Machine Learning for Multiomics: Identifying Interacting DNA Elements through NYU HPC

Anni Zheng
az1932@nyu.edu

Catherine Hua
jh6780@nyu.edu

Lyric Li
rl3754@nyu.edu

Abstract

Enhancers play a critical role in regulating gene expression, often working with other enhancers to influence transcription. Despite their importance, the mechanisms by which enhancers interact remain poorly understood, with limited evidence supporting the existence of significant interactions. Recent studies have demonstrated that enhancer effects generally combine multiplicatively and exhibit rare interactions that are challenging to detect. To advance this understanding, a fully functional processing pipeline and analytical framework was developed under the NYU High Performance Computing (HPC) environment. This framework integrates the generalized linear model (GLM) for imputing gene expression with the Poisson regression model for analyzing associations between enhancer accessibility and gene expression at the single-cell level, enabling the observation and analysis of interactions among enhancer pairs. Lastly, the project studied enhancer pair activities specific to one cell type, CD14-Mono, which were identified based on the model's results and showed that these activities may be saturated and non-linear.

1 Introduction and Background

Quantifying the activity of a gene and the nearby DNA regions that potentially impact its expression has always been a key research question. A single gene may have multiple enhancers, yet how they collectively regulate gene expression remains poorly understood. Recent advancements in DNA sequencing technologies allow researchers to measure the activity of genes as well as the activity of nearby DNA regions that may influence gene expression. For example, machine learning methods were employed to identify DNA regions linked to gene activity and quantify their impact. For instance, a GLM framework (GLiMMIRS) developed to interrogate enhancer effects was applied to PBMC(Peripheral Blood Mononuclear Cell) datasets, examining 46,166 enhancer pairs and their associated genes. The results indicated limited evidence for epistatic-like interactions, suggesting that enhancer interactions are either rare or difficult to detect with current methodologies.

Despite this progress, understanding enhancer–gene interactions remains limited by the inability to fully capture the non-linear actions of DNA regions, as current models fail to adequately represent interaction effects among active and inactive DNA regions separately. To address this gap, a comprehensive pipeline should be build to based on enhancer activities to capture the combinatorial interactions among active and inactive enhancer pairs.

Building on prior methodologies, this project integrates the GLiMMIRS framework and a Poisson regression model, which associates enhancer chromatin accessibility with gene expression. This pipeline specifically compares the interactions of active and inactive DNA regions. In other words, the model enables the detection of non-linear enhancer interactions by examining specific combinations of enhancer activity at a genome-wide scale within the NYU HPC environment, analyzing enhancer pair activities in a specific cell type from the PBMC dataset.

Preprint. Under review.

2 Related Works

The exploration of enhancer activity and its regulatory effects on gene expression has been a central topic in genomic research. Below are some significant contributions that have informed the development of our project.

2.1 SCARlink: Single-Cell Enhancer Mapping

SCARlink introduced a gene-level regulatory model that predicts single-cell gene expression and links enhancers to target genes using scRNA-seq and scATAC-seq data. By employing regularized Poisson regression constrained with non-negative coefficients, SCARlink models the positive regulatory effects of chromatin on gene expression and maps enhancer–gene links at the genomic tile level, enabling the resolution of enhancer–gene associations in a cell-type-specific manner [2]. This model provides key insights for developing the generalized linear modeling frameworks used in our work.

2.2 SCENT: Statistical Enhancer-Gene Mapping

SCENT employs a nonparametric statistical method to model associations between enhancer chromatin accessibility and gene expression in multimodal single-cell datasets. It constructed 23 cell-type-specific enhancer–gene maps enriched for causal variants across 1,143 diseases and traits [3]. The SCENT framework emphasizes the scalable construction of enhancer–gene maps, which are critical for understanding the function of noncoding variants in disease-relevant tissues. Additionally, it serves as a steppingstone by providing inputs for our model, which identifies active and inactive enhancer combinations or pairs and evaluates the significance of their interactions.

2.3 GLiMMIRS: Generalized Linear Models for Enhancer Interactions

Enhancer interactions and their collective regulation of transcription have been investigated using the GLiMMIRS framework (i.e., a generalized linear model) [4]. This model analyzes enhancer effects from single-cell datasets and evaluates interactions among enhancer pairs, with limited evidence found for synergistic or redundant interactions. This work raises questions about the significance of enhancer interactions and served as the primary motivation for testing this hypothesis by developing a pipeline to simulate and analyze enhancer activities.

