A Multi-Modal Foundation Model Across Species for Interpreting Gene Functions

Tianyu Liu, Gefei Wang, Yu Li, Wengong Jin, Hongyu Zhao
Yale University, Broad Institute, The Chinese University of Hong Kong, Northeastern University
tianyu.liu@yale.edu

Abstract

Artificial Intelligence shows impressive performances in computational biology, especially in modelling DNA sequences and processing biomedical text annotations. To interpret the contributions of modality representations in learning and predicting patterns in genomics and genetics, we develop DNACLIP and train it with paired DNA sequences and text descriptions from over 300,000 genes across 24 species, to model text and DNA sequences jointly and perform cross-species gene functional analysis. Through extensive benchmarking analysis, we show the unique contributions of aligned gene embeddings and text embeddings in various downstream applications, including gene clustering, gene annotation, disease risk prediction, function prediction, perturbation prediction, and expression prediction, etc. We also use DNACLIP to discover disease-specific gene programs from atlas data. Finally, we discuss the dominant areas of modality-specific embeddings and provide guidelines for users to select embeddings based on their requirements.

1 Introduction

DNA sequences are the foundation for heredity and evolution, mainly based on the functions of transcripts or genes [9]. Specifically, the region with genetic signals plays an important role in determining cellular structure, function, and fate, through transcription and translation [21, 31]. Therefore, interpreting the function of genes is essential for us to understand the meaningful context behind the complex biological sequences and processes, and address representative tasks such as genome annotation [27], gene representation [17, 18], gene-phenotype association [10], and others.

In this work, we develop a multi-modal foundation model, named DNACLIP, by leveraging the pre-trained information from both gLMs and LLMs, to generate better sequence embeddings for gene representation. By collecting the DNA sequences and corresponding functional annotations from NCBI [25] and UniProt [8], we build a large-scale text-sequence dataset to pre-train DNACLIP. Based on a contrastive learning framework known as CLIP [23], we learn a new sequence embedding enriched with text annotation, and a new text embedding enriched with information about biological sequences, and align their embeddings into a joint space for bridging the gap between modalities. Such alignment relationship can be naturally used to generalize functional annotation for unseen genes or poorly annotated genes in different species. Compared with the classical sequence alignment algorithm BLAST (basic local alignment search tool) [3], our method takes the overall dependency into consideration, and can also learn the change of sequences with low similarity. Finally, we demonstrated that incorporating gene representations from DNACLIP with other domain-specific models can also empower their performances in specific tasks, such as predicting perturbation effects and gene expression levels based on single-cell transcriptomics. We believe that our method is an effective way at the stage of building foundation models for DNA sequence modeling and further unlocks the power of using deep learning to study genomes.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

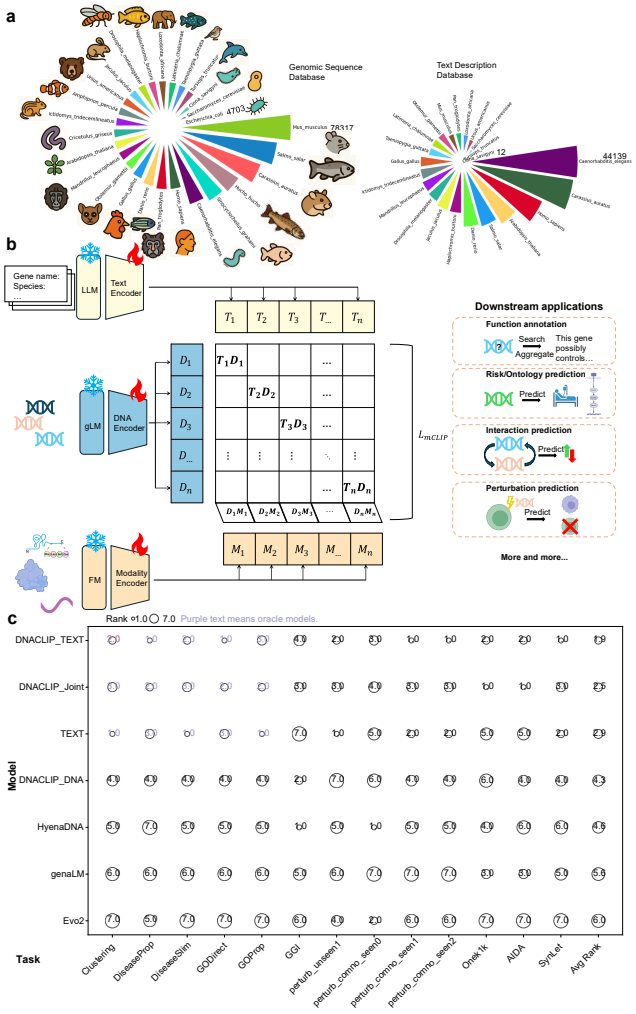


Figure 1: Overview of large-scale DNA-TEXT paired datasets and model architecture. (a) Statistics of DNA sequences and text annotations of genes across different species. We highlight the largest and smallest number. (b) Model architecture of DNACLIP. Here we train a CLIP model [23] for modalities including text (T) and DNA sequence (D), and it also accepts other modalities (M) as inputs, and the combination of modalities is determined based on ablation studies. DNACLIP supports various downstream applications, including function annotation, disease risk/gene ontology prediction, gene-level interaction prediction, and perturbation prediction, etc.

2 Results

Dataset Construction. We first constructed a multi-modal gene (DNA sequence)-annotation (text) dataset to train DNACLIP. Our DNA sequences are extracted from 24 species covering mammals, birds, fish, microorganisms, and others. The illustration of data distribution and statistics is summarized in Figure 1 (a). To collect information on DNA sequences, we downloaded the gtf/gff3 files to access genome information of each gene, including position in the chromosome, starting site, and ending site from Ensembl [13]. We then extracted the sequences from the reference genomes of these species. To collect information on gene annotation, such as gene name and functional summary, we retrieved the gene information from NCBI [26] based on authorized API. Through our analysis of the data structure, we identified that Mus musculus has the largest number of identified genes and Escherichia coli has the smallest. Also, we found that the labeling of genes from different species has a largely imbalanced distribution. The genes of some species lack functional annotation and exhibit the characteristics of low-resource data.

