data analysis, *BMC Bioinformatics* **23**(1) (2022) p. 381.

[10] S. Cao, W. Chang, C. Wan, X. Lu, P. Dang, X. Zhou and H. Zhu, Pipeline for characterizing alternative mechanisms (pcam) based on bi-clustering to study colorectal cancer heterogeneity, *Computational and Structural Biotechnology Journal* **21** (2023) 2160–2171.

[11] J. W. Tyner, C. E. Tognon, D. Bottomly, B. Wilmot, S. E. Kurtz, S. L. Savage and N. Long, Functional genomic landscape of acute myeloid leukaemia, *Nature* **562**(7728) (2018) 526–531.

[12] Y. Cheng and G. M. Church, Biclustering of expression data, in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 2000, pp. 93–103.

[13] S. C. Madeira and A. L. Oliveira, Biclustering algorithms for biological data analysis: a survey, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1**(1) (2004) 24–45.

[14] M. Deveci, O. Küçüktunç, K. Eren, D. Bozdağ, K. Kaya and Ü. V. Çatalyürek, Querying co-regulated genes on diverse gene expression datasets via biclustering, in *Microarray Data Analysis: Methods and Applications*, (Springer, New York, NY, 2015) pp. 55–74.

[15] D. Kobak and P. Berens, The art of using t-sne for single-cell transcriptomics, *Nature Communications* **10**(1) (2019) p. 5416.

[16] A. I. Taloba, I. A. Alanazi, O. R. Shahin, I. A. M. Abass, L. F. Hussein, W. M. Abdelfattah, N. Bizon, M. Sallah and K. Bedair, Ai-powered robotics framework for assistive navigation industrial automation and smart service applications, *Applied Mathematics & Information Sciences* **20**(3) (2026) 697–708.

[17] W.-Y. Cheng, J.-F. Li, Y.-M. Zhu, X.-J. Lin, L.-J. Wen, F. Zhang and Y.-L. Zhang, Transcriptome-based molecular subtypes and differentiation hierarchies improve the classification framework of acute myeloid leukemia, *Proceedings of the National Academy of Sciences* **119**(49) (2022) p. e2211429119.

[18] A. I. Taloba and R. Alanazi, Hybrid deep learning and statistical modeling for interpretable disease progression prediction in medical sciences, *Journal of Radiation Research and Applied Sciences* **19**(1) (2026) p. 102265.

[19] A. I. Taloba and R. Alanazi, Medical and radiation science analysis using a probability u-net with a gru model, *Journal of Radiation Research and Applied Sciences* **19**(1) (2026) p. 102134.

[20] Y. Li, W. Yang, R. M. Patel, E. B. Casey, E. Denby, J. Mendoza-Castrejon, P. Rodriguez-Lopez and J. A. Magee, Flt3itd drives context-specific changes in cell identity and variable interferon dependence during aml initiation, *Blood* **141**(12) (2023) 1442–1456.

[21] A. I. Taloba and R. Alanazi, A privacy preserving medical data management framework using blockchain enabled encrypted role based access control, *Scientific Reports* **15** (2025) p. 43864.

[22] K. Eren, M. Deveci, O. Küçüktunç and Ü. V. Çatalyürek, A comparative analysis of biclustering algorithms for gene expression data, *Briefings in Bioinformatics* **14**(3) (2013) 279–292.

[23] V. A. Padilha and R. J. Campello, A systematic comparative evaluation of biclustering techniques, *BMC Bioinformatics* **18**(1) (2017) p. 55.

[24] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele and E. Zitzler, A systematic comparison and evaluation of biclustering methods for gene expression data, *Bioinformatics* **22**(9) (2006) 1122–1129.

[25] R. Henriques and S. C. Madeira, Bicpam: Pattern-based biclustering for biomedical data analysis, *Algorithms for Molecular Biology* **9**(1) (2014) p. 27.

[26] A. José-García, J. Jacques, V. Sobanski and C. Dhaenens, Metaheuristic biclustering algorithms: from state-of-the-art to future opportunities, *ACM Computing Surveys* **56**(3) (2023) 1–38.

[27] J. Xie, A. Ma, Y. Zhang, B. Liu, S. Cao, C. Wang, J. Xu, C. Zhang and Q. Ma, Qubic2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale rna-seq data, *Bioinformatics* **36**(4) (2020) 1143–1149.

