LENS: LLM-BASED ENRICHMENT OF NESTED SUBCLUSTERS

Alisha Saboowala[1†], **Ping Wu**[2], **Yogesh Pandit**[1], **David Richmond**[1], **Jan-Christian Huetter**[1], **Avtar Singh**[2], **Vladimir Ermakov**[1†]

ABSTRACT

Recent studies show that large language models (LLMs) can accurately annotate curated gene sets. However, real-world biological data are often noisy, complicating clustering and requiring manual curation. We evaluate the ability of LLMs to refine noisy, geometry-derived gene clusters into biologically meaningful components. We introduce LENS (LLM-based Enrichment of Nested Subclusters), a hybrid framework that combines geometric clustering with LLM-based reasoning to decompose gene clusters into interpretable, nested, and overlapping biological submodules. By leveraging biological information encoded in LLM training corpora, LENS provides a scalable complement to traditional enrichment approaches and improves biological specificity in complex, high-dimensional datasets.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong performance in biological reasoning tasks such as gene set annotation. Prior work shows that LLMs can annotate curated gene sets with high accuracy (Wang et al., 2025; Hu et al., 2024; Yuan et al., 2025). However, these approaches typically produce single annotations and do not capture nested biological structure. Here, we evaluate the ability of LLMs to identify biologically meaningful gene sets under noisy conditions. We apply this approach to published perturbation datasets (Replogle et al., 2022; Funk et al., 2022), showing how geometrically derived gene clusters can be refined via LLM reasoning into nested, interpretable biological components (e.g., processes, pathways, cellular components, or protein complexes). "Nested" denotes hierarchical refinement while allowing overlapping membership to capture genes participating in multiple biological contexts. While Gene Set Enrichment Analysis (GSEA) is widely used, it is limited by predefined gene set libraries (e.g., MSigDB) (Subramanian et al., 2005). We introduce LENS (LLM-based Enrichment of Nested Subclusters), a hybrid framework that combines geometric structure in high-dimensional data with biological knowledge encoded in LLMs. In doing so, LENS improves the efficiency and specificity of biological interpretation and allows LLMs to suggest alternative gene programs that extend beyond well-documented gene sets in curated databases.

2 METHOD

2.1 IMPLEMENTATION

LENS combines geometric structure in high-dimensional experimental data with LLM-based biological reasoning. Input perturbation data are first clustered using a user-specified geometric method (e.g., Leiden (Traag et al., 2019) or HDBSCAN (Campello et al., 2015)), allowing LENS to operate on precomputed clusters appropriate to the dataset and experimental context. From these clusters, LENS extracts gene lists and applies LLM-based annotation to identify biologically meaningful structure, as illustrated in Figure 1. In this work, we use the `gpt-4o` model with temperature set to zero to reduce stochasticity (see Appendix A.2 for model evaluation). Each gene list is provided to the LLM via a structured prompt, which returns annotated gene subclusters. LENS permits overlapping subclusters, allowing genes to participate in multiple biological contexts.

[†]Correspondence: saboowaa@gene.com, ermakovv@gene.com
[1]Computational Sciences-Center of Excellence, Genentech, South San Francisco, CA, USA
[2]Department of Cell and Tissue Genomics, Genentech, South San Francisco, CA, USA

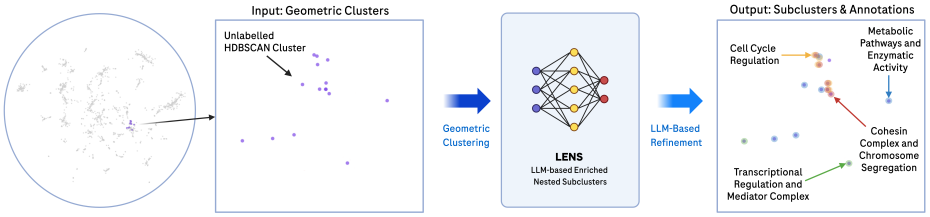

Figure 1: Illustrative example of LENS applied to the Replogle et al. dataset, refining geometric clusters into nested gene subclusters.

