Contrastive Learning for Gene Set Enrichment Analysis Post-Processing

Leonardo P. A. Biral [1] **Sandeep Dave** [2]

Abstract

Gene Set Enrichment Analysis (GSEA) is one of the most frequently used tools in computational biology, but often returns hundreds to thousands of significant pathways, which makes evaluation time-consuming and limits interpretability. Current GSEA post-processing methods cluster pathways by pairwise gene set overlap, but these approaches fail on the >80% of pathway pairs that share no genes and largely ignore textual annotations. We introduce gtCLIP, the first contrastive learning framework for GSEA post-processing. gtCLIP aligns gene set embeddings from a gene set foundation model with pathway descriptions encoded by PubMedBERT in a shared embedding space, enabling clustering of GSEA result sets into communities of biologically-related pathways. Our key methodological contribution is a soft-target contrastive objective that preserves cross-modal alignment and incorporates gene set overlap, placing biologically related pathways near each other in the embedding space. We evaluated gtCLIP on held-out GSEA results from five blood cancer cohorts, achieving cross-modal retrieval Recall@5 (R@5) of 59.8% and 51.3% on validation and test pathways respectively. On downstream clustering, gtCLIP attained 92.4% mean NES sign coherence as well as 3.3-fold higher within-cluster gene set overlap and 3.9-fold higher silhouette scores compared to the strongest overlap-based baseline. Ablations confirmed the contributions of the soft-target loss, PubMedBERT's biomedical text pretraining, combined pathway title-description input, and foundation encoder fine-tuning. gtCLIP is open-source and available on HuggingFace at `DaveLab/gtCLIP`.

[1]Department of Computational Biology & Bioinformatics, Duke University, Durham, NC, United States [2]Department of Medicine, Duke University, Durham, NC, United States. Correspondence to: Leonardo P. A. Biral <leonardo.biral@duke.edu>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

1. Introduction

Gene Set Enrichment Analysis (GSEA) is one of the most widely used tools in transcriptomics (Subramanian et al., 2005). By testing whether predefined gene sets show concordant differences between two conditions, GSEA quantifies pathway enrichment. However, a single run may return hundreds to thousands of significant results. These large result sets arise because closely related pathways share genes and describe overlapping biological themes, making interpretation time-consuming, subjective, and difficult to reproduce.

Several tools aim to reduce this redundancy. EnrichmentMap (Merico et al., 2010) builds a network of pathways connected by Jaccard Index (JI) or Szymkiewicz–Simpson (SS) gene-overlap similarity and applies clustering to identify related groups. REVIGO reduces redundancy among Gene Ontology terms via semantic similarity (Supek et al., 2011). The current state-of-the-art method, GeneSetCluster 2.0, extends GeneSetCluster 1.0's relative risk (RR) and hierarchical clustering approach for community detection (Ortega-Legarreta et al., 2025; Ewing et al., 2020). While effective, these methods share key limitations: they weigh all genes equally, assign zero similarity to pathway pairs with no gene overlap regardless of potential underlying relatedness, weakly exploit the rich textual information in pathway titles and descriptions if at all, and do not provide an embedding function that generalizes to unseen pathways without recomputing all pairwise similarities.

Contrastive learning provides a natural solution. Inspired by CLIP (Radford et al., 2021), we present gtCLIP (gene set-text CLIP), a framework aligning gene set representations from Gene Set Foundation Model (GSFM) (Clarke et al., 2025) with pathway descriptions encoded by PubMedBERT (Gu et al., 2021). Once trained, gtCLIP maps any gene set or textual description into a shared embedding space, enabling fast, reusable clustering that captures both gene set and semantic similarity. To our knowledge, gtCLIP is the first contrastive model developed for GSEA post-processing.

Our key methodological contribution is a soft-target contrastive loss. Standard CLIP uses InfoNCE loss which effectively uses the identity matrix as the target distribution, forcing each gene set to match only its paired description. However, pathway gene sets partially overlap, providing a direct measure of biological similarity. We blend identity

targets with a SS matrix controlled by a hyperparameter λ, encouraging biologically similar pathways to occupy nearby regions of the embedding space. An auxiliary reconstruction loss additionally trains a decoder to recover gene membership from the shared embedding.

We evaluate gtCLIP zero-shot on GSEA results from five hematological malignancy RNA-seq cohorts, comparing clustering quality against three overlap-based baselines and TF-IDF (Sparck Jones, 1972) at multiple granularities in both gene and text space. We additionally benchmark against GeneSetCluster 2.0 on an external COVID-19 result set. Finally, we summarize clusters via LLM-based annotation rather than the word cloud-based approach used by existing methods. gtCLIP is open-source and available on HuggingFace at `DaveLab/gtCLIP`.

2. Methods

2.1. Pathway corpus

We compiled a pathway corpus of 12,612 gene sets aggregating multiple pathway databases including GO Biological Process (GOBP), Cellular Component (GOCC), and Molecular Function (GOMF), Kyoto Encyclopedia for Genes and Genomes (KEGG), Reactome, WikiPathways, BioCarta, Pathway Interaction Database (PID), and additional curated gene set collections (Liberzon et al., 2011). Each pathway entry consists of a list of unique gene symbols and text composed of the pathway's title and description. Before training, we remove any mention of the pathway's source database to minimize database-dependent bias in the text embedding. We observed extremely sparse pairwise gene overlap, with 84.2% of all pathway pairs sharing zero genes and non-zero overlaps concentrating at low SS coefficients (Figure A1).

2.2. gtCLIP architecture

gtCLIP is a contrastive model mapping gene set embeddings from GSFM to pathway text embeddings from PubMedBERT via multi-layer perceptron (MLP) heads that project each foundation embedding into a shared latent space of dimension d (Figure 2a). These foundation encoders are described in more detail in the Appendix. We fine-tune all GSFM layers and the last 10 of 12 PubMedBERT layers. We also train the projection heads, temperature parameter τ, and an auxiliary gene set decoder MLP that predicts gene membership from the projected gene embedding, helping retain biological signal in the shared space.

At inference, gene sets are passed through GSFM and the gene projection head while pathway text is passed through PubMedBERT and its projection head. Both outputs are ℓ_2-normalized before computing cosine similarity. The cosine similarity matrix is used to construct a k-nearest neighbors (kNN) graph from which we generate interpretable clusters of related pathways using community detection (Figure 2b).

