Sharif University of Technology
september 2025
Department of physics

Feature Selection vs. Shapley Values: Concordance in Identifying Marker Genes in Gene Networks

Amirhossein Movahedi, Setia Bikdeli, Narges Khorshidi

Abstract

Gene marker identification is a fundamental challenge in single-cell data, where thousands of genes need to be examined to distinguish specific patterns of each cell type. Classical feature selection methods provide effective statistical tools for dimensionality reduction and identification of distinct genes, but they often fail to consider the context of biological networks. In this study, we present a framework based on game theory and Shapley values to quantify the contribution of each gene to the differentiation of cell clusters in PBMC3K data. By comparing Shapley scoring and classical feature selection, the degree of convergence between statistical significance and network role in discovering marker genes was investigated. The proposed pipeline includes preprocessing of single-cell data, cluster identification with Louvain algorithm, and construction of presence/absence matrix to calculate Δ-Shapley specific for each cluster. These values are then integrated into protein-protein interaction networks (STRING) and, using the neighborhood centrality criterion, genes that are both cluster-specific and biologically central to the network are identified. The results obtained are consistent with known marker genes such as IL7R, CD14, and NKG7, and also introduce new candidates with strong network effects. Pathway enrichment analysis also confirms the involvement of these genes in immune processes. Our findings show that combining Shapley values with network centrality provides a powerful strategy for identifying marker genes and builds a bridge between statistical feature selection and functional interpretation in gene networks.

1 Introduction

Usually, methods such as feature selection are used to analyze genetic data and find gene markers. In network science and the study of agent-based networks in game theory, Shapley analysis and various approaches can be used. In the article (Gametheoreticcentrality:anovel approachtoprioritizediseasecandidate genesbycombiningbiologicalnetworks withtheShapleyvalue 2020), the author claimed that these approaches can be used to study gene networks, especially in autism, and identify genes of high importance.

This motivated us to use these methods to compare the results of feature selection on blood data (PBMC-3k). And we did this: First, we read the “PBMC-3k” data; we discarded poor-quality cells and low-numbered genes, and normalized the reading value of each cell so that the cells could be compared with each other. Then, we performed PCA on the genes that had the most changes between cells and divided them into eight clusters (approximately different cell types) based on the neighborhood of the cells with the Louvain algorithm. We kept the label of each cluster in the cluster-label column for later use.

Next, we created a binary “present/absent” matrix for each gene; it is enough to see if the raw count of that gene in each cell is zero or more. We applied this matrix twice: once only on the cells of the same cluster (case group) and once on the rest of the cells (control group). Using the Shapley game formula, we calculated the information contribution of each gene in both groups and calculated their difference;

we call this difference Δ-Shapley. If Δ is large, it means that the gene is almost specific to that cluster, for example, IL7R is specific to CD4 T cells (Figure 1).

Now, to complete the network look, we sent the list of all genes in a cluster to the STRING database and obtained their protein-protein interaction network (Figure 2). For each node or vertex of this network, we assign the Δ-Shapley weight of the same gene and create a "weighted neighborhood centrality" score: we add up the weights of itself and its neighbors and apply them to the degree of each node. This score indicates that a gene is not only specific to the cluster, but also plays an important role in the biological network of that cluster.

Finally, for each cluster, we select the five genes with the highest centrality score as the final marker. The final result is a table that lists the cluster name (e.g., "T CD4" or "B") and its five marker genes (e.g., IL7R, CCR7, CD14, MS4A1, etc.) for each cluster (last figure...). This table can be used for biological interpretation, comparison with other papers, or design of subsequent experiments; because the selected genes, in addition to being specific, play a key role in the protein interaction network.

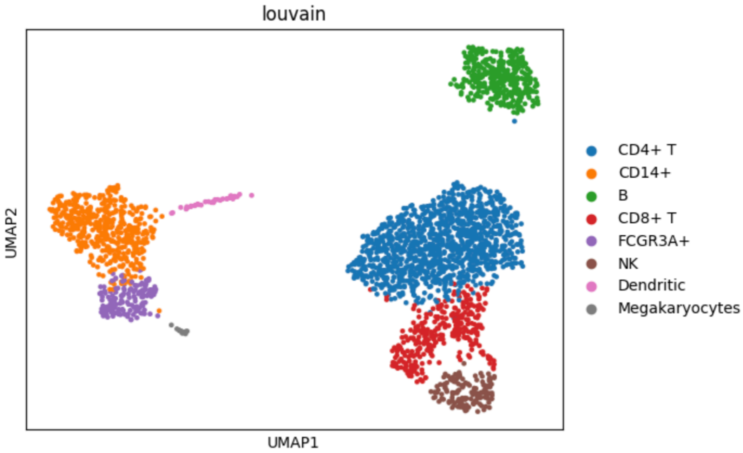


Figure 1: 8 clusters with Louvain clusters.

2 Data preprocessing

In this part of the project, we first load the blood data of "PBMC-3k" cells. Our goal is to clean and prepare the data for more detailed analysis. In the first step, we remove genes that are seen in fewer than three cells, because we have little information about them. Then we calculate metrics for cell quality, such as the number of expressed genes and the percentage of mitochondrial genes. Since cells with a high percentage of mitochondrial genes are usually of low quality, we discard those with more than 5 percent mitochondrial expression. We also filter out cells with very few or very many genes—which are likely dead or multicellular cells.

Next, we normalize the total read value of each cell to make the cells comparable, and then take the logarithm to make the data distribution more balanced. From all the genes, we select those with the highest variance between cells, as these genes are likely to play a greater role in differentiating cell types. We then perform dimensionality reduction using PCA and divide the cells into eight clusters based on their neighborhoods using the Louvain algorithm, representing different types of immune cells. Finally, we store the label of each cluster in a column called "cluster-label" (Figure 1) to use in the next steps of the analysis.

