Estimating functions and effects of enhancers on cellular regulation

A Project Report

submitted by

SAMIKSHA MAURYA

in partial fulfilment of the

requirements for the award of the

degree of

MASTER OF TECHNOLOGY

(MT23251)

COMPUTATIONAL BIOLOGY

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY
DELHI

NEW DELHI- 110020

Certificate

This is to certify that the thesis titled *"Estimating functions and effects of enhancers on cellular regulation"* being submitted by **Samiksha Maurya** to the Indraprastha Institute of Information Technology, Delhi, for the award of the Master of Technology in Department of Computational Biology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

July, 2025

Dr. Vibhor Kumar
Thesis Supervisor
Associate Professor
Department of Computational Biology
Indraprastha Institute of Information Technology, Delhi
New Delhi 110020

Acknowledgement

I extend my sincere gratitude to Dr. Vibhor Kumar for his invaluable guidance and support during my M.Tech thesis. His expertise, encouragement, and constructive feedback have been crucial in shaping my work, inspiring me to strive for excellence and overcome challenges. Dr. Kumar's commitment to creating a supportive research atmosphere and his confidence in my capabilities have been incredibly inspiring. It has been a privilege to work under his guidance, and I am grateful for his significant time and contributions to my academic journey.

I am also grateful to Jaidev Sharma, a PhD student under Dr. Vibhor Kumar, whose unwavering support and invaluable assistance played a pivotal role in completing my thesis. His guidance and encouragement were essential throughout the process, and I am truly grateful for his input.

Furthermore, I extend my thanks to my friends for their continuous encouragement and support. Additionally, I acknowledge the contributions of the broader research community, whose insights shared through publications, conferences, and online platforms have been indispensable in shaping my research.

Special thanks to the institution for providing the resources and environment necessary for this project's success.

Lastly, I am profoundly thankful to my family for their unwavering love and support, which has been a constant source of strength throughout.

Samiksha Maurya

Abstract

Enhancers and their non-coding counterparts, enhancer RNAs (eRNAs), are crucial regulators of gene expression, directing cellular identity, developmental programs, and disease states. The functional implications of enhancer perturbation, whether performed via knockout or knockdown, are still incompletely characterized because not all perturbations produced measurable changes in gene expression.

To address this, we developed ePerturbDB (http://reggen.iiitd.edu.in:1210/), a manually curated, open-access repository of 83,743 enhancer and eRNA perturbation records collected from experimental literature in diverse biological contexts. ePerturbDB provides users with the ability to compare user query genomic loci with available validated perturbations and interrogate the associated genes and ontology terms to infer possible regulatory implications, thereby enabling both functional annotation and translational therapies.

Simultaneously, we implemented a machine learning methodology predicting gene-pathway-enhancer association based on enhancer activity profiles to find enhancers that regulate genes within defined biological pathways. These predictions provide insight into enhancer biology and collectively serve as a tool to explore the enhancer-mediated gene regulation and support the foundation of designing functional studies in the contexts of genomics and precision medicine.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Gene expression regulation is a foundation on which cell identity, development and adaptive responses to environmental stimuli are organized. Enhancers are one of the special forms of such distal regulatory DNA sequences as part of the cis-regulatory elements encoded within the genome, and play an important role in the regulation of the transcription of the target gene [1]. Unlike promoters, which are typically found just upstream of the target genes, enhancers may act on very long stretches of genomic DNA, and are also independent of orientation in their effects [2]. Enhancers make long-range interactions between gene promoters possible via chromatin looping systems by providing a molecular scaffold that assembles transcription factors (TFs), co-activators and chromatin remodeling complexes [3].

Active enhancers are characterized by transcription of bidirectional, non-coding RNAs known as enhancer RNAs (eRNAs) [5]. Despite the fact that the exact mechanics behind eRNAs remain unclear, accumulating evidence suggests that they may exert positive reinforcement of enhancerpromoter interactions, or change local chromatin and change transcriptional output [6]. This has resulted in the use of the occurrence of eRNAs as a proxy readout of enhancer activity in genome-wide functional analyses [7].

1.1 Background

Enhancers were first discovered to be DNA regions that might raise proximal genes' transcriptional output. Their crucial functions in coordinating developmental processes, identifying cell lineage, and assisting in the molecular basis of numerous disorders have been demonstrated by subsequent research [8,9]. These regulatory elements are frequently linked to certain epigenomic signatures, including DNase I hypersensitive sites and histone modifications like H3K27ac and H3K4me1, which together identify areas of accessible and transcriptionally permissive chromatin [10].

Producing enhancer RNAs (eRNAs), a family of bidirectionally transcribed non-coding RNAs derived from enhancer loci, is a hallmark of transcriptionally active enhancers [5]. A growing body of research links eRNAs to important regulatory processes like looping of enhancer-promoter, the RNA polymerase II recruitment and other transcriptional machinery components, and the modification of local chromatin architecture, even though their stability and functional significance vary [6,11]. Despite these discoveries, the exact role of eRNAs in function remains unclear, especially in in vivo systems [12]. Targeted perturbation methods like RNA interference (RNAi) and CRISPR interference (CRISPRi) have been extensively used to clarify the molecular functions of enhancers and eRNAs. These methods provide causal evidence for enhancer

functionality in various biological contexts and enable the systematic dissection of enhancer-mediated gene regulation [13,14].

Enhancers use chromatin modifiers, co-activators, and transcription factors to control the transcription of genes. Depending on the species, tissue, and cell state, their activity varies. To find active enhancers, several biochemical assays have been created, including ChIP-seq, DNase-seq, and ATAC-seq, which profile histone modifications such as H3K27ac and H3K4me1 as well as chromatin accessibility [50]. Direct proof of enhancer activity is provided by functional assays like Massively Parallel Reporter Assays (MPRA) and luciferase-based methods.

The production of short, bidirectionally transcribed non-coding RNAs known as enhancer RNAs (eRNAs) is a crucial indicator of active enhancers [5,7,4]. Enhancer-promoter looping, transcriptional activation, and chromatin remodeling are all closely linked to eRNAs. As a result, their existence is now commonly used as a stand-in for enhancer activity. Additionally, research has documented the presence of "primed" or "poised" enhancer areas designated. As a result, their existence is now commonly used as a stand-in for enhancer activity. In addition, research has documented the presence of "primed" or "poised" enhancers, which are areas characterized by incomplete or repressive histone changes before complete activation[51].

Functional characterisation of enhancers and eRNAs is still scarce, despite improvements in enhancer discovery. For confirming enhancer function and comprehending their regulatory roles in gene expression, experimental perturbation approaches like RNA interference (RNAi), CRISPR interference (CRISPRi), and CRISPR-Cas9 have become essential [13,14].

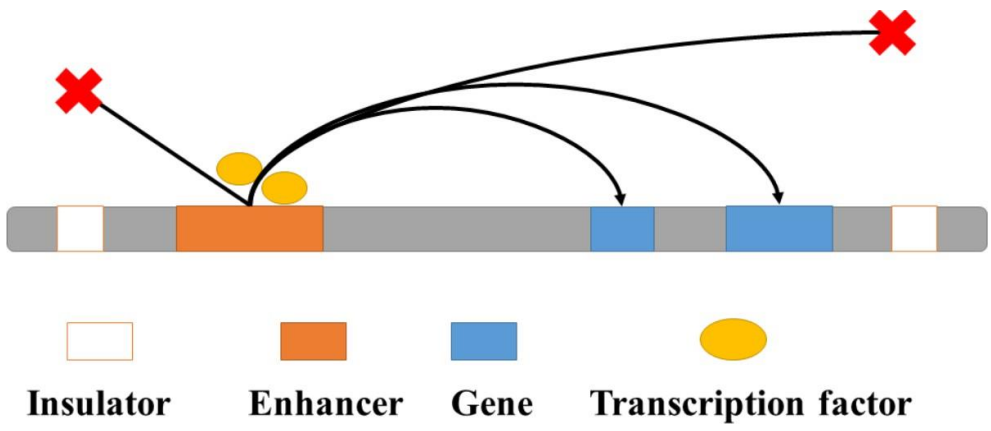


Figure 1.1: Enhancers as cis-regulatory elements

Understanding the intricacy and accuracy of gene regulation depends heavily on enhancer combinatorics. Instead of operating independently, several enhancers frequently work together to control a single gene. They can do this in redundant, additive, or synergistic ways to precisely control gene expression in both space and time [55,56]. Multiple transcriptional inputs can be integrated by cells into this multilayer regulation, which enhances the flexibility and robustness of gene expression programs during development [57]. During development and differentiation, when accurate expression thresholds are crucial for determining the correct cell fate, combinatorial enhancer activity is particularly significant. Furthermore, research has demonstrated that certain enhancer elements within a combinatorial set can be disrupted, resulting in developmental abnormalities and gene misregulation, highlighting their functional significance. In order to decipher gene regulation and direct the development of targeted therapeutics, it is crucial to investigate the combinatorial logic of enhancers.

1.2 Challenges in Functional Annotation of Enhancers

The discovery of potential enhancers across the entire genome has been made possible by developments in high-throughput genomic technologies like ChIP-seq, ATAC-seq, and emerging RNA sequencing. Nevertheless, converting these discoveries into useful annotations is still quite difficult [15]. The biological effects of enhancer or enhancer RNA (eRNA) disruption are not well-represented in centralized databases, despite intensive profiling efforts. This limits the translational potential of these networks and prevents a thorough knowledge of gene regulatory networks [6,12].

Both biological and technical variables hamper the functional prediction of enhancers and eRNAs. One major drawback is the absence of direct annotation techniques. Because enhancers and eRNAs do not yield stable, readily measurable products like protein-coding genes do, it is challenging to deduce their function from sequence or expression data alone [16]. The majority of current identification techniques rely on indirect markers, such as transcription factor binding patterns, chromatin accessibility, and histone modifications (e.g., H3K27ac), which imply regulatory potential but do not validate functionality, especially in vivo [10,12].

The high context specificity of enhancer and eRNA function is another significant barrier. These regulatory components frequently function differently depending on the cell type, tissue, or situation [4,7]. Certain environments or developmental stages may be the sole times an enhancer is active, whereas other times it is inert [17]. Because of this temporal and geographical heterogeneity, it isn't easy to extrapolate enhancer function across biological settings, and effective interpretation requires high-resolution, cell-type-specific information [18].

Mapping enhancers to their target genes is a challenging task. Enhancers can control genes hundreds of kilobases away, and they may choose to target a more distant gene instead of the closest one [2,3]. To establish enhancer–promoter connectivity, this long-range regulation defies linear genome annotation and necessitates integrative approaches like perturbation-based methods

(like CRISPR interference) or chromatin conformation capture techniques (like Hi-C, Capture-C) [3,13,14,19]. Even with these methods, if the resulting interactions are not carefully analyzed, they may be unclear [20].

Enhancer and eRNA function experimental validation remains a hurdle. Low-throughput and labor-intensive assays include luciferase reporter systems, CRISPR-based deletions, and RNA interference [13,14,21]. There is a significant discrepancy between predicted and functionally verified elements since it is impossible to experimentally confirm every candidate, even though the human genome may include hundreds of thousands of putative enhancers [21].

Furthermore, there is ongoing debate on the biological function of eRNAs. According to some research, eRNAs actively regulate genes by stabilizing transcriptional machinery or promoting enhancer–promoter looping [5,6,11], but other studies suggest that they might only be consequences of enhancer activity [22]. This uncertainty casts doubt on the necessity of eRNAs in regulatory systems and makes it more difficult to plan and evaluate functional studies that target them [12,22]. Finally, another level of complexity is introduced by the phenomena of combinatorial control and regulatory redundancy. Multiple enhancers working together or compensating for one another regulate the majority of genes [23]. It can be challenging to determine the precise role of individual enhancers or the eRNAs that are linked to them since the disruption of a single enhancer may not produce a detectable phenotypic effect [24].