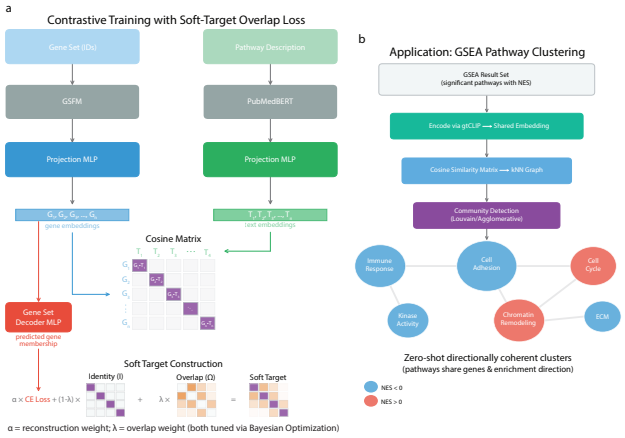

Figure 1. **gtCLIP architecture & implementation** **(a)** Gene sets and pathway text are encoded by GSFM and PubMedBERT and projected into a shared embedding space. The soft target matrix blends the identity and SS matrices weighted by λ. Gene set reconstruction loss is weighted by α. **(b)** Significant pathways from a GSEA result set are encoded with the trained gtCLIP model and clustered, producing biologically-coherent communities.

2.3. Soft-target contrastive loss

InfoNCE treats each gene set-pathway text pair as a unique match. This ignores that biologically similar pathways will likely have overlapping gene sets. We modify the target distribution by incorporating pairwise gene set SS.

For a mini-batch of β gene set-text pairs, let $\mathbf{I}$ be the $\beta \times \beta$ identity matrix and $\mathbf{\Omega}$ be the $\beta \times \beta$ zeroed-diagonal SS matrix. The SS of two gene sets A and B and the soft-target matrix are defined in Equations 1 and 2 respectively.

$$\text{overlap}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{1}$$

$$\mathbf{P} = (1 - \lambda) \cdot \mathbf{I} + \lambda \cdot \mathbf{\Omega} \tag{2}$$

where $\lambda \in [0, 1]$ is the overlap weight hyperparameter. Each row of $\mathbf{P}$ is then normalized to sum to 1, yielding a probability distribution. The contrastive loss is computed as the symmetric Kullback-Leibler (KL) divergence between the softmax of the cosine similarity logits and the soft target distribution:

$$\mathcal{L}_{\text{gene}\to\text{text}} = \left[D_{\text{KL}}\left(\text{softmax}\left(\frac{\mathbf{S}}{\tau} \right) \middle\| \mathbf{P} \right) \right] \tag{3}$$

$$\mathcal{L}_{\text{text}\to\text{gene}} = \left[D_{\text{KL}}\left(\text{softmax}\left(\frac{\mathbf{S}^\top}{\tau} \right) \middle\| \mathbf{P}^\top \right) \right] \tag{4}$$

$$\mathcal{L}_{\text{cont}} = \frac{1}{2}\left[\mathcal{L}_{\text{gene}\to\text{text}} + \mathcal{L}_{\text{text}\to\text{gene}}\right] \tag{5}$$

where $\mathbf{S} = \mathbf{G}\mathbf{T}_{\text{emb}}^{\top}$ is the $\beta \times \beta$ cosine similarity matrix between gene embeddings $\mathbf{G}$ and text embeddings $\mathbf{T}_{\text{emb}}$, and τ is the temperature parameter. The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cont}} + \alpha \cdot \mathcal{L}_{\text{recon}} \tag{6}$$

where $\mathcal{L}_{\text{recon}}$ is the binary cross-entropy between the gene set decoder output and the multi-hot encoded gene set membership and α is the reconstruction weight. Both λ and α are tuned via Bayesian hyperparameter optimization. Training details are provided in the Appendix.

2.4. GSEA-aware data splitting

To ensure zero-shot evaluation, all pathways appearing in any evaluation GSEA result set were fully excluded from contrastive training. We split GSEA results from five hematological malignancy RNA-seq cohorts into validation (3 datasets) and test (2 datasets) sets, yielding 8,943 training, 1,797 validation, and 1,872 test pathways. GSEA post-processing metrics were computed on each result set separately and averaged within each split. We additionally evaluate on the same COVID-19 GSEA result set used by Ortega-Legarreta et al. (2025). Full dataset descriptions and splitting details are provided in the Appendix.

2.5. Baselines & evaluation

We compared gtCLIP to three overlap-based baselines used in existing GSEA post-processing methods: JI, SS, and RR as well as TF-IDF, whose frequency-based gene and term weighting may better capture subtler pathway differences (Sparck Jones, 1972). Each baseline was evaluated using both gene set and text-derived distance matrices across the same agglomerative clustering granularities k as gtCLIP.

We evaluate on two tasks: (1) cross-modal retrieval (R@k and mean reciprocal rank (MRR)) and (2) GSEA result set clustering quality, measured by normalized enrichment score (NES) sign coherence (the fraction of each cluster sharing the same NES sign), within-cluster gene set overlap, and silhouette score. All metrics are reported as averages with 95% confidence intervals over 5 random seeds. Formal metric definitions are provided in the Appendix.

2.6. Cluster summarization

Existing GSEA post-processing methods label clusters using AutoAnnotate, which selects the most frequent words from pathway names (Oesper et al., 2011; Kucera et al., 2016). However, such labels often miss multi-word concepts and overemphasize generic terms. Instead, we generate cluster labels via a zero-shot GPT-5.2 API call (OpenAI, 2025), prompting with the pathway names and top-10 TF-IDF terms for each cluster. These labels are used solely for visualization and have no influence on training or clustering. The prompt template and comparisons of word cloud versus LLM labels are in the Appendix.

2.7. Ablation studies

We performed four ablations to validate design choices: (1) replacing the soft-target loss with InfoNCE, (2) replacing PubMedBERT with general-domain BERT (Reimers & Gurevych, 2019), (3) using only pathway titles or only descriptions as the text input, and (4) freezing both foundation encoders and training only the projection heads (Figure A2).

3. Results

3.1. Cross-modal retrieval performance

gtCLIP achieved strong cross-modal retrieval on held-out pathways with R@5 of 59.8 $\pm$ 1.9% on validation and 51.3 $\pm$ 1.2% on test pathways. MRRs were 42.1 $\pm$ 1.7% and 36.0 $\pm$ 1.3% respectively. The Spearman ρ between pairwise cosine similarity in gene embedding space and gene set overlap was 0.16 $\pm$ 0.0 and 0.21 $\pm$ 0.0 on validation and test sets, indicating the embedding captures gene-level similarity despite the high sparsity of pairwise overlap.

3.2. gtCLIP learns a biologically meaningful embedding

To visualize the learned embedding space, we projected pathway embeddings into two dimensions using UMAP (McInnes et al., 2020).

