

Hard and Soft GMM-EM for Overlapping Protein Complex Detection Across Five Disease-Specific PPI Networks

Anif Hanifa Setianingrum^{1,2*}, Wisnu Ananta Kusuma³,
Agus Buono⁴, Karlisa Priandana⁵

^{1,3,4,5}Doctoral Program in Computer Science, School of Data Science Mathematic and Informatics,
IPB University, Bogor, Indonesia

²Informatic Engineering, Faculty Science and Technology, University of Islamic Syarif
Hidayatullah, Jakarta, Indonesia

³Bioinformatics Study Program, Faculty of Mathematics and Natural Sciences,
IPB University Indonesia

ABSTRACT

Detecting overlapping protein complexes in disease-specific protein–protein interaction (PPI) networks remains a significant challenge, as multifunctional proteins may participate in more than one biological module. This study compares conventional hard and probabilistic soft assignments within a common Gaussian Mixture Model–Expectation Maximization (GMM–EM) framework across five disease-specific PPI networks. Proteins were represented by seven topology-based centrality features, and the optimal number of components was determined using an automatically detected Bayesian Information Criterion (BIC) elbow. Hard assignment used the highest posterior probability, whereas soft assignment retained memberships at $\tau = 0.10, 0.20$, or 0.30 . Clusters were evaluated against CORUM using Jaccard, Precision, Recall, F1-score, and Overlap Coefficient. Performance was network-dependent: ALPS achieved the highest Jaccard ($0.0852\text{--}0.0903$), Precision ($0.0934\text{--}0.0979$), and F1-score ($0.1555\text{--}0.1638$); leukemia the highest Recall/Overlap ($0.6375\text{--}0.6386$); breast cancer the largest mean intersection ($16.62\text{--}16.64$); HCC showed intermediate performance; and lung cancer the lowest Jaccard/F1 but greatest membership uncertainty, with 16.1% retaining a second posterior probability ≥ 0.10 . The soft assignment increased mean Recall in leukemia ($0.4333\text{--}0.4601$) and lung cancer ($0.3799\text{--}0.4030$). Aggregate F1 improved from 0.315 to 0.321 (Wilcoxon $p=0.0038$), and permutation testing showed a mean F1 gain of 0.018 ($10,000$ permutations, $p=0.012$). Soft assignment mainly preserves secondary membership information, with benefits depending on network structure. Limitations include static STRING networks, topology-only features, CORUM-based validation, unrecorded runtime, and no experimental verification.

Keywords: *Hard-Soft Clustering, Gaussian Mixture Model Expectation Maximization, Overlapping protein complexes, Diseases-specific PPI networks.*

Open Access

Type: Research Paper

Article:

Received: April 12, 2026

Revised: March 16, 2026

Accepted: July 13, 2026

Issued: October 30, 2026

© Setianingrum et al. (2026).



This is an open-access article
under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license

*Correspondence Address:

anif_h_anif@apps.ipb.ac.id

1. INTRODUCTION

Understanding disease mechanisms increasingly relies on the analysis of protein-protein interaction (PPI) networks, in which interacting proteins form complexes that serve as fundamental functional units in cellular processes [1], [2], [3], [4], [5]. Compared with isolated gene-level analysis, PPI network analysis enables the identification of coordinated molecular modules associated with biological functions, signaling pathways, and disease mechanisms. Protein complexes regulate essential cellular activities, including signal transduction, gene expression, immune responses, and apoptosis. Therefore, identifying protein complexes from PPI networks is an important strategy for understanding the molecular basis of human diseases.

However, many proteins perform multiple biological functions and participate in several protein complexes simultaneously, creating overlapping functional modules in PPI networks. Proteins such as CASP8, TRAF2, and BCL2 participate in multiple biological processes and signaling pathways, including apoptosis, immune regulation, and NF- κ B signaling. This overlapping organization increases network complexity and challenges conventional clustering methods that assign proteins to discrete communities, thereby complicating the identification of biologically meaningful protein complexes [6], [7], [8].

Several overlapping protein-complex detection and soft-clustering methods have shown promise, but some remain computationally demanding or depend strongly on external biological annotations, limiting scalability to large disease-specific PPI networks [9], [10], [11]. Protein-complex detection has also moved toward graph representation learning. [12], for example, reported GCN-mean-shift F-measures of 0.616 on Human PPI and 0.506 on Mouse PPI, outperforming several conventional methods in those benchmark

settings. These results demonstrate the potential of learned graph representations while remaining dependent on dataset and evaluation design.

Recent reviews by [13] and [14] describe representation-learning approaches including DeepWalk, Node2vec, GCN, GAT, GraphSAGE, graph autoencoders, heterogeneous graph embeddings, Transformers, BERT, and ESM. These methods learn latent protein representations for classification, link prediction, or clustering, but commonly require downstream steps to obtain explicit protein-complex memberships. Reported values such as SDNN-PPI AUC 0.9960, AttentionEP F-score 0.8262/AUC 0.9793, and PPII-AEAT F1 0.92 were obtained from different datasets and tasks and therefore are not direct head-to-head comparisons.

In contrast, the present study employs an interpretable topology-feature-based probabilistic GMM-EM framework in which posterior membership probabilities are estimated directly from centrality-based features, enabling explicit and controlled comparison between hard and probabilistic soft protein-complex assignments.

Among probabilistic approaches, Gaussian Mixture Model-Expectation Maximization (GMM-EM) models heterogeneous feature distributions and estimates posterior memberships that can represent potential protein-complex overlap [15] [16], [17]. In contrast, hard methods such as MCODE and MCL assign each protein to a single cluster [18], [19]. Other probabilistic and representation-learning models, including NOCD, variational graph autoencoders, and multi-label GCNs, also address overlapping communities. GMM-EM is used here because hard and soft assignments are derived from the same fitted likelihood-based mixture model, enabling a controlled comparison while modeling covariance among correlated topological features. It is therefore adopted as a transparent probabilistic framework

rather than claimed as universally superior to graph-based or neural alternatives.

Soft clustering addresses exclusive assignment by retaining secondary posterior memberships for a protein [20],[21] thereby representing membership uncertainty and potential functional overlap. This is relevant to multifunctional proteins that may participate in several processes or complexes. However, multiple memberships do not necessarily imply more accurate protein-complex reconstruction or biological multifunctionality. Their value must therefore be assessed through consistent statistical evaluation and validation against established reference complexes.

To improve comparability across studies, this work applies one topology-aware GMM-EM framework to five disease-specific PPI networks: lung cancer, breast cancer, leukemia, hepatocellular carcinoma, and ALPS. OMIM-associated genes were integrated with STRING interactions and predictions were validated against CORUM. Hard and probabilistic soft assignments used the same fitted model, topology features, and model-selection procedure, so differences primarily reflect retention of secondary posterior memberships across networks with different structural characteristics.

The study contributes: (1) a controlled comparison of hard and probabilistic soft GMM-EM assignments using identical topology features and model selection; (2) cross-disease analysis of network heterogeneity and posterior-membership uncertainty; and (3) consistent CORUM validation. Rather than introducing a new GMM-EM algorithm, the study tests when secondary posterior memberships add information beyond conventional hard assignment.

2. METHODS

The step-by-step research methodology shown in Figure 1 is described as follows:

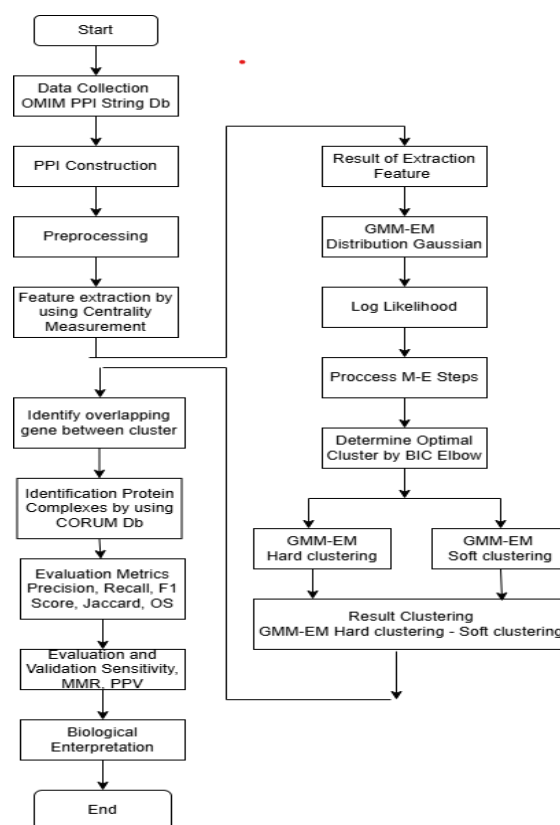


Figure 1. The step-by-step research methods

The methodology was designed to compare hard and probabilistic soft assignments within the same Gaussian Mixture Model-Expectation Maximization (GMM-EM) framework. Disease-associated genes from OMIM were mapped to STRING to construct disease-specific PPI networks, which were preprocessed by removing self-loops, duplicate interactions, invalid protein identifiers, and low-confidence interactions. Seven topology-based features were extracted and standardized before clustering, while the optimal number of Gaussian components was determined using AIC/BIC analysis with an automatically detected BIC elbow. The same fitted GMM-EM model generated both hard and soft assignments, enabling the effect of retaining secondary membership probabilities to be evaluated under controlled conditions. Predicted clusters were validated against CORUM reference complexes using Jaccard Index, Precision, Recall, F1-score, and Overlap Coefficient. By keeping data sources, preprocessing, feature extraction, model

selection, and validation procedures consistent across diseases, the workflow reduces methodological confounding and supports reproducible comparison across PPI networks with different structural characteristics.

2.1 Dataset

This study investigated five disease-specific PPI networks: autoimmune lymphoproliferative syndrome (ALPS), breast cancer (BCR), lung cancer (LCR), hepatocellular carcinoma (HCC), and leukemia (LEU). These diseases were selected to represent networks with different biological characteristics and levels of structural complexity, while ensuring that disease-associated genes could be mapped to valid STRING protein identifiers. Disease-associated genes were obtained from OMIM and mapped to STRING to retrieve their corresponding protein-protein interactions.

STRING interactions with combined score ≥ 0.4 were retained to balance reliability and network coverage. Self-interactions, duplicate pairs, inconsistent identifiers, and unmapped proteins were removed; duplicate protein pairs were represented by one undirected edge. All available connected and disconnected components were retained, and missing interactions were not imputed.

Following preprocessing, each disease-specific PPI network was represented as an undirected, weighted graph, $G = (V, E)$, where V denotes the set of protein nodes and E denotes the set of STRING-supported protein-protein interactions.

The same preprocessing and feature-extraction procedure was applied to all five diseases to support comparability across networks. CORUM was used later as a curated computational reference for biological validation; therefore, agreement with CORUM was interpreted as consistency with previously characterized protein complexes rather than as

experimental verification or evidence of de novo complex discovery.

2.2 Network Topology Feature Extraction

Each protein node was characterized using seven complementary topological features: degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, clustering coefficient, PageRank centrality, and subgraph centrality. These features were selected in advance to capture different aspects of network organization.

Degree centrality describes local connectivity; betweenness identifies bridge proteins occurring on shortest paths; closeness reflects network accessibility; eigenvector centrality and PageRank characterize global influence; clustering coefficient measures local neighborhood cohesiveness; and subgraph centrality captures participation in higher-order closed-walk structures. Together, these features provide a topology-aware representation of local, global, bridging, neighborhood, and higher-order structural properties.

Let $G = (V, E)$ denote an undirected weighted PPI network, where V is the set of proteins and E is the set of protein interactions. For each protein node, the seven topological descriptors were calculated from the disease-specific network. For path-based metrics such as betweenness and closeness centrality, STRING combined scores were used as edge weights. The resulting feature matrix was normalized before clustering to reduce scale differences among centrality measures.

For each protein $v_i \in V$, seven topology-based features were computed as follows:

In equation (1), Degree centrality is defined as:

$$C_D(v_i) = \frac{\deg(v_i)}{N - 1} \quad (1)$$

where $\deg(v_i)$ denotes the number of edges connected to node v_i , and N represents the total number of nodes in the network.

Betweenness centrality is formulated in Equation (2)

$$C_B(v_i) = \sum_{s \neq v_i \neq t} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \quad (2)$$

where σ_{st} denotes the total number of shortest paths from node s to node t , and $\sigma_{st}(v_i)$ denotes the number of those paths that pass through node v_i .

Closeness centrality is shown in equation (3)

$$C_C(v_i) = \frac{N-1}{\sum_{j=1}^N d(v_i, v_j)} \quad (3)$$

where $d(v_i, v_j)$ denotes the shortest path distance between nodes v_i and v_j .

Eigenvector centrality was expressed in equation (4)

$$C_E(v_i) = \frac{1}{\lambda} \sum_{j=1}^N A_{ij} C_E(v_j) \quad (4)$$

where a_{ij} is the adjacency matrix element and λ is the largest eigenvalue of the network.

The clustering coefficient is formulated in Equation (5)

$$C_i = \frac{2 \times e_i}{k_i(k_i-1)} \quad (5)$$

where e_i represents the number of edges among the neighbors of node v_i , and k_i is the degree of node v_i .

The PageRank is defined in Equation (6), where $M(v_i)$ denotes the set of nodes linking to v_i , k_j^{out} is the out-degree of node v_j , N is the total number of nodes, and α is the damping factor that controls the balance between link-following and random transitions.

$$C_P(v_i) = \frac{1-\alpha}{N} + \alpha \sum_{v_j \in M(v_i)} \frac{C_P(v_j)}{k_j^{out}} \quad (6)$$

The subgraph centrality of a node v_i quantifies its participation in all closed walks within the network. As shown in equation (7), where $(A^l)_{ii}$ represents the number of closed walks of length l starting and ending at node v_i , A is the adjacency matrix, and $l!$ is a normalization factor.

$$C_S(v_i) = \sum_{l=0}^{\infty} \frac{(A^l)_{ii}}{l!} \quad (7)$$

The seven resulting centrality features were standardized using z-score normalization implemented with StandardScaler to reduce scale differences before GMM fitting.

2.3 Gaussian Mixture Model with Expectation–Maximization (GMM-EM)

GMM-EM was used to model the distribution of standardized topological feature vectors as a mixture of latent Gaussian components. Each component was interpreted as a candidate protein complex or functional module, whereas the posterior probability of a protein belonging to each component represented its membership strength. The EM algorithm iteratively estimated component responsibilities during the Expectation step and updated mixture weights, means, and covariance matrices during the Maximization step until convergence of the log-likelihood.

$$X = \{x_1, x_2, \dots, x_N\}$$

$$x_i =$$

[degree, betweenness, closeness, eigenvector

, clustering coefficient, pagerank, subgraph]

$$p(x_i) = \sum_{k=1}^K \pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k) \quad (8)$$

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x_i | \mu_j, \Sigma_j)} \quad (9)$$

A protein was classified as overlapping when its posterior membership probability was greater than or equal to the predefined threshold τ in at least two clusters:

$$\sum_{k=1}^K I(\gamma_{ik} \geq \tau) \geq 2 \quad (10)$$

Such proteins are regarded as multi-functional proteins that potentially

participate in multiple protein complexes and biological pathways simultaneously. Sensitivity analyses were performed using $\tau=0.10$, 0.20, and 0.30. Thus, overlap was not inferred using a separate clustering algorithm but directly from posterior probabilities generated by the same fitted GMM-EM model. This design enables direct comparison between hard and soft assignment while holding the feature representation, fitted mixture model, and model-selection procedure constant.

To improve reproducibility, GMM-EM was implemented using a full covariance matrix, k-means initialization, five independent initializations ($n_{\text{init}} = 5$), a maximum of 100 EM iterations, and a fixed random state of 42. The implementation also used a predefined convergence tolerance $= 1 \times 10^{-3}$, and covariance regularization $= 1 \times 10^{-6}$. The full covariance specification allows correlations among the seven topology-derived features to be represented within each Gaussian component.

2.4 AIC and BIC Elbow-Based Model Selection

Candidate numbers of mixture components ranged from $K = 2$ to $K = 20$. For each candidate value of (K), the GMM-EM model was fitted and evaluated using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), both of which balance model fit against model complexity. The final number of components was selected using the BIC elbow rather than subjective visual inspection alone.

The BIC elbow was automatically identified using the maximum perpendicular-distance criterion relative to the line connecting the first and last BIC points. This approach allows the elbow point to be determined consistently by identifying when additional Gaussian components result in relatively small BIC improvements compared with the added model complexity. AIC was retained as

complementary model selection information, whereas the automatically detected BIC elbow determined the final number of components (K) used for both hard and soft assignments.