1.3 Related Work

For the annotation and functional investigation of enhancer elements, several reputable databases offer useful information. A carefully maintained database called EnhancerDB contains human enhancers that have been experimentally confirmed, with a focus on those linked to disease settings. It is a trustworthy resource for finding enhancers that functional assays support [25]. Enhancer Atlas provides genome-wide enhancer annotations using high-throughput datasets like ChIP-seq, spanning a variety of tissues and animals. This framework offers comparative genomics investigations and makes it easier to investigate tissue-specific regulatory elements [26].

Enhancers are identified by the FANTOM database using Cap Analysis of Gene Expression (CAGE), which detects the bidirectional transcription of enhancer RNAs (eRNAs). It has made a substantial contribution to our knowledge of the ubiquitous transcriptional activity linked to enhancers [4]. Furthermore, KnockTF gathers information on transcription factor perturbations and the accompanying changes in downstream gene expression. It enables the functional interpretation of transcriptional regulatory networks by incorporating perturbation experiments [28]. Together, these databases provide a fundamental framework for studying transcription factor-mediated gene regulation in various biological contexts, enhancer identification, and tissue-specific functional annotation.

CHAPTER - 2

Database for Experimentally Perturbed Enhancers

We created ePerturbDB, a comprehensive and integrative database focused on experimentally perturbed enhancers and enhancer RNAs (eRNAs), to facilitate systematic research of enhancer function. The database combines enhancer annotations from well-known genomic repositories with knockout and knockdown experimental data that have been carefully selected from the literature. Enhancer function analysis and prediction modeling of regulatory effects are made possible by ePerturbDB, which aggregates and normalizes data from many experimental and computational sources.

2.1 Assembly of a Reference Enhancer Dataset

Developing a high-confidence, harmonized reference collection of human enhancers was a crucial step in building ePerturbDB. Enhancer coordinates from several publicly accessible databases were combined and categorized into primary and secondary sources based on their novelty, methodological soundness, and dependence on experimental validation.

2.1.1 Primary Databases

Primary databases were chosen due to their strong reliability in identifying enhancers and providing direct experimental evidence.

FANTOM Enhancer Database: The FANTOM project, which finds enhancers using Cap Analysis of Gene Expression (CAGE) signals, provided the enhancer coordinates for the human genome (GRCh38/hg38) [4]. These enhancers are strong indicators of transcriptionally active enhancer regions and are distinguished by their bidirectional transcriptional activity

VISTA Enhancer Browser: The VISTA Enhancer Browser, which offers experimentally verified enhancers in the human and mouse genomes, is where the data were obtained. In vivo reporter experiments have provided high-confidence functional confirmation for the enhancers in this dataset[53].

2.1.2 Secondary Databases

Resources that include enhancer data from several sources or make use of computational predictions along with high-throughput epigenomic profiling are examples of secondary databases.

EnhancerDB: This database, which includes entries from FANTOM and VISTA, compiles enhancer data from 41 human tissues. EnhancerDB is a useful supplementary reference for tissue-specific enhancer studies since it is an integrative and cumulative resource [25].

EnhancerAtlas: The BED file containing the enhancer data for the hg19 genome assembly was downloaded. EnhancerAtlas supports extensive cross-tissue and cross-species regulation research by compiling enhancers found by a variety of high-throughput experiments (such as ChIP-seq) [26].

TCEA Super Enhancer Database: The Tissue-specific Cis-regulatory Element nomenclature (TCEA) database included the super-enhancer areas for 86 human cell and tissue types. Clusters of enhancers linked to genes essential for cell identity identify these areas[52].

ENdb2: This database includes human enhancers with experimental validation. The hg19 genome assembly format's Download portion yielded BED-format files, which added another layer of experimentally supported enhancer data [54].

Based on the size of their enhancer collections, the availability of experimental validation, and the presence of multi-tissue annotations, four important databases - FANTOM, VISTA, EnhancerDB, and EnhancerAtlas were given priority to create a coherent enhancer reference dataset. When combined, these archives provide a broad and representative collection of enhancer regions from various experimental platforms and biological circumstances. Genomic coordinates (chromosome, start, and end) and related annotations were extracted from BED-formatted data for every source. These files provided the framework for adding enhancer function annotations and integrating perturbation data into ePerturbDB.

2.1.3 Data Collection and Standardization

All BED files were processed in a common curation pipeline to ensure the uniformity and easy integrations of multiple types of data. each of the datasets was sorted first in terms of chromosomal location using the bedtools sort function. original datasets were bio-informatics genomics coordinates BED format with chromosome names, source-specific genomic span and labels integration. Standardization Due to the utilization of multiple genome assemblies (mainly hg19 and hg38), genomic coordinates were standardized to a common reference. All participating datasets were facilitated into GRCh38 (hg38) by UCSC LiftOver to the human genome assembly. This normalizing step removes any variations in the positions of the data within each individual

record, thus ensuring comparability after normalization becomes simpler to perform and achieve with reliable data.

We used the bedtools multiinter function to calculate the overlap of enhancer intervals across the four major areas to find high-confidence enhancer regions. Location of enhancers, likely to be functional regulatory regions, was retained as a service as a consensus location when they were confidentially backed by two or more databases. Each consensus enhancer had a unique internal identification (e.g., enhrx-1 or enhrx-2) to allow consistent referencing within databases. The identifiers that denote databases where sources are drawn were also attached to these consent items. Where possible, also the makers of enhancer-target gene connections in the original resources were added. This standardized reference set of enhancers led to the later curation of the enhancer perturbation data in the literature into ePerturbDB.

2.2 Literature Mining and Curation of Knockout Perturbations

ePerturbDB relied on the manual, systematic curation of enhancer and eRNA perturbation in the scientific literature as a significant section. A literature mining pipeline was employed to generate a large-confidence collection of empirically determined perturbation events. A handpicked set of specific words (enhancer knockout, enhancer knockdown, eRNA knockout, eRNA knockdown, enhancer deletion, CRISPRi enhancer, and shRNA eRNA) was utilized to search PubMed and Google Scholar. The primary focus was on functional genomics studies, using gene-editing or gene-silencing approaches, to modulate measurable enhancer or eRNA activity.

The selection of studies was set by the criteria of experimental rigor, relevance, and availability of data on perturbations. Preference was given to the publications utilizing popular methods of perturbations such as CRISPR-Cas9, small interfering RNA (siRNA), CRISPRi, and short hairpin RNA (shRNA). A large amount of metadata was extracted from the Methods, Results, and Supplementary Information sections of each article selected. Among the well-chosen data that were included were genomic coordinates of the perturbed enhancer or eRNA, type of perturbation (eRNA or enhancer), mode of perturbation, tissue or cell line, and whether the perturbation resulted in a meaningful change in gene expression, as well as whether the effect on the target gene was direct or indirect. In the cases where only perturbation sequences (e.g., sgRNA or shRNA) were reported, the target genomic loci were identified using the UCSC BLAT tool. To validate the mapped coordinates, we applied a tool named UCSC LiftOver and, in case of need, transferred it to the genome assembly named GRCh38 (hg38). It was under these conditions that it was possible to reliably combine the perturbation information with the enhancer reference set and also ensure consistency in the coordinate representation.

2.2.1 Enhancer-to-Gene Association and Functional Annotation

We performed functional annotation by pairing the closest gene to each enhancer once we standardised the curated enhancer perturbation assays to the GRCh38 (hg38) genome assembly. To obtain likely enhancer-gene regulatory associations, we identified the nearest transcription start site (TSS) using the bedtools closest program to all enhancer loci. Each paired mapping had its gene name, TSS coordinates, chromosomal localization, and location of the enhancer on the genome relative to the TSS annotated. This approach allowed identifying possible enhancer+gene combinations, particularly when the original publication did not highlight a particular gene target.

Our genomic features included simple yet informative characteristics of the enhancers, like length in base pairs, distance to nearest TSS, etc, along with interactions on a gene level. These features are shown to influence transcript production and open up chromatin, mainly related to enhancer behaviour and regulatory significance. They increased the interpretability of the dataset prepared according to the curated data and allowed performing machine learning-based classification tasks. On the whole, these annotations facilitate (i) the understanding of the regulatory potential of enhancers, (ii) prioritization of enhancers to be followed up in the laboratory, and (iii) the improvement of downstream models.

2.3 Augmentation of the Dataset via Genomic Intersection

At the end of the literature curation process, we built a dataset of 38,266 distinct enhancer perturbation events, and they were all given a unique internal identifier. To enrich and validate this dataset, we used genomic intersection with the reference enhancer set, which was earlier assembled using public databases (EnhancerAtlas, FANTOM5, VISTA, and EnhancerDB).

Second, the curated dataset was used to examine enhancer loci that overlapped with annotated enhancer regions in the reference collection using bedtools intersect. By this process of intersection, this increased the number of enhancer entries by 75,977 records. These were then also combined with the initial literature-curated set based on their internal IDs, resulting in a final merged set of 83,744 enhancer perturbation records. To support thorough research and the follow-up use, all enhancer records in the complete version of ePerturbDB were complemented by a vast list of metadata. The genomic information in this included a unique enhancer identifier, chromosomal position, start and end coordinates, the version of the genome assembly, and, among others, internally assigned keys. A source annotation that was retained included original enhancer identities and whether the enhancer was classified as an enhancer or eRNA.

To record the experimental context, the following information was recorded: The sequences that were being targeted (such as sgRNA, shRNA, and siRNA), the perturbation method being employed, the effects of deletion or knockdown observed, and whether the effects on gene

regulation were direct or indirect. Additional biological context was provided by the information on cell lines, types of tissues, and gene names produced by proximity-based TSS connection and the literature. Quantitative variables such as p-values, Z-scores, CRISPRi scores, β-values and log2 fold change (log 2FC) were provided when available to point to the statistical and functional significance. All the original study names, Digital Object Identifier (DOI), PubMed ID (PMID), and the planned enhancer group name were retained together with the place of each publication of the data. The coordinate system was tested and was standardized to the hg38 reference assembly to ensure compatibility among the sets. Missing or conflicting metadata were curated manually, and all unnecessary entries were eradicated to ensure consistency and dependability.

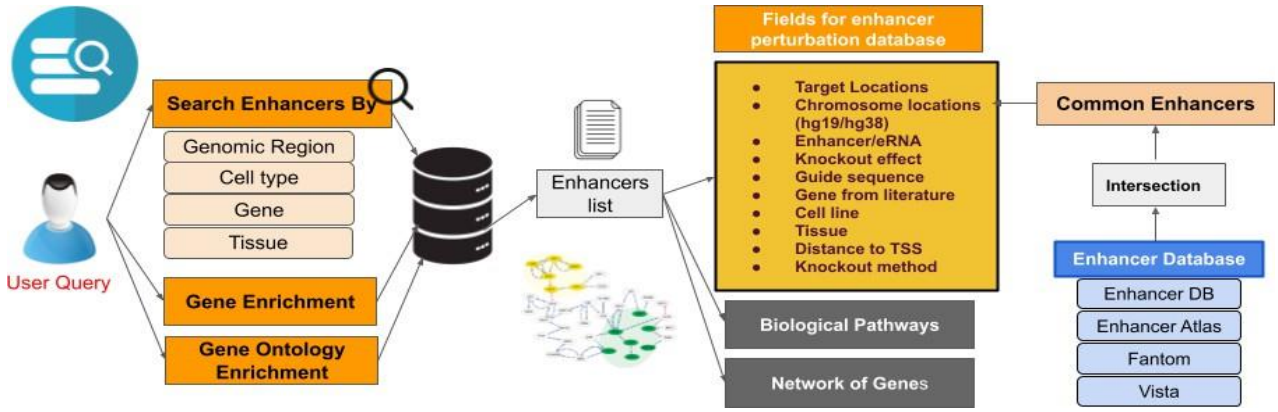


Figure 2.1: Graphical abstract of the database - ePerturbDB website.