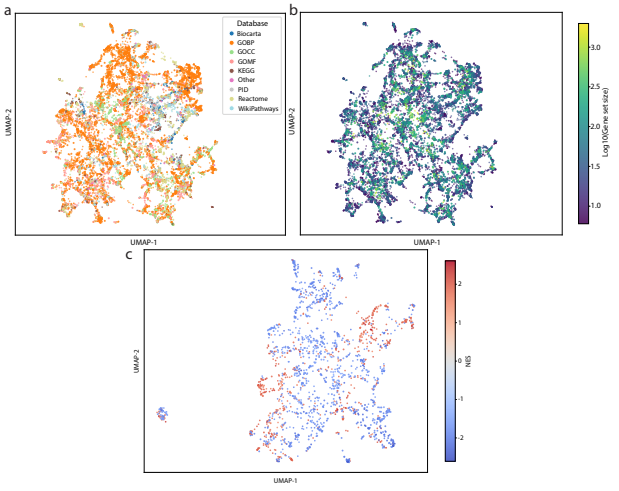

Figure 2. **Embedding space visualization** UMAP projections of all pathway embeddings colored by **(a)** database, **(b)** $\log_{10}$ gene set size, and **(c)** GSEA-evaluated pathways colored by NES value, showing separation of up- and down-regulated pathways.

Coloring by pathway database and gene set size revealed thorough intermixing rather than database or size-specific clusters, indicating gtCLIP learns biologically driven representations rather than database or scale artifacts (Figure 2a,b). When coloring GSEA-evaluated pathways by NES value, positively and negatively enriched pathways separated spatially, suggesting the embedding captures structure correlated with experimental outcomes despite gtCLIP never being trained on NES values (Figure 2c)

3.3. gtCLIP outperforms baselines

In gene space, gtCLIP achieved the highest NES sign coherence at all values of k on the validation set, reaching $95.7 \pm 1.3\%$ at k=25, significantly exceeding the strongest baselines at each metric ($p<0.01$; Figure 3a,c). gtCLIP also achieved significantly higher validation within-cluster gene set overlap and silhouette scores across all k values ($p<0.001$), indicating the learned embedding produces more biologically homogeneous and well-separated clusters than alternative methods. On the test set, gtCLIP NES sign coherence was competitive, but not uniformly the best, reflecting how this metric can be sensitive to small directionally heterogeneous clusters. Importantly, however, gtCLIP continued to yield significantly higher within-cluster gene overlap and silhouette scores at all values of k ($p<0.01$), suggesting the shared embedding more reliably captures shared biology.

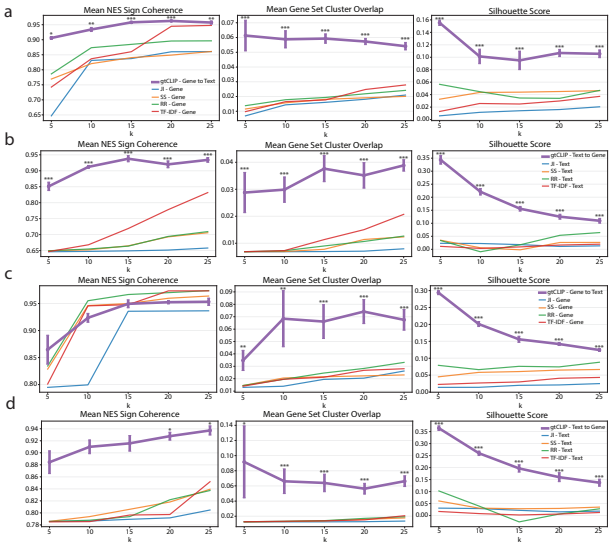

Figure 3. **Baseline comparison on GSEA post-processing** NES sign coherence, within-cluster gene set overlap, and silhouette score across cluster granularities ($k = 5, 10, 15, 20, 25$). **(a)** Validation, gene-space baselines. **(b)** Validation, text-space baselines. **(c)** Test, gene-space baselines. **(d)** Test, text-space baselines.

The advantage of gtCLIP was even more pronounced in text space (Figure 3b,d). Each baseline significantly underperformed gtCLIP in all metrics and granularities on the validation split ($p<0.001$). On the test set, gtCLIP maintained high NES sign coherence, significantly outperforming baselines at k=20, 25 ($p<0.05$) while also improving within-cluster overlap and silhouette scores across all values of k ($p<0.05$). These results highlight gtCLIP's ability to encode biologically meaningful structure into the text embedding space, enabling pathway clustering using descriptions alone. This is particularly useful for newly described pathways or custom gene sets. gtCLIP's generalizability extended to an external COVID-19 GSEA dataset, achieving 1.35× higher within-cluster gene set overlap than GeneSetCluster 2.0 in a fully zero-shot setting (Figure A3). Cluster network visualizations on held-out test datasets are shown in Figures A4 and A5.

4. Discussion

We presented gtCLIP, a contrastive learning framework for GSEA post-processing that aligns gene set and text representations in a shared embedding space via a soft-target loss incorporating gene set overlap coefficients. The resulting clusters exhibit high NES sign coherence, within-cluster gene overlap, and silhouette scores relative to four baselines across all granularities tested.

A central advantage is zero-shot generalization: gtCLIP embeds any pathway without recomputing $O(n^2)$ pairwise similarities as demonstrated on held-out hematological malignancy and COVID-19 GSEA results. The soft-target loss is the key methodological contribution. Our ablation confirmed it achieves comparable retrieval to InfoNCE while significantly improving gene set overlap correlation, indicating it specifically embeds biological similarity structure in the latent space. UMAP analyses and cluster visualizations further indicate gtCLIP clusters pathways by biology rather than database or gene set size artifacts.

The main limitation is evaluation scope: while the COVID-19 analysis demonstrates cross-disease generalization, validation on solid tumors and non-cancer phenotypes would strengthen the claims. Additionally, while our baselines reflect the methods used by existing GSEA post-processing tools, comparison against stronger learned representations such as alternative foundation model embeddings or concatenated encoder features without contrastive alignment would further isolate gtCLIP's contributions.

Promising extensions include incorporating additional modalities such as protein interaction networks or pathway topology into the contrastive objective and applying gtCLIP to multi-condition GSEA results where cluster consistency across experimental conditions is valuable. The shared embedding space may also enable cross-modal retrieval as a discovery tool, for example retrieving candidate gene sets from a natural language description of a biological process. gtCLIP is open-source and available on HuggingFace at `DaveLab/gtCLIP`.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3292500.3330701.

Clarke, D. J. B., Marino, G. B., and Ma'ayan, A. A gene set foundation model pre-trained on a massive collection of diverse gene sets. *bioRxiv*, pp. 2025.05.30.657124, June 2025. doi: 10.1101/2025.05.30.657124. Preprint; posted June 2, 2025.

Ewing, E., Planell-Picola, N., Jagodic, M., and Gomez-Cabrero, D. GeneSetCluster: a tool for summarizing and integrating gene-set analysis results. *BMC Bioinformatics*, 21(443), 2020. doi: 10.1186/s12859-020-03784-z.