The resulting hard and soft clustering outputs were evaluated against curated CORUM reference protein complexes using Jaccard Index, Precision, Recall, F1-score, and Overlap Coefficient. Sensitivity, Positive Predictive Value (PPV), and Maximum Matching Ratio (MMR) may additionally be reported where applicable. Because hard and soft outputs were derived from the same fitted GMM-EM model, this evaluation provides a controlled comparison of assignment strategies rather than a comparison between unrelated clustering algorithms.

CORUM was treated as a computational biological reference or gold-standard source rather than as experimental confirmation. Consequently, CORUM-supported matches indicate consistency with previously characterized protein complexes and should not be interpreted as experimental verification or *de novo* discovery of novel complexes. Functional enrichment using GO, KEGG, or Reactome was not part of the current computational pipeline and is therefore considered a complementary validation strategy for future work rather than evidence used to define the reported clusters.

Robustness of soft assignment was examined using the three posterior-membership thresholds ($\tau=0.10$), (0.20), and (0.30), while the STRING confidence criterion was explicitly fixed at a combined score ≥ 0.4 for primary network construction. This standardized procedure enabled hard and probabilistic soft GMM-EM assignments to be evaluated consistently across the five disease-specific PPI networks.

3. RESULTS AND DISCUSSION

3.1 Feature Extraction

The disease-specific PPI networks were constructed from disease genes from OMIM and STRING databases in protein interaction data, then analyzed using hard and soft GMM-EM assignment strategies. The networks were visualized using Python-based graph layouts and Cytoscape to examine structural organization and interaction density. Table 1 summarizes the number of nodes, edges, and reference protein complexes used in the analysis.

Table 1. DATA SET PPI (5) DISEASES

Disease Name	Number of Edges	Sum of Nodes	Protein Complexes
Leukemia	35.615	6.391	2.895
Breast Cr	25.072	5.037	2.750
Lung Cr	133.454	15.454	3.168
ALPS	622	150	406
HCC	50.151	8.599	3.119

The selected k values, which are used for all subsequent analyses, are summarized in Table 2. The plot for lung cancer, for example, showed a clear elbow at k=8, while HCC showed its elbow at k=7. The ALPS network, having a more ambiguous BIC curve, was determined to have an optimal k =9. Table 2 summarizes the optimal k value for each disease as follows:

Table 2. K Optimal for each disease

Disease Name	Network Name	Optimal Cluster
Breast cancer	BCR	6
Lung cancer	LCR	8
Leukemia	LEU	7
HCC	HCC	7
ALPS	ALPS	9

Table 3 highlights substantial topological differences among the five disease-specific PPI networks. ALPS is the smallest and most locally cohesive, with the highest clustering coefficient (0.669) and normalized degree connectivity (0.0557). In

contrast, LCR is the largest, has the lowest clustering coefficient (0.348), and the highest average closeness (0.395), suggesting weaker local compartmentalization and less distinct cluster boundaries.

Subgraph centrality increases markedly from ALPS to LCR, reflecting more numerous and complex closed walks in larger networks. Because this measure is strongly affected by network size and topology, it is not interpreted as direct evidence of protein-complex overlap; soft-clustering utility is assessed using posterior probabilities, entropy, membership margins, and CORUM validation.

Table 4 further shows that ALPS has the highest average degree, betweenness, eigenvector centrality, clustering coefficient, and PageRank, consistent with a compact, locally cohesive network.

The larger BCR, HCC, LEU, and LCR networks show lower normalized degree, betweenness, eigenvector, clustering-coefficient, and PageRank values than ALPS, but higher closeness. LCR has the highest mean closeness (0.395363), followed by HCC (0.316542), LEU (0.298022), and BCR (0.283181). This pattern is consistent with broader interaction neighborhoods and potentially weaker local cluster separation.

Subgraph centrality rises from 1.505×10^4 in ALPS to 1.2715×10^{21} in LCR, indicating greater structural complexity in larger networks. Because this measure alone cannot establish multifunctionality or overlap, it is interpreted jointly with posterior probability, entropy, cluster density, connectivity ratio, and CORUM matching. 10^4 10^{21}

Table 3. Network Characteristics of Five Disease-Specific PPI Networks

Description	ALPS	BCR	HCC	LEU	LCR
Network Size	Very Small	Large	Very Large	Large	Extremely Large
Modularity	High	Moderate	Moderate	Moderate-High	Low

Description	ALPS	BCR	HCC	LEU	LCR
Clustering Coefficient	High (0.669)	Moderate (0.543)	Moderate (0.485)	Moderate (0.519)	Low (0.348)
Closeness Centrality	Low–Moderate (0.194)	Moderate (0.283)	High (0.317)	Moderate–High (0.298)	Very High (0.395)
Degree Connectivity	Very High (0.0557)	Low (0.0020)	Low (0.0014)	Low (0.0017)	Very Low (0.0011)
Subgraph Complexity	Very Low (1.5×10^4)	Moderate (2.9×10^{10})	High (7.8×10^{12})	Very High (3.2×10^{14})	Extremely High (1.3×10^{21})
Overlap Potential	Low	Moderate–High	High	Very High	Very High
Expected Cluster Separation	High	Moderate	Moderate	Low-Moderate	Low
Suitability for Soft Clustering	Moderate	High	High	Very High	Very High

Table 4. Feature Extraction of Centrality Measurement – Five (5) Diseases

Disease	Total Genes	Avg. Degree	Avg. Betweenness	Avg. Closeness	Avg. Eigenvector	Avg. Clustering Coefficient	Avg. PageRank	Avg. Subgraph Centrality
ALPS	150	0.055660	0.013566	0.194221	0.032991	0.669077	0.006667	1.5053×10^4
BCR	5,037	0.001977	0.000807	0.283181	0.004753	0.542636	0.000199	2.9379×10^{10}
HCC	8,599	0.001357	0.000473	0.316542	0.003692	0.485009	0.000116	7.8160×10^{12}
LEU	6,391	0.001744	0.000640	0.298022	0.004107	0.518882	0.000156	3.1686×10^{14}
LCR	15,454	0.001117	0.000225	0.395363	0.002915	0.348192	0.000065	1.2715×10^{21}

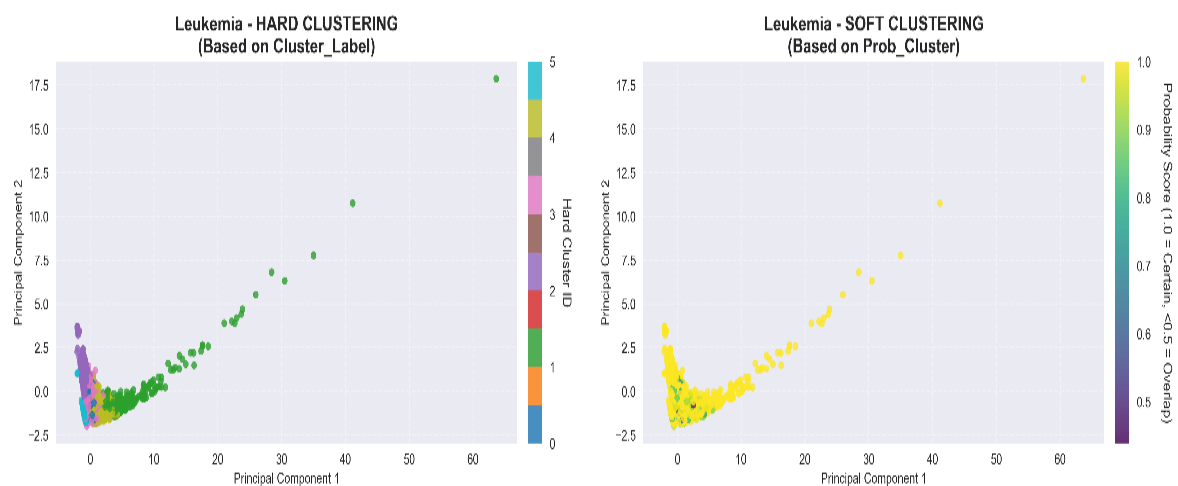


Figure 2. Comparison of GMM-EM hard clustering (left) and GMM-EM soft clustering (right) in the leukemia