3 Methods

Coalitional Game Theory and Shapley Value

We model gene–set scoring as a cooperative (coalitional) game on a finite player set N (genes) with characteristic function $v : 2^N \longrightarrow \mathbb{R}$. The Shapley value assigns to each player $i \in N$ its average marginal contribution across all permutations:

$$\phi_i(v) = \sum_{T \subseteq N : i \notin T} \frac{|T|! \left(|N| - |T| - 1\right)!}{|N|!} \left(v(T \cup \{i\}) - v(T)\right). \tag{1}$$

Interpretation. $\phi_i(v)$ is the expected gain added by gene i when joining a random coalition; it is fair (symmetric, efficient) and widely used for attribution.

Microarray Game on a Binary Presence Matrix

Let $B \in \{0,1\}^{|N| \times |S|}$ encode presence/absence of a feature (e.g., abnormal expression or loss-of-function mutation) for gene $g_i \in N$ in sample $s_j \in S$; $B_{ij} = 1$ means the feature is present. For a coalition $T \subseteq N$, define the unanimity game for a subset $M \subseteq N$ as

$$u_M(W) = \begin{cases} 1, & \text{if } M \subseteq W, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The *microarray game* [**?**] aggregates, over samples, whether all genes in T co-occur:

$$v^*(T) = \frac{1}{|S|} \sum_{j \in S} u_{M_j}(T), \qquad M_j = \{\, i \in N : \ B_{ij} = 1 \,\}. \tag{3}$$

For this game, the Shapley value admits a closed form:

$$\phi_i(v^*) = \sum_{j \in S} \frac{R_{ij}}{|S|}, \qquad R_{ij} = \begin{cases} 0, & B_{ij} = 0, \\ \dfrac{1}{|M_j|}, & B_{ij} = 1, \end{cases} \tag{4}$$

i.e., gene i receives $\frac{1}{|M_j|}$ credit from sample j iff present there. *Computationally*, (4) is $O(|N||S|)$, avoiding the exponential cost of (1).

Game-Theoretic Neighborhood-Based Relevance on a Gene Graph

Let $G = (N, E)$ be a gene network (e.g., PPI/GRN). For prior node weights $k \in \mathbb{R}^N_{\geq 0}$ and neighborhood of a set $N_T(E) = \{\, j \in N : \ \exists\, g \in T \text{ with } (j, g) \in E \,\}$, define the *neighborhood game* by

$$v_E^k(T) = \sum_{j \in T \cup N_T(E)} k_j. \tag{5}$$

This class yields a polynomial-time Shapley value [**?**]:

$$\phi_i\left(v_E^k\right) = \sum_{j \in N_i(E) \cup \{i\}} \frac{k_j}{\deg_j(E) + 1}, \tag{6}$$

where $N_i(E)$ is the (open) neighborhood of i and $\deg_j(E) = |N_j(E)|$. *Interpretation.* A neighbor j contributes more when it is locally "rare'' (small degree), emphasizing connectors to low-degree modules.

Combined Game: Injecting Expression-Based Evidence into Topology

To couple expression-based specificity with topology, we use the microarray game to derive per-gene evidences $\big(\phi_1(v^*), \ldots, \phi_{|N|}(v^*)\big)$ and plug them as priors into the network game by setting

$$\phi_i(v^*) \to k_i \tag{7}$$

Applying (6) with these k yields the *game-theoretic centrality* that reflects both within-sample co-occurrence and network context.

Cluster-Specific Scoring and Δ-Shapley

For a target cell cluster C and its complement $\overline{C}$, we compute Shapley vectors on present/absent matrices restricted to each group, $\phi^{(C)}(v^*)$, $\phi^{(\overline{C})}(v^*)$, and define the specificity contrast

$$\Delta\text{-Shapley}_i \;=\; \phi_i^{(C)}(v^*) - \phi_i^{(\overline{C})}(v^*). \tag{8}$$

Large positive (8) indicates gene i is specific to C. We then form network-aware scores via (7)–(6) inside C to prioritize markers that are both specific and topologically central.

4 Discussion

In this section, we use a concept called Shapley in game theory to examine the role of each gene in cluster differentiation. First, we construct a binary matrix in which for each gene in each cell we simply check whether it is expressed or not; that is, if its value is greater than zero, it is considered “present” and otherwise “absent.” We construct this matrix separately for the “case” (cells in a specific cluster) and “control” (the rest of the cells).
Then, using the “microarray-based Shapley” formula, we calculate for each gene how much it is present among the cells and what its information contribution is. This calculation is similar to the case where genes are considered participants in a game and we are looking for each gene’s fair share of the overall expression.
Next, we calculate the Shapley difference between the case and control groups—which we call the Δ-Shapley. The higher this value, the more specifically the gene is expressed in that cluster. For example, the IL7R gene, which has a high Δ-Shapley in the CD4 T cluster, can be considered a marker of this cell type. Finally, we enter the genes with the highest Shapley into the STRING database to measure their inter-protein relationships and draw a gene network of them. Larger nodes in this graph are genes that have contributed more to the clustering of cells.

In this figure(Figure2), we see a network of the top 10 genes in terms of Shapley weight. These genes contributed the most information to the distinction between clusters. We sent them to the STRING database to see if there were any interactions or functional relationships between their protein products. Nodes represent genes, and the larger they are, the higher the Shapley weight of that gene—meaning that the gene plays a stronger role in cell type recognition. The edges you see between nodes represent known or predicted interactions in STRING. For example, in this graph, genes like LTB, CD48, and PTPRCAP are connected to each other and likely function in a common biological pathway. In contrast, genes like RBM3 or SRSF5 have no connection with each other and appear isolated.