Gu, Y., Tinn, R., Cheng, H., et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. doi: 10.1145/3458754.

Kucera, M., Isserlin, R., Arkhangorodsky, A., and Bader, G. D. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. *F1000Research*, 5: 1717, 2016. doi: 10.12688/f1000research.9090.1.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12): 1739–1740, 2011. doi: 10.1093/bioinformatics/btr260.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G. D. Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE*, 5(11):e13984, 2010. doi: 10.1371/journal.pone.0013984.

Oesper, L., Merico, D., Isserlin, R., and Bader, G. D. WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. *Source Code for Biology and Medicine*, 6:7, 2011. doi: 10.1186/1751-0473-6-7.

OpenAI. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/, December 2025. Accessed 2026-03-01.

Ortega-Legarreta, A., Maillo, A., Mouzo, D., et al. GeneSetCluster 2.0: a comprehensive toolbox for gene set analysis combining features of previous tools. *BMC Bioinformatics*, 26:219, 2025. doi: 10.1186/s12859-025-06249-3.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 01 1972. ISSN 0022-0418. doi: 10.1108/eb026526. URL https://doi.org/10.1108/eb026526.

Subramanian, A., Tamayo, P., Mootha, V. K., et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005. doi: 10.1073/pnas.0506580102.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7):e21800, 2011. doi: 10.1371/journal.pone.0021800.

A. Appendix

A.1. Dataset description

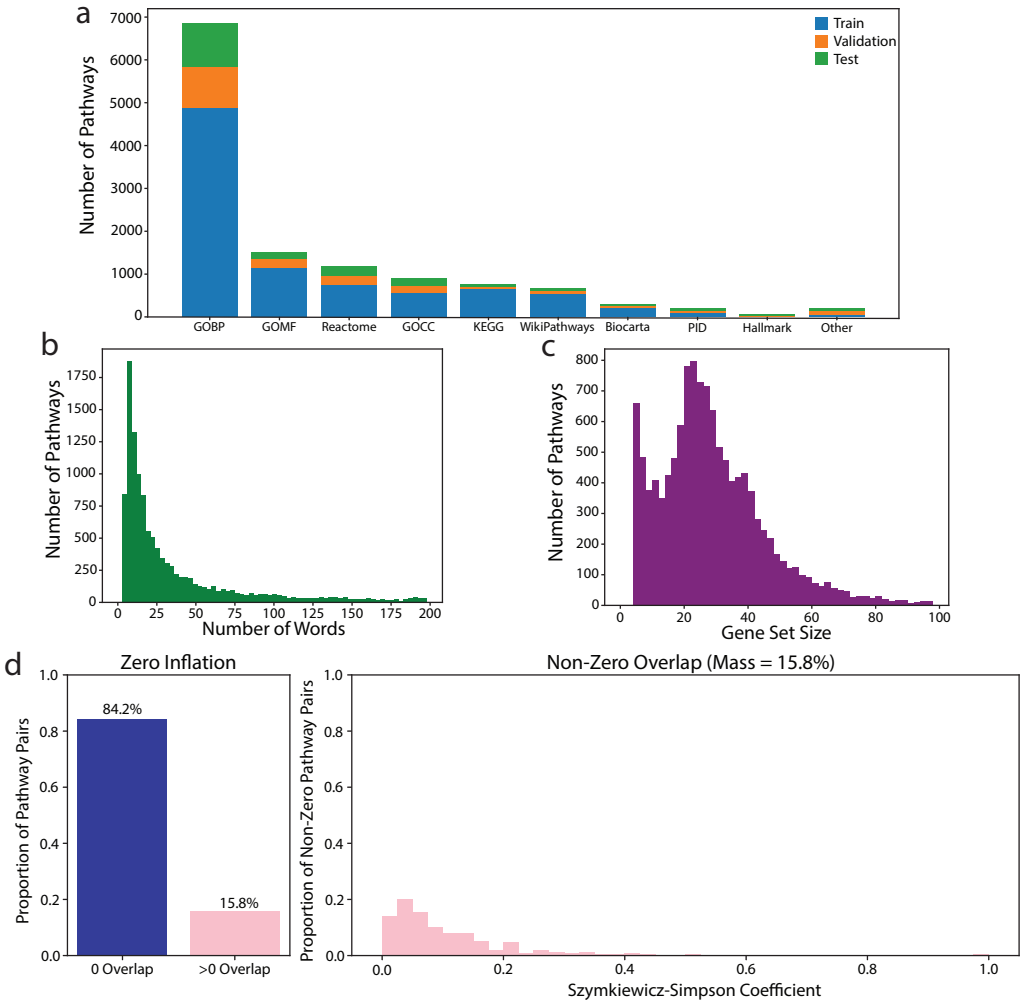

Figure A1. **Pathway corpus & overlap sparsity**
(a) Number of pathways per top database plus a catch-all Other category. **(b)** Distribution of number of words per pathway. **(c)** Distribution of gene set sizes. **(d)** Distribution of pairwise gene SS coefficients. Most pathway pairs have zero overlap (left) and non-zero overlaps concentrate at low coefficients (right).

A.2. Foundation encoders

Gene set embeddings were generated by GSFM, an open-source pretrained gene set foundation model available on HuggingFace at `maayanlab/gsfm`. GSFM takes a set of genes as input and produces a 256-dimensional embedding vector that captures gene-gene co-occurrence patterns learned from large-scale gene set libraries. Pathway text embeddings were obtained from PubMedBERT, a text encoder available on Huggingface at `pritamdeka/S-PubMedBert-MS-MARCO`. PubMedBERT is a BERT model pretrained on a large dataset of PubMed abstracts and biomedical text, producing a 768-dimensional embedding vector. This domain-specific pretraining allows PubMedBERT to more comprehensively encode biological language than a standard English language text encoder.

A.3. GSEA-aware data splitting

To evaluate gtCLIP's ability to embed pathways it has never encountered during training, we designed a GSEA-aware splitting protocol. We obtained GSEA results from five internal hematological malignancy RNAseq datasets: Extranodal

marginal zone lymphoma (EMZL), CD30+ cutaneous T-cell lymphoma (CTCL), primary cutaneous B-cell lymphoma (PCBCL), high-grade B-cell lymphoma (HGBL), and T-cell acute lymphoblastic leukemia (T-ALL). All pathways appearing in any of these result sets were fully excluded from contrastive training to prevent leakage. Each GSEA has a designated reference population in which enriched pathways carry a negative NES (Table A1).

The GSEA result sets were divided into validation (EMZL, CTCL, PCBCL) and test (HGBL, T-ALL) splits. The remaining pathways were used exclusively as the train set. For contrastive learning evaluation, the unique set of pathways in the validation and test splits were used, yielding 8,943 training, 1,797 validation, and 1,872 test pathways. GSEA post-processing evaluation was performed on each validation and test GSEA result set separately and performance was averaged. Additionally, we evaluate gtCLIP on the same COVID-19 GSEA result set used in the GeneSetCluster 2.0 paper (Ortega-Legarreta et al., 2025).