2.2 METRICS

Hypergeometric Test and Biological Relevance Score (BRS). To evaluate clustering performance, we assess enrichment for geometric clusters and LLM-derived subclusters using the hypergeometric test (Boyle et al., 2004). Similar to established enrichment analyses (Celik et al., 2022), we use the CORUM database as the reference gene set universe, which provides a curated collection of mammalian protein complexes (Tsitsiridis et al., 2022). The resulting enrichment p-value reflects the likelihood of observing the overlap between a cluster gene set and a reference complex by chance. We define the Biological Relevance Score (BRS) as:

$$\mathrm{BRS} = -\log_{10}\left(\min_{g\in\mathcal{G}} p_g\right), \tag{1}$$

where p_g denotes the enrichment p-value for reference gene set g, and $\mathcal{G}$ is the set of reference gene sets. Higher BRS values indicate stronger enrichment against known protein complexes, providing a quantitative measure of biological specificity.

Precision, Recall, and F1. Gene set clusters are additionally evaluated using precision, recall, and F1 score, computed at the gene level against reference CORUM complexes. For each reference complex, metrics are reported for the LENS-derived subcluster that maximizes F1. For a LENS-derived subcluster and reference complex, true positives correspond to genes shared between the two sets. Precision is defined as the fraction of genes in the subcluster that belong to the reference complex, while recall is defined as the fraction of reference complex genes recovered by the subcluster.

3 EXPERIMENTS

3.1 SYNTHETIC DATA

To evaluate the robustness of LENS, we conducted synthetic experiments using curated CORUM protein complexes with controlled levels of noise. In each experiment, one or more reference complexes were augmented with randomly sampled genes and provided as input to LENS. Performance was assessed by selecting, for each reference complex, the LENS-derived subcluster that maximized F1 score and computing precision, recall, and F1 for that subcluster (mean $\pm$ 95% bootstrap confidence intervals).

Experiment 1: Noise robustness and dependence on functional annotation depth. We first evaluated the ability of LENS to recover individual CORUM complexes under increasing levels of noise. We restricted analysis to 5-gene complexes and stratified them by CORUM functional annotation depth, measured by the number of PubMed references supporting curated functional annotations (`functions_pmid`). Within this set, complexes were split into top and bottom prominence groups. For each complex, an increasing number of random genes were added before applying LENS; recovery performance decreases monotonically with noise but remains substantially higher for highly annotated complexes (Figure 2A). Figure 2B shows that the number of inferred subclusters increases with noise and saturates at approximately 15–20 clusters. At high noise levels, inferred subclusters increasingly correspond to broader biological programs (e.g., cell cycle or transcriptional regulation), reflecting aggregation of the target complex with functionally related genes rather than loss of core complex members.

Experiment 2: Robustness across complex sizes under proportional noise. Next, we assessed whether recovery performance depends on complex size when signal-to-noise ratio is controlled. We selected highly annotated CORUM complexes across multiple size buckets and added random genes in proportion to complex size (10× the number of complex genes). As shown in Figure 2C–D, precision, recall, and F1 scores are largely consistent across size buckets, indicating that recovery performance is governed primarily by relative noise level rather than absolute complex size.

Experiment 3: Recovery of multiple complexes under noise. Finally, we evaluated LENS in settings containing multiple CORUM complexes simultaneously. We selected the top N most functionally annotated complexes, combined their gene sets, and added random genes at a fixed 5× proportional noise ratio. As N increased from 1 to 19, recall remained relatively high while precision decreased, leading to a gradual reduction in F1 score (Figure 2E). The precision drop reflects expansion of inferred clusters to include additional functionally related genes. Figure 2F shows that the fraction of complexes recovered with recall ≥ 0.5 remains above 70% until the inferred subcluster count approaches the saturation regime, after which recovery performance declines.