2.4. Final Database Assembly

The definitive version of ePerturbDB consists of high-quality, empirically confirmed enhancer/eRNA perturbation events that have been processed with rigorous error checking, extensively literate annotation, and systematic incorporation of gene-level relationships. All genomic coordinates in the database are based on the GRCh38 (hg38) human genome assembly, which ensures compatibility with reference annotation and existing methods of genomic analysis. All enhancer entries are precisely mapped, linked to proximal gene information, and put into context using extensive experimental metadata, including the biological context, detected regulatory outcomes, and the type and method of interference. This rigorous system of annotation makes functional interpretation of enhancer activity possible and is also useful in describing gene regulatory pathways in much more detail. ePerturbDB supports numerous applications in

functional genomics, epigenetics, and the regulation of genes. A few of these include the classification and prediction of enhancer and eRNA functionality, the reconstruction of gene regulatory networks, meta-analyses of CRISPR-based functional genomics screens and the investigation of disease-pertinent regulational changes in particular cell forms or physiological contexts. ePerturbDB a unified, genome-scale resource, which integrates annotations of annotations found in well-known enhancer databases with elaborate experimental evidence picked up in the literature. It offers robust support to the study of enhancer biology, disentangle the relationship between enhancers and genes, and support data-driven, scalable discovery in regulatory genomics with close attention to data curation, genomic normalization, and quality control.

2.5 Functionality and Architecture of ePerturbDB

The ePerturbDB is an interactive database software based on web applications developed using Streamlit and capable of investigating, analyzing, and sharing enhancer perturbation data sets. The platform aims to facilitate exploration, analysis, and distribution of selected enhancer perturbation datasets. Instead, with its easy-to-use interface, a researcher may access standardized experiments: enhancer knockdown and knockout experiments in multiple tissues, cell lines, and methods conducted by multiple labs.

The structure of the platform architecture consists of multiple modules that are divided into independent pages or tabs and offer exclusive analytical tools in order to enable enhancer-oriented research. Some of them are Enhancer Pathway and Gene Enrichment Analysis, Coordinate-Based Enhancer Query, Advanced Search Tools, Statistical Visualization, Enhancer Identification, and Data Submission. Together, these modules help one perform comprehensive genomic analyses, such as visualizing the statistical trends and doing the pathway-level functional enrichment to find the changed enhancers and gauging the connections between genes. Each of the modules is then explained in this paper, their intention, methodology, and their suitability to enhancer-centered genomics studies.

2.5.1 Enhancer Identification

Users can look for enhancers or eRNAs using a variety of biological and experimental parameters thanks to the flexible and intuitive interface offered by the ePerturbDB Enhancer Identification module. A structured, form-based query method that makes it easier to efficiently retrieve perturbation data from the underlying database makes this functionality available.

To obtain information on a certain enhancer or eRNA entry, users can conduct searches by providing a unique Enhancer ID. Furthermore, there are filters to narrow down searches by chromosomal count, allowing for locus-specific analysis of the human genome. By choosing the preferred regulatory element type, users can also differentiate between transcribed enhancer RNAs (eRNAs) and canonical enhancer elements. The tool allows users to query items containing

targeting sequences, such as siRNA, shRNA, or sgRNA, as part of their perturbation experiments, facilitating sequence-based retrieval. Users can separate records with statistically proven perturbation effects by using filters based on experimental significance. Additionally, by enabling selection from the main enhancer databases included in ePerturbDB, such as FANTOM, VISTA, EnhancerDB, and EnhancerAtlas, the module offers source-specific queries.

Users can investigate perturbation-specific effects by using filters based on perturbation approaches (e.g., CRISPR/Cas9, CRISPR interference [CRISPRi], and shRNA) to account for methodological heterogeneity. Researchers can focus on enhancer perturbations that produce detectable biological effects by applying an additional filter for phenotypic effects. Context-dependent investigation of enhancer function and gene regulation is made possible by the ability to refine searches by entering certain tissue types, cell lines, or gene names.

When a search query is run, the system provides a list of enhancer or eRNA entries that match, together with pertinent metadata such as biological context, perturbation method, targeted sequences, and genomic coordinates. In addition to being viewable through the web interface, the findings can be downloaded for offline examination in CSV format.

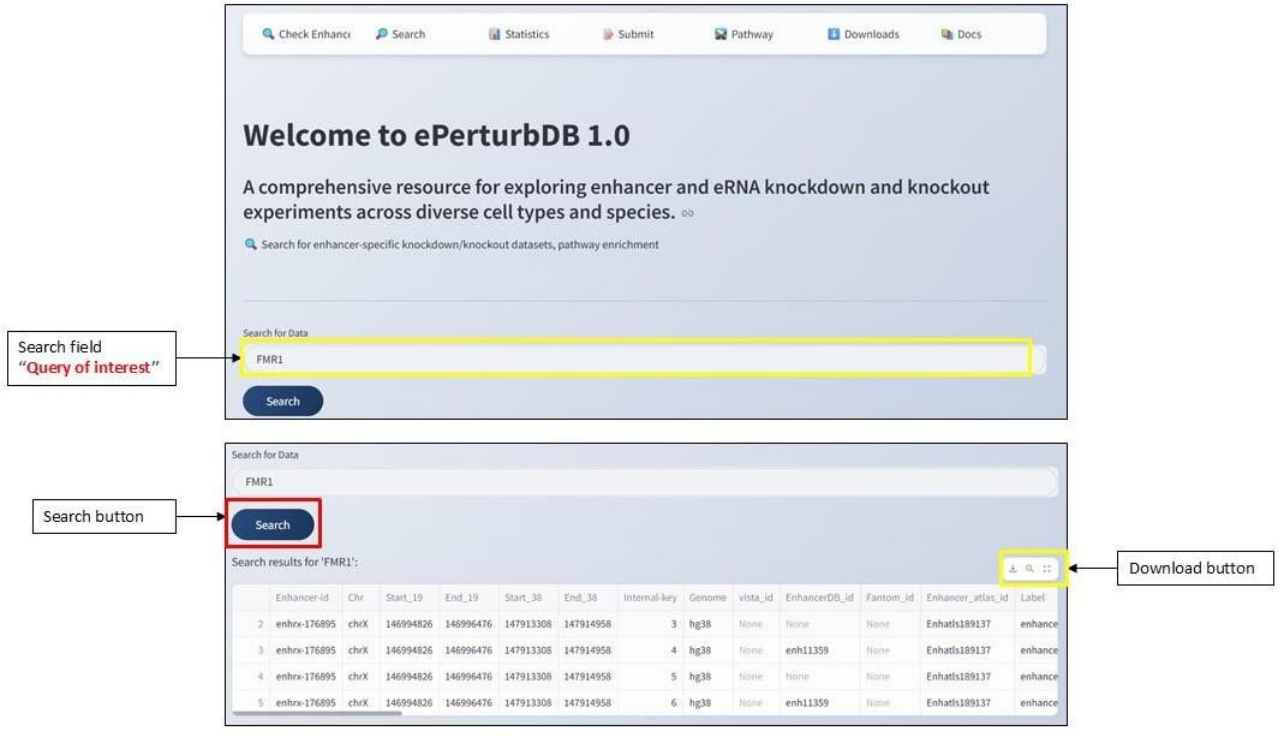
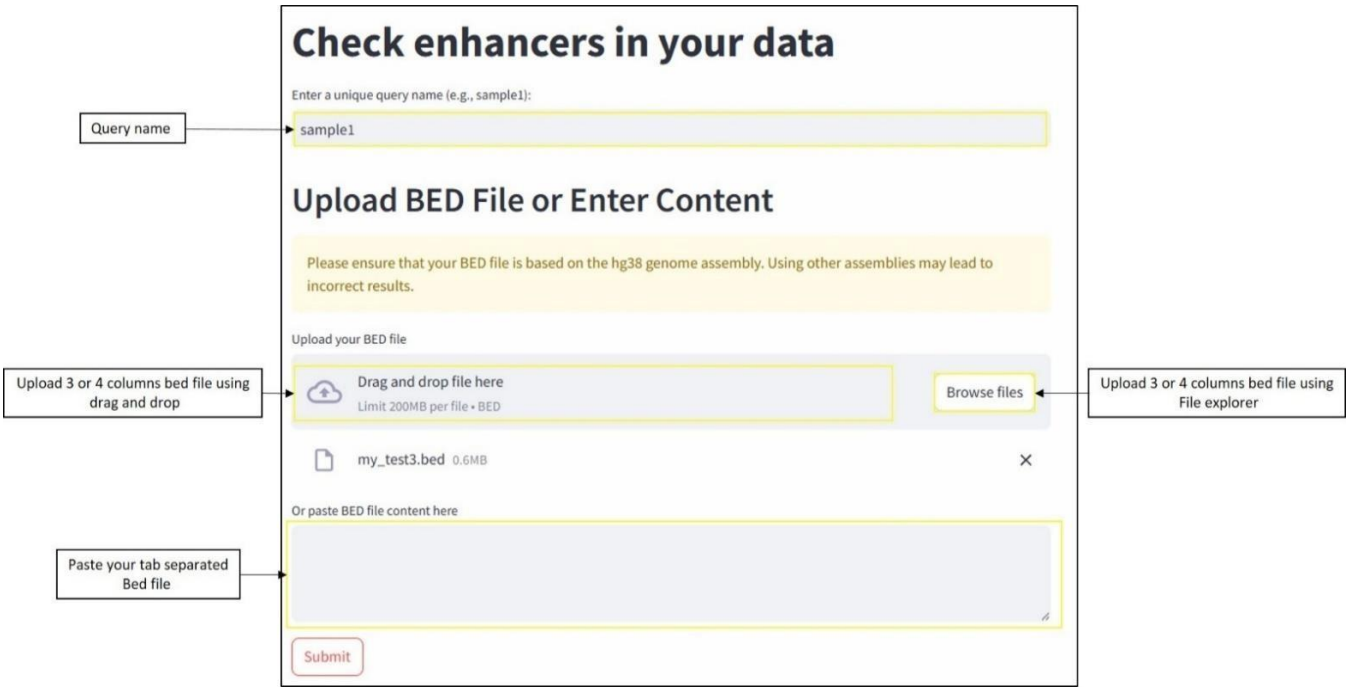