Table A1. Summary of hematological malignancy GSEA result sets.

Set	Split	Reference	# Pathways
EMZL	Gastric vs. ocular site of presentation	Gastric	1857
CTCL	Topical vs. cytotoxic disease therapy	Topical	1731
PCBCL	Leg-type vs. MZL	Leg-type	2307
HGBL	MYC/BCL6 DH vs. MYC/BCL2 DH + TH	MYC/BCL6 DH	1501
T-ALL	Adult vs. child + young adult	Adult	1579

Abbreviations: double-hit (DH), MYC/BCL2/BCL6 triple-hit (TH).

A.4. Hyperparameter optimization

Hyperparameters were optimized with a 70 trial Optuna study maximizing two objectives: R@5 and gene-overlap Spearman ρ (Table A2) (Akiba et al., 2019). We selected our final model configuration from the Pareto front of R@5 vs. gene-overlap Spearman ρ (Table A4).

Table A2. Training configuration & computational details.

Parameter	Value
Maximum epochs	100
Early stopping patience	3 epochs
Best epoch (avg. across seeds)	32 ± 8
Optimizer	AdamW
LR schedule	Cosine annealing with linear warmup
Number of seeds	5
Hardware	NVIDIA A100 40GB
Training time (avg. across seeds)	47.2 ± 10.6 minutes
Framework	PyTorch 2.9.0

A.5. Evaluation metrics

A.5.1. Cross-modal retrieval

For each gene set in the evaluation split, we retrieve the top-k most similar text descriptions (gene-to-text) and vice-versa (text-to-gene) using cosine similarity in the shared embedding space. We report R@k at $k = 1, 5, 10$ and MRR, averaged across both retrieval directions. We additionally report the Spearman correlation ρ between pairwise cosine similarity in the gene embedding space and pairwise gene set overlap coefficients.

A.5.2. GSEA post-processing

For each held-out GSEA result set, we embed all significant (p_{adj} <0.05) pathways using the trained gtCLIP model, construct a cosine similarity matrix, and cluster using agglomerative clustering at multiple granularities ($k = 5, 10, 15, 20, 25$). We note that while gtCLIP embeddings can also be clustered using the Louvain algorithm, we use agglomerative clustering

for our comparison to baselines to ensure we are consistently evaluating the same number of clusters for gtCLIP and the baselines. We report the following clustering quality metrics averaged across datasets within each split:

NES sign coherence: for each cluster C_k, NES sign coherence is the fraction of pathways sharing the majority NES sign:

$$\text{Coherence}(C_k) = \frac{\max\big(|\{i \in C_k : \text{NES}_i > 0\}|,\ |\{i \in C_k : \text{NES}_i < 0\}|\big)}{n_k} \tag{7}$$

The reported mean NES sign coherence is the weighted average across clusters, using cluster size n_k as weights.

Within-cluster mean gene set overlap: for a set of clusters $\{C_1, \ldots, C_K\}$, we compute the mean within-cluster SS as follows. For each cluster C_k containing $n_k \geq 2$ pathways, we calculate all $\binom{n_k}{2}$ pairwise overlap coefficients using Equation 1. The reported mean gene set cluster overlap is the grand mean over all pairs pooled across all clusters:

$$\overline{\text{overlap}} = \frac{\displaystyle\sum_{k=1}^{K} \sum_{(i,j) \in C_k} \text{overlap}(A_i, A_j)}{\displaystyle\sum_{k=1}^{K} \binom{n_k}{2}} \tag{8}$$

This pair-weighted formulation naturally penalizes methods that assign many pathways to a single incoherent cluster, since large clusters contribute quadratically more pairs.

Silhouette score: quantifies cluster separability by comparing, for each sample, its mean distance to other members of its own cluster (cohesion) with its mean distance to members of the nearest neighboring cluster (separation). Scores range from -1 to $+1$, where values near $+1$ indicate tight, well-separated clusters and values near 0 indicate overlapping cluster boundaries. We compute silhouette scores using cosine distance in the embedding space, which measures the angular similarity between vectors.

A.6. Ablation studies

Soft-target loss vs. InfoNCE (Figure A2a) Replacing the soft-target loss with standard InfoNCE did not have a significant impact on retrieval metrics (R@5: 59.8 $\pm$ 1.9% vs. 58.2 $\pm$ 1.1% on validation; p=0.12), but significantly reduced ρ with a mean validation set performance of 0.06 $\pm$ 0.0 (p=3.74e-10). This trend was preserved in the test set with the soft-target loss model achieving significantly higher ρ than the InfoNCE-trained model (p=2.57e-11). This confirmed that using the soft-target loss preserves cross-modal retrieval while more effectively encoding the biological similarity between pathways, which is the primary goal of GSEA post-processing.

PubMedBERT vs. BERT (Figure A2b) Replacing PubMedBERT with general-domain BERT reduced retrieval performance across all metrics. Validation set R@5 dropped significantly from 59.8 $\pm$ 1.9% to 50.6 $\pm$ 1.5% (p=1.36e-5), consistent with PubMedBERT's stronger encoding of biomedical terminology. This significant gap was maintained on the test set (p=1.70e-6), demonstrating domain-specific pretraining provides a generalizable advantage. ρ (not shown) did not significantly differ between the two models, indicating the gene encoder captures overlap structure independently of the text encoder's domain specificity. This suggests the text encoder contributes primarily to cross-modal alignment.

Text input ablation (Figure A2c) Using only pathway names or descriptions significantly reduced retrieval performance compared to the combined input. On the validation set, the full model achieved R@5 of 59.8 $\pm$ 1.9% compared to 29.6 $\pm$ 1.6% (name only) and 30.4 $\pm$ 1.5% (description only) with all differences significant on the test set as well (p<1e-8). This indicates pathway names and descriptions provide complementary, potentially additive information with names offering concise identifiers and descriptions providing richer context, both of which are important for high quality textual encoding.

Encoder unfreezing (Figure A2d) Keeping both encoders fully frozen significantly reduced validation set R@5 from 59.8 $\pm$ 1.9% to 44.3 $\pm$ 2.6% (p=2.31e-6) with a similar reduction on the test set (p=4.98e-7). ρ was not significantly different between the frozen vs. fine-tuned encoders on the validation set, but was significantly higher in the fine-tuned model on test set pathways (p=0.015). This performance gap justifies our decision to fine-tune the pretrained encoders.