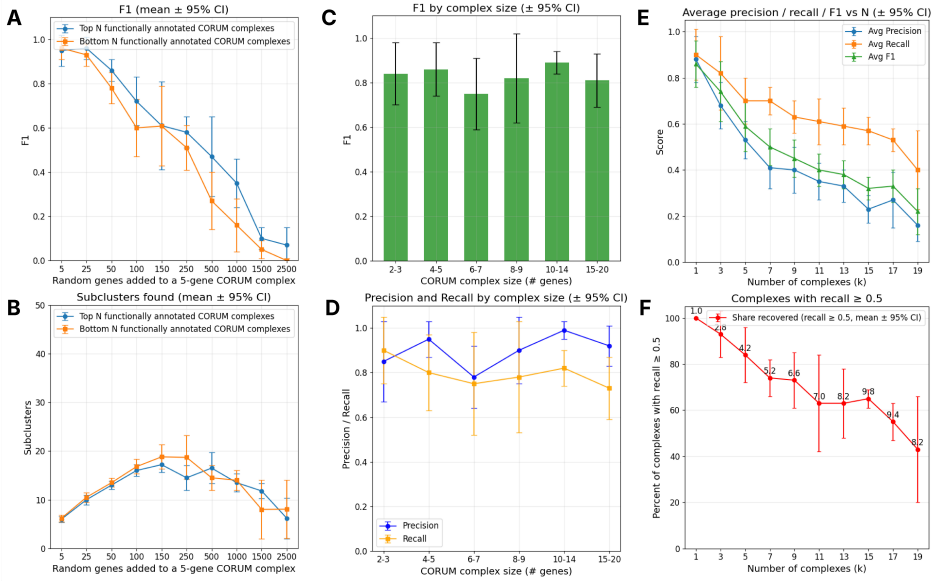

Figure 2: Evaluation of LENS in synthetic experiments with controlled noise and task complexity. (A) F1 score as a function of the number of added random genes for 5-gene CORUM complexes, stratified by functional annotation depth (top vs bottom CORUM annotation groups). (B) Number of inferred LLM subclusters as a function of added random genes for the same experiment as in (A), showing saturation at approximately 15–20 clusters. (C) F1 score across CORUM complex size buckets under proportional noise (10× complex size), restricted to highly annotated complexes. (D) Precision and recall corresponding to (C), demonstrating similar recovery performance across complex sizes when signal-to-noise ratio is controlled. (E) Average precision, recall, and F1 score as a function of the number of combined CORUM complexes (N), with random genes added at a fixed 5× proportional noise ratio. (F) Fraction of complexes recovered with recall ≥ 0.5 in the multi-complex setting, showing sustained recovery until subcluster saturation is reached. Values show mean ± 95% bootstrap confidence intervals.

3.2 Real-world Perturbation Datasets

To evaluate LENS on real biological data, we applied it to two perturbation datasets with existing geometric cluster annotations: genome-scale CRISPR perturbations (Replogle et al., 2022) and optical pooled screening of essential human genes (Funk et al., 2022). In both cases, LENS was applied to precomputed geometric clusters and evaluated using the Biological Relevance Score (BRS).

Genome-scale CRISPR perturbations (Replogle et al.). The Replogle et al. dataset consists of CRISPRi-based Perturb-seq profiles across 2.5 million K562 cells, embedded and clustered using HDBSCAN with manual annotation by the original authors. Applying LENS to these clusters produces refined subclusters with increased biological specificity. As shown in Figure 3B, 85.9% of geometric clusters exhibit higher BRS when considering the maximum score among LENS-derived subclusters, indicating improved biological relevance after LLM-based refinement.

Optical pooled screening of essential genes (Funk et al.). The Funk et al. dataset profiles phenotypes of 5,072 essential genes in HeLa cells, clustered using Leiden after dimensionality reduction and annotated by domain experts. LENS refines these geometric clusters into subclusters that correspond more closely to protein complexes and pathways. Figure 3A illustrates representative examples in which LENS-derived subclusters achieve higher BRS and more specific biological labels than the original cluster annotations. Across the full dataset, 93.2% of geometric clusters show increased BRS after LENS refinement (Figure 3C).