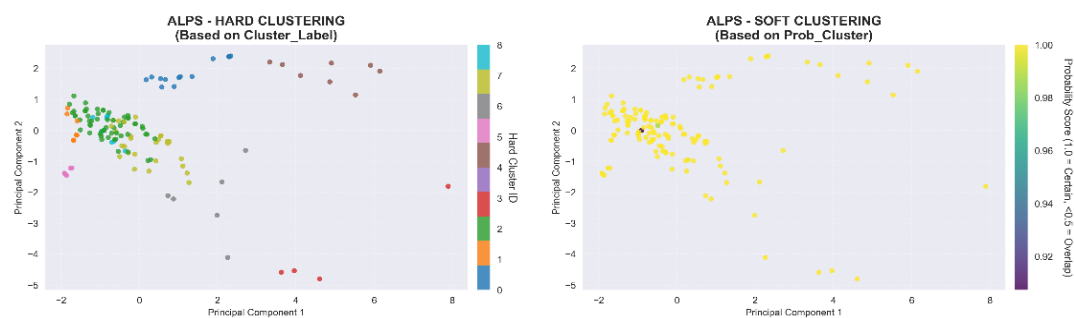


Figure 3. Comparison of GMM-EM hard clustering and GMM-EM soft clustering in the ALPS

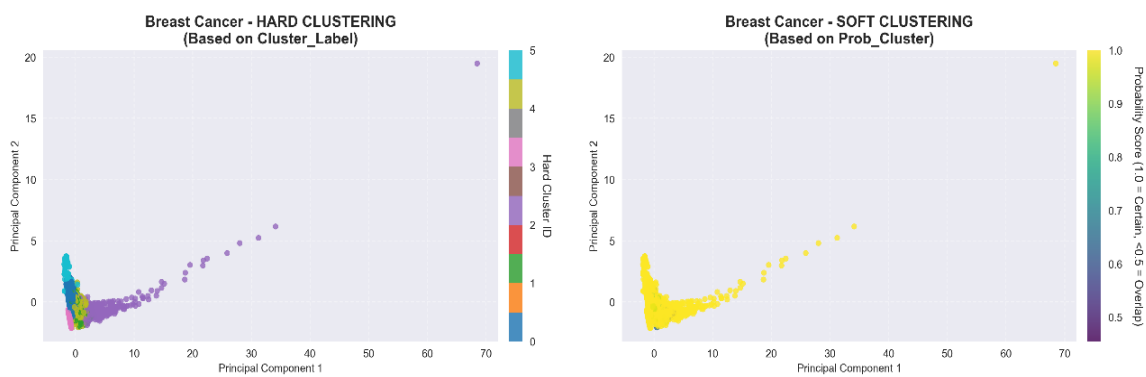


Figure 4. Comparison of GMM-EM hard clustering and GMM-EM soft clustering in the breast cancer

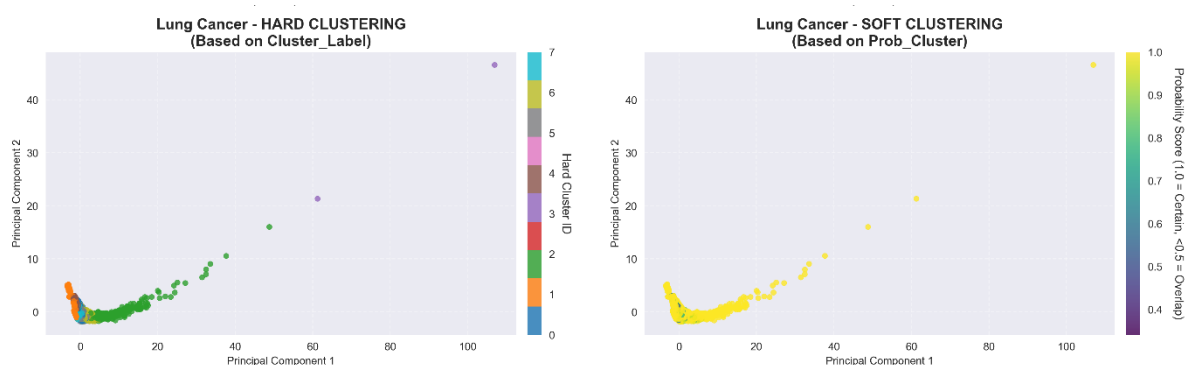


Figure 5. Comparison of GMM-EM hard clustering and GMM-EM soft clustering in the lung cancer

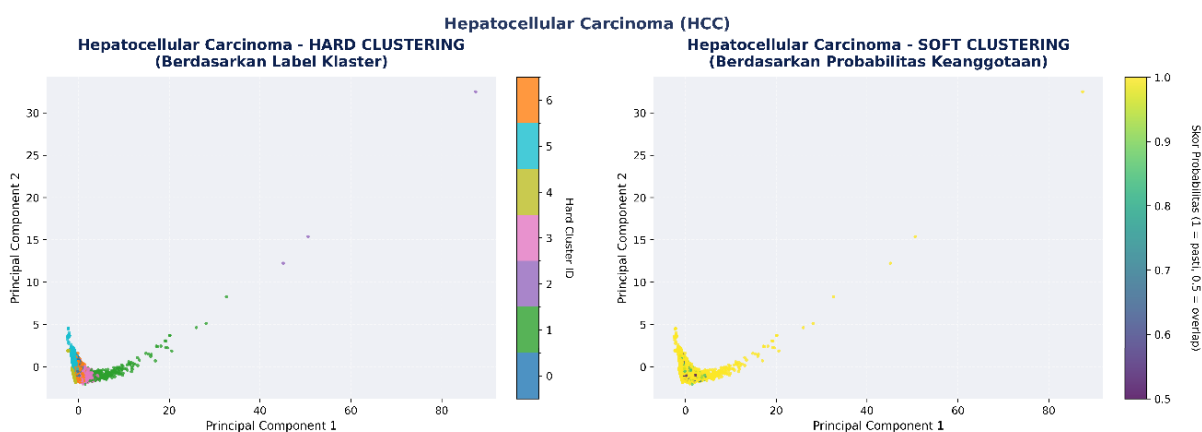


Figure 6. Comparison of GMM-EM hard clustering and GMM-EM soft clustering in the HCC

3.2 The Comparative analysis of GMM-EM hard and soft clustering across five disease-specific networks.

Figures 2–6 present two-dimensional PCA projections of the hard and soft GMM-EM clustering results for leukemia, ALPS, breast cancer, and lung cancer. PCA reflects separation in feature space, not PPI topology or proof of overlap, it is therefore interpreted with posterior probabilities,

normalized entropy, probability margins, subgraph density, and connectivity ratio.

ALPS shows heterogeneous subgraph cohesion: Cluster 1 is fully connected (density=1.000), Cluster 0 is large (n=58) but sparse (density=0.055), and Cluster 4 has no internal interactions. This confirms that feature-space clustering can differ from direct PPI cohesion, so biological relevance is evaluated separately against CORUM.

For ALPS, hard and soft assignments are nearly identical: mean maximum posterior

probability is 0.9987, normalized entropy is 0.0021, and 98.0% of proteins have maximum posterior ≥ 0.99 . Memberships are therefore highly confident and well separated, leaving little practical difference between assignment strategies.

Thus, GMM-EM membership reflects similarity in centrality features rather than direct PPI connectivity alone, requiring separate biological validation.

Leukemia shows approximately 97% hard-soft agreement, with mean maximum posterior 0.982 and entropy 0.025. Dominant assignments remain stable, while soft clustering retains secondary membership information that may indicate participation in multiple complexes.

BCR has generally low cluster density, although Clusters 2 and 5 are fully connected. CORUM matches include CCR4-NOT, CENP-A/NAC-CAD, and the Spliceosome, supporting biological interpretation but not de novo complex discovery.

BCR also shows approximately 97% hard-soft agreement with high membership confidence. Its main soft-clustering benefit is retention of secondary memberships for proteins potentially associated with multiple functional modules, which still require independent biological evidence.

LCR clusters vary strongly in cohesion: Clusters 2, 4, and 3 are highly connected, whereas Cluster 1 has very low density and connectivity. Some components therefore reflect centrality-feature

similarity more than direct interaction cohesion, requiring CORUM validation.

LCR has the highest clustering uncertainty, with mean maximum posterior 0.947 and entropy 0.063, indicating greater secondary-membership ambiguity and potential overlap that requires biological confirmation.

Overall, soft-clustering utility is network-dependent: ALPS changes little, whereas BCR, leukemia, HCC, and especially LCR show greater membership uncertainty. Any advantage should therefore be supported by statistical and biological validation.