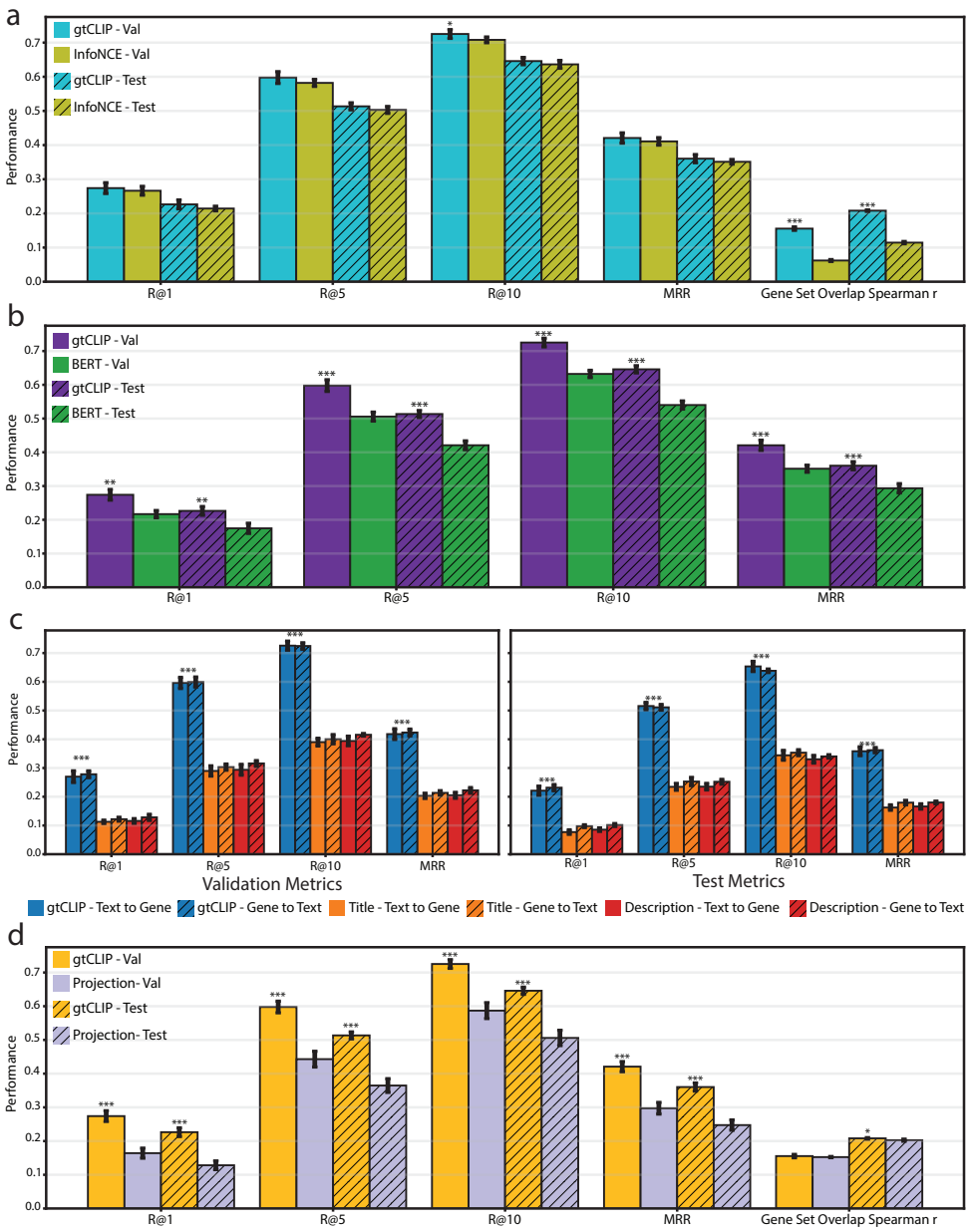

Figure A2. **Ablation comparisons**
(a) Soft-target loss vs. InfoNCE. **(b)** PubMedBERT vs. BERT-base. **(c)** Text input: full vs. name only vs. description only. **(d)** Encoder unfreezing: full model vs. projection-only. * indicates p¡0.05, ** indicates p¡0.01, *** indicates p¡0.001.

A.7. Comparison with GeneSetCluster 2.0 on COVID-19 GSEA data

To evaluate gtCLIP's transferability to unseen GSEA result sets, we applied the trained gtCLIP model to the same COVID-19 GSEA result set used to evaluate GeneSetCluster 2.0, directly comparing gtCLIP's performance to that of GeneSetCluster 2.0 (Ortega-Legarreta et al., 2025). Three GSEA result sets were obtained from a COVID-19 RNAseq study comparing patients at acute infection (119 significant pathways), 3 months post-infection (109 pathways), and 6 months post-infection (11 pathways) to healthy controls. The union of these result sets included 225 unique GOBP pathways which was reduced to 122 after GeneSetCluster 2.0 filtering.

GeneSetCluster 2.0 was run with default parameters, producing 8 clusters from 122 pathways. Gene-to-text gtCLIP embeddings were generated from the provided gene sets. Pathway titles were not provided so the text-to-gene embeddings

were generated solely from pathway descriptions. We also averaged the gene-to-text and text-to-gene embeddings to produce a mean gtCLIP embedding. We clustered these embeddings via Louvain community detection, tuning kNN such that each embedding method produced 8 clusters for consistent comparison to GeneSetCluster 2.0. We only report mean within-cluster gene set overlap as we could not obtain silhouette scores for the GeneSetCluster 2.0 clustering.