3 Dataset

The dataset used for this study comprises three key files: atac_matrix, rna_matrix, and meta_data, each serving a distinct purpose in the analysis pipeline. These files provide the foundation for mapping enhancer activity, gene expression, and cell metadata, enabling a comprehensive exploration of enhancer-gene interactions.

1. The atac_matrix file contains chromatin accessibility data for each enhancer region across all cells in the dataset. This matrix provides a quantitative measure of enhancer activity, with values ranging from 0(no activity) to 1(full activity).
2. The rna_matrix file contains single-cell RNA sequencing (scRNA-seq) data, capturing gene expression levels for each gene across all cells. Expression values range from 0 to 5585, representing the raw count of transcripts per gene.
3. The meta_data file includes additional information about each cell, such as: nUMI: Total number of unique molecular identifiers, reflecting sequencing depthcelltype: The biological classification of each cell. For this project, we used celltype CD14-mono for analysis, and percent.mito: The percentage of mitochondrial gene expression, often used to filter out low-quality cells.

To prepare the data for analysis, a filtering step was applied to both the rna_matrix and the atac_matrix that rows (genes in rna_matrix and enhancers in atac_matrix) were retained only if at least 5% of cells had non-zero counts for the respective gene or enhancer. This ensures that only genes and enhancers with meaningful variability and activity across cells are included.

4 Methodology

4.1 Methods

SCENT (single-cell enhancer target gene mapping): models the causal relationship between enhancers and genes in a single cell for both common and rare diseases. This algorithm uses Poisson regression, due to the sparsity of of RNA and ATAC data, and bootstrap-based significance testing to obtain empirical p-values, thus mapping enhancers to target genes. The algorithm also ensures the output gene-peak-peak pairs pass a threshold with a non-zero proportion $> 5\%$ before apply the Poisson regression. Given the selected significant gene-peak-peak pairs, we were able to perform multiple testing correction on p-values and filter for all enhancer-gene pairs with an FDR(False Discovery Rate) < 0.1.

With the obtained enhancer-gene pairs as one of the inputs of the epistasis model, we first filtered RNA, ATAC and metafile based on the celltype(the example is CD14-Mono, but all other celltypes are able to produce desired outputs as well). To test out the individual activity of each enhancer-gene pairs in cells, we created three models: cells10 where enhancer2 is inactive, regardless of enhancer1; cells01 where enhancer1 is inactive, regardless of enhancer2; cells11 where both enhancers are active. All three conditions run against cells00 where both enhancers are inactive to detect possible interactions and activities.

$$\begin{aligned}
&\text{General Form:}\quad E_i \sim \text{Poisson}(\lambda_i)\\
&\qquad\qquad\qquad\log(\lambda_i) = \beta_0 + \beta_{\text{enhancer}} X_{\text{enhancer}} + \beta_{\text{mito}} X_{\text{mito}} + \beta_{\text{UMI}} X_{\text{UMI}}\\
\text{For the three models:}\ &\log(\lambda_{i10}) = \beta_0 + \beta_{\text{enhancer1}} X_{\text{enhancer1}} + \beta_{\text{mito}} X_{\text{mito}} + \beta_{\text{UMI}} X_{\text{UMI}}\\
&\log(\lambda_{i01}) = \beta_0 + \beta_{\text{enhancer2}} X_{\text{enhancer2}} + \beta_{\text{mito}} X_{\text{mito}} + \beta_{\text{UMI}} X_{\text{UMI}}\\
&\log(\lambda_i 11) = \beta_0 + \beta_{\text{enhancer1}} X_{\text{enhancer1}} + \beta_{\text{enhancer2}} X_{\text{enhancer2}} + \beta_{\text{mito}} X_{\text{mito}} + \beta_{\text{UMI}} X_{\text{UMI}}
\end{aligned}$$

The models return the values of intercepts, beta estimates and corresponding p-values. For the p-values that were below 0.1, an additional bootstrapping was applied to improve confidence.