Figure 7 integrates PPI topology, dominant GMM component, and membership probability: nodes are proteins, edges are interactions, positions follow network topology, color indicates the dominant component, and node size may represent centrality. The view reveals dense modules and bridge proteins while retaining the need for biological validation.

Soft memberships can be visualized through node borders, halos, or secondary-color rings. High-confidence proteins show one dominant assignment, whereas substantial second posterior probabilities highlight candidate secondary or overlapping memberships. Such proteins may be multifunctional, but this interpretation remains a hypothesis until supported by CORUM, pathway analysis, literature, or experimental evidence.

Table 5. Posterior-membership diagnostics across the five disease-specific PPI networks

Disease	n	K	Mean max posterior (95% CI)	Normalized entropy (95% CI)	Mean top-2 margin	Max posterior ≥ 0.99	Second posterior ≥ 0.10
ALPS	150	9	0.9987 [0.9967, 1.0000]	0.0021 [0.0001, 0.0053]	0.9974	98.0%	0.0%
BCR	5,037	6	0.9893 [0.9878, 0.9909]	0.0152 [0.0136, 0.0169]	0.9790	92.2%	3.1%
HCC	8,599	7	0.9770 [0.9754, 0.9785]	0.0295 [0.0280, 0.0312]	0.9546	84.1%	7.2%
LEU	6,391	7	0.9821 [0.9804, 0.9837]	0.0251 [0.0233, 0.0269]	0.9645	87.2%	5.5%
LCR	15,454	8	0.9467 [0.9449, 0.9484]	0.0629 [0.0614, 0.0646]	0.8963	64.9%	16.1%

Example of Network-Level PPI Visualization with Dominant Cluster and Posterior Membership

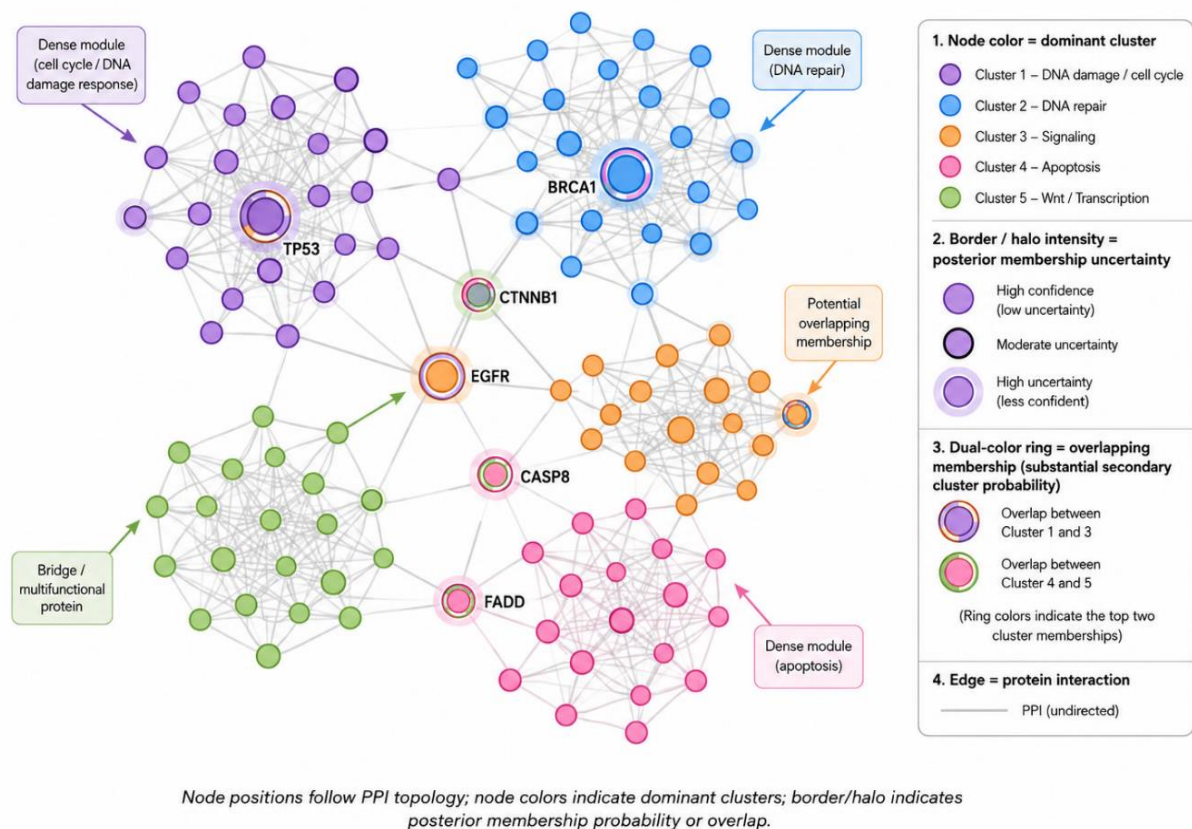


Figure 7. The PPI topology, node color represents the dominant cluster, and either node border, size, transparency, or a secondary color represents posterior membership probability or overlapping membership

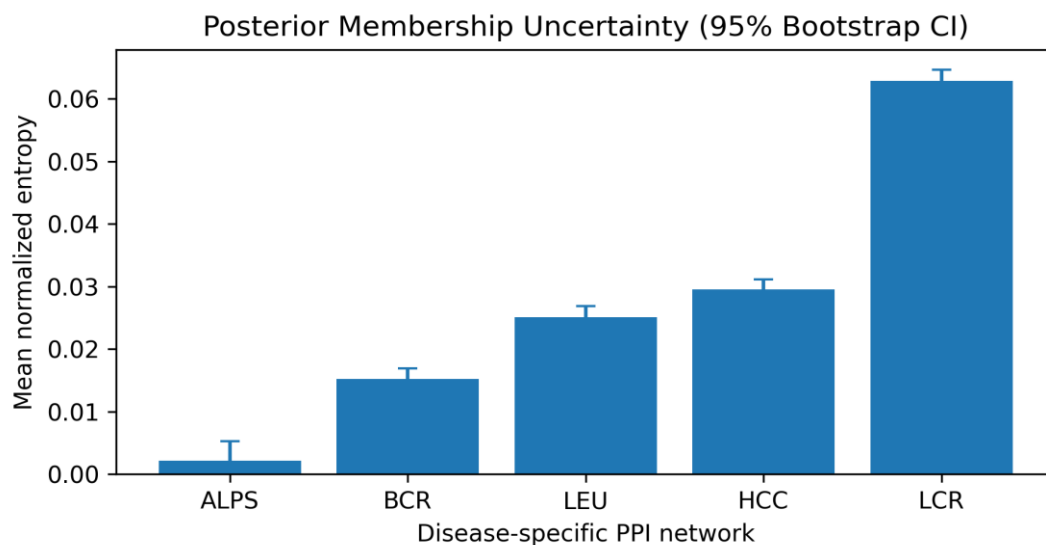


Figure 8. Mean normalized posterior entropy with 95% bootstrap confidence intervals.

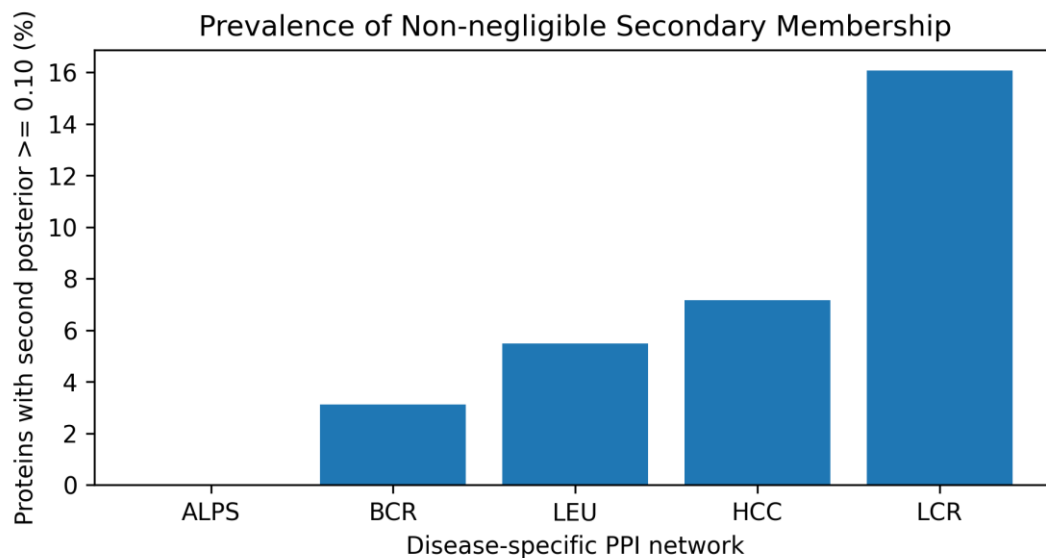


Figure 9. Percentage of proteins with a second posterior membership probability of at least 0.10.