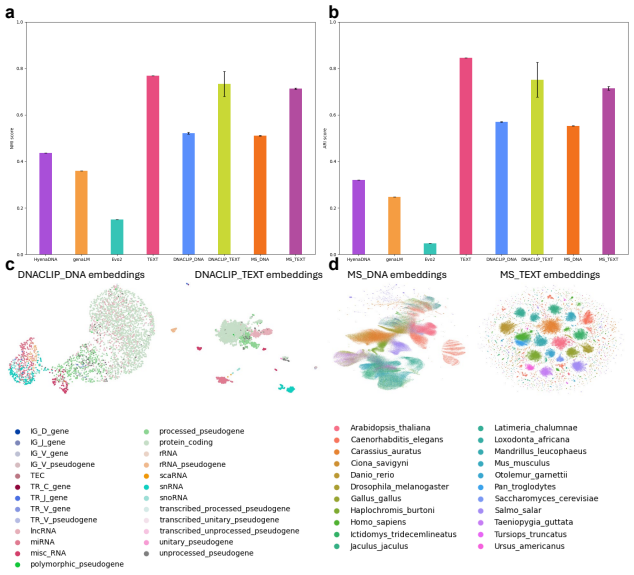


Figure 2: Evaluation results of gene function clustering. We report the standard deviation (SD) for trained baseline methods with five different random seeds (HyenaDNA and TEXT are from pre-trained models and thus their SD is 0). (a) NMI scores across gene embeddings from different methods based on functional labels. (b) ARI scores across gene embeddings from different methods based on functional labels. (c) UMAP visualization of gene embeddings from DNACLIP colored by functional labels. (d) UMAP visualization of gene embeddings from MS_DNACLIP colored by species.

Method Overview. The DNA sequence embeddings of genes are generated based on HyenaDNA [20], which has a long context window and shows strong performance across some baseline analyses. We also include other genomic language models (gLMs), such as genaLM [12] and Evo2 [5], as baseline methods. The text embeddings of gene annotations are generated based on the OpenAI embedding model [1]. The most basic form of training DNACLIP is to learn a sequence projector and a text projector based on gene pairs from two modalities. Moreover, DNACLIP can also accept input from more than two modalities, which serves as a flexible multi-modal integration and alignment framework. The overview of DNACLIP is illustrated in Figure 1 (b), with the corresponding downstream applications including gene-level functional annotation, disease risk gene prediction, gene ontology prediction, and perturbation prediction, etc.

DNACLIP better encodes and represents gene functions. By training DNACLIP, the sequencing data can leverage the denoised gene function information from text embeddings and thus the embeddings from DNA sequences can be improved over the raw sequence embeddings generated by HyenaDNA or other gLMs. To demonstrate this, we downloaded the classes of human gene functions (such as protein-encoding and non-coding RNA) from [28] and performed clustering analysis based on the gene embeddings in the testing set. Higher clustering metrics, such as normalized mutual information (NMI) and Adjusted Rand Index (ARI) [22], mean better gene representations.

Figures 2 (a) and (b) show our clustering performances with gene embeddings from different sources, annotated with NMI and ARI scores. The results from DNACLIP are repeated with runs from five different random seeds, to investigate the robustness. For the embeddings from DNA sequences, we found that DNACLIP achieved the best performances, with clear improvement over other baselines. Furthermore, as we expected, the text embeddings had the highest clustering metrics, followed by the results from the text projector. However, embeddings from the text projector did not lose much information, shown by the slight decrease in clustering scores, but its dimension was reduced (3072 to 128) and thus DNACLIP could save resources for generating gene representations efficiently. Including multi-species information did not affect the clustering performance. Since the human genes are better studied and have better annotation quality, including text descriptions across different species did not lead to an obvious performance drop, which implies that the functional annotation