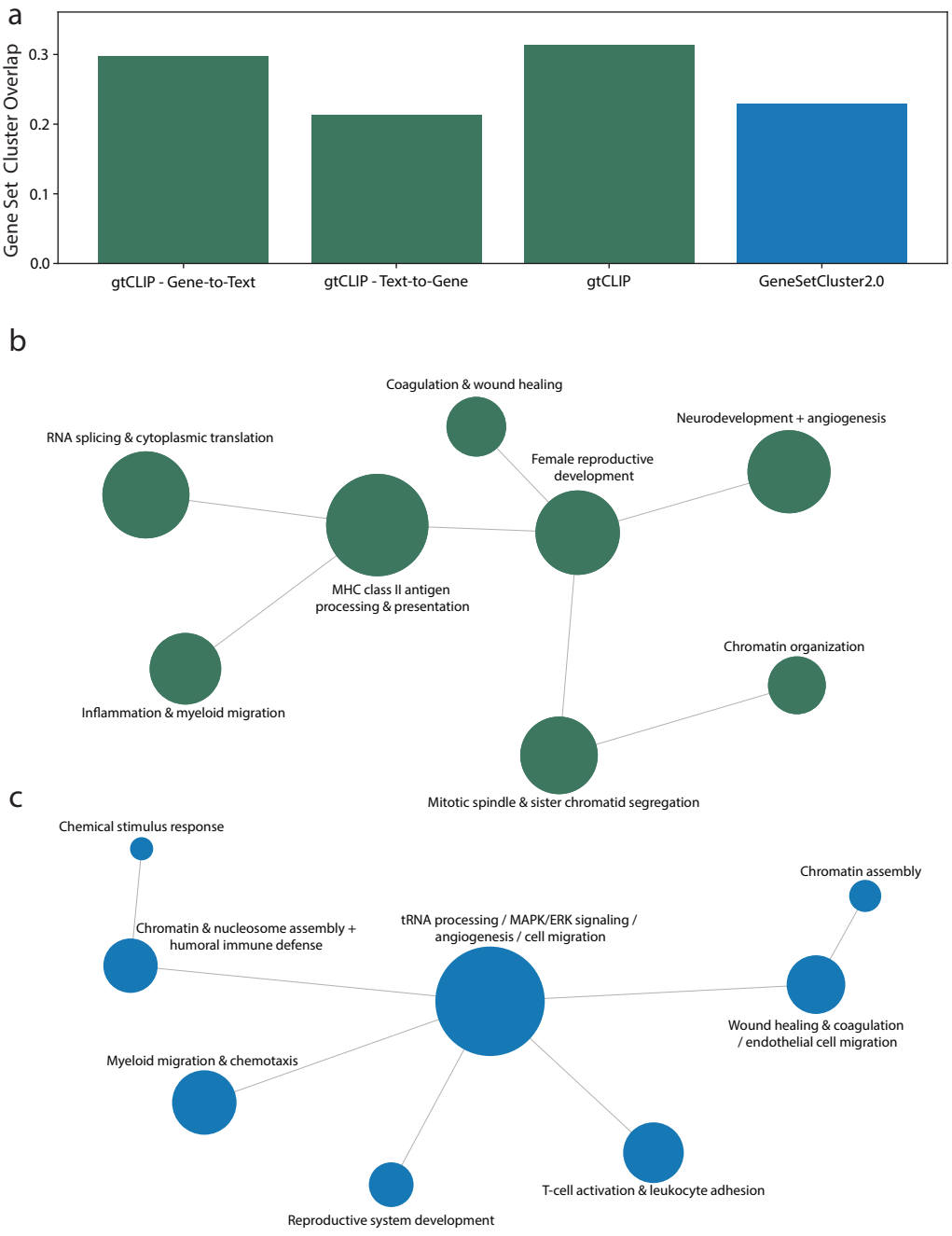

Figure A3. **Comparison of gtCLIP to GeneSetCluster 2.0**
(a) Mean gene set cluster overlap for gtCLIP gene-to-text, text-to-gene, mean, and GeneSetCluster 2.0. Cluster networks from **(b)** mean gtCLIP embeddings and **(c)** GeneSetCluster 2.0. Node size is proportional to cluster size.

We observe gtCLIP with averaged embeddings achieves the highest performance with a mean gene set cluster overlap of 0.31 followed by gtCLIP gene-to-text (0.29), GeneSetCluster 2.0 (0.23), and gtCLIP text-to-gene (0.21) (Figure A3a). The averaged gtCLIP embeddings achieve 1.35× higher within-cluster overlap than GeneSetCluster 2.0. Performance was

identical across random seeds so no error bars are shown. We suspect gtCLIP text-to-gene performance had the lowest performance because pathway titles were missing, limiting the semantic information that was encoded.

The resulting gtCLIP cluster network produces 8 balanced clusters with coherent biological themes such as RNA splicing and cytoplasmic translation, MHC class II antigen processing and presentation, and inflammation and myeloid migration (Figure A3b). GeneSetCluster 2.0 also produces 8 clusters but assigns 50 of 122 pathways (41%) to a single heterogeneous cluster labeled "tRNA processing / MAPK/ERK signaling / angiogenesis / cell migration" (Figure A3c). This cluster contains 4 distinct biological themes and consequently has the lowest within-cluster overlap (0.12), dragging down GeneSetCluster 2.0's overall performance.

Beyond overall performance, the two methods differ markedly in cluster balance. gtCLIP's 8 clusters are more balanced in size, ranging from 15 to 47 pathways with a median of 29 whereas GeneSetCluster 2.0's clusters range from 2 to 50 with a median of 13. Comparing the largest cluster from each method, gtCLIP's (n=47) achieves a within-cluster overlap of 0.28, which is 2.3× higher than the overlap of GeneSetCluster 2.0's largest cluster. This demonstrates gtCLIP maintains biological coherence at scale rather than aggregating unrelated pathways into a single catch-all group. Conversely, GeneSetCluster 2.0 produces three clusters with fewer than five pathways, which offer little interpretive value since they are too small to reveal meaningful, higher-level biological themes. gtCLIP avoids both extremes, producing clusters large enough to be informative yet cohesive enough to represent coherent biology.

A.8. Cluster summarization

A.8.1. LLM PROMPT

```
{
  "System": "For each cluster, analyze the pathways and top-10 TF-IDF terms and generate a
     title that best summarizes the cluster.",
  "Clusters": {
    "Cluster 0": {
      "Pathway names": ["GOBP_HISTONE_MODIFICATION",
        "REACTOME_CHROMATIN_MODIFYING_ENZYMES", ...],
      "Top-10 TF-IDF terms": ["histone", "transcription",
        "complex", "methylation", ...]
    },
    "Cluster 1": { ... },
    ...
  }
}
```

A.8.2. COMPARISON WITH AUTOANNOTATE-STYLE WORD FREQUENCY LABELING

AutoAnnotate is the standard tool for labeling pathway clusters in Cytoscape-based EnrichmentMap workflows (Merico et al., 2010; Kucera et al., 2016). It generates cluster labels by selecting the most frequent non-stopword terms across pathway names within each cluster via the WordCloud app (Oesper et al., 2011). To enable a direct comparison, we reimplemented this word-frequency approach: for each cluster, we tokenized all pathway names, removed standard English stopwords and common uninformative terms (e.g., "pathway," "regulation," "process," "cell," etc.), and selected the top 3 most frequent remaining words.

Table A3 compares the resulting labels in T-ALL. The word-frequency approach produces labels that are often generic or misleading such as "signaling stimulus binding" (cluster 9). This is because high-frequency words like "signaling" dominate multiple functionally distinct clusters (clusters 6, 7, 8, 9, 10). The LLM-generated labels instead capture the specific biological theme, such as "Development / signaling + starvation response" or "Cytokine receptor signaling (IL6-JAK-STAT)." This shows our LLM-based methodology produces far more descriptive and meaningful annotations than current frequency-based approaches such as AutoAnnotate.

Table A3. Cluster annotations and enrichment scores.