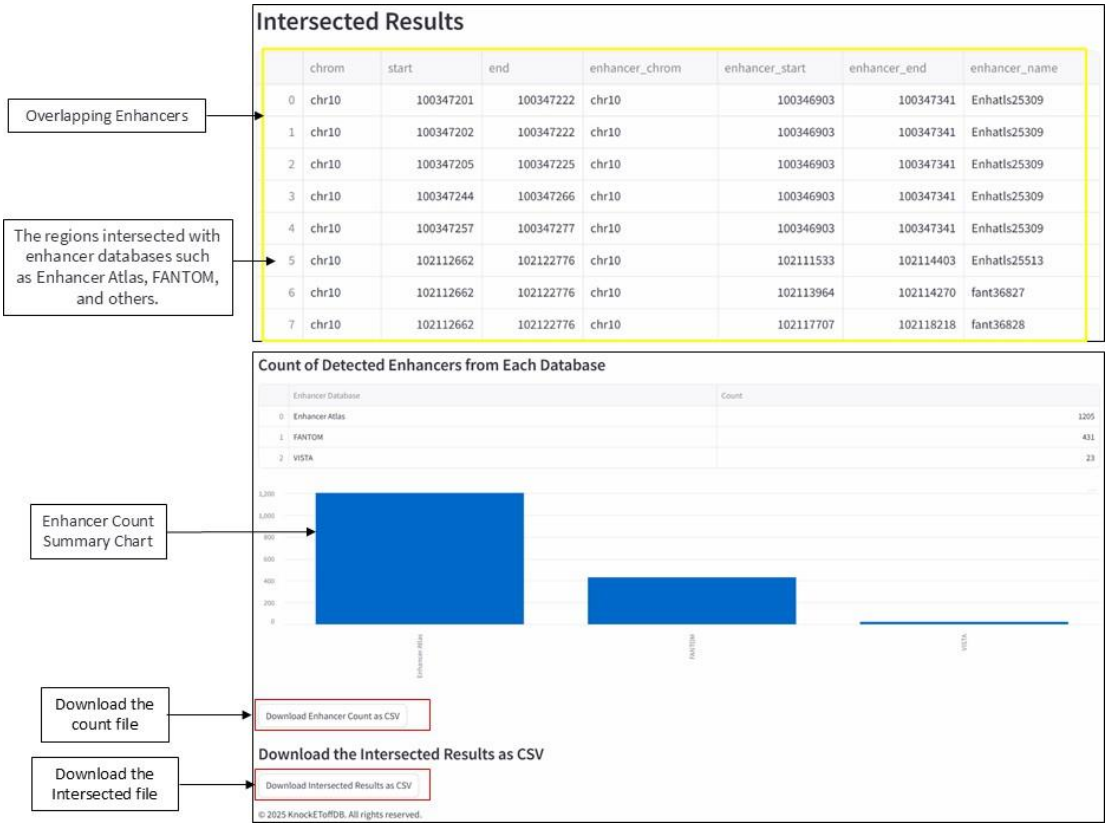
Figure 2.2: ePerturbDB – Interface of Enhancer Database

2.5.2 Check Enhancers (by Genomic Region)

Through intersections with reference enhancer datasets, the Check Enhancers module allows users to examine if certain genomic regions of interest match previously characterized enhancers. Researchers can submit their genomic coordinates for comparison against approved enhancer collections thanks to this capability, which accepts input in BED format.

By giving their query a title, users can start an analysis. They can then upload a BED file or just enter genomic area data into the text input field. In order to find overlaps with enhancer annotations from well-known resources like FANTOM, EnhancerAtlas, and other integrated databases inside ePerturbDB, the input data are internally transformed into CSV format and processed utilizing computational techniques subsequent to submission. A comprehensive list of overlapping enhancer entries, complete with genomic locations and source database annotations, is the output. A clear description of the findings is also provided via a visual summary chart that shows the distribution of overlapping enhancers per reference database. To aid with validation, downstream analysis, or integration with other genomic workflows, the enhancer list and summary visualization can be downloaded in easily navigable forms.





	chrom	start	end	enhancer_chrom	enhancer_start	enhancer_end	enhancer_name
0	chr10	100347201	100347222	chr10	100346903	100347341	Enhatls25309
1	chr10	100347202	100347222	chr10	100346903	100347341	Enhatls25309
2	chr10	100347205	100347225	chr10	100346903	100347341	Enhatls25309
3	chr10	100347244	100347266	chr10	100346903	100347341	Enhatls25309
4	chr10	100347257	100347277	chr10	100346903	100347341	Enhatls25309
5	chr10	102112662	102122776	chr10	102111533	102114403	Enhatls25513
6	chr10	102112662	102122776	chr10	102113964	102114270	fant36827
7	chr10	102112662	102122776	chr10	102117707	102118218	fant36828

	Enhancer Database	Count
0	Enhancer Atlas	1205
1	FANTOM	431
2	VISTA	23

Figure 2.3: Result of Check Enhancers page of ePerturbDB

2.5.3 Search by Biological Criteria

With the use of a collection of specialized search options offered by the Search by Biological Criteria module, users can locate enhancer perturbation records by using important biological metadata. Using biologically relevant filters like gene name, cell line, tissue type, or genomic area, this functionality facilitates targeted enhancer study. After executing a targeted query, each sub-module produces a structured table with enhancer items that fit the user-specified specifications.

i. Search by Genomic Region:

Users can enter genomic coordinates using this sub-module by entering the BED content into the designated text box or by uploading a file in BED format. The system produces a list of enhancer perturbation records that intersect with the selected regions after processing the input. The findings support additional research into region-specific enhancer activity by including pertinent enhancer annotations and metadata.

ii. Search by Gene Name:

Users can access enhancer perturbation data linked to certain genes of interest using this sub-module. The system searches the curated database using the gene name that users enter, returning enhancer entries that are either experimentally reported to regulate the specified gene or genes. Each matched enhancer's genomic coordinates, perturbation technique, regulatory element type (enhancer or eRNA), and biological context (tissue or cell line employed in the study) are all included in the output that is produced. Quantitative metrics like p-values, $\log_2$ fold change, and phenotypic impacts are also included when accessible.

iii. Search by Cell Line:

Users can investigate enhancer or eRNA perturbation events unique to a particular cell line with this module. Users can access enhancer information associated with perturbation experiments carried out in a particular cellular setting by inputting the name of the relevant cell line. Enhancer coordinates, perturbation method, related genes, biological effects, and available quantitative metrics (e.g., $\log_2$FC, p-values) are all included in the output. This tool makes it easier to analyze enhancer function and regulatory dynamics according to cell type.

iv. Search by Tissue/Cell Type:

With the use of this module, users can query enhancer perturbation data according to particular cell lineages or tissue types. Users can access enhancer or eRNA records related to research carried out in a particular biological setting by inputting the type of tissue or cell of interest. Tissue-specific analysis of regulatory elements is made possible by the output, which comprises genomic coordinates, perturbation techniques, target genes, and pertinent biological results.

v. Predict Target Genes: With the help of this module, users can find probable target genes that are situated within a specific genomic range (for example, ±5 kb) of enhancer areas that they have specified. The technology predicts neighboring genes based on TSS proximity after users upload a BED file or paste genomic coordinates. The following list of anticipated target genes supports the functional interpretation of enhancer loci provided by the user.

2.5.4 Database Statistics

An extensive summary of the variety and an overview of experimental data curated within ePerturbDB can be found on the Database Statistics page. The database currently has 14,080 unique enhancer entries and 83,743 enhancer perturbation records. While proximity-based annotation has mapped 4,498 distinct adjacent genes to enhancer loci, manual curation of the literature has found 604 target genes that have been empirically confirmed. The information represents a wide range of biological contexts and includes 30 different tissue and cell types. The

methodological diversity is demonstrated by the distribution of perturbation approaches, which include CRISPR-Cas9 for 58,646 entries, CRISPR interference (CRISPRi) for 24,938, and combined CRISPRi/CRISPRn for 44. 41 siRNA, 33 siRNA + LNA, 20 shRNA, 14 LNA ASOs, 3 shRNA + CRISPRi, 2 siRNA + GapmeR, and 2 TALEN entries are also included in the dataset. These figures highlight ePerturbDB's depth, experimental diversity, and biological scope, providing a solid basis for analyzing enhancer activity in a variety of cellular and molecular settings.

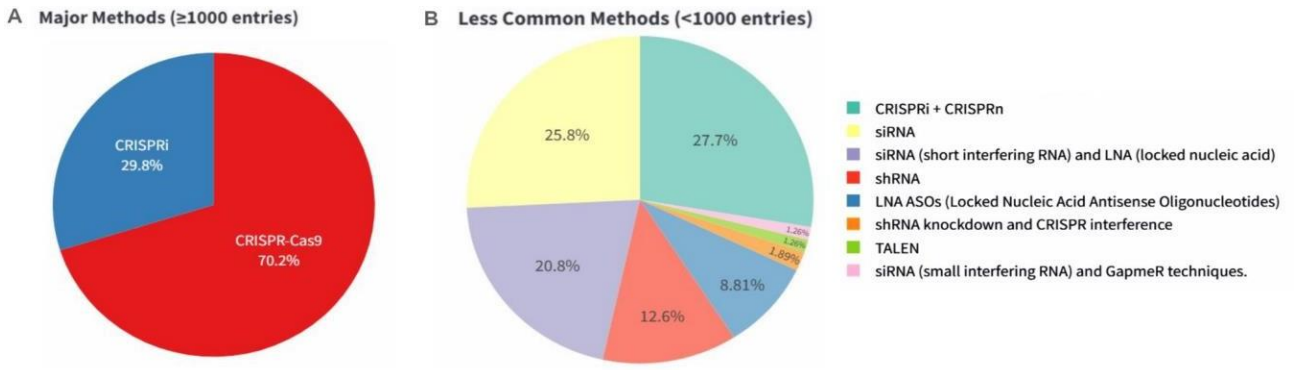


Figure 2.4: Experimental method used for perturbing the enhancers and eRNA using the major and the less commonly used methods

Method of Knockout	Count
CRISPR-Cas9	58646
CRISPRi	24938
CRISPRi + CRISPRn	44
siRNA	41
siRNA (short interfering RNA) and LNA (locked nucleic acid)	33
shRNA	20
LNA ASOs (Locked Nucleic Acid Antisense Oligonucleotides)	14
shRNA knockdown and CRISPR interference	3
siRNA (small interfering RNA) and GapmeR techniques.	2
TALEN	2

Table 2.1: Count of the method used for the knockout/knockdown of the enhancers/eRNA in ePerturbDB

2.5.5 Submit New Data

Users can add their own experimentally acquired enhancer or eRNA perturbation data to ePerturbDB through the Submit page, which serves as an interactive community contribution platform. This feature helps keep the database a dynamic resource and encourages collaborative data augmentation. A structured submission form that is intended to collect important experimental metadata is presented to users. The genomic coordinates of the affected area (in human genome assembly format), related or anticipated target genes, the experimental technique (e.g., CRISPR-Cas9, CRISPRi, shRNA, siRNA), and any phenotypic effects that were noticed after the perturbation are all necessary pieces of information. To guarantee traceability and source legitimacy, authors must also include bibliographic information for the study, such as DOI, PubMed ID (PMID), or journal citation. This community-driven submission process advances studies in functional genomics and regulatory biology while bolstering ePerturbDB's long-term objective of acting as an extensive, current library for enhancer perturbation data.

2.5.6 Enhancer Pathway & Gene Enrichment Analysis

The Enhancer Pathway & Gene Enrichment Analysis module finds enriched genes and related biological pathways, allowing users to assess the biological significance of enhancer regions. This characteristic is especially helpful for comprehending how enhancer perturbations affect regulation in a functional genomic setting. Genomic coordinates in BED format are uploaded or pasted by users to start the analysis. To act as a personalized reference set for enrichment comparison, a backdrop file can be optionally supplied; if not, the tool will fall back on a pre-established reference dataset. Important input parameters are the number of top genes to report (minimum of 5, default is 20) and the distance criterion (default: 5000 base pairs) for determining enhancer–gene closeness.