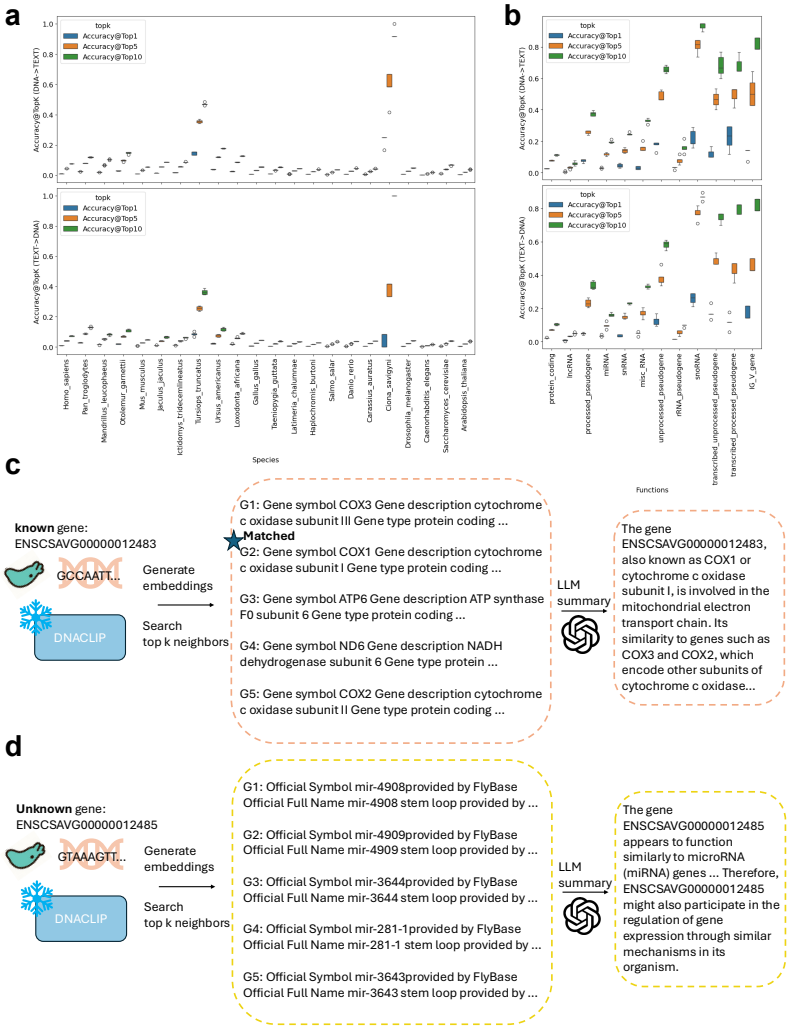


Figure 3: Results of functional annotation with DNACLIP. (a) Retrieval accuracy at different levels across different species based on the training and validation datasets. (b) Retrieval accuracy at different levels across different gene functions based on the testing dataset. (c) An example of using DNACLIP and LLMs to reproduce known gene functions with DNA sequence embeddings. (d) An example of using DNACLIP and LLMs to annotate unknown gene functions with DNA sequence embeddings.

based on the similarity retrieval for under-explored genes is practical. Furthermore, we visualize the gene embeddings with UMAP [19] in Figures 2 (c) (only testing set) and (d) (all paired genes), which shows that embeddings from both projectors can have clear separation for most genes colored by functional annotation and species.

The related hyper-parameters in DNACLIP training are tuned to the best performance, as discussed in the Methods section and Extended Data Figures 4 (a)-(c). Here we found that higher temperature values in contrastive learning and dimension could yield a better sequence projector to distinguish genes from different functional groups. Reducing the temperature could generate better gene embeddings from the text projector. We also performed ablation studies to investigate the contributions of different modalities and training frameworks. According to Extended Data Figures 5 (a) (NMI score) and (b) (ARI score), fine-tuning the gLM (using HyenaDNA as the backbone model) does not improve the ability of embeddings in representing gene functions, and having protein embeddings generated by ESMC of corresponding genes also does not improve the clustering performance.

Annotating unknown gene functions with similarity-based retrieval. As we previously mentioned, one of the basic but powerful functions of contrastive learning models is retrieval. The contribution of DNACLIP serves as an annotation tool for low-resource gene annotation information in the under-explored species' genomes. To evaluate the capacity of DNACLIP in sequence-text retrieval, we considered two directions, which include searching the corresponding text description based on sequence (DNA→TEXT), and searching the corresponding sequence based on text description (TEXT→DNA), across different species and gene functions. We also selected Accuracy@TopK (K$\in (1, 5, 10)$), which computes the accuracy based on top-K candidates ranked by similarity for the given sample, to perform evaluation.

Figure 3 (a) shows the retrieval results across different species based on training and validation datasets, which demonstrates that DNACLIP can match accurately in species such as Ciona savigyni and Tursiops truncatus. Moreover, we also observed imbalanced performances across different species, for example, the annotation accuracy in Homo sapiens remains at a relatively low level across different K. To explain this result, we also show the retrieval results based on the testing dataset in Figure 3 (b), which only contains genes from Homo sapiens but with different functional annotations. We ranked the functional annotations based on the number of genes in this functional class (as shown in the figure, we have the largest number of genes with protein coding information and the smallest number of genes belonging to the category IG_V), while as the number of genes decreases, the performance increases. Therefore, the functional similarity of genes makes it impossible to capture 100% of the correspondence between sequence and text description, which implies that DNACLIP learns the correct biological patterns by keeping the within-group similarity in DNA sequence modelling and does not overfit. Such result is also observed in CLIP models trained with other biomedical data, shown in [6], with a similar level of accuracy.

To provide a sanity check, we show an example of reproducing the known gene function annotation by selecting the gene *ENSCSAVG00000012483* (also known as *COX1*), which has the text description provided by NCBI. Here we searched its top 5 neighbors in the space of text embeddings with DNACLIP, shown in Figure 3 (c). This figure shows that we successfully capture the matched text annotation for this gene in its neighborhood list and we further summarized the descriptions from these candidates with LLMs such as GPT-4o [15]. GPT-4o also provided the correct functional description for this gene and linked this gene with other genes, such as *COX3* and *COX2*, which encode other subunits of cytochrome c oxidase, suggesting its role in catalyzing the reduction of oxygen to water, a critical step in cellular respiration.

To illustrate the functionality of DNACLIP in annotating unseen genes, we selected the gene *ENSCSAVG00000012485* in Ciona savigyni, encoded it with DNA sequences to generate gene representations, and searched its top 5 neighbors in the space of text embeddings, shown in Figure 3 (d). The discovered candidates in Drosophila are all in the same class, known as microRNA, and contain functional annotations including biological mechanisms that they involve. By summarizing these candidates with GPT-4o [15], we can produce the functional summary of this gene. MicroRNA samples typically function by binding to complementary sequences in target messenger RNAs (mRNAs) to inhibit their translation or promote their degradation, often as part of the RNA-induced silencing complex (RISC). Therefore, the gene *ENSCSAVG00000012485* might also participate in the regulation of gene expression through similar mechanisms in its organism. Interestingly, it has been shown that synteny and colinearity patterns in Ciona are much more similar to those found in the clades of Caenorhabditis and Drosophila [14], so the annotations of DNACLIP may have genome-level interpretability.

3 Discussion and Conclusion

Here we pre-train DNACLIP with DNA sequences and text description pairs of genes from different species. The initial representations of DNA sequences are generated by genomic language models, while the initial representations of texts are generated by large language models. The paired sets of vectors are fed into the CLIP model and transformed into modality-specific embeddings based on sequence and text encoders as well as the modality-alignment-specific loss functions. Using a series of ablation studies, we demonstrate the better performance of our framework that can balance the quality of the embeddings generated by the two encoders to represent the function of the gene.

References

[1] Openai embedding models. `https://platform.openai.com/docs/guides/embeddings/embedding-models`. Accessed: 2025-07-20.

[2] Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

[3] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[5] Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pages 2025–02, 2025.