This graph helps us determine whether genes that were statistically significant are actually significant in biological networks. In the next steps, we will look at other versions of this image with different settings to better identify hidden biological patterns.

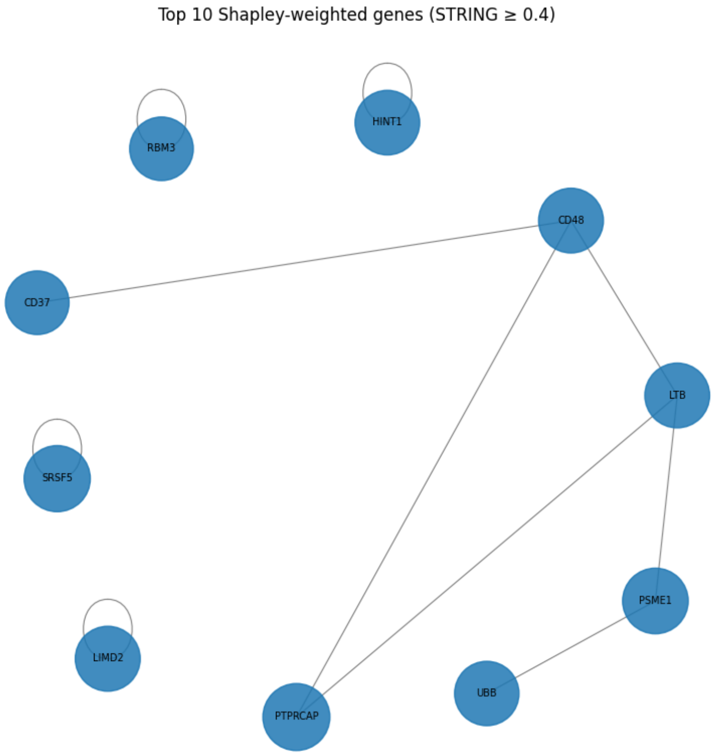


Figure 2: Top Ten Gene Network.

Then, for a more detailed analysis, instead of the top 10 genes, we selected the top 15 percent of genes in terms of Shapley—that is, about 30 to 40 genes. We then used the STRING database with a confidence threshold of 0.3 to find functional relationships between these genes.

In Figure3 (Figure 3), we have displayed the graph as a regular circular grid. The nodes are genes, their size indicates their Shapley weight, and their color indicates the level of connectivity; genes with more than two connections are red, and the rest are blue. This simple view helps us understand which genes are at the center of the network interactions—for example, the genes HSP90AA1 or CD48, LTB.

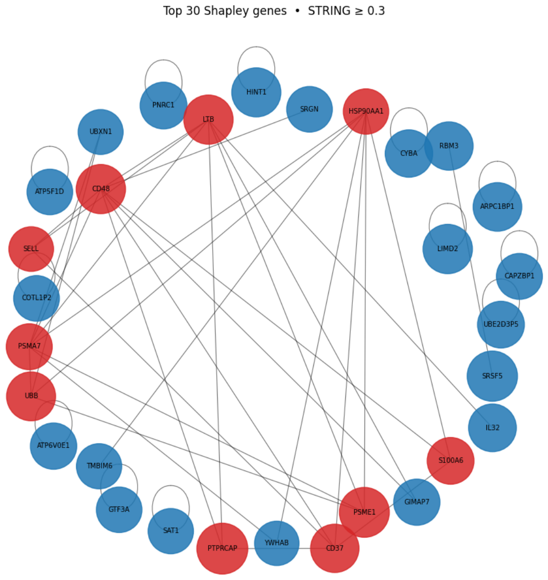


Figure 3: We are 30% sure that the two proteins are connected..

But in the next image(Figure 4), we have used the spring-layout layout, which brings connected nodes closer together and pushes unrelated nodes outward. This representation is more natural and better shows the internal structure of the network. For example, a central cluster is formed by genes such as SAT1, PNRC1, and UBE2D3P5, which have a lot of interactions. In contrast, genes such as RBM3 or SRSF5, which have less interactions, are placed on the periphery.

The goal of these two representations is to see whether genes that are statistically significant are also functionally related. In the next steps, we can compare these networks for specific clusters, either by changing the Shapley or STRING threshold, and see how the patterns change.

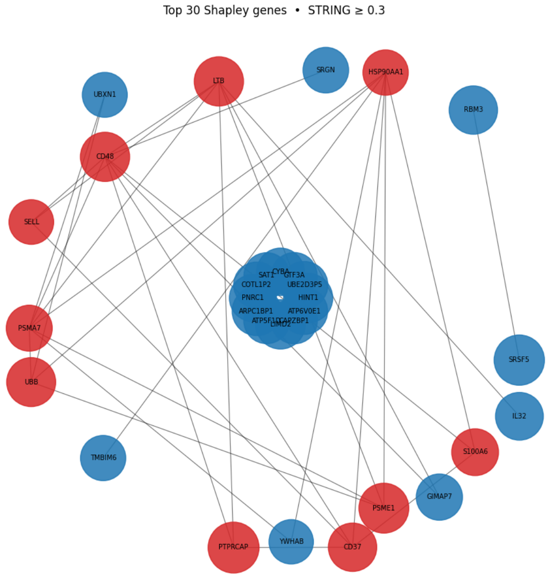


Figure 4: We are 30% sure that the two proteins are connected..