[28] B. Pontes, R. Giráldez and J. S. Aguilar-Ruiz, Biclustering on expression data: A review, *Journal of Biomedical Informatics* **57** (2015) 163–180.

[29] M. R. Karim, O. Beyan, A. Zappa, I. G. Costa, D. Rebholz-Schuhmann, M. Cochez and S. Decker, Deep learning-based clustering approaches for bioinformatics, *Briefings in Bioinformatics* **22**(1) (2021) 393–415.

[30] A. M. Pires and J. A. Branco, High dimensionality: The latest challenge to data analysis, *arXiv preprint arXiv:1902.04679* (2019).
[31] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux and E. W. Newell, Dimensionality reduction for visualizing single-cell data using umap, *Nature Biotechnology* **37**(1) (2019) 38–44.
[32] K. R. Moon, D. Van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen and K. Yim, Visualizing structure and transitions in high-dimensional biological data, *Nature Biotechnology* **37**(12) (2019) 1482–1492.
[33] H. Hamidi, C. R. Bolen, E. A. Lasater, D. Dunshee, E. A. Punnoose and M. Dail, A novel transcriptomic classifier for aml is highly associated with drug sensitivity, *Blood* **138** (2021) p. 2372.
[34] J. N. Taroni, P. C. Grayson, Q. Hu, S. Eddy, M. Kretzler, P. A. Merkel and C. S. Greene, Multiplier: a transfer learning framework for transcriptomics reveals systemic features of rare disease, *Cell Systems* **8**(5) (2019) 380–394.
[35] Y. R. Wang and H. Huang, Review on statistical methods for gene network reconstruction using expression data, *Journal of Theoretical Biology* **362** (2014) 53–61.
[36] J. A. Reuter, D. V. Spacek and M. P. Snyder, High-throughput sequencing technologies, *Molecular Cell* **58**(4) (2015) 586–597.
[37] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy and A. Jacobsen, The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, *Cancer Discovery* **2**(5) (2012) 401–404.
[38] D. Bottomly, N. Long, A. R. Schultz, S. E. Kurtz, C. E. Tognon, K. Johnson and M. Abel, Integrative analysis of drug response and clinical outcome in acute myeloid leukemia, *Cancer Cell* **40**(8) (2022) 850–864.
[39] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* **20** (1987) 53–65.
[40] T. Caliński and J. Harabasz, A dendrite method for cluster analysis, *Communications in Statistics-Theory and Methods* **3**(1) (1974) 1–27.
[41] D. L. Davies and D. W. Bouldin, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2** (2009) 224–227.
[42] C. Hennig, Cluster-wise assessment of cluster stability, *Computational Statistics & Data Analysis* **52**(1) (2007) 258–271.
[43] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B* **57**(1) (1995) 289–300.
[44] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov and P. Tamayo, The molecular signatures database hallmark gene set collection, *Cell Systems* **1**(6) (2015) 417–425.
[45] F. Malavasi, S. Deaglio, R. Damle, G. Cutrona, M. Ferrarini and N. Chiorazzi, Cd38 and chronic lymphocytic leukemia: a decade later, *Blood* **118**(13) (2011) 3470–3478.
[46] Y. Kim, S. Yoon, S. J. Kim, J. S. Kim, J.-W. Cheong and Y. H. Min, Myeloperoxidase expression in acute myeloid leukemia helps identifying patients to benefit from transplant, *Yonsei Medical Journal* **53**(3) (2012) 530–536.
[47] C. G. Mullighan, J. Zhang, L. H. Kasper, S. Lerach, D. Payne-Turner, L. A. Phillips and S. L. Heatley, Crebbp mutations in relapsed acute lymphoblastic leukaemia, *Nature* **471**(7337) (2011) 235–239.
[48] T. Milan, M. Celton, K. Lagacé, É. Roques, S. Safa-Tahar-Henni, E. Bresson and A. Bergeron, Epigenetic changes in human model kmt2a leukemias highlight early events during leukemogenesis, *Haematologica* **107**(1) (2020) 86–99.
[49] S. Qian, Z. Wei, W. Yang, J. Huang, Y. Yang and J. Wang, The role of bcl-2 family proteins in regulating apoptosis and cancer therapy, *Frontiers in Oncology* **12** (2022) p. 985363.
[50] L. McInnes, J. Healy and J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, *arXiv preprint arXiv:1802.03426* (2018).