[6] Weiqing Chen, Pengzhi Zhang, Tu N Tran, Yiwei Xiao, Shengyu Li, Vrutant V Shah, Hao Cheng, Kristopher W Brannan, Keith Youker, Li Lai, et al. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, pages 1–15, 2025.

[7] Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature biomedical engineering*, 9(4):483–493, 2025.

[8] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1): D204–D212, 2015.

[9] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[10] Qile Dai, Geyu Zhou, Hongyu Zhao, Urmo Võsa, Lude Franke, Alexis Battle, Alexander Teumer, Terho Lehtimäki, Olli T Raitakari, Tõnu Esko, et al. Otters: a powerful twas framework leveraging summary-level reference data. *Nature Communications*, 14(1):1271, 2023.

[11] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.

[12] Veniamin Fishman, Yuri Kuratov, Aleksei Shmelev, Maxim Petrov, Dmitry Penzar, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53 (2):gkae1310, 2025.

[13] Peter W Harrison, M Ridwan Amode, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, et al. Ensembl 2024. *Nucleic acids research*, 52(D1):D891–D899, 2024.

[14] Matthew M Hill, Karl W Broman, Elia Stupka, William C Smith, Di Jiang, and Arend Sidow. The c. savignyi genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution. *Genome Research*, 18(8):1369–1379, 2008.

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[16] Tianyu Liu, Tianqi Chen, Wangjie Zheng, Xiao Luo, and Hongyu Zhao. scelmo: Embeddings from language models are good learners for single-cell data analysis. *bioRxiv*, pages 2023–12, 2023.

[17] Tianyu Liu, Yuge Wang, Rex Ying, and Hongyu Zhao. Muse-gnn: Learning unified gene representation from multimodal biological graph data. *Advances in neural information processing systems*, 36:24661–24677, 2023.

[18] Tianyu Liu, Tinglin Huang, Yingxin Lin, Rex Ying, and Hongyu Zhao. Unicorn: Towards universal cellular expression prediction with an explainable multi-task learning framework. *bioRxiv*, pages 2025–01, 2025.

[19] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018.

[20] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

[21] Helen Pearson. Genetics: what is a gene? *Nature*, 441(7092), 2006.

[22] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[24] Adityanarayanan Radhakrishnan, Cathy Cai, Barbara A Weir, Christopher Moy, and Caroline Uhler. Synthetic lethality screening with recursive feature machines. *Cancer Research*, 84 (6_Supplement):897–897, 2024.

[25] Eric W Sayers, Jeffrey Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Ryan Connor, Michael Feldgarden, Anna M Fine, Kathryn Funk, Jinna Hoffman, et al. Database resources of the national center for biotechnology information in 2025. *Nucleic acids research*, 53(D1): D20–D29, 2025.

[26] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, et al. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020:baaa062, 2020.

[27] Lincoln Stein. Genome annotation: from sequence to biology. *Nature reviews genetics*, 2(7): 493–503, 2001.

[28] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

[29] Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with sclambda. *bioRxiv*, 2024.

[30] 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*, pages 1–9, 2025.

[31] Patricia J Wittkopp and Gizem Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69, 2012.

[32] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[33] Jeffrey Zhong, Lechuan Li, Ruth Dannenfelser, and Vicky Yao. Benchmarking gene embeddings from sequence, expression, network, and text models for functional prediction tasks. *bioRxiv*, pages 2025–01, 2025.

A Appendix

A.1 Problem Definition.

Here we consider a set of multi-modal datasets $\mathcal{D} = (X^1, X^2, .., X^n)$, where each modality comes from the same identity and the index is paired. For a set of genes from different species, we define the sequence modality as X^D and text annotation modality as X^T. We aim to align these modalities by training a set of projectors P^D and P^T, and the projectors take original modalities as inputs and generate embeddings in the aligned space. Such embeddings can also be used to detect neighbors and work as novel gene embeddings for handling gene-level tasks.

A.2 Dataset Construction.

We first collect the DNA sequences of genes of different species from NCBI and Ensembl, and we select the reference genomes with known location information of genes (chromosome, starting site, and ending site) to extract the paired DNA sequence of each gene. For the same gene, we collect text descriptions from NCBI and UniProt to cover gene functions and protein functions, by following the methods described in GenePT [7]. Finally, we generate embeddings of text descriptions based on the OpenAI text embedding model by following methods described in GenePT and scELMo [16]. We then split the training, validation, and testing data for model training and deployment. DNA sequences are embedded based on HyenaDNA [20] (1 million length context window).

A.3 Model Construction.

We create the set of projects based on the variation of Contrastive Language-Image Pre-Training (CLIP) [23], known as Contrastive Language-Sequence Pre-Training. The idea of CLIP is to maximize the probability of matching the embeddings from the same sample but different modalities, based on contrastive learning. Here, we take embeddings from DNA sequences and embeddings from text descriptions as inputs, and train the CLIP to have two modality-specific projectors. We also test different variations of CLIP, including default CLIP, SigLIP [32] and proposed similarity-penalized CLIP.

Considering n samples in our datasets, we have the produced embeddings from two encoders as $s_i^D = P^D(X_i^D)$ and $s_i^T = P^T(X_i^T)$ for the i_{th} pair, and the temperature t for contrastive learning. The default CLIP loss is defined as:

$$\mathcal{L}_{mCLIP} = -\frac{1}{2n}\sum_{i=1}^{n}(\log\frac{e^{ts_i^D s_i^T}}{\sum_{j=1}^{n} e^{ts_i^D s_j^T}} + \log\frac{e^{ts_i^D s_i^T}}{\sum_{j=1}^{n} e^{ts_j^D s_i^T}}),$$

where the first $\log$ term computes the DNA to TEXT retrieval result and the second $\log$ term computes the TEXT to DNA retrieval result.

SigLIP replaced the method to compute probability with Sigmoid. The default SigLIP loss is defined as:

$$\mathcal{L}_{SigLIP} = -\sum_{i=1}^{n}\sum_{j=1}^{n}\log\frac{1}{1+e^{z_{ij}(ts_i^D s_j^T + b)}},$$

where b is a trainable parameter and $z_{ij} = 1$ when $i = j$ (positive pairs), else $z_{ij} = 0$ (negative pairs).