These parameters specify the mapping of enhancers to genes and direct the enrichment scoring process. A real-time status tracker shows the processing progress when the user configures the parameters and presses the "Run Analysis" button to start the analysis. Using databases like Reactome, KEGG, and WikiPathways, the technique maps neighboring genes to enhancer areas and carries out gene set enrichment. The "Analysis Results" page displays the findings, which comprise: A pie chart that summarizes the distribution of genes, a list of enriched genes with rankings, and Pathway enrichment results for each phrase that include statistical significance values (such as p-values). Every result can be downloaded in CSV format, allowing for additional investigation or connection with different analytical workflows.

This module supports downstream applications in functional annotation, pathway analysis, and hypothesis creation for experimental validation by making it easier to interpret enhancer perturbation data biologically. The output comprises route enrichment data from Reactome,

KEGG, and WikiPathways, an ordered list of genes enriched based on enhancer proximity, and a pie chart showing gene distribution. The matching p-value for each enriched pathway is shown, offering statistical evidence of the enhancer-associated genes' functional significance in particular biological processes.

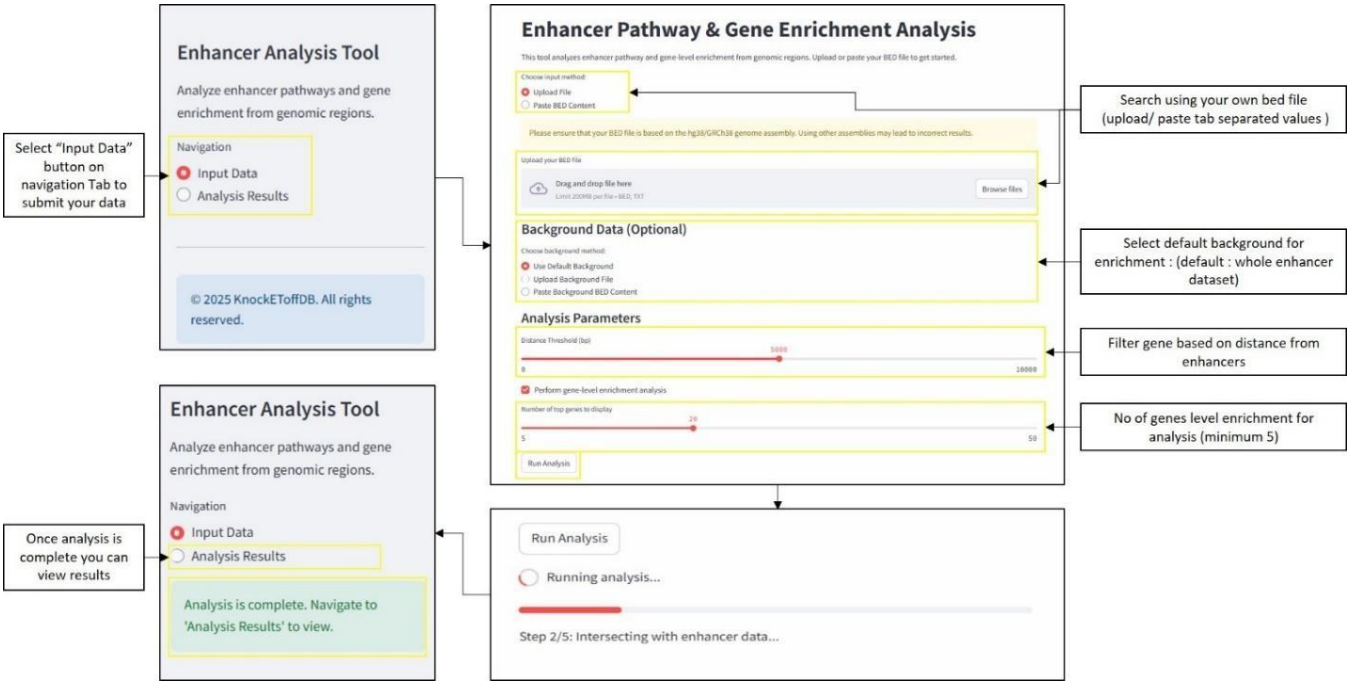
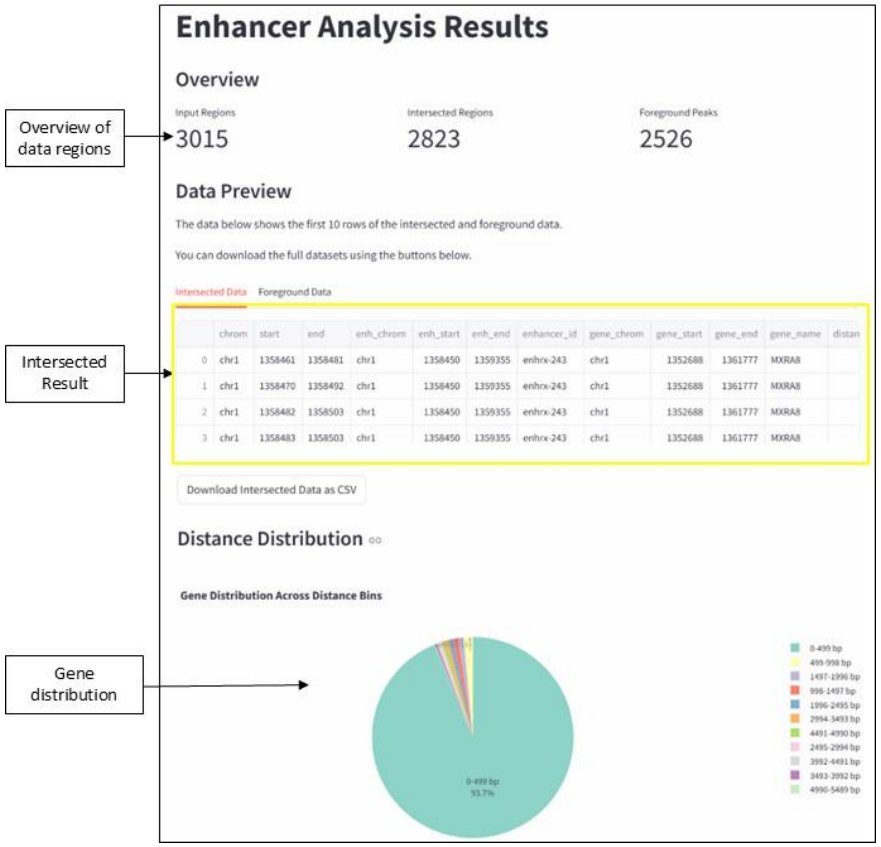
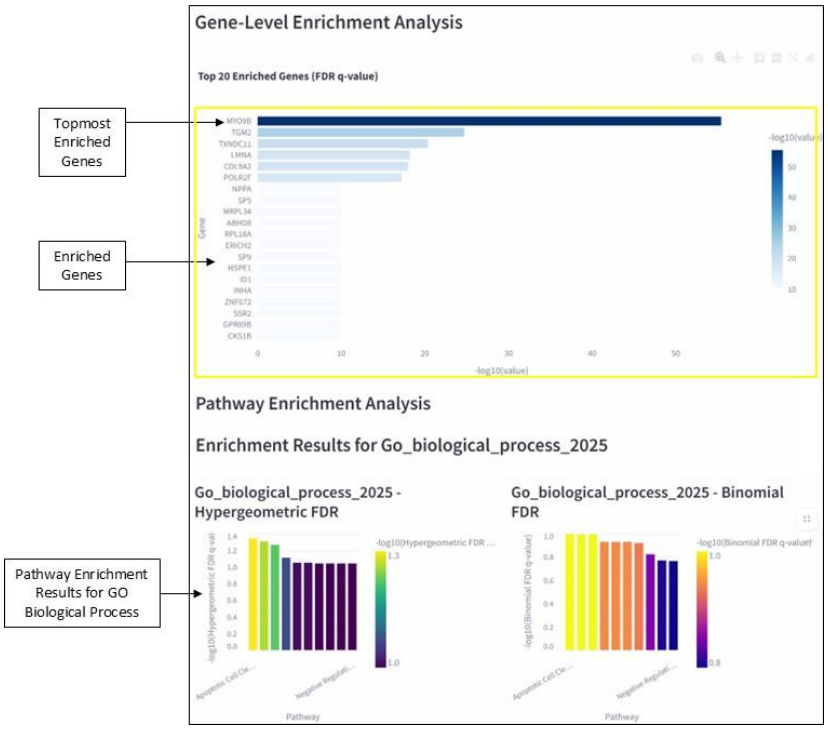


Figure 2.5: Enhancer Pathway & Gene Enrichment Page of ePerturbDB


Enhancer Analysis Results
Overview
Input Regions
3015
Intersected Regions
2823
Foreground Peaks
2526
Data Preview
The data below shows the first 10 rows of the intersected and foreground data.
You can download the full datasets using the buttons below.
Intersected Data
Foreground Data
Download Intersected Data as CSV
Distance Distribution
Gene Distribution Across Distance Bins
Overview of data regions
Intersected Result
Gene distribution


Gene-Level Enrichment Analysis
Top 20 Enriched Genes (FDR q-value)
-log10(value)
Pathway Enrichment Analysis
Enrichment Results for Go_biological_process_2025
Go_biological_process_2025 - Hypergeometric FDR
Go_biological_process_2025 - Binomial FDR
Pathway
Topmost Enriched Genes
Enriched Genes
Pathway Enrichment Results for GO Biological Process

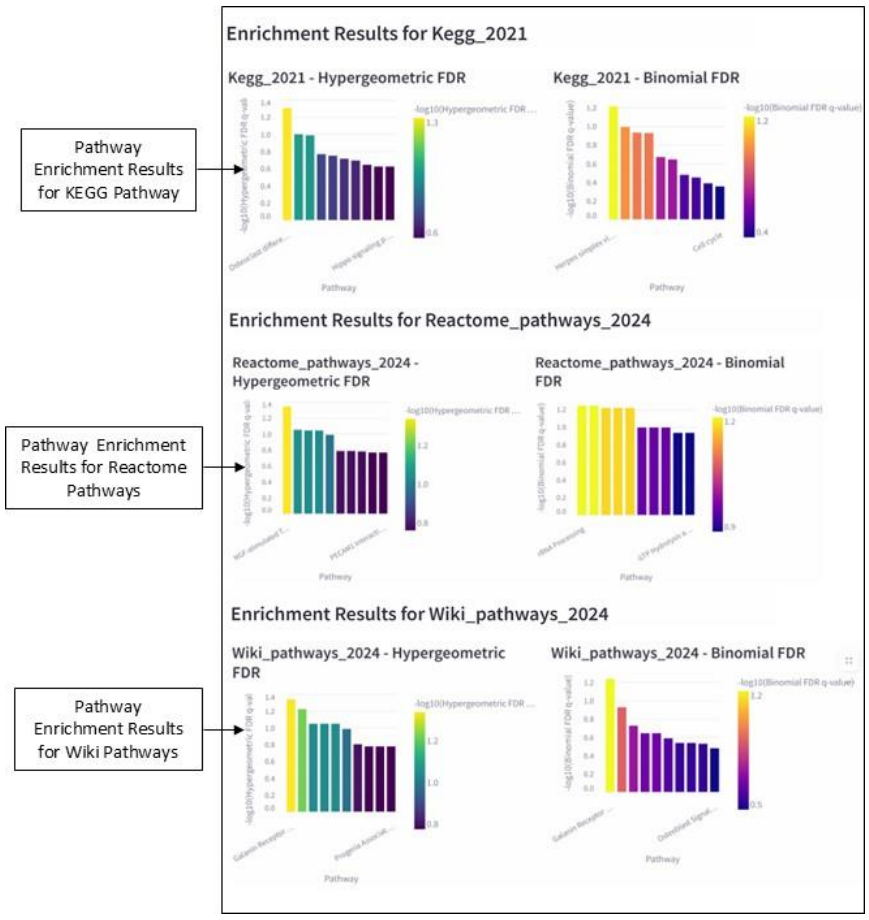


Figure 2.6: Result of Enhancer Analysis of Pathway & Gene Enrichment page of ePerturbDB

2.6 Generation of Cell Type-Specific Enhancer Regions Using ChIP-Atlas

We used the ChIP-Atlas platform to find enhancer regions specific to cell types, concentrating on the H3K27ac histone modification track, a reliable indicator of active enhancers. All cell types under the H3K27ac track were selected for the study to guarantee comprehensive biological coverage using the ChIP-Atlas Peak Browser. High-confidence peaks were retained by applying a threshold score of 50, which produced BED files that corresponded to several tissue classifications (e.g., Adipose, Blood, Bone, etc.).