5 Pipeline

5.1 Workflow

The pipeline first retrieves GENCODE annotations from an external database with the latest updates and reads in expression and peak matrices from a 10X multiome experiment. Input RNA and ATAC matrices are filtered to include gene expressions with a minimum of 5% non-zero counts. The output CSV files, containing peak names and gene names, along with the GENCODE annotations, are combined to annotate genes and peaks. This annotation serves as an intermediate step for processing genes and peaks with their names, start, and end information, and saves the outputs in `.bed` format to ensure compatibility with `bedtools`.

`bedtools window` identifies proximity or overlaps between two datasets—in this case, genes and peaks—using a 500,000 base pair search window. This ensures the inclusion of biologically relevant pairs, including distant yet functionally significant ones. Additionally, using the RNA matrix, highly variable genes for epistasis testing are identified with the `FindVariableFeatures` functionality (MVP method) implemented in `Seurat`. The mean-variance plot method divides genes into bins based on their mean expression, and for each bin, the variance is standardized into a z-score to classify highly variable genes.

The obtained highly variable genes, along with the previously generated gene names and gene-peak pairs, are used as inputs to filter and split gene-peak pairs. By integrating all three input files and converting formats, the script produces 32 filtered gene-peak CSV files to downscale computational and memory requirements. After finalizing the gene-peak pairs, the pipeline runs the `SCENT` method to identify potential enhancer-gene pairs for the same gene. This step incorporates false discovery control by computing the expected proportion of rejected null hypotheses that are actually true. The null hypothesis is rejected when the adjusted p-value falls below a specified level, ensuring the false discovery rate is controlled at that level. The output enhancer pairs are then passed into the epistasis model, along with the three input datasets, to filter active cells based on three conditions: enhancer1 active, enhancer2 active, and both active.