Similarity-penalized CLIP added one penalty term to restrict the learned embeddings to have the same gene-gene interaction strength of each modality. The loss function is defined as:

$$L_{SimPenaltyCLIP} = \mathcal{L}_{mCLIP} + \frac{1}{n}\sum_{i=1}^{n}(s_i^D(s_i^D)' - X_i^D(X_i^D)')^2 + \frac{1}{n}\sum_{i=1}^{n}(s_i^T(s_i^T)' - X_i^T(X_i^T)')^2.$$

Based on our ablation results, we scale the default CLIP model with genes from different species and functional groups.

Regarding hyper parameters, we select the best learning rate, dimension of latent space, and temperature by tuning the model based on different conditions.

A.4 Similarity-based Gene Retrieval.

Since many genes from some species do not have gene annotation, by using our trained DNACLIP to generate gene embeddings of less-annotated genes, we can retrieval the given genes with closet annotated genes based on similarity (e.g., top-5 genes), and we call LLMs such as GPT-4o to summarize the description of these genes as the renewed functional annotations. This method is a novel gene annotation method, especially for low-resource species.

A.5 Demonstration of the Functionality of Gene Embeddings.

By training DNACLIP, we can generate gene embeddings from both sequence space and text description space. These gene embeddings may have task-specific preferences, and understanding these preferences is crucial for making recommendations based on users' requirements. Additionally, they can serve as a benchmarking analysis to demonstrate the strength of the proposed model and framework. Here we consider three different tasks, including disease gene prediction, gene ontology prediction, gene-gene interaction prediction, gene functional clustering, and perturbation prediction. The first three tasks are described and formed in [33], and we use the same dataset to perform evaluation for both text and sequence embeddings. Disease gene prediction means we predict genes that cause disease due to mutation in hereditary diseases, and these diseases are selected from Mendelian inheritance disorders. Gene ontology prediction means we predict genes involved in certain Gene Ontology (GO) pathways. Gene-gene interaction means we predict whether a pair of genes interacts in the biological process or not. Gene functional clustering means we annotate each gene based on its functional class from Geneformer [28] and perform clustering to check if genes with the same functional annotation are co-localized or not. The perturbation effect prediction task is adapted from scLAMBDA [29], which leverages gene embeddings from text descriptions to predict the perturbation effect based on control stage single-cell transcriptomic profiles. Similarly, we can also predict the cell viability for cancer cell lines with a random forest regressor from gene expression profiles and gene embeddings [24]. To predict cell-type-specific gene expression levels, we modify UNICORN [18] with choices from more gene embeddings.

We note that text embeddings might benefit from knowledge leakage for the prediction and clustering tasks, as the text descriptions of the given gene might contain information of diseases or pathway-related information, and thus the sequence embeddings are more important there and text embeddings work as a quality control tool, which means a well-trained model should produce text embeddings which perform well for these tasks.

A.6 Extraction and analysis of gene programs.

By leveraging the trained gene embeddings generated by the sequence encoder, we can utilize distance algorithm to construct the similarity matrix of different genes, and use clustering algorithm to annotate gene cluster (programs) to perform gene enrichment analysis. Here we select the genes from a pre-defined set and extract different gene programs to run Gene Ontology Enrichment Analysis (GOEA) [4, 2, 11] and represent the gene clusters based on associated gene pathways. We also utilize GeneAgent [30] to help us summarize the major functions of gene programs produced by DNACLIP. GeneAgent accepts a list of genes as the input and outputs the functional summary of these genes as the output.

A.7 Model Evaluation.

To evaluate DNACLIP and the produced gene embeddings, we consider both training-aware evaluation and task-specific evaluation. For the training-aware evaluation, we compute the CLIP loss based on the validation dataset to ensure that our model converges to the optimized stage with the lowest validation loss. For task-specific evaluation, we consider different metrics based on the selected tasks. For disease gene prediction and gene ontology prediction, since they are both gene-level binary classification problems, we use AUROC and AUPRC as metrics [22]. For the gene-gene interaction prediction task, since it is a paired-genes-level binary classification problem, we still use

AUROC and AUPRC as metrics. For the gene functional clustering task, we use traditional clustering metrics, including NMI and ARI [22]. For the perturbation prediction task, we use the Pearson Correlation Coefficient (PCC) and Mean Squared Error (MSE) between observed and predicted expression profiles after mean expression correction for evaluation. For the expression prediction task, we select the Pearson Correlation Coefficient (PCC) and Mean Squared Error (MSE) between the observed and predicted expression profiles for evaluation.