Table 6. Holm-adjusted Wilcoxon p-values for paired hard-soft CORUM Best-Match metrics

Disease	Paired clusters	Jaccard	Recall	Precision	F1-score	Overlap Coef.
ALPS	9	1.0000	1.0000	1.0000	1.0000	1.0000
BCR	6	1.0000	1.0000	1.0000	1.0000	1.0000
HCC	7	1.0000	1.0000	1.0000	1.0000	1.0000
LCR	8	0.1953	0.9375	0.1953	0.1953	0.9375
LEU	6	1.0000	0.4688	1.0000	1.0000	0.4688

Table 7. Representative CORUM Best-Match results illustrating hard-soft behavior

Disease/Cluster	CORUM complex	Hard shared	Soft shared	Hard Recall	Soft Recall	Hard Precision	Soft Precision
ALPS C4	DISC (FAS-FADD-CASP8-CFLAR)	4/4	4/4	1.0000	1.0000	0.5000	0.5000
BCR C0	Spliceosome	43/143	45/143	0.3007	0.3147	0.0202	0.0208
BCR C5	CENP-A/NAC-CAD	10/13	10/13	0.7692	0.7692	0.0980	0.0980
HCC C2	p53 homotetramer	1/1	1/1	1.0000	1.0000	0.3333	0.3333
LCR C3	p53 homotetramer	1/1	1/1	1.0000	1.0000	0.5000	0.5000
LEU C1	HCF-1 complex	8/18	9/18	0.4444	0.5000	0.0345	0.0360
LEU C2	Spliceosome	38/143	39/143	0.2657	0.2727	0.0156	0.0157

3.3 Posterior Membership Uncertainty and Statistical Validation

Hard and soft outputs were derived from the same fitted GMM: the hard label is the component with the highest posterior probability, while the soft output retains the complete posterior vector. Thus, the dominant component is identical by construction; the additional information in soft clustering is secondary probability

mass, summarized by maximum posterior probability, normalized entropy, and the top-two probability margin.

Posterior diagnostics show a clear uncertainty gradient. ALPS is most sharply separated (mean maximum posterior 0.9987; normalized entropy 0.0021; 0% with second posterior ≥ 0.10), whereas LCR is most uncertain (0.9467; 0.0629; 16.1%). BCR is relatively well separated, with LEU and HCC intermediate. These measures

quantify competition between Gaussian components without assuming that network size alone proves biological overlap. Kruskal–Wallis tests confirmed differences among diseases for normalized entropy ($H=4542.44$, $p<0.001$, $\varepsilon^2=0.127$) and maximum posterior probability ($H=4815.53$, $p<0.001$, $\varepsilon^2=0.135$). These results establish network-dependent membership uncertainty, not superiority of soft clustering. Table 5 confidence intervals were obtained from 1,000 nonparametric bootstrap resamples within each disease.

3.4 Paired Statistical Comparison of GMM-EM Hard and Soft CORUM Best Matches

For the statistical comparison, corresponding GMM component labels were paired within each disease, and CORUM Best-Match Jaccard, Recall, Precision, F1-score, and Overlap Coefficient values were compared between hard and soft assignments using two-sided Wilcoxon signed-rank tests. The Holm correction was applied to control for multiple comparisons across the five evaluation metrics within each disease. Because the sample sizes were limited to the relatively small number of Gaussian components in each disease-specific network, effect magnitudes and confidence intervals were interpreted alongside the adjusted p-values.

No Holm-adjusted p-value reported in Table 6 was below the significance threshold of 0.05. Therefore, the available cluster-level CORUM Best-Match results do not provide sufficient statistical evidence to establish the superiority of soft assignment over hard assignment within any individual disease network. Effect sizes were small or mixed. BCR Recall increased from 0.4813 to 0.4836 ($\Delta=+0.0023$; 95% bootstrap CI approximately -0.00001 to $+0.0070$). LEU Recall increased from 0.4333 to 0.4601 ($\Delta=+0.0267$; CI $+0.0023$ to $+0.0570$), but remained non-significant after Holm correction (adjusted $p=0.4688$).

LCR Recall increased from 0.3799 to 0.4030 while Jaccard, Precision, and F1-score decreased slightly; ALPS was essentially unchanged. Soft assignment therefore preserves secondary membership information while producing only modest, network-dependent changes in CORUM recovery.

3.5 CORUM-Based Validation and Interpretation of Low Precision and Jaccard Scores

Based on the advanced evaluation metrics across five disease-specific PPI networks, probabilistic soft clustering generally demonstrated improved performance compared with conventional hard clustering, although the magnitude of improvement depended on the characteristics of each network. For ALPS, hard and soft clustering produced identical results, with a Sensitivity (S_n) of 0.2549, PPV of 0.1333, Accuracy of 0.1844, and MMR of 0.3583, indicating that soft assignment did not provide additional benefit in this network. For BCR, soft clustering increased Sensitivity from 0.4718 to 0.4808, Accuracy from 0.1042 to 0.1045, and MMR from 0.5549 to 0.5659, although PPV slightly decreased from 0.0230 to 0.0227. HCC showed the most consistent improvement across all metrics, with Sensitivity increasing from 0.4835 to 0.5019, PPV from 0.0143 to 0.0148, Accuracy from 0.0832 to 0.0863, and MMR from 0.5646 to 0.5798. The largest improvement was observed in LCR, particularly in Sensitivity, which increased from 0.5294 to 0.5552, and MMR, which increased from 0.6000 to 0.6208, although PPV slightly decreased from 0.0119 to 0.0115. In LEU, soft clustering also improved Sensitivity from 0.4720 to 0.4859, Accuracy from 0.0846 to 0.0857, and MMR from 0.5590 to 0.5730, with only a very small decrease in PPV. Overall, the mean Sensitivity increased from 0.4423 for hard clustering to 0.4557 for soft clustering, while the mean MMR increased from

0.5274 to 0.5396 and the mean Accuracy increased from 0.1072 to 0.1081, whereas the mean PPV remained nearly unchanged. These findings indicate that the primary advantage of soft clustering lies in improving the ability to capture protein-complex members and enhancing the correspondence between predicted complexes and CORUM reference complexes, rather than consistently increasing precision. Therefore, the benefit of probabilistic soft assignment appears to be network-dependent and is more pronounced in more complex networks such as HCC, LCR, and LEU.

3.6 Biological Interpretation in the Context of Previous Studies

ALPS is an immune regulatory disorder primarily associated with *FAS* mutations and impaired apoptosis, with complex genetic variation and penetrance that may contribute to ALPS-like disorders, autoimmunity, and potentially carcinogenesis. *FAS* variants and increased DNTs are associated with autoimmune diseases and may contribute to carcinogenesis, suggesting that impaired *FAS*-mediated apoptosis extends beyond inherited immune disorders. [22]

In breast cancer (BCR), the human CCR4–NOT complex, particularly the CNOT1 subunit, acts as a transcriptional repressor of ER α - and RXR-mediated nuclear receptor signaling, thereby modulating target gene expression and influencing cancer-related molecular pathways [23]

Accordingly, these matches support biological plausibility rather than novel-complex claims. Myeloid neoplasms frequently harbor spliceosome mutations that cause distinct abnormal splicing patterns. RNA-Seq analysis of 1,258 patients and 63 healthy donors identified 17,300 dysregulated alternative splicing events, common mis-spliced transcripts, and splicing factor-dependent neojunctions, with reduced intron retention particularly in

patients carrying splicing factor mutations. [24]

Genomic profiling has identified distinct HCC molecular subclasses with different genetic alterations and clinical characteristics. Aggressive HCC is associated with TP53 inactivation, oncogenic pathway activation, and stemness-related features, whereas less aggressive tumors often harbor CTNNB1 mutations and WNT pathway activation. These subclasses may help stratify patients according to their likelihood of benefiting from specific targeted therapies. [25]

Lung cancer is a leading cause of cancer-related mortality, with NSCLC accounting for the majority of cases. β -Catenin, encoded by *CTNNB1*, is a key regulator of NSCLC pathogenesis through the Wnt/ β -catenin signaling pathway.