Together, these results are consistent with the hypothesis that LENS refines geometry-derived clusters into biologically specific and interpretable subclusters across diverse perturbation modalities.

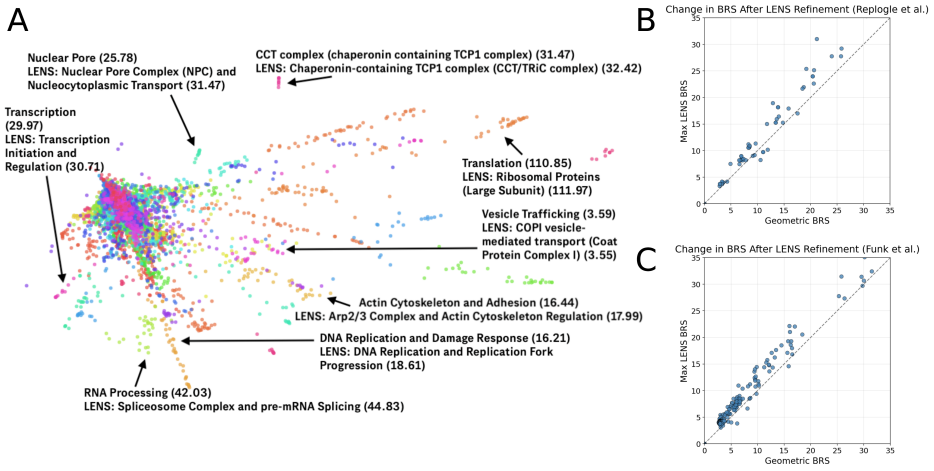

Figure 3: LENS refinement improves biological relevance in real-world perturbation datasets. (A) Representative examples from the Funk et al. optical pooled screening dataset. For each geometric cluster, the original annotation reported by Funk et al. is shown alongside the best-matching LENS-derived subcluster (selected by maximum Biological Relevance Score, BRS). LENS-derived subclusters consistently yield more specific biological labels and higher BRS values. (B) Replogle et al. dataset: change in BRS following LLM-based refinement. Each point represents a geometric cluster, plotted by its original BRS (x-axis) and the maximum BRS among LENS-derived subclusters (y-axis). (C) Funk et al. dataset: analogous analysis showing changes in BRS after LENS refinement for gene-level phenotypic clusters derived from optical pooled screening. Positive values indicate increased biological specificity relative to the original geometric clusters.

4 Conclusion

LENS integrates geometric clustering with LLM-based biological reasoning to refine gene clusters into interpretable, nested submodules. We validate this framework using synthetic experiments and large-scale perturbation datasets, demonstrating robust recovery of core biological structures and improved biological specificity under noisy conditions. Future work will focus on more systematic analysis of how LENS-derived subclusters depend on clustering resolution and input perturbations, as well as on examining genes that are weakly associated with or unassigned to any subcluster, which may highlight under-characterized or novel biological roles. Beyond single-gene perturbations, LENS may be applied to other modalities, such as chemical perturbations or cell-type–specific responses, further expanding its utility for scalable biological discovery.

REFERENCES

Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched GO terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004. doi: 10.1093/bioinformatics/bth456.

Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Joerg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):1–51, 2015. doi: 10.1145/2733381.

Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Nathan H. Lazar, Rahul Mohan, Conor Tillinghast, Tommaso Biancalani, Marta Fay, Berton A. Earnshaw, and Imran S. Haque. Biological cartography: Building and benchmarking representations of life. *bioRxiv*, pp. 2022.12.09.519400, 2022. doi: 10.1101/2022.12.09.519400.

Luke Funk, Kuan-Chung Su, Jimmy Ly, David Feldman, Avtar Singh, Brittania Moodie, Paul C. Blainey, and Iain M. Cheeseman. The phenotypic landscape of essential human genes. *Cell*, 185 (24):4634–4653, 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.10.017.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T. Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *Nature Methods*, 22(1):54–65, 2024. doi: 10.1038/s41592-024-02525-x.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575, 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.05.013.