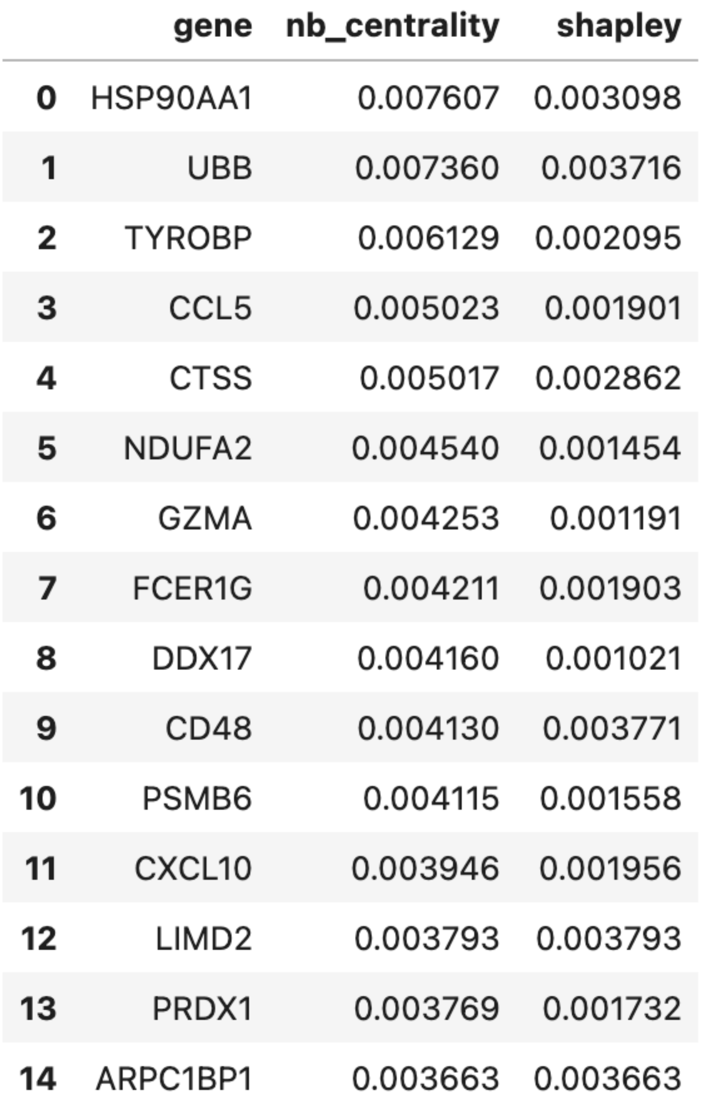
neighbourhood centrality

Next, instead of focusing solely on the Shapley value of each gene, we used a network measure called neighbourhood centrality. The idea is that genes that have a high Shapley contribution themselves or their close neighbors are likely to play a more important role in the biological structure of the network.

To calculate this measure for each gene (node), we divided the sum of its own Shapley weight and its neighbors by the number of edges of each node (i.e., degree+1) and normalized it. This gives a centrality score for each gene that takes into account both statistical value and network position.

The ranking results show that genes such as TYROBP and HSP90AA1, UBB, CD48 are at the top. Interestingly, some genes such as ARPC1BP1 or LIMD2 may not have a very high Shapley weight, but they have a high score due to their presence in a central area of the network(Figure 5).

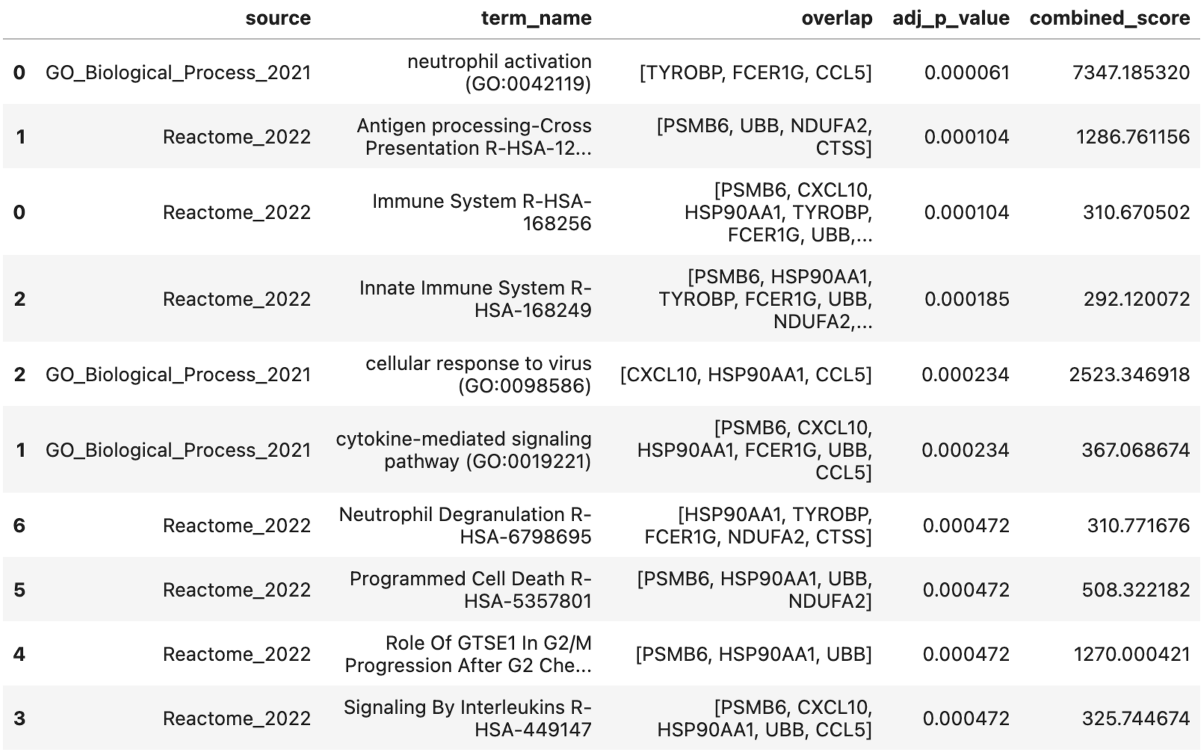
This combination of Shapley and network structure helps us identify important genes not only based on expression, but also in terms of their functional position in the network of interactions.