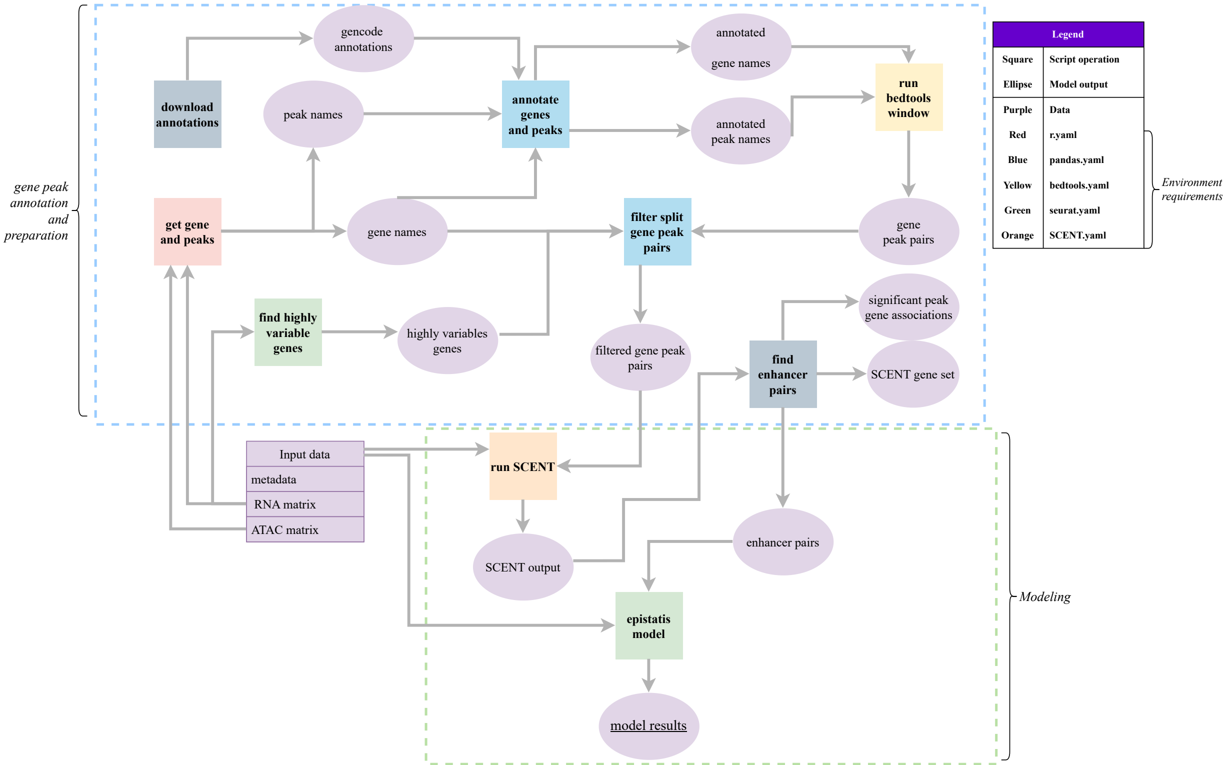


Figure 1: Pipeline workflow and outputs

5.2 HPC Integration

NYU-IT HPC Greene cluster was utilized to execute the project in our system, replacing the original LSF session scheduler. The Greene cluster operates on a Linux operating system across its individual nodes. Resource management and job scheduling were handled by the Slurm software system.

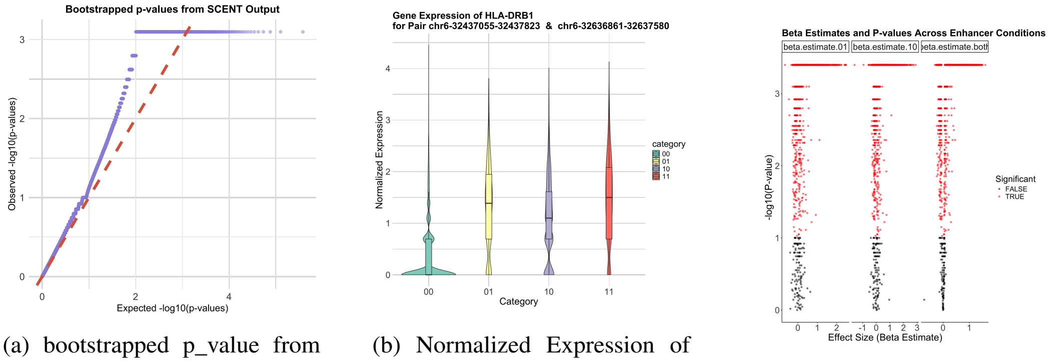
The Snakemake pipeline, a hybrid of Python and shell scripting, was employed for workflow management. Inputs and outputs were described as required by the project. The pipeline's initiation commands were stored in the run_snakemake.sh file. This script included ‘SBATCH‘ commands for handling job and error files, as well as the ‘module load‘ command to activate the integrated Snakemake environment within the HPC system. A custom folder was created for package management and job-specific isolation to address the unavailability of some necessary packages within the HPC environment, allowing for efficient package discovery and management.

The pipeline could be initiated using the following script:

```
snakemake —latency-wait 60 —forceall —use-conda —jobs 32 —cluster
’sbatch —time=48:00:00 —mem=32G —cpus-per-task=8 -o out.%J.txt -e
err.%J.txt’
-p $(echo results/epistasis_models/epistasis_models_01_10_1_{1..32}.csv
results/epistasis_models/epistasis_models_01_10_2_{1..32}.csv
results/epistasis_models/epistasis_models_11_00_{1..32}.csv)
```

6 Results and Analysis

To validate the significance of enhancer-gene pairs identified by the SCENT model, we performed bootstrap-based significance testing and visualized the results using a Quantile-Quantile (QQ) plot (Figure 2). The QQ plot compares the distribution of observed -log10(p-values) to the expected -log10(p-values) under the null hypothesis. In the QQ plot, the observed values (black dots) largely align with the expected values (red diagonal line) in the lower p-value range, indicating that the model outputs generally conform to expectations under the null. However, deviations from the diagonal line are observed at higher -log10(p) values, suggesting the presence of statistically significant enhancer-gene associations that exceed the null hypothesis, supporting the hypothesis of biologically meaningful enhancer-gene associations.