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. 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.

V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. doi: 10.1038/s41598-019-41695-z.

George Tsitsiridis, Ralph Steinkamp, Madalina Giurgiu, Barbara Brauner, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Research*, 51(D1):D539–D545, 2022. doi: 10.1093/nar/gkac1015.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. GeneAgent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(1):45–53, 2025. doi: 10.1038/s41592-025-02748-6.

Ying Yuan, Xing-Yue Monica Ge, Aaron Archer Waterman, Tommaso Biancalani, David Richmond, Yogesh Pandit, Avtar Singh, Russell Littman, Jin Liu, Jan-Christian Huetter, and Vladimir Ermakov. HypoGeneAgent: A hypothesis language agent for gene-set cluster resolution selection using Perturb-seq datasets. *arXiv preprint arXiv:2509.09740*, 2025. doi: 10.48550/arXiv.2509.09740.

LLM USAGE DISCLOSURE

Large language models were used to assist with manuscript wording, editorial review, and limited research support, including code drafting and background summarization. All code, scientific content, analyses, and conclusions were verified and validated by the authors, who assume full responsibility for this work.

A Appendix

A.1 LENS Prompt

```
System: "You are a bioinformatics expert. Analyze the given list of genes
and identify one or more general groups of related genes. These groups can
include (but are not limited to): cellular components (complexes, organelles,
subcellular structures), biological pathways (signaling pathways, metabolic
pathways, regulatory pathways), biological functions (molecular functions,
biological processes), disease associations, or any other biologically
meaningful groupings. Groups can overlap - the same gene may belong to
multiple groups. Return a JSON array where each element is an object with
two fields: 'group_name' and 'genes'."

Human: "Given the following list of genes, identify one or more general
groups of related genes (such as cellular components, pathways, biological
functions, disease associations, etc.) and which genes from this list are
related to each group. Groups can overlap - the same gene may appear in
multiple groups if it participates in multiple biological contexts. Return
a JSON array of objects, where each object has 'group_name' and 'genes'
fields. You may return multiple groups if the gene list suggests multiple
distinct biological associations.

Genes: {gene_list}"
```