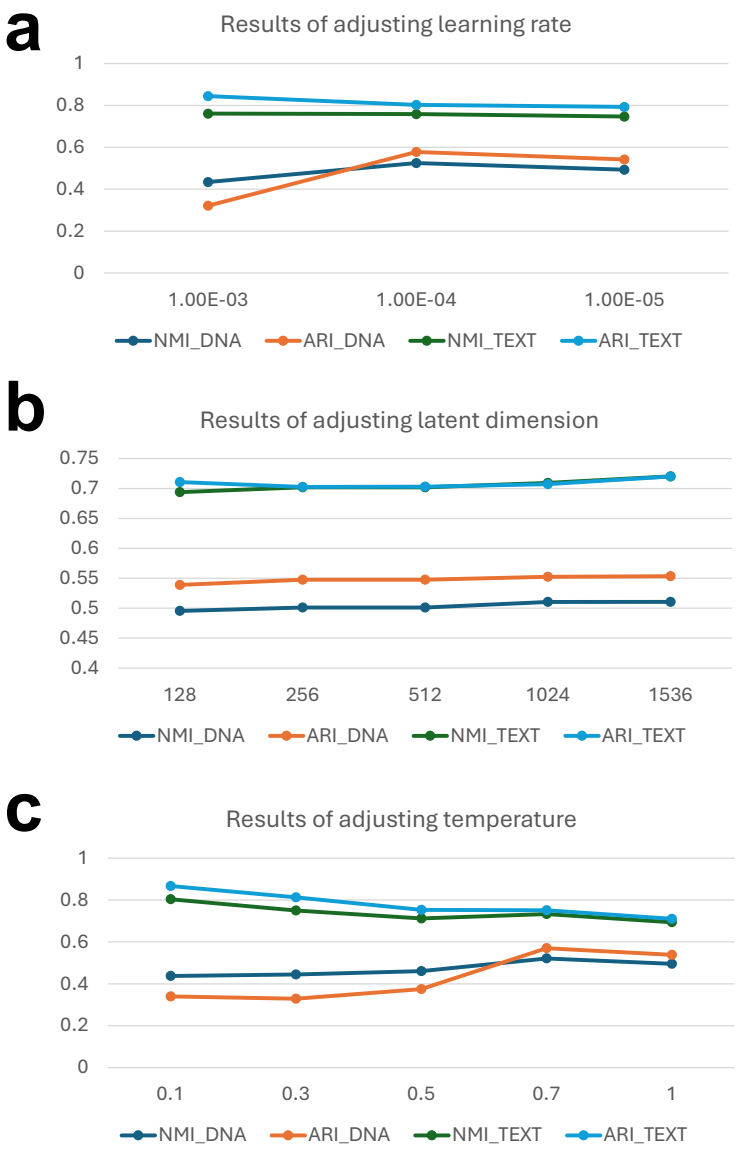
A.8 Baseline Methods.

The baseline methods included in our project are HyenaDNA, Evo2 [5], GenaLM [12], and GenePT. HyenaDNA is a genomic language model pre-trained with DNA sequences from the human reference genome, and it is based on the Hyena architecture. Evo2 is an extension of HyenaDNA and is trained with multi-species DNA sequences. Evo2 also has a larger scale (7B). GenaLM is also a genomic language model pre-trained with DNA sequences from multiple species. We use the multi-species mode of GenaLM. We also include GenePT, which takes functional descriptions of genes and embeds them with an OpenAI embedding model to access gene embeddings.

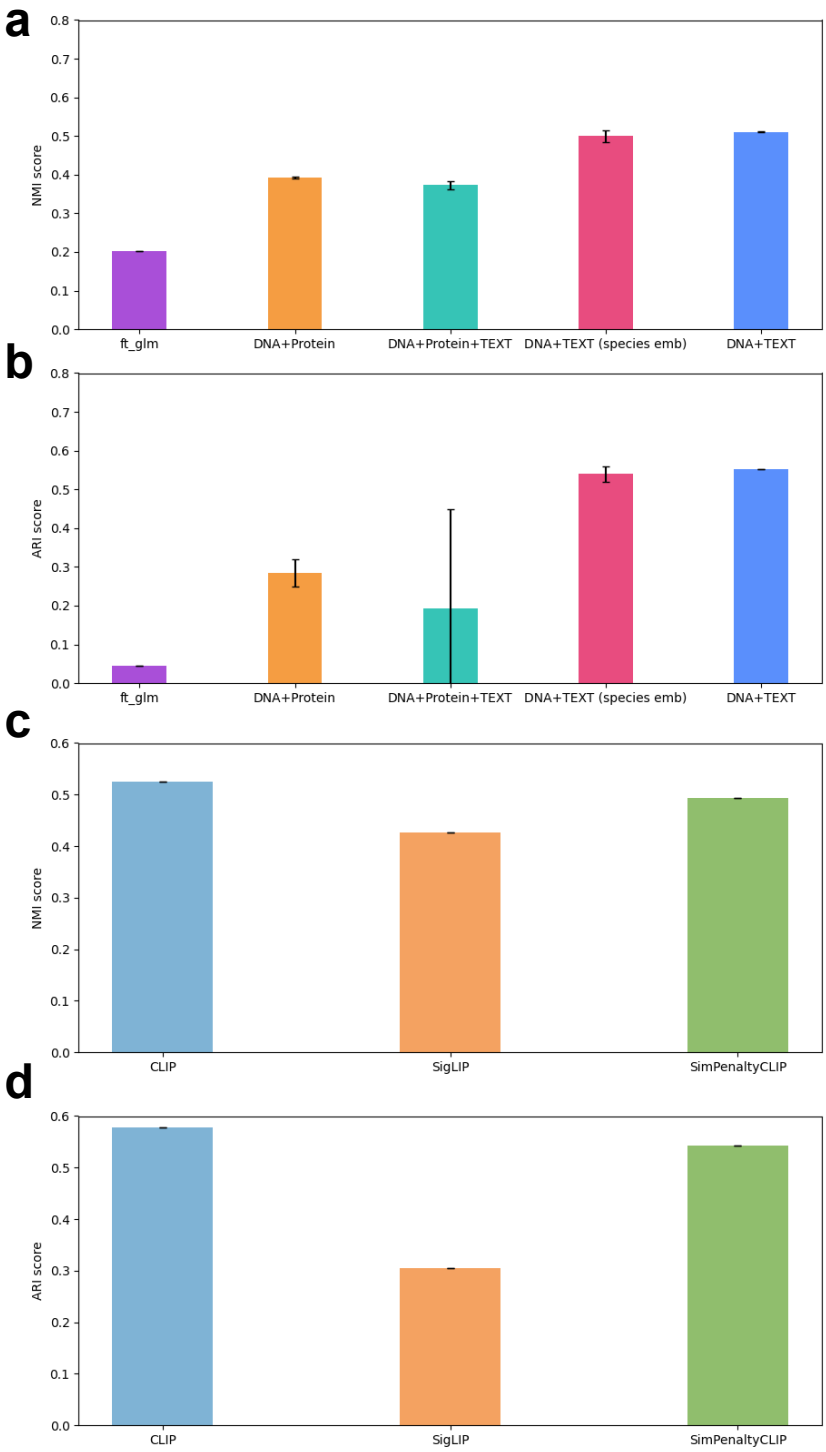
A.9 Code Availability

We will release our codes after peer review.