	gene	nb_centrality	shapley
0	HSP90AA1	0.007607	0.003098
1	UBB	0.007360	0.003716
2	TYROBP	0.006129	0.002095
3	CCL5	0.005023	0.001901
4	CTSS	0.005017	0.002862
5	NDUFA2	0.004540	0.001454
6	GZMA	0.004253	0.001191
7	FCER1G	0.004211	0.001903
8	DDX17	0.004160	0.001021
9	CD48	0.004130	0.003771
10	PSMB6	0.004115	0.001558
11	CXCL10	0.003946	0.001956
12	LIMD2	0.003793	0.003793
13	PRDX1	0.003769	0.001732
14	ARPC1BP1	0.003663	0.003663

Figure 5: Top 15 by neighbourhood score.

For analysis, we selected the top 5% of genes based on network centrality from all genes. Our goal was to see in which biological pathways or cellular functions these central genes are involved.
For this, we used the Enrichr database and searched two important databases:
Reactome for biological pathways and GO Biological Process for biological functions.
The enrichment results showed that these genes are significantly involved in immune processes. For example, the ”neutrophil activation” pathway from the GO database was found to have a very low p-value (0.00061) and a very high combined score. Also, important pathways from Reactome such as:
-Antigen processing – cross-presentation
-Innate immune system
-Cytokine-mediated signaling
all overlap with our selected genes. This means that genes that were at the center of the network according to Graph and Shapley are actually active in key immune pathways.
This agreement between statistical, network, and biological analyses demonstrates the validity of the findings and the purposefulness of the selections—especially genes like UBB and HSP90AA1, PSMB6, and CD48, which were shared across multiple key pathways.Figure 6



	source	term_name	overlap	adj_p_value	combined_score
0	GO_Biological_Process_2021	neutrophil activation (GO:0042119)	[TYROBP, FCER1G, CCL5]	0.000061	7347.185320
1	Reactome_2022	Antigen processing-Cross Presentation R-HSA-12...	[PSMB6, UBB, NDUFA2, CTSS]	0.000104	1286.761156
0	Reactome_2022	Immune System R-HSA-168256	[PSMB6, CXCL10, HSP90AA1, TYROBP, FCER1G, UBB,...	0.000104	310.670502
2	Reactome_2022	Innate Immune System R-HSA-168249	[PSMB6, HSP90AA1, TYROBP, FCER1G, UBB, NDUFA2,...	0.000185	292.120072
2	GO_Biological_Process_2021	cellular response to virus (GO:0098586)	[CXCL10, HSP90AA1, CCL5]	0.000234	2523.346918
1	GO_Biological_Process_2021	cytokine-mediated signaling pathway (GO:0019221)	[PSMB6, CXCL10, HSP90AA1, FCER1G, UBB, CCL5]	0.000234	367.068674
6	Reactome_2022	Neutrophil Degranulation R-HSA-6798695	[HSP90AA1, TYROBP, FCER1G, NDUFA2, CTSS]	0.000472	310.771676
5	Reactome_2022	Programmed Cell Death R-HSA-5357801	[PSMB6, HSP90AA1, UBB, NDUFA2]	0.000472	508.322182
4	Reactome_2022	Role Of GTSE1 In G2/M Progression After G2 Che...	[PSMB6, HSP90AA1, UBB]	0.000472	1270.000421
3	Reactome_2022	Signaling By Interleukins R-HSA-449147	[PSMB6, CXCL10, HSP90AA1, UBB, CCL5]	0.000472	325.744674

Figure 6: Biological significance of the top five percent

Genes enriched in the ”neutrophil activation” pathway:
-TYROBP
-FCER1G
-CCL5
This means that these three genes were among the top 5% of genes in terms of centrality, and are significantly involved in this important biological pathway.

5 Result

In this section, instead of analyzing the entire data, we performed centrality analysis separately for each cluster. That is, for each cell type (such as CD4+ T, B-cell, NK, etc.), we examined which genes play a more key role within the interaction network of the same cluster.
To do this, we first created a binary matrix for each cluster that shows the expression of genes in those cells. Then, using Shapley weights, we calculated the information weight of each gene in the same cluster. These weights were entered into the STRING database to build the interaction network, and finally, using the centrality formula, we calculated a network score for each gene.
The final output is a list of the top 5 genes for each cluster.

-This list can be a basis for finding specific markers for each cell type or used to select biological targets in further research.

cluster	**top1**	**top2**	**top3**	**top4**	**top5**
B	S100A8	S100A9	FCN1	AIF1	PTPRCAP
CD14+	CD7	CD3G	CD3D	CD3E	NOSIP
CD4+ T	CD2	AQP3	CD3D	CD3G	CD3E
CD8+ T	CD79B	HLA-DMA	HLA-B	HLA-F	HLA-E
Dendritic	GZMB	PRF1	SRGN	VIM	CD7
FCGR3A+	GZMA	GZMK	LYAR	NXT1	CST7
Megakaryocytes	PTPRCAP	EEF1A1	H3F3B	PFN1	CD3G
NK	PTPRCAP	CD68	TIMP1	CFD	RHOC

Table 1: Top-5 genes per cluster (white background).