Following download, each BED file was preprocessed into a standard four-column BED format, including an annotation for the cell type, start position, end position, and chromosome. Each peak's cell type metadata was extracted from the ChIP-Atlas annotation and entered into the fourth column. Using this column label as the identifier, the tissue-level BED files were divided into

distinct cell-type-specific BED files to aid in downstream analysis. To provide traceability and orderly storage, a defined naming scheme was implemented.

Then, bedtools intersect was used to intersect each cell type-specific BED file with the previously assembled reference enhancer set. This stage made it possible to identify probable cell-type-specific enhancers by identifying enhancer regions that overlapped with known enhancers and were active. An extensive collection of enhancer BED files tailored to individual cell types, each marked with exact genomic coordinates, was the result. These datasets are useful for functional annotation specific to cell types and may easily be combined with knockdown or perturbation data to further explore regulatory mechanisms.

2.7 Clinical Study of the Database

2.7.1 Case Study 1: Application in Triple-Negative Breast Cancer (TNBC) cell line

We carried out a case study by using epigenomic data of the patients via the PDX models, i.e., patient-derived xenograft of TNBC (triple-negative breast cancer), demonstrating therapeutic applicability and the utility of the ePerturbDB database. The aim was to examine the intersection between enhancer activity as demonstrated by real profiles in tumors and enhancer perturbations with experimentally confirmed perturbations in ePerturbDB.

H3K27ac ChIP-seq data describing enhancer activity of the entire genome with the help of histone acetylation profile were downloaded from TNBC PDX tumor models (GEO accession IDs: GSM6133016, GSM6133017, and GSM6133018). Such peak locations were then uploaded to the ePerturbDB platform via the check enhancers interface to locate any overlapping experimentally perturbed enhancers or enhancer RNAs (eRNAs). A lot of overlapping enhancer entries were identified in the search.

Findings from the PDX Model:

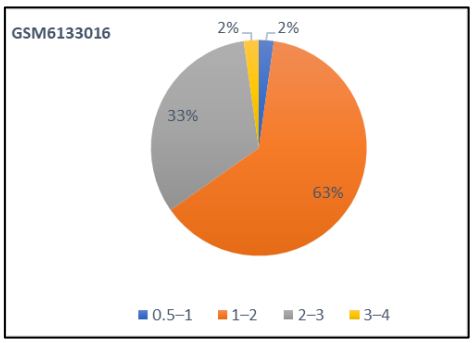
In sample 4272-TG5 (GEO: GSM6133016), we identified 321 overlapping enhancers that corresponded to ePerturbDB entries that had been experimentally validated. These enhancer regions were supported by perturbation data across four individual breast cancer studies, with 2,299 sgRNAs, and led predominantly by CRISPR-based methodology. Among these perturbations, notably a large part was occasioned by a research published by Lewis et al. [32] and Fei et al. [35]. In the case of CRISPRi screening, Lewis et al. screened the breast cancer cell line and identified that the enhancement of expression caused significant changes in gene expression. According to our research, 37 percent of sgRNAs with GSM6133016, 19 percent with GSM6133017, and 59 percent with GSM6133018 were associated with changes in target gene expression at 13. Our research is consistent with this claim. Indeed, our results are consistent with

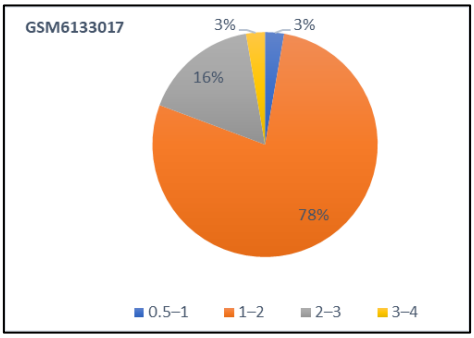
the clinical heterogeneity of TNBC by suggesting intense patient-specific enhancer use and different levels of perturbation susceptibility.

Functional Enrichment of Targeted Enhancers

To examine the biological relevance of the coinciding enhancer regions, we carried out the gene ontology enrichment analysis using the GREAT tool. The input was formed by the genomic loci of enhancer regions in TNBC PDX samples. Enrichment analysis of GO keywords found those that are functionally significant in terms of tube formation, cell junction organisation, and epithelial formation. Such processes have been known to be crucial in terms of carcinogenesis, metastasis, and remodeling of tissues in breast cancer. These findings confirm the hypothesis that the enhancers identified in ePerturbDB are not only physiologically functional on the disease-specific background but also experimentally validated.

The findings of the enhancement of perturbation overlaps in PDX models of TNBC are represented in the section below. Specifically, findings of a study on Lewis et al. [32] that had experimental outcomes that aligned with the in silico predicted enhancer regions in the TNBC models were exemplified in a pie chart that represented the scale of the reported fold changes in the expression of various target genes due to enhancer or eRNA perturbation. Furthermore, a bar plot was drawn indicating the number of enhancer perturbation events (sgRNAs or eRNA knockdowns) of four independent breast cancer studies overlapped with the estimated regions of enhancers within the PDX models of three TNBC patients. Moreover, the functional enrichment analysis of genes located near the disrupted enhancers revealed crucial biological processes, enhanced within the overlapping areas, e.g., the processes of tumor development and epithelial cell organization. Such discoveries support their functional applicability in breast cancer biology since they emphasize that the consistent availability of predicted active enhancers in therapeutic TNBC specimens instigates a common acknowledgment of predicted dynamics with experimentally validated data in the perturbations.





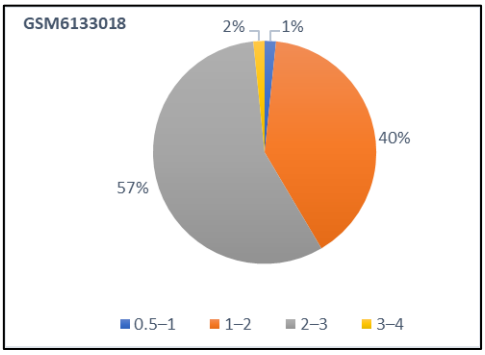


Figure 2.7: Distribution of fold change of target genes by the enhancer and eRNA perturbation overlapped with enhancers predicted reported by Lewis et al.[32]

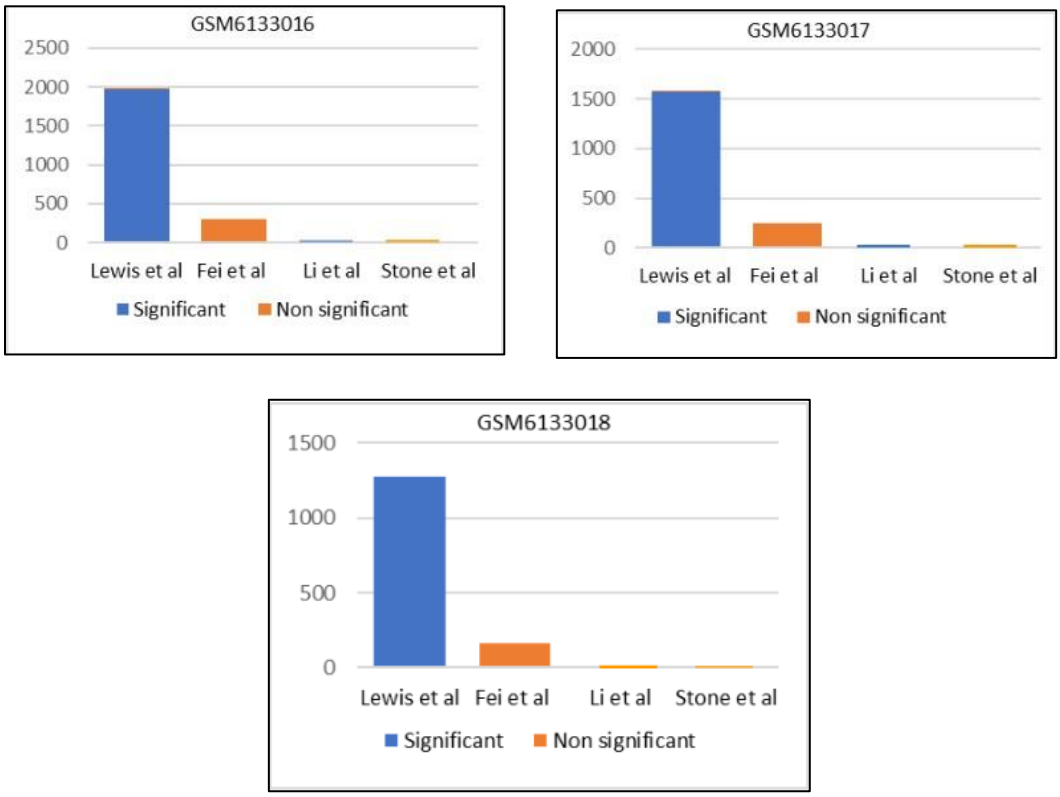
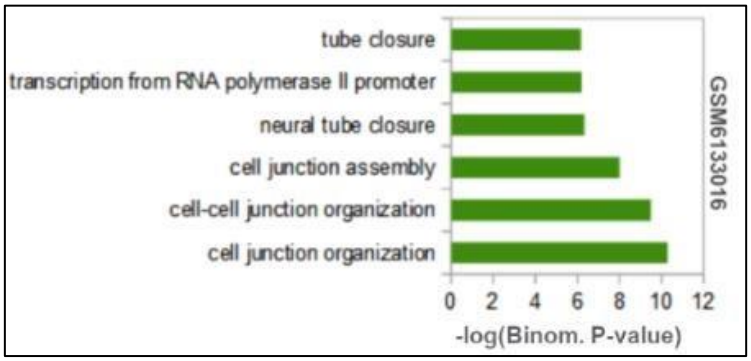
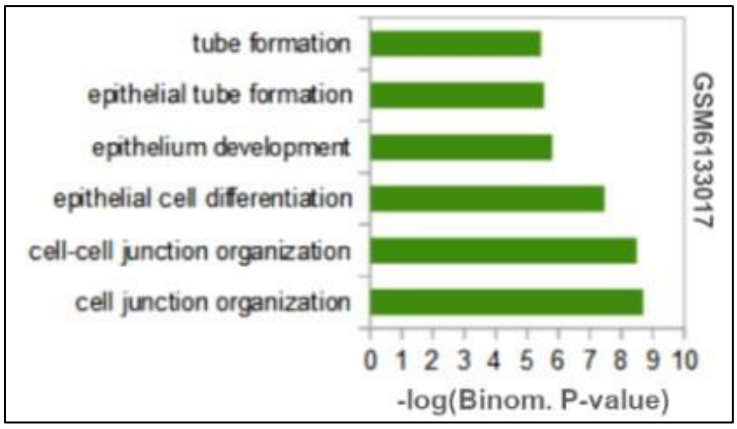