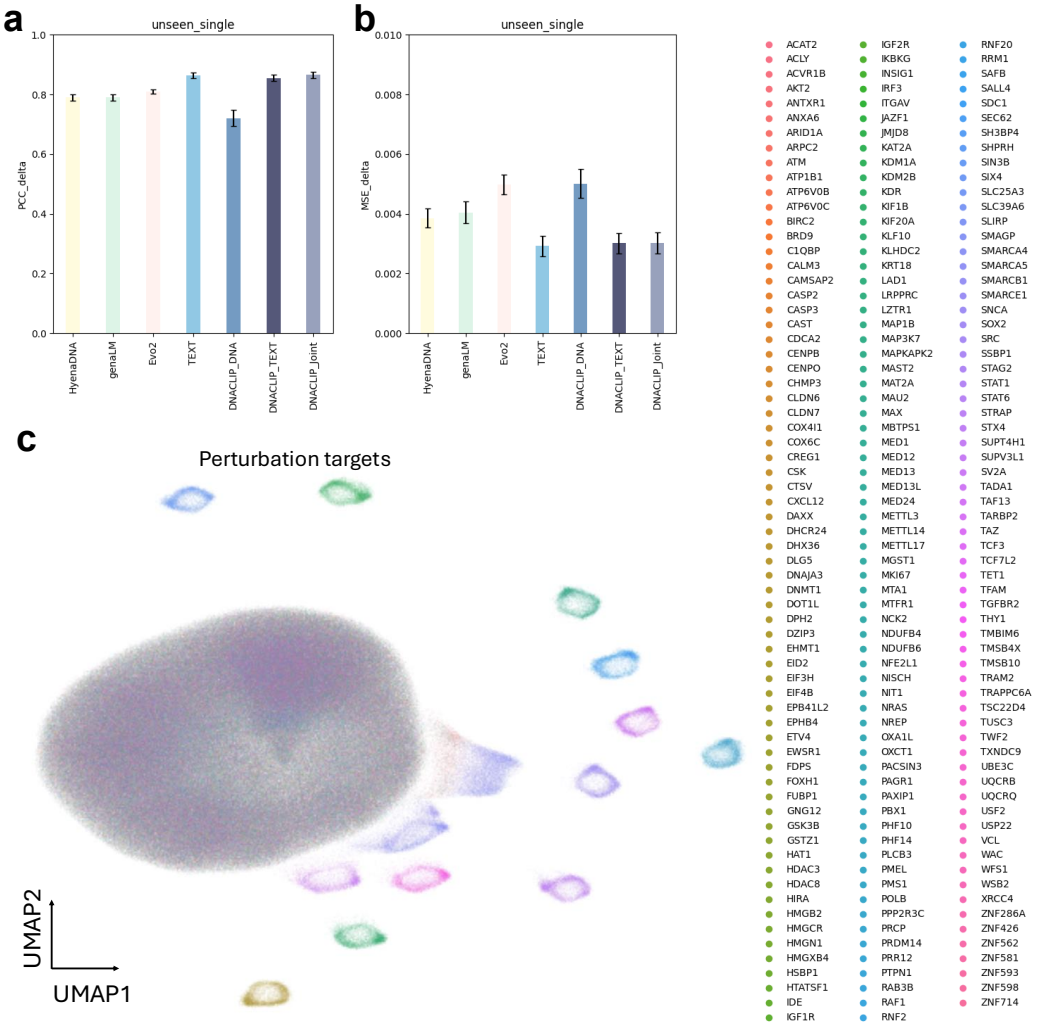
A.10 Supplementary figures



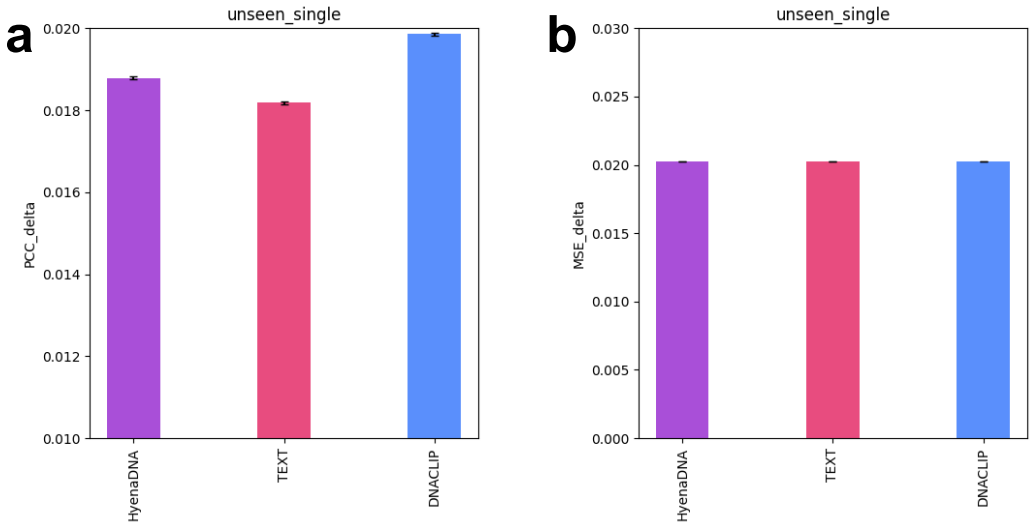
Extended Data Fig. 4: Clustering performances based on different hyper-parameters. (a) Relationship between learning rate and clustering metrics. (b) Relationship between latent dimension and clustering metrics. (c) Relationship between temperature and clustering metrics.