analogy

with the PBMC3k dataset from Scanpy First we normalize the raw counts with CPM, which stands for "counts per million." CPM rescales each cell so that the total counts in that cell sum to one million; this corrects for different sequencing depths between cells. After that, we binarize the data at CPM ≥ 1, meaning we mark a gene as present (1) in a cell if its CPM is at least 1, and absent (0) otherwise. We then load a STRING edge-list from a local TSV file. STRING is a well-known database of protein–protein interactions; an "edge-list" is just a two-column table of interacting gene pairs; and TSV means a "tab-separated values" text file. We standardize gene names, then keep only genes that (1) appear in the STRING network, (2) are expressed in at least 10% of cells, and (3) are not common housekeeping genes. On this reduced set we build the interaction graph and implement the neighborhood step exactly as in the article (a gene's score equals its own weight divided by degree+1, plus the same term from each neighbor).
Cells are clustered with a standard Scanpy pipeline (normalize, log-transform, pick highly variable genes, PCA, nearest neighbors, Leiden clustering). We then replace cluster indices with the fixed labels, CD4+ T, CD14+, B, CD8+ T, FCGR3A+, NK, Dendritic, and Megakaryocytes, so all outputs use real names.

For each cluster, we split the binarized matrix into "cluster" versus "rest." We compute the microarray Shapley weights for each split, pass them through the neighborhood scoring on the STRING graph, and take the absolute difference as a cluster-specific score. Sorting that score gives the top markers per cluster, which we save in both long (cluster, gene, rank) and wide (cluster with top1–top5) formats. Because we work on genes that are both expressed and present in the interaction graph, the results emphasize genes that are specific to each cluster and supported by their network context.Table 1

Our marker lists largely match standard PBMC biology, but several rows are labeled with the wrong cell type because the fixed label order was attached to arbitrary Leiden indices. For example, the row you labeled “B” has S100A8/S100A9/FCN1/AIF1, classic CD14^{+} inflammatory monocyte genes, while the row labeled “CD8+ T” contains CD79B and HLA class genes, which is clearly the B-cell cluster. Likewise, the rows labeled “Dendritic” and “FCGR3A+” are dominated by cytotoxic markers (GZMB/PRF1 and GZMA/GZMK/CST7), pointing to NK/CD8 cytotoxic T subsets. In other words, the gene sets we obtained are biologically sensible and close to what the other article shows.

there are method differences that make our lists diverge from the other paper’s pictureTable 2. The other figure is typically driven by differential expression or curated “canonical” markers, while the article of our focus ranks genes with a network-aware score: each gene’s presence weight (the microarray Shapley value) is propagated over the STRING interaction graph and down-weighted by node degree. That design favors genes that are both specific to a cluster and embedded in supportive neighborhoods, and it can nudge ranks toward neighbors of canonical markers or temper very high-degree hubs. It’s reasonable, for instance, to see HLA genes riding with B-cell markers or to see CD79A/B outrank MS4A1 if MS4A1’s degree penalty is stronger in the subgraph you used. These shifts don’t contradict the biology; they reflect the article’s goal of emphasizing context-supported markers rather than pure abundance.

Cell Type	**Marker Genes**
Naïve CD4+ T	IL7R, CCR7
CD14+ Mono	CD14, LYZ
Memory CD4+ T	IL7R, S100A4
B	MS4A1
CD8+ T	CD8A
FCGR3A+ Mono	FCGR3A, MS4A7
NK	GNLY, NKG7

Table 2: Known markers for PBMCs.

Preprocessing choices also push results away from the other article in predictable ways. We (i) binarized expression at CPM ≥ 1 so the method sees “present/absent” rather than graded expression, (ii) restricted the universe to genes that are in STRING and expressed in ≥10% of cells, and (iii) filtered obvious housekeeping genes. Those filters reduce noise and prevent zero-degree genes from dominating, but they also remove some low-coverage or weakly connected canonical markers the other paper may still list. Given that constraint, it makes sense that your top-5 sets are slightly different while still clearly identifying the same biological compartments (CD14^{+} monocytes, B cells, NK/CD8 cytotoxic, T-cell subsets).

Putting it together: once the clusters are renamed by what their markers say (rather than by index), our implementation results line up with the other paper at the level of cell types, and the remaining rank differences are exactly what we’d expect from a network-weighted method operating on a STRING-supported gene universe.

Our consensus labels now reflect what the top genes actually indicate for each cluster (CD14^{+} with S100A8/S100A9/FCN1; B cells with CD79B and HLA; CD8/NK with GZMA/GZMK or GZMB/PRF1). That matches standard PBMC biology and the other paper at the cell-type level.

cluster_id	cluster_original	cluster_consensus	top1	top2	top3	top4	top5
0	CD4+ T	CD4+ T	CD2	AQP3	CD3D	CD3G	CD3E
1	CD14+	CD4+ T	CD7	CD3G	CD3D	CD3E	NOSIP
2	B	CD14+	S100A8	S100A9	FCN1	AIF1	PTPRCAP
3	CD8+ T	B	CD79B	HLA-DMA	HLA-B	HLA-F	HLA-E
4	FCGR3A+	CD8+ T	GZMA	GZMK	LYAR	NXT1	CST7
5	NK	FCGR3A+	PTPRCAP	CD68	TIMP1	CFD	RHOC
6	Dendritic	NK	GZMB	PRF1	SRGN	VIM	CD7
7	Megakaryocytes	Dendritic	PTPRCAP	EEF1A1	H3F3B	PFN1	CD3G

Table 3: Cluster relabeling by consensus and top-5 marker genes per cluster (white background).