Figure 2.8: No. of sgRNA or eRNA knockdown perturbation experiments overlapped with predicted enhancers in three triple-negative breast cancer patients from different studies





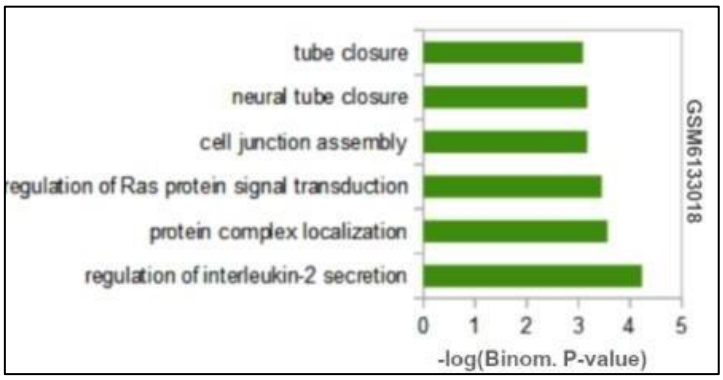


Figure 2.9: Enriched pathways for genes to the perturbed enhancers in TNBC cell lines overlapped with predicted enhancers in four studies

2.7.2 Case Study 2: Analysis of eRNA-Linked Survival Data from the TCEA Breast Cancer Cohort

The second case study of breast cancer data of The Cancer Epigenome Atlas (TCEA) was conducted to gain additional evidence on the capability and translational relevance of the ePerturbDB platform. To obtain utility insight into disease-relevant regulatory elements, our study aimed to define whether enhancer RNAs (eRNAs) originally linked with patient survival in breast cancer are captured in the enhancer perturbation dataset across ePerturbDB. The analysis focused on 327 eRNA regions, which had been annotated on the TCEA cohort with the help of the survival properties, including the hazard ratios and Cox proportional hazards (overall survival) value. Specific attention was paid to the eRNAs associated with the BRCA gene, which is actively known to modulate the process of breast cancer and its clinical outcome.

The eRNA areas with statistically significant survival associations (as evidenced by low p-values or high hazard ratios) were given high priorities, whereas those that are not statistically significant (as evidenced by high p-values or low hazard ratios) may be given a lower priority after ranking by clinical significance. To make sure that it is compatible, the eRNA genomic coordinates (using hg19 genome assembly) were transformed, by using the UCSC LiftOver tool, to hg38. To be used in subsequent database queries, the curated and standardized list of eRNA was transformed into a file in the BED format that contained the coordinates of chromosomes, start, and end.

Using the ePerturbDB query interface, the system identified 75 overlapping entries that matched the targeted locations after feeding in the prepared BED file. These were entries of three independent studies, which were curated within the database, that were sgRNA-targeted perturbations that overlapped with the loci of interest in eRNA. The experiments applied the methods of CRISPR/Cas9 and CRISPR interference. Surprisingly, Wang et al. [33] were the origin of a significant proportion-56 of 75 sgRNAs, or about 74%. Conversely, reports by Fei et al. [35] and Kelly et al. [27] had argued great biological relevance to the regulatory relevance of these loci by reporting that many sgRNAs that target the same or overlapping eRNA regions had effects that were meaningful by functional understanding. It was found by the elimination of duplicate entries to further refine the analysis and retain only non-overlapping areas of enhancers compatible with unique sgRNA hits. The number of perturbation events in each enhancer site was given as a barplot, and based on it, it was possible to define hotspot enhancers for which a set of studies was carried out. Revealed recurrent sites of enhancers are more likely to be functionally relevant and were defined as potential candidates to be further tested.

The nearby genes in the regions of the intersected eRNA were then checked with the ePerturbDB Gene Enrichment module. This functional annotation marked PPP1R15B as a significantly enriched gene that can be regulated by the overlapping eRNAs. Further, the platform provided statistical results reflecting quantitative evidence of the biological values of such regulatory relationships, including the p-value of the gene ontology (GO) enrichment. The ePerturbDB case

study illustrates how, in combination with experimental evidence on perturbations, ePerturbDB can create integrated evidence of perturbation across experiments with clinical transcriptome data to elucidate eRNA loci with regulatory importance and survival relevance in cancer. The database allows for identification of therapeutic implementable regulatory elements through the identification of eRNAs with strong clinical links and overlapping them with experimentally validated enhancer perturbations. The possibility of steering novel therapeutic opportunities in breast cancer and other disorders is exhibited by the identification of salient genes such as PPP1R15B and over-represented biological functions of the platform.

The findings had opened up a couple of new insights into the relationships between experimentally validated enhancer perturbations in ePerturbDB and clinically relevant eRNAs in the TCEA breast cancer cohort. To emphasize the point that these enhancer perturbations used in the present research were based on three individual studies, a pie chart was plotted displaying the number and proportions of matching enhancer elements across the studies dependent upon their research of origin: Fei et al. [35] Kelly et al. [27], and Wang et al. [33]. In order to depict the rate of perturbations per enhancer, localized to specific chromosomal regions, the density of sgRNAs associated with enhancers across multiple genomic locations was shown. Also, an overview of the source articles used in intersecting the confounding enhancer perturbations displays that only Fei et al. and Kelly et al. were implying a significant phenotypic effect, although Wang et al. had the highest number of co-occurring entries. Further enrichment analysis led to the identification of the top genes around the disrupted enhancer regions, which were of possible regulatory effects. These enhanced areas that overlapped were ultimately used to do a gene ontology (GO) enrichment analysis to identify the top biological processes and pathways that might be affected by these breast cancer-associated eRNA-associated enhancers.

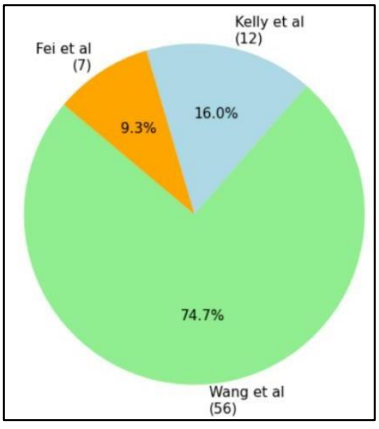


Figure 2.10: Intersected enhancer perturbation distribution reported from 3 studies (Fei et al.[35], Kelly et al.[27] and Wang et al.[33]) with TCEA-BRCA-eRNA

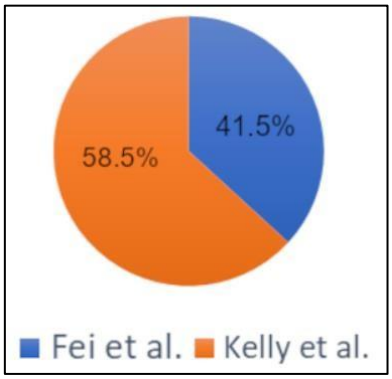


Figure 2.11: Significant knockout effect of the overlapped enhancer perturbation distribution in the corresponding study

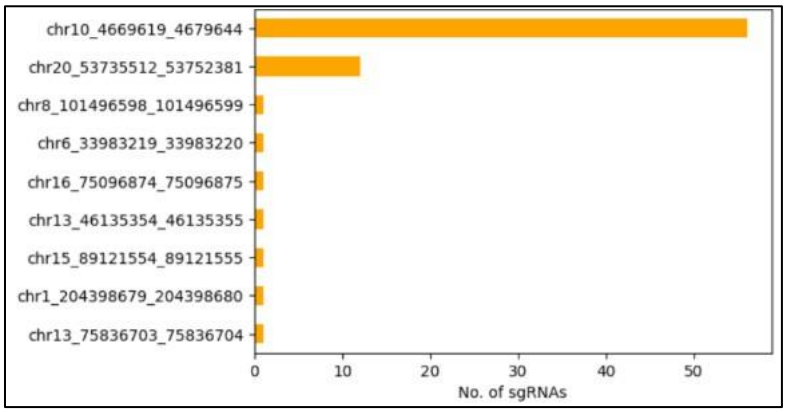


Figure 2.12: Number of sgRNAs present for each enhancer chromosome location found in the overlapped eRNA associated with Breast survival (in TCEA database)

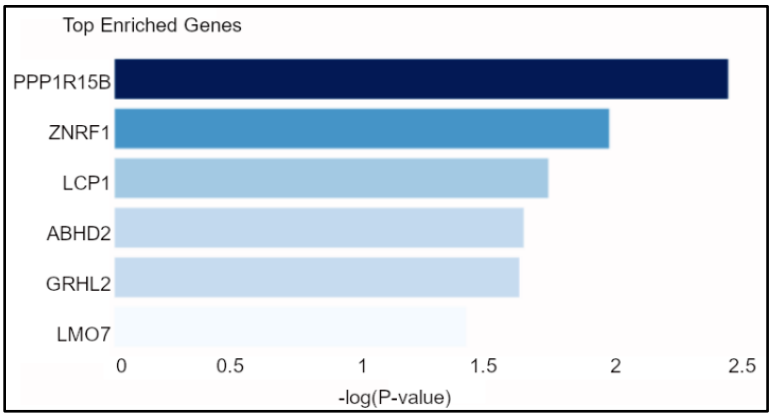


Figure 2.13: Gene enrichment for perturbed enhancers overlapped with TCEA-BRCA-eRNA.

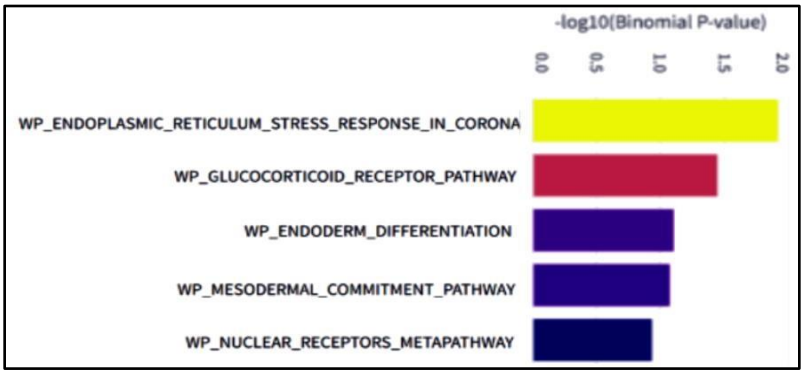


Figure 2.14: Gene ontology enrichment for perturbed enhancers overlapped with TCEA-BRCA-eRNA.

CHAPTER - 3

Enhancer Function Prediction Using Machine Learning Model

An essential first step in comprehending gene regulation networks and their involvement in cellular functions and disease pathways is the functional characterisation of enhancers. This chapter introduces machine learning (ML)-based techniques for forecasting the regulatory role of enhancers and the genes they target. The first method infers enhancer–gene–pathway relationships by using transcription factor (TF) binding data close to gene promoters. The approach seeks to determine which enhancer regions are functionally associated with genes inside particular biological pathways by including enhancer activity signals, such as peak scores from epigenomic tests. This formulation gives priority to enhancer–gene interactions related to signaling, development, or metabolism and permits a pathway-level understanding of enhancer regulation.

The second strategy leverages machine learning trained on CRISPR-based functional genomics data to construct enhancer–gene connections to validate the enhancer perturbation data curated in ePerturbDB. This method predicts whether perturbing a particular enhancer would impact gene expression by using genomic and epigenomic characteristics of enhancer regions, such as histone modifications, chromatin accessibility, and distance to transcription start sites.

3.1 Enhancer Function Prediction using Transcription Factor Binding and Promoter Activity

The functional significance of enhancers in gene regulation was computationally predicted using a machine learning-based system that combines chromatin accessibility data with enhancer and promoter annotations. By measuring regulatory activity across many tissues, this method seeks to pinpoint enhancer areas that are most likely to be involved in transcriptional control, especially through their interactions with tissue-specific promoters.