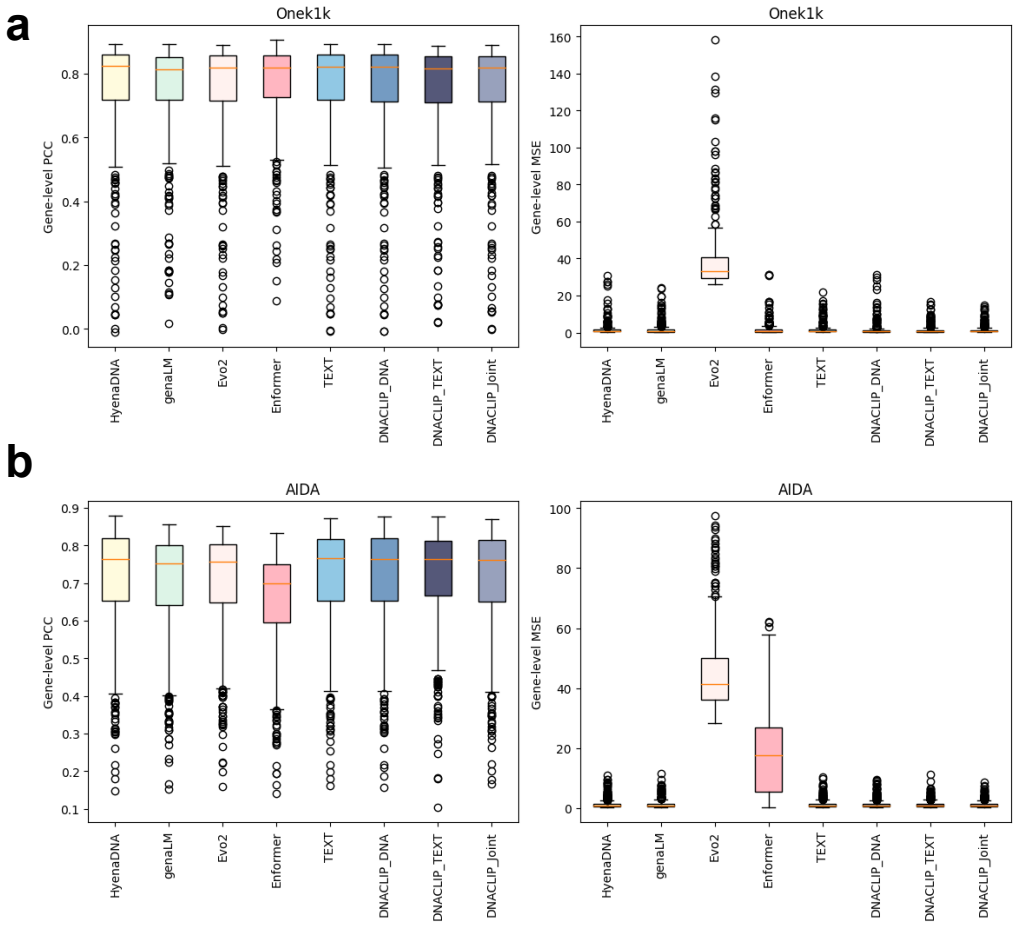
Extended Data Fig. 5: Clustering performances based on different components. (a) NMI scores across different model inputs and training strategies. (b) ARI scores across different model inputs and training strategies. (C) NMI scores across different model architectures. (b) ARI scores across different model architectures.



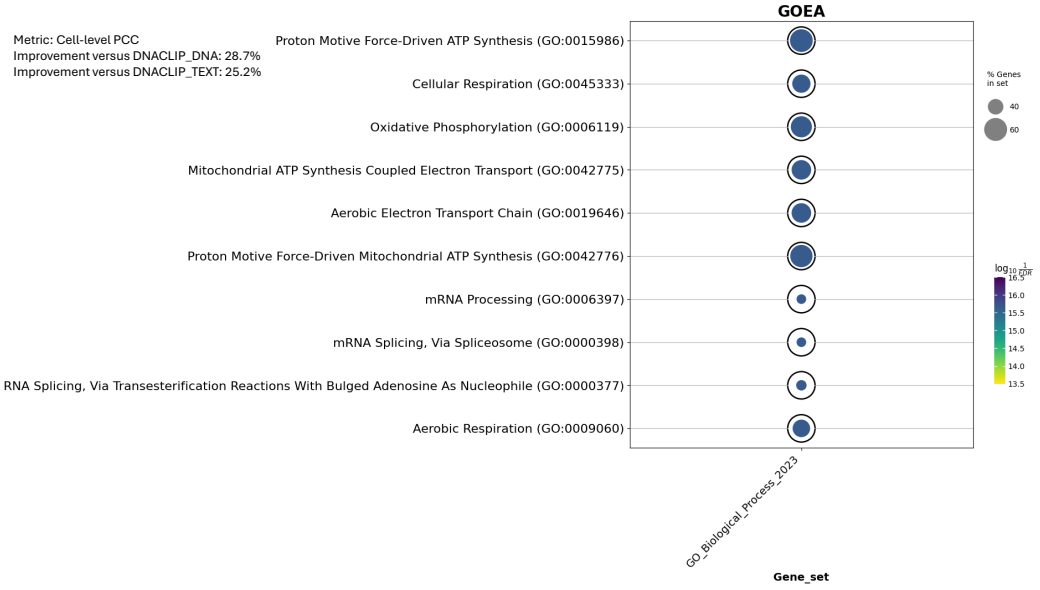
Extended Data Fig. 6: Perturbation predictions from more datasets. (a) Evaluation results with PCC_delta across different methods based on the Adamson dataset. (b) Evaluation results with on MSE_delta across different methods based on the Adamson dataset. (c) UMAP visualization of the generation performance based on the ARC Challenge dataset colored by perturbation targets.



Extended Data Fig. 7: Perturbation predictions from the HCT116 perturbed dataset. (a) Evaluation results with PCC_delta across different methods. (b) Evaluation results with on MSE_delta across different methods.



Extended Data Fig. 8: Gene-level evaluation results based on different datasets. (a) Gene-level PCC and MSE scores across different baseline methods based on the Onek1k dataset. (b) Gene-level PCC and MSE scores across different baseline methods based on the AIDA dataset.



Extended Data Fig. 9: GOEA results of specific gene clusters. We only select the top 10 pathways to show ranked by their enrichment scores. Other information such as FDR and proportion is annotated in this figure. We also report the ratio of prediction improvement introduced by the joint mode of DNACLIP versus other modes.