We also created the following figures: per-cluster bar charts of the top differences, a heatmap of the union of top-5 genes, and a quick check of the graph's degree distribution, which you can see below.Figure7
Each panel in the Figure 7 lists the five genes with the highest Δ score for that cluster (cluster vs. rest), so these are our most cluster-specific, network-supported markers.

So in summary, PBMC3k data were CPM normalized and binarized (CPM≥1) and only genes present in STRING and ≥10% of cells were kept; hubs were attenuated with degree-dependent scoring.

-For each cluster, the Shapley score of the microarray was calculated and plotted on the STRING graph; the "cluster and rest" difference Δ determined the specific markers.

-The "top 5 genes for each cluster" output was obtained and was mostly consistent with standard PBMC biology (CD14$^+$ with S100A8/A9/FCN1, B with CD79B/HLA, NK/CD8 with GZMA/GZMK or GZMB/PRF1).

-A few naming inconsistencies were due to the binding of fixed labels to arbitrary Leiden indices; names were corrected by consensus relabeling based on the actual markers.

-The bar graphs of each cluster, the heatmap of the top-5 community, and the examination of the degree distribution confirm the specificity of the markers and the sparsity of the graph.

So finally, by modifying the labels, the obtained cell types are aligned with the reference table of PBMCs and the different ranks are exactly what is expected from a STRING-based network weighting method.

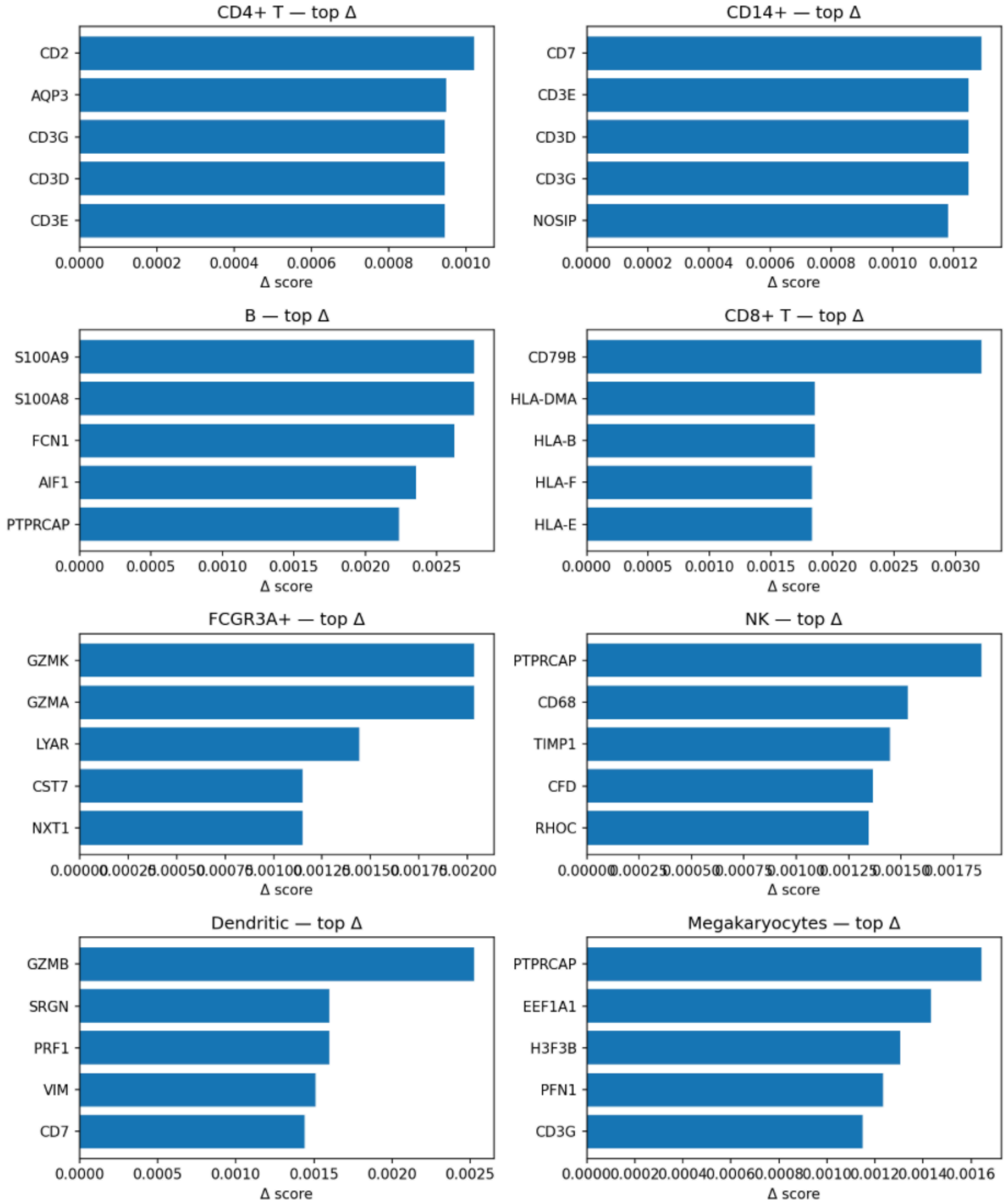


Figure 7: per-cluster bar charts

Finally, we use another chart for another comparison.dotplot Figure 8

Our dot plot shows, for each cluster (row) and each marker gene (column), the dot size as a percentage of expressing cells and the color as the mean expression. So we see both “coverage” and “intensity” together, something that is not possible with a simple bar graph.