This review summarizes its involvement in EMT, the tumor microenvironment, nuclear translocation, EGFR-TKI resistance, and its potential as a prognostic biomarker. [26]

3.7 Computational and Visualization Considerations

The implementation standardizes seven centrality features, uses random_state=42 and n_init=5, and selects components from BIC profiles (ALPS=9, BCR=6, HCC=7, LCR=8, LEU=7). With full covariance and fixed feature dimension, GMM cost grows mainly with network size and K, while posterior storage scales as $n \times K$.

This theoretical complexity is not an empirical benchmark because runtime, peak memory, hardware, and iteration counts were not recorded reproducibly. Future analyses should record centrality-extraction time, GMM fitting time, EM iterations, and peak memory under a common environment. Figures 2–7 are PCA feature-space projections rather than PPI-network visualizations. Figures 6–8 add posterior uncertainty information, while complete topology-level visualization requires original edge lists to position nodes by

interactions and highlight secondary memberships.

3.8 Overall Synthesis, Baseline Scope, and Limitations

The analyses support a network-dependent rather than universally superior interpretation of soft GMM-EM. ALPS shows minimal posterior uncertainty and unchanged hard-soft CORUM metrics; BCR changes only slightly; leukemia has the clearest Recall increase; HCC is mixed; and LCR combines the highest uncertainty with higher Recall but slightly lower Jaccard, Precision, and F1-score. No Holm-adjusted cluster-level Best-Match test reaches $p < 0.05$.

The principal contribution of soft assignment is probabilistic representation of uncertainty discarded by hard argmax labels. This is most evident in LCR, where 16.1% of proteins retain a second posterior ≥ 0.10 , and least evident in ALPS. Whether these memberships represent true multifunctionality or overlapping complexes remains a biological hypothesis requiring CORUM, pathway, literature, or experimental support.

Competitive ranking against MCODE, MCL, ClusterONE, or other baselines requires harmonized node universes, thresholds, preprocessing, and metric aggregation; existing project outputs do not satisfy these conditions for a safe direct numerical ranking.

Overall, topology is associated with different posterior uncertainty across diseases. Soft assignment can modestly improve recovery of selected CORUM proteins but not consistently Precision, Jaccard, or F1-score. Biologically recognizable DISC, Spliceosome, HCF-1, CENP-A/NAC-CAD, and p53-related matches support meaningful substructure, while novel-discovery, scalability, and superiority claims require further validation.

3.9 Comparative Evaluation Metrics GMM-EM based hard clustering and soft clustering

Hard and soft GMM-EM outputs were evaluated against CORUM using Best Match per Cluster, Top-5 Matches per Cluster, and All Positive Overlapped Protein Complexes. Jaccard, Precision, Recall, F1-score, and Overlap Coefficient were interpreted jointly because broad predicted components can recover substantial portions of a reference complex while still yielding low Precision and Jaccard. Absolute values therefore remain essential when judging reconstruction quality.

For leukemia soft clustering, representative Best-Match complexes include the Spliceosome, HCF-1, BRAFT, and SCAMP1-SCAMP3. Cluster 3 has the highest Recall/Overlap Coefficient for BRAFT (0.6923), Cluster 7 reaches 0.6667 for SCAMP1-SCAMP3, and Cluster 2 matches HCF-1 with Jaccard=0.0432, Precision= 0.0457, Recall=0.4444, and F1=0.0829. Spliceosome-associated clusters recover reference proteins but retain low Jaccard and Precision.

Thus, leukemia soft clustering supports partial recovery of known CORUM complexes rather than high-precision reconstruction of complete complexes.

ALPS hard and soft clustering are identical on average (Jaccard=0.205, Recall=0.677, Precision=0.231, F1=0.315, Overlap Coefficient=0.688). The strongest FAS-FADD-CASP8-CASP10 match reaches Jaccard=0.500, Recall=1.000, Precision=0.500, F1=0.667, and Overlap=1.000, confirming recovery without a soft-assignment advantage.

LEU shows modest numerical gains: mean Jaccard rises from 0.0161 to 0.0168, Recall from 0.3841 to 0.4020, Precision from 0.0171 to 0.0177, F1 from 0.0315 to 0.0328, and Overlap from 0.3841 to 0.4020. These changes improve sensitivity but remain small and must be interpreted with the paired statistical tests.

BCR also improves only slightly: Sensitivity rises from 0.4719 to 0.4809, Accuracy from 0.1043 to 0.1045, and MMR from 0.5549 to 0.5659, while PPV remains comparable; the Spliceosome match recovers 45 instead of 43 proteins. This supports marginally increased coverage rather than broad superiority.

Across diseases, soft GMM-EM is not universally superior to hard assignment. ALPS is essentially unchanged, whereas leukemia, BCR, HCC, and especially LCR show more secondary probabilistic membership. Gains in leukemia and BCR are modest and often coexist with low absolute Precision and Jaccard. The primary advantage is therefore preservation of uncertainty and potential multifunctional memberships; statistical validation, external baselines, biological support, and reproducible scalability analysis remain necessary to establish when this information improves protein-complex identification. This interpretation emphasizes that probabilistic membership should complement, rather than replace, conventional cluster labels when uncertainty is substantial, especially when independent biological validation is available for secondary memberships.

CONCLUSION

This study evaluated a topology-aware Gaussian Mixture Model-Expectation Maximization (GMM-EM) framework for detecting candidate overlapping protein complexes in five disease-specific PPI networks (ALPS, BCR, LEU, HCC, and LCR), applying identical features, model-selection procedures, and fitted models to both hard and soft assignments. The benefit of soft assignment was network-dependent. ALPS showed highly confident memberships and nearly identical hard-soft results, whereas larger heterogeneous networks showed greater posterior uncertainty, which was highest in LCR, where 16.1% of proteins retained a second posterior probability ≥ 0.10 . LEU and BCR showed modest gains in the recovery of selected CORUM proteins, but these gains were not

consistently accompanied by higher Precision, Jaccard, or F1-score, and paired Wilcoxon signed-rank tests with Holm correction showed that hard-soft differences did not reach statistical significance at the $p < 0.05$ threshold within any individual disease network.

Therefore, soft GMM-EM should be regarded as a complementary approach rather than a universal replacement for hard clustering. Its main contribution is the preservation of membership uncertainty and secondary memberships that are lost under strict hard assignment, which may help prioritize candidate multifunctional proteins in networks with substantial uncertainty, whereas networks with well-separated memberships may be adequately represented by hard assignment. CORUM-based validation identified several recognizable structures, including DISC-associated proteins in ALPS, Spliceosome and HCF-1 complexes in leukemia, CENP-A/NAC-CAD and CCR4-NOT-related structures in breast cancer, and p53-related matches in HCC and lung cancer; however, low Precision and Jaccard values indicate that several predicted components were broader than their reference complexes. These matches should therefore be interpreted as computational evidence of consistency with known complexes rather than as experimental confirmation or discovery of novel complexes, and shared complexes across networks should not be interpreted as evidence of disease comorbidity.

This study is limited by its reliance on static STRING-based networks, topology-derived features, and CORUM as the principal reference, as well as by the absence of experimental validation and standardized runtime and scalability benchmarking. Future work should compare GMM-EM with established methods such as MCODE, MCL, and ClusterONE on larger independent benchmark datasets under harmonized protocols; strengthen statistical evaluation using more matched complexes, confidence

intervals, and effect-size estimates; and integrate multi-omics, dynamic, and tissue-specific PPI data, together with functional enrichment and experimental verification, to determine whether secondary memberships reflect genuine multifunctional protein-complex participation.

DECLARATIONS

Author Contributions: The authors explicitly outline their individual contributions to the research and manuscript development. The first author conducted the primary research, collected and processed disease-specific protein–protein interaction (PPI) data, performed feature extraction, implemented hard and soft GMM-EM clustering, analyzed and validated the experimental results, and drafted the manuscript. The second author provided principal research supervision, guidance on bioinformatics methodology, and critical review of the biological interpretation and validation of protein complexes. The third author contributed to the conceptual framework, computational methodology, statistical analysis, and critical evaluation of the clustering results. The fourth author provided methodological guidance, assisted in the evaluation and interpretation of the experimental results, and contributed to manuscript review and refinement. All authors reviewed and approved the final version of the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest: The authors declare that they have no known financial or personal relationships that could have influenced the research presented in this manuscript.