ID	n	Word Frequency	GPT-5.2 Label	NES
0	91	complex transcription binding	Chromatin remodeling / histones / transcription control	+1.86
1	62	spindle organization microtubule	Mitotic spindle / microtubules / cytokinesis	+1.82
2	70	metabolic biosynthetic lipid	Lipid & glycan biosynthesis / membrane metabolism	-1.77
3	161	immune activation differentiation	Interferon + MHC / complement (immune activation)	-1.80
4	108	membrane vesicle granule	ER / vesicle membranes + trafficking	-1.79
5	161	cycle dna replication	Cell cycle + DNA replication / E2F-G2M	+2.01
6	56	adhesion signaling matrix	Integrins + ECM	-1.65
7	46	signaling receptor stat	Cytokine receptor signaling (IL6-JAK-STAT)	-1.76
8	26	signaling apoptotic peptidase	Apoptosis + protease/peptidase regulation	-1.72
9	122	signaling stimulus binding	Development / signaling + starvation response	-1.58
10	83	signaling production receptor	Innate inflammation: TLR/NLR + IL6 + NF-κB	-1.77
11	50	up.v1 dn.v1 late.v1	Perturbation/oncogenic signatures (MYC, KRAS/TNFA, RNAi sets)	-1.63
12	30	metabolic oxygen species	Redox & heme metabolism	-1.87
13	30	transmembrane transport transporter	Transmembrane transporters / pH & ion transport	-1.75
14	76	mrna complex splicing	RNA processing: splicing + ribosome + NMD / starvation	+1.79
15	12	migration chemotaxis leukocyte	Leukocyte chemotaxis / migration	-1.67

A.9. Cluster visualization of held-out test datasets

To demonstrate gtCLIP's usage as a GSEA post-processing method, we applied Louvain community detection (kNN=10, resolution=1.0) to the two test GSEA result sets and visualized the resulting cluster networks. For T-ALL, gtCLIP produced 16 clusters with clear biological interpretability (Figure A4). Upregulated clusters in pediatric/adolescent T-ALL included cell cycle and DNA replication (n=161), chromatin remodeling and histone modification (n=91), and mitotic spindle assembly (n=62), consistent with the expected higher proliferation in younger cases. Pathways enriched in adults formed coherent immune-related clusters including leukocyte chemotaxis (n=12), interferon/complement signaling (n=161), and IL6–JAK–STAT signaling (n=46). All but three clusters had NES sign coherence greater than 0.97. The RNA processing cluster (n=76) exhibited the lowest coherence at 0.74, consistent with the heterogeneity of RNA processing pathways that span transcriptional and translational regulation.

For HGBL, gtCLIP recovered many analogous higher-order themes (Figure A5). Adaptive immunity and T-cell differentiation emerged as a distinct cluster, while extracellular matrix (ECM) and epithelial–mesenchymal transition (EMT) themes were prominent. Shared categories across both diseases, such as leukocyte chemotaxis, cytokine signaling, integrin-mediated adhesion, and proteasome-mediated antigen processing, were consistently recovered as coherent clusters. These test result sets highlight the utility of gtCLIP embedding-based clustering in GSEA post-processing.

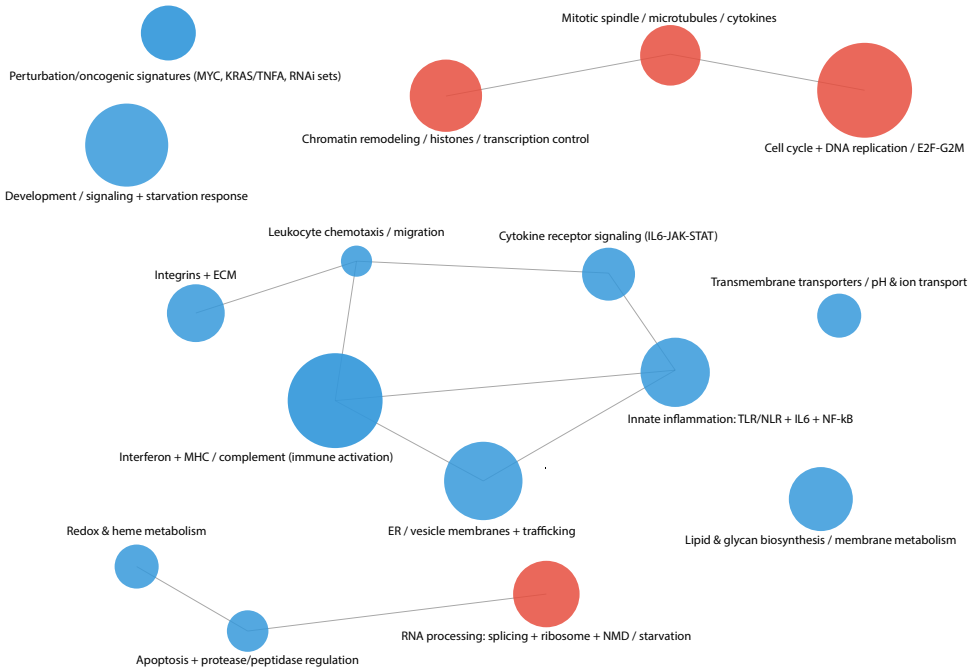

Figure A4. **Cluster network on held-out T-ALL dataset**
Network visualization of gtCLIP-derived pathway clusters for T-ALL. Node size is proportional to cluster size; blue indicates NES < 0 and red indicates NES > 0.

Figure A5. **Cluster network on held-out HGBL dataset**
Network visualization of gtCLIP-derived pathway clusters for HGBL. Node size is proportional to cluster size; blue indicates NES < 0 and red indicates NES > 0.

Table A4. Selected gtCLIP configuration.

Component	Value
Gene set encoder (GSFM)	
Base model	`maayanlab/gsfm`
Input dimension	256
Layers unfrozen	4 (of 4 total)
Projection MLP depth	1
Projection output dim	128
Text encoder (PubMedBERT)	
Base model	`pritamdeka/S-PubMedBert-MS-MARCO`
Input dimension	768
Layers unfrozen	10 (of 12 total)
Projection MLP depth	2
Projection hidden dim	256
Projection output dim	128
Shared space	
Embedding dimension d	128
Temperature τ initialization	0.0154
Temperature parameterization	$\tau = \text{clamp}(\exp(\log \tau), 0.01, 1.0)$
Gene set decoder	
Architecture	2-layer MLP ($128 \to 256 \to \lvert V \rvert = 19400$)
Activation	GELU
Output	Per-gene logit (BCE loss)
Learning rates	
Projection heads	3.35×10^{-4}
PubMedBERT layers	5.67×10^{-5}
GSFM layers	2.04×10^{-5}
Regularization	
Weight decay	1.31×10^{-6}
Dropout	0.15
Loss weights	
Overlap weight λ	0.15
Reconstruction weight α	0.274
Training	
Batch size	64
Warmup steps	450
Mixed precision (AMP)	Yes