(a) bootstrapped p_value from SCENT output

(b) Normalized Expression of Genes for Enhancer Pair

(c) Interaction Effects

Figure 2: Model outputs

To further investigate the combinatorial effects of enhancer pairs on gene expression, we categorized cell states based on the activity of two enhancers for each enhancer-gene pair. The categories are as follows:

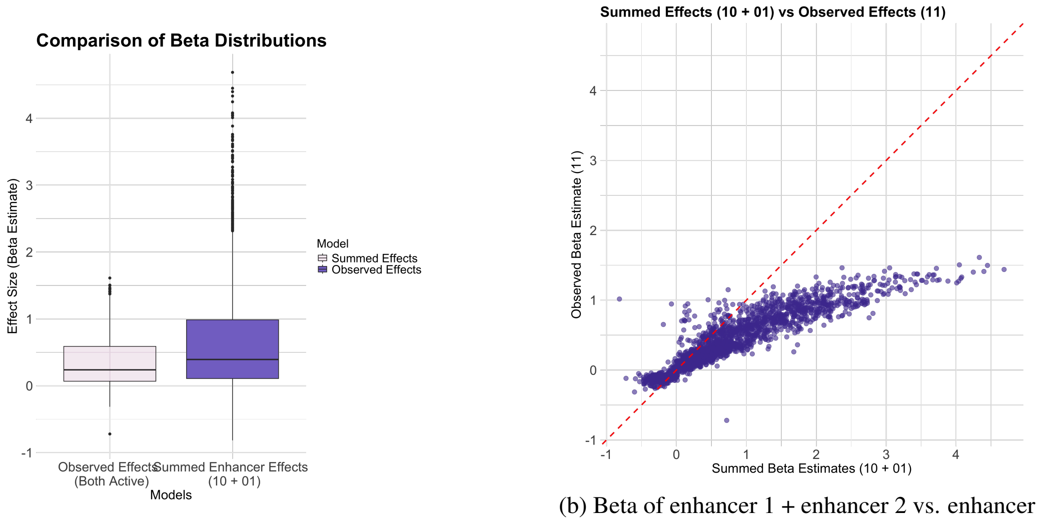
- 10: Enhancer 1 active, Enhancer 2 inactive.
- 00: Both enhancers inactive.
- 01: Enhancer 2 active, Enhancer 1 inactive.
- 11: Both enhancers active.

Take the normalized expression of HLA-DRB1 for the enhancer pair chr6-32437055-32437823 and chr6-32636861-32637580 as an example. This sharp increase in the 11 category suggests a non-linear or possibly synergistic relationship between the two enhancers, where their combined activation significantly enhances gene expression compared to single enhancer activity.

After getting valid gene-peak-peak pairs, we applied the epistasis model to evaluate the significance and strength of enhancer activity effects under different conditions, we compared the beta estimates and p-values across three contrasts 01 vs 00, 10 vs 00, and 11 vs 00. Figure 4 presents the volcano plot, which visualizes the beta estimates (effect size) on the x-axis and -log10(p-values) on the y-axis, with significant values ($p < 0.05$) highlighted in red. For all three comparisons, a subset of enhancer-gene pairs shows positive beta estimates, indicating a positive association with gene expression when enhancers are active. Beta estimates for 11 vs 00 (both enhancers active) tend to be larger compared to 01 vs 00 and 10 vs 00, suggesting an additive or synergistic effect when both enhancers are active comparing to only one active

After confirming the effect of interaction would be more than only one enhancer active, we then analyze to determine whether the combined effect of single enhancer activations (10 and 01) equals or exceeds the effect of both enhancers being active simultaneously (11), we compared their beta estimates using both a boxplot and a scatterplot.

The boxplot indicates that the summed effect of 10 + 01 is consistently larger than the observed beta values for 11. The median and interquartile range of beta_sum_enhancers exceed those of beta_estimate_both. This suggests that the interaction between enhancers under the 11 condition is not strictly additive and may exhibit dampened or redundant effects. We gain similar conclusion from the scatterplot. The majority of points fall below the red dashed line, meaning that the beta estimates for 11 are generally lower than the summed values of 10 + 01. This deviation suggests a sub-additive interaction between enhancers when both are active, supporting the hypothesis that enhancer interactions may not be purely additive and could involve regulatory limitations or saturation effects.



(a) Beta comparison boxplot

(b) Beta of enhancer 1 + enhancer 2 vs. enhancer both

Figure 3: Comparison of beta distributions and scatterplot for enhancer effects.

7 Discussion and Future Work

There were some limitations within the NYU HPC environment in which we had built our pipeline upon. The standard HPC job limit of 48 hours often constrained pipeline execution, especially for larger datasets. Due to the absence of GPU acceleration in the GLM and SCENT models, testing was limited to specific PBMC cell types. Larger cell types with extensive datasets frequently exceeded the time limit, causing incomplete runs.