Generative Artificial Intelligence: The authors declare that generative artificial intelligence was not used in the scientific research process, including data analysis and the formulation of conclusions. If used, AI-assisted tools were limited to language editing and proofreading. The authors take full responsibility for the manuscript's content.

Ethics Approval and Informed Consent:

Not applicable. This study utilized publicly available biological databases and did not involve human participants, animals, or confidential personal data.

Data Availability: The datasets used in this study were obtained from publicly accessible databases, including STRING and CORUM. Supporting research data and computational results are available from the corresponding author upon reasonable request.

ACKNOWLEDGEMENTS

This research was supported by an internal grant from Informatic Engineering, Faculty of Science and Technology, University of Islamic Syarif Hidayatullah. The study was conducted at the School of Data Science, Mathematics, and Informatics, IPB University, which served as the host institution and provided academic and facility support that enabled this work

REFERENCES

- [1] G. Liu, B. Liu, A. Li, X. Wang, J. Yu, and X. Zhou, “Identifying Protein Complexes With Clear Module Structure Using Pairwise Constraints in Protein Interaction Networks,” *Front. Genet.*, vol. 12, no. August, pp. 1–18, 2021, doi: 10.3389/fgene.2021.664786.
- [2] G. Tsitsiridis *et al.*, “CORUM: the comprehensive resource of mammalian protein complexes-2022,” *Nucleic Acids Res.*, vol. 51, no. D1, pp. D539–D545, 2023, doi: 10.1093/nar/gkac1015.
- [3] F. Liang *et al.*, “Recent advances and challenges in protein complex model

- accuracy estimation,” *Comput. Struct. Biotechnol. J.*, vol. 23, no. January, pp. 1824–1832, 2024, doi: 10.1016/j.csbj.2024.04.049.
- [4] N. Zhao, T. Wu, W. Wang, L. Zhang, and X. Gong, “Review and Comparative Analysis of Methods and Advancements in Predicting Protein Complex Structure,” *Interdiscip. Sci. - Comput. Life Sci.*, vol. 16, no. 2, pp. 261–288, 2024, doi: 10.1007/s12539-024-00626-x.
- [5] A. G. Gasic, A. Sarkar, and M. S. Cheung, “Understanding protein-complex assembly through grand canonical maximum entropy modeling,” *Phys. Rev. Res.*, vol. 3, no. 3, 2021, doi: 10.1103/PhysRevResearch.3.033220.
- [6] M. Mustafa *et al.*, “Apoptosis: A Comprehensive Overview of Signaling Pathways, Morphological Changes, and Physiological Significance and Therapeutic Implications,” *Cells*, vol. 13, no. 22, pp. 1–29, 2024, doi: 10.3390/cells13221838.
- [7] H. Oda, A. Annibaldi, D. L. Kastner, and I. Aksentijevich, “Genetic Regulation of Cell Death: Insights from Autoinflammatory Diseases,” *Annu. Rev. Immunol.*, vol. 43, no. 1, pp. 313–342, 2025, doi: 10.1146/annurev-immunol-090222-105848.
- [8] H. Mao, X. Zhao, and S. C. Sun, “NF- κ B in inflammation and cancer,” *Cell. Mol. Immunol.*, vol. 22, no. 8, pp. 811–839, 2025, doi: 10.1038/s41423-025-01310-w.
- [9] Z. Zhang, J. Song, J. Tang, X. Xu, and F. Guo, “Detecting complexes from edge-weighted PPI networks via genes expression analysis,” *BMC Syst. Biol.*, vol. 12, no. Suppl 4, 2018, doi: 10.1186/s12918-018-0565-y.
- [10] J. Wang, J. Liang, W. Zheng, X. Zhao, and J. Mu, “Protein complex detection algorithm based on multiple topological characteristics in PPI networks,” *Inf. Sci. (Ny)*, vol. 489, pp. 78–92, 2019, doi: 10.1016/j.ins.2019.03.015.
- [11] R. Wang, G. Liu, and C. Wang, “Identifying protein complexes based on an edge weight algorithm and core-attachment structure,” *BMC Bioinformatics*, vol. 20, no. 1, 2019, doi: 10.1186/s12859-019-3007-y.
- [12] N. Zaki, H. Singh, and E. A. Mohamed, “Identifying Protein Complexes in Protein-Protein Interaction Data Using Graph Convolutional Network,” *IEEE Access*, vol. 9, pp. 123717–123726, 2021, doi: 10.1109/ACCESS.2021.3110845.
- [13] E. Alvarez-Mamani, R. Dechant, C. A. Beltran-Castañón, and A. J. Ibáñez, “Graph embedding on mass spectrometry- and sequencing-based biomedical data,” *BMC Bioinformatics*, vol. 25, no. 1, pp. 1–19, 2024, doi: 10.1186/s12859-023-05612-6.
- [14] J. Cui, S. Yang, L. Yi, Q. Xi, D. Yang, and Y. Zuo, “Recent advances in deep learning for protein-protein interaction: a review,” *BioData Min.*, vol. 18, no. 1, 2025, doi: 10.1186/s13040-025-00457-6.
- [15] L. Zhu, J. Lei, L. Klei, B. Devlin, and K. Roeder, “Semisoft clustering of single-cell data,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 2, pp. 466–471, 2019, doi: 10.1073/pnas.1817715116.
- [16] M. Chimienti *et al.*, “The use of an unsupervised learning approach for characterizing latent behaviors in accelerometer data,” *Ecol. Evol.*, vol. 6, no. 3, pp. 727–741, 2016, doi: 10.1002/ece3.1914.
- [17] H. Kim, S. W. Hyun, G. Hoogenboom, C. H. Porter, and K. S. Kim, “Fuzzy union to assess climate suitability of annual ryegrass (*Lolium multiflorum*), Alfalfa (*Medicago sativa*) and Sorghum (*Sorghum bicolor*),” *Sci. Rep.*, vol. 8, no. 1, pp. 1–15, 2018, doi: 10.1038/s41598-018-28291-3.
- [18] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, “An efficient algorithm for large-scale detection of protein families,” 2002, doi: 10.1093/nar/30.7.1575.
- [19] G. D. Bader and C. W. V. Hogue, “An automated method for finding molecular complexes in large protein interaction networks,” *BMC Bioinformatics*, vol. 4, 2003, doi: 10.1186/1471-2105-4-2.
- [20] T. Nepusz, H. Yu, and A. Paccanaro, “Detecting overlapping protein complexes in protein-protein interaction networks,” *Nat. Methods*, vol. 9, no. 5, pp. 471–472, 2012, doi: 10.1038/nmeth.1938.

- [21] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, "CFinder: Locating cliques and overlapping modules in biological networks," *Bioinformatics*, vol. 22, no. 8, pp. 1021–1023, 2006, doi: 10.1093/bioinformatics/btl039.
- [22] F. Consonni, E. Gambineri, and C. Favre, "ALPS, FAS, and beyond: from inborn errors of immunity to acquired immunodeficiencies," *Ann. Hematol.*, vol. 101, no. 3, pp. 469–484, 2022, doi: 10.1007/s00277-022-04761-7.
- [23] G. S. Winkler, K. W. Mulder, V. J. Bardwell, E. Kalkhoven, and H. T. M. Timmers, "Human Ccr4-Not complex is a ligand-dependent repressor of nuclear receptor-mediated transcription," *EMBO J.*, vol. 25, no. 13, pp. 3089–3099, 2006, doi: 10.1038/sj.emboj.7601194.
- [24] C. E. Hershberger *et al.*, "HHS Public Access," vol. 35, no. 4, pp. 1108–1120, 2021, doi: 10.1038/s41375-020-1002-y.Complex.
- [25] N. Goossens, X. Sun, and Y. Hoshida, "Molecular classification of hepatocellular carcinoma: potential therapeutic implications," *Hepatic Oncol.*, vol. 2, no. 4, pp. 371–379, 2015, doi: 10.2217/hep.15.26.
- [26] L. Peng, R. Gao, L. Han, T. Fu, and C. Bian, "Roles of β -catenin protein in non-small cell lung cancer (Review)," *Oncol. Lett.*, vol. 30, no. 6, pp. 1–11, 2025, doi: 10.3892/ol.2025.15311.