The dot pattern is consistent with the reference table: B cells light up the MS4A1 and CD79A/B genes with large, dark dots; CD14⁺ monocytes have strong signals for CD14, LYZ, and the inflammatory granules S100A8/A9; NK/CD8 are bright for the cytotoxic complex GZMA, GZMK, GZMB, PRF1, as well as NKG7/GNLY; CD4⁺ T cells signal for IL7R/CCR7 (usually with moderate generality but good intensity, indicating naïve/memory).

This agreement shows that our cell type naming is also confirmed by the raw expression data; that is, the markers selected as top by the network-based method (Δ-Shapley + STRING) are also “on” in the expression distribution in exactly the same clusters.

In addition to confirmation, the dot plot also shows heterogeneity: where the dot is small but bold, the gene is very strongly expressed in a small subpopulation (activation/subtype). This is the reason for some differences with the table of “classical markers” only.

As a result, dotplot plays the role of a “bridge” between network analysis and known biology: our proposed markers are both valid in terms of network specificity and consistent in terms of expression pattern with the reference panel of PBMCs.

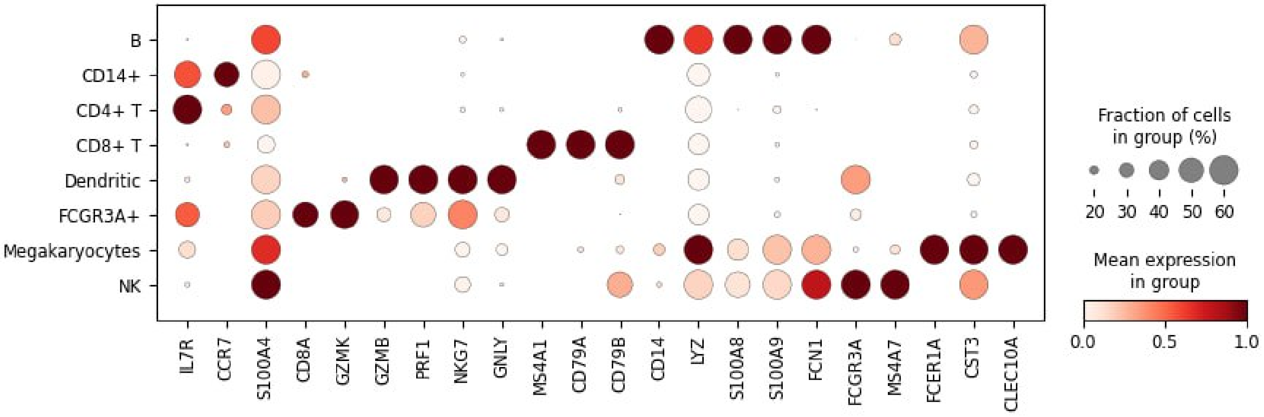


Figure 8: dotplot

Our method selects markers that are supported by both expression specificity and network texture by computing Δ-Shapley (cluster) and injecting it into the STRING network. This combination reduces the effect of noise by normalizing by degree and prioritizes more functional genes than feature selection alone. DotPlots provide independent control for naming accuracy and specificity and explain apparent discrepancies with markers. Overall, network-based Δ-Shapley—alone or in conjunction with feature selection—provides an interpretable, robust, and practical framework for marker prioritization and final reporting.

Ultimately, our proposed method provides both “biological validity” and “statistical robustness,” and its output is truly usable for marker selection and cluster naming—so it can be used for analyses.

6 Programming process

The raw data, preprocessing steps, Shapley and graph calculations, and all the steps performed are located in our GitHub group.

https://github.com/AmirHossein-Movahedi/single-cell/blob/main/top-fixed.ipynb

References

1-https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03693-1

2.https://www.biorxiv.org/content/10.1101/2021.02.01.429207v3

3.https://string-db.org/
4.https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131928

5.https://www.youtube.com/watch?v=uvyG9yLuNSE

6.https://mathewchamberlain.github.io/SignacX/

7.https://pmc.ncbi.nlm.nih.gov/articles/PMC4822726/

8. https://scholar.google.com/scholar_lookup?title=Coalitional%20Game%20Theory%20Facilitates%20Identification%20of%20Non-Coding%20Variants%20Associated%20With%20Autism&journal=Biomed%20Inform%20Insights&volume=11&pages=1-6&publication_year=2019&author=Sun%2CMW&author=Gupta%2CA&author=Varma%2CM&author=Paskov%2CKM&author=Jung%2CJY&author=Stockham%2CNT

9.https://scholar.google.com/scholar_lookup?&title=Investigation%20of%20Post-Transcriptio
20Gene%20Regulatory%20Networks%20Associated%20with%20Autism%20Spectrum%20Disorders%20by%20MicroRNA%20Expression%20Profiling%20of%20Lymphoblastoid%20Cell%20Lines&journal=Genome%20Med&volume=2&issue=4&publication_year=2010&author=Sarachana%2CT&author=Zhou%2CR&author=Chen%2CG&author=Hu%2CVW

10.https://pmc.ncbi.nlm.nih.gov/articles/PMC6055932/

11.https://scholar.google.com/scholar_lookup?title=An%20overview%20of%20recent%20applications%20of%20Game%20Theory%20to%20bioinformatics&journal=Inf%20Sci&volume=180&issue=22&pages=4312-22&publication_year=2010&author=Moretti%2CS&author=Athanasios%2CVV

12.https://www.sciencedirect.com/science/article/pii/S0888754318302118

13.https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-03134-1

14.https://www.nature.com/articles/s41592-025-02624-3

15.https://www.sc-best-practices.org/preprocessing_visualization/feature_selection.html

16.https://bioconductor.org/books/3.13/OSCA.basic/feature-selection.html

17.https://www.sciencedirect.com/science/article/abs/pii/S1568494623009584
18. https://satijalab.org/seurat/articles/pbmc3k_tutorial.html