Additionally, large input files demanded significant resources, leading to prolonged execution times (e.g., over 12 hours for certain jobs). This delay hindered debugging processes.

Moreover, certain tasks required older Python methods, which were unavailable in the default HPC environment. Although some packages were successfully installed locally, cluster execution occasionally failed to locate them. To address this, specific methods were sourced from open-source Python GitHub repositories without altering the original code. Fortunately, most required older packages were either compatible with newer versions or not essential for our tasks, allowing the pipeline to execute without critical errors.

Not only that, the tool, snakemake, which we had built our pipeline upon was also unstable. Note that snakemake internally generates a directed acyclic graph (DAG) to plan workflow execution. However, the execution was sometimes unstable, either tolerating incomplete rule outputs or halting prematurely. This required modifications to ensure all rules contributed to the final outputs, overcoming issues with skipped or incomplete execution due to faulty DAG structures.

The pipeline and analysis have only been tested on one cell type out of 30 due to limited time and resources. For future work, it would be beneficial to test all cell types to gain a more accurate and comprehensive conclusion. Furthermore, the current analysis focused on identifying pairwise enhancer interactions using generalized linear models. Future work could explore more sophisticated models, such as non-linear models (e.g., neural networks or tree-based methods) to capture complex enhancer synergies.

8 Conclusion

A pipeline was developed to integrate the previous GLM and Poisson models for mapping enhancer-gene pairs and assessing the significance of their interactions compared to the absence of interactions. Additionally, a closed-up analysis on the CD14-Mono cell type revealed that enhancer interactions have a synergistic effect may not be purely additive and could involve regulatory constraints or saturation effects.

References

[1] Tian, J., Lou, J., Cai, Y., Rao, M., Lu, Z., Zhu, Y., Zou, D., Peng, X., Wang, H., Zhang, M., Niu, S., Li, Y., Zhong, R., Chang, J., & Miao, X. (2020) Risk SNP-Mediated Enhancer-Promoter Interaction Drives Colorectal Cancer through Both FADS2 and AP002754.2. *Cancer Research* **80**(9):1804-1818. doi: 10.1158/0008-5472.CAN-19-2389. Epub 2020 Mar 3. PMID: 32127356.

[2] Mitra, S., Malik, R., Wong, W., Rahman, A., Hartemink, A.J., Pritykin, Y., Dey, K.K., & Leslie, C.S. (2024) Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nature Genetics* **56**(4):627-636. doi: 10.1038/s41588-024-01689-8. Epub 2024 Mar 21. Erratum in: *Nature Genetics* **56**(6):1319. doi: 10.1038/s41588-024-01805-8. PMID: 38514783; PMCID: PMC11018525.

[3] Sakaue, S., Weinand, K., Isaac, S., Dey, K.K., Jagadeesh, K., Kanai, M., Watts, G.F.M., Zhu, Z.; Accelerating Medicines Partnership® RA/SLE Program and Network; Brenner, M.B., McDavid, A., Donlin, L.T., Wei, K., Price, A.L., & Raychaudhuri, S. (2024) Tissue-specific enhancer-gene maps from multimodal single-cell data identify causal disease alleles. *Nature Genetics* **56**(4):615-626. doi: 10.1038/s41588-024-01682-1. Epub 2024 Apr 9. PMID: 38594305; PMCID: PMC11456345.

[4] Zhou, J., Guruvayurappan, K., Toneyan, S., Chen, H.V., Chen, A.R., Koo, P., & McVicker, G. (2023) Analysis of single-cell CRISPR perturbations indicates that enhancers act multiplicatively and provides limited evidence for epistatic-like interactions. *bioRxiv* 2023.04.26.538501; doi: https://doi.org/10.1101/2023.04.26.538501.

Acknowledgments

We thank our mentors Kushal Dey (Memorial Sloan Kettering Cancer Center), Karthik Guruvayurappan (Memorial Sloan Kettering Cancer Center), and Aahlad Manas Puli (New York University) for their guidance and valuable insights throughout this project. We also acknowledge the support of the NYU High Performance Computing (HPC) for providing computational resources necessary for our analysis.