A.2 LLM Evaluation for LENS

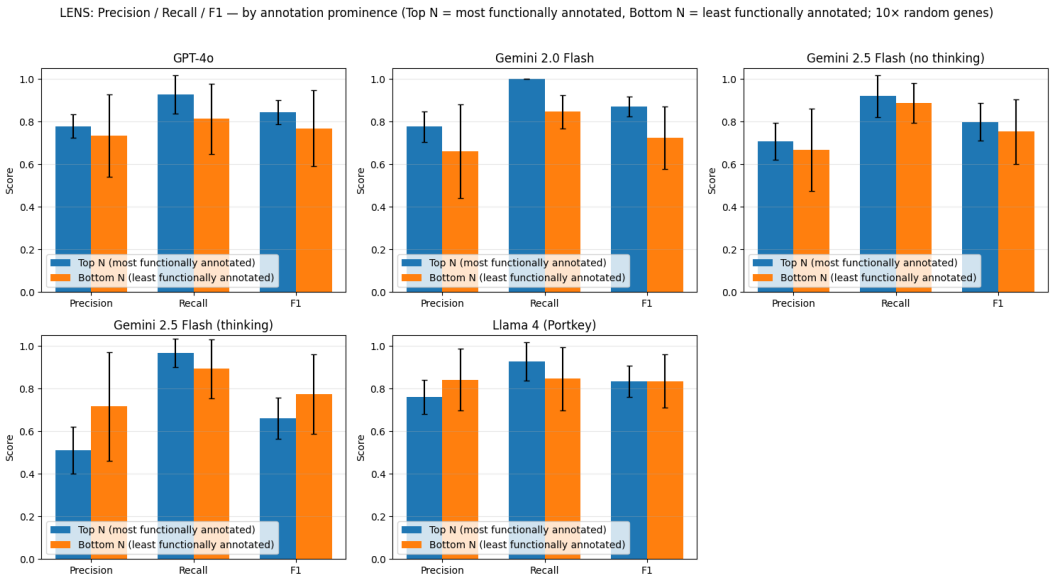

Figure 4: Performance of LENS across large language models stratified by CORUM functional annotation depth. Precision, recall, and F1 score are shown for recovery of single CORUM complexes augmented with proportional noise (10× random genes), comparing the top N most functionally annotated complexes and the bottom N least functionally annotated complexes (ranked by the number of PubMed references supporting curated functional annotations, `functions_pmid`). Results are aggregated over 10 trials per model; error bars indicate $\pm$ 1 standard deviation.

A.3 LENS RESULT FOR REPLOGLE ET AL. DATASET

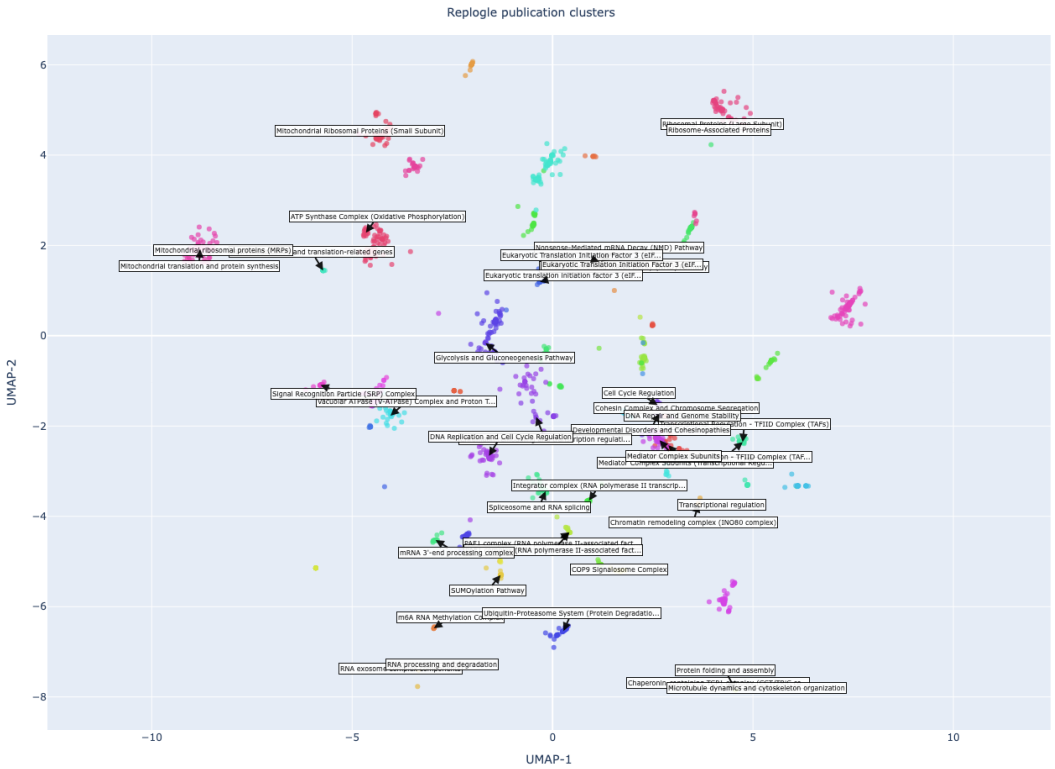

Figure 5: Replogle et al. dataset overview (LENS output). Global UMAP embedding of geometric clusters derived from Perturb-seq gene-level perturbation profiles. Annotations shown correspond to LENS-derived gene set subclusters containing at least five genes and exhibiting $\geq 70\%$ overlap with reference gene sets from MSigDB or CORUM.