The well-known database of human enhancers, the VISTA Enhancer Browser, provided experimentally verified enhancer coordinates. The dataset, which is based on the GRCh37/hg19 human genome assembly, contains enhancers with confirmed tissue-specific activation. Across many circumstances and tissue types, promoter regions are made up of promoter peaks that aggregate transcription start sites. After being formatted in BED format, the two datasets were concatenated to create a single genomic annotation file. This combined file had 618,693 genomic intervals with a unique region identifier for each entry, along with the corresponding chromosome, start, and end coordinates. To model tissue-specific enhancer activity, a binary one-hot encoded matrix was created.

3.1.1 Quantification of Enhancer and Promoter Activity via Epigenomic Signals

We used the ChIP-Atlas database's histone modification profiles to infer the chromatin-level activity and regulatory potential of enhancer and promoter sites. A large collection of ChIP-seq datasets encompassing a variety of transcription factors and histone marks may be found in this repository. Active histone marks, particularly H3K27ac and H3K4me1, which are recognized markers of enhancer and promoter activity, were the main focus of our investigation. A total of 5,613 ChIP-seq peak data representing various human tissues and cell types were downloaded in BigWig (.bw) format. These files include genome-wide signal intensity profiles that show histone mark enrichment and chromatin accessibility.

Conversion of BigWig to BedGraph Format

The UCSC Genome Browser utilities' bigWigToBedGraph utility was used to convert each BigWig file into BedGraph format so that quantitative analysis could be performed. Genomic intervals and the corresponding signal intensity levels are represented by four-column data entries in the generated BedGraph files. The level of histone modification enrichment at particular genomic locations is reflected in these signals.

After conversion, bedtools map, which calculates summary statistics for overlapping genomic areas, was used to map the BedGraph files onto the combined enhancer-promoter BED file. The overlapping intervals in the BedGraph files were used to compute mean signal values for each enhancer or promoter area. By quantitatively expressing the enrichment of histone marks and, consequently, the regulatory capacity of the corresponding genomic sites, these mean values functioned as activity scores. Enhancers were categorized according to their tissue-specific regulatory potential using these calculated activity scores as features in a supervised machine learning framework. We were able to methodically assess enhancer function and activity concerning promoter proximity and chromatin dynamics thanks to our integrative approach.

The bedtools map tool was used to statistically assess the enrichment of histone modification peak signals over regions of enhancer and promoter. With the use of this program, quantitative signal data, like those in BedGraph files, can be projected onto user-specified genomic intervals. Using ChIP-seq data for active histone marks (e.g., H3K27ac, H3K4me1), the BedGraph files in this study were generated from ChIP-Atlas and included 5,613 distinct sample identifiers (SRX/DRX accession numbers) from various human tissues and cell types.

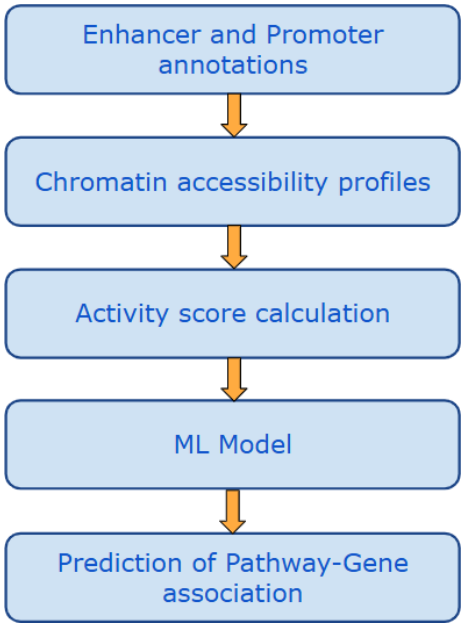


Figure 3.1: Workflow of the Prediction of Enhancer-Pathway-Gene Association

A BED-format file including 618,693 enhancer and promoter regions made up the reference set. To calculate the mean signal value for each overlapping genomic interval, each BedGraph file was mapped onto this reference using bedtools map with the summary statistic set. Histone modification intensities across regulatory sites could be aggregated using this method, resulting in a continuous score that represented chromatin accessibility and possible regulatory activity.

While low or nonexistent signals suggested either repression or inactivity under the specified biological context, regions with high average scores showed significant enrichment of histone marks, which are often linked with active enhancers or promoters. Following mapping, the score outputs from all 5,613 histone datasets were combined into a single data matrix, with each column denoting the mean signal score from a particular histone modification experiment (sample ID) and each row representing a genomic region (enhancer or promoter).

3.1.2 Construction of Enhancer-Promoter Score Matrix

After signal aggregation, the calculated mean signal values were extracted from the last column by processing the output files from Bedtools Map. A thorough enhancer-promoter activity matrix including 5613 columns (sample-specific histone signal scores) and 618,693 rows (genomic areas) was created by combining these values column-wise. We filtered this matrix to keep only non-disease-associated ChIP-seq datasets to guarantee biological relevance and prevent confounding effects from pathological samples. This was accomplished by using a metadata-based filtering technique that eliminated sample IDs linked to medical problems using keyword-based matching. 2,256 high-confidence, non-disease-associated sample IDs were left over after this phase. The final

dimensions of the filtered matrix were 618,693 × 2,256. The input feature set for the creation of supervised machine learning models was this matrix. Utilizing histone modification signal patterns to forecast enhancer activity, tissue selectivity, and regulatory capacity in a way that is both statistically sound and biologically interpretable was the aim.

3.1.3 Machine Learning Pipeline

The goal was to develop a machine learning (ML) model that could forecast gene-pathway association using enhancer peak scores as predictive factors. The primary objective of the model is to ascertain if enhancer regions are functionally linked to genes within specific biological pathways, such as metabolic, developmental, or signaling pathways, based on their chromatin activity profiles. To ascertain if a certain enhancer is associated with any gene in a given route, a binary classification model was trained for each pathway. To enable dependable training and interpretation, the modeling framework was implemented in R using a set of popular machine learning tools and dependencies. These included SHAPe for xgboost, RandomForest, caret, E1071, ROCR, Mltest, Sperrorest, and Glmnet.

Dataset used

Three crucial input files provided the training data for the classification model. First, the set of genes involved in different biological pathways was defined using a Pathway Genes file in GMT format. The names of the pathways and the related gene lists were listed in each row of this file. Second, quantitative feature values (such as peak scores) for enhancer and promoter regions were obtained using the Peak Scores matrix, which was calculated from histone modification profiles. This matrix recorded enhancer activity across various tissues and experimental settings. Lastly, enhancers and promoters were mapped to the closest genes based on experimental annotation or genomic proximity using a Union Peak-Gene Mapping file in BED format.

Pathway-Based Genomic Region Labeling

The gene list linked to each biological pathway identified in the GMT file was extracted. Enhancer areas associated with these genes were chosen as positive examples for the classification model's training using the Peak-Gene Mapping file. The input features were obtained by retrieving the corresponding feature vectors for these enhancers from the Peak Scores matrix. An equal number of negative samples, enhancer/promoter regions unrelated to any pathway-associated gene, were chosen at random to prevent classification bias brought on by data imbalance. Additionally, these negative samples were classified as class 0 and linked to their peak score features. As a result, a balanced dataset with an equal number of positive (class 1) and negative (class 0) samples was created for each pathway.

Thus, a binary-labeled dataset with pathway membership as the target variable and enhancer peak scores as features made up the final input to the machine learning model.

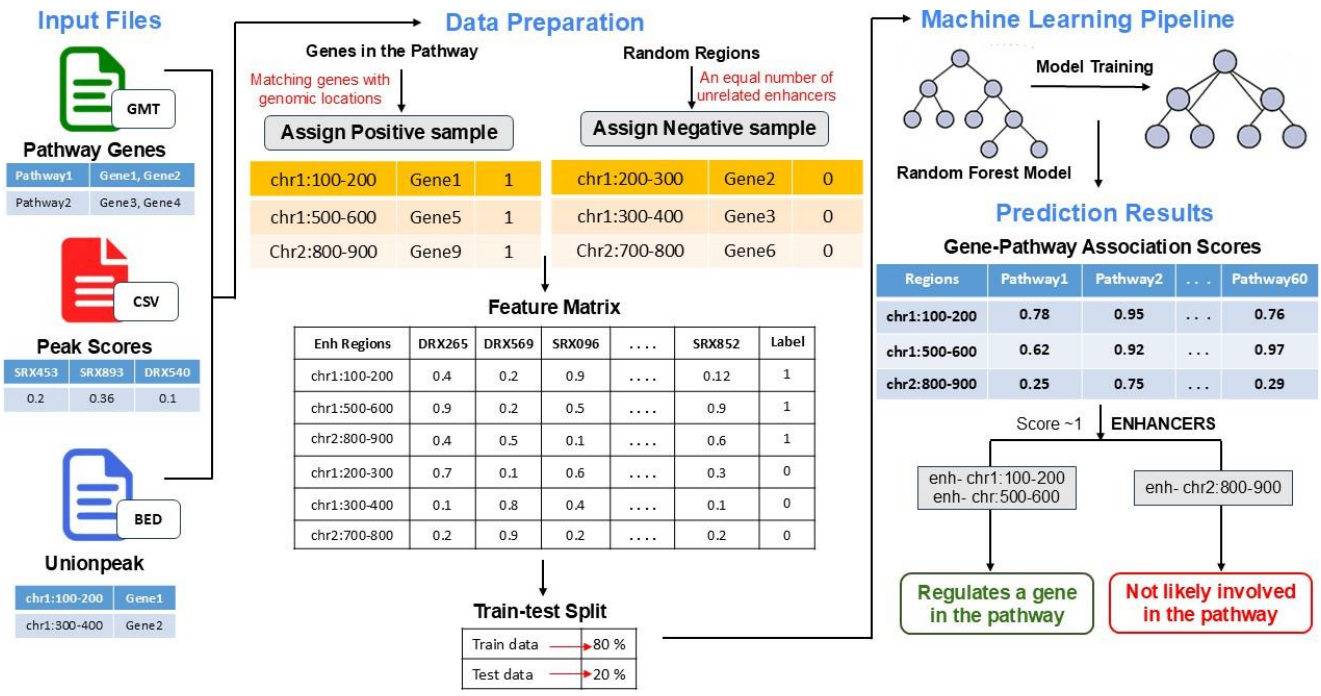


Figure 3.2: Overview of the ML model implementation of Enhancer-Gene Association

3.1.4 Model Building and Training

A Random Forest algorithm was chosen for the classification task because of its robustness to high-dimensional input, ability to model nonlinear relationships, and resistance to overfitting. Each pathway's dataset was split into training (80%) and testing (20%) subsets at random to provide enough learning data and an objective evaluation set. In addition to preventing overfitting on the training data, this partitioning assisted in evaluating the generalizability of the model.

The R package randomForest was used to train the model. A forest of 200 decision trees was built for each model to guarantee predictive stability and lower variance between runs. The binary class labels showed whether each region was associated with the biological pathway being studied, and the enhancer peak scores served as input features for the model. Following that, the trained model was assessed on the held-out test set using performance metrics that were calculated using ROCR and associated tools, including precision, accuracy, recall, F1-score, and AUC-ROC score.

3.1.5 Model Prediction and Evaluation

For every input instance, the machine learning model produces a probability score that indicates the likelihood that a certain enhancer is functionally linked to a gene that is a part of a particular biological pathway. These probabilistic scores, which continually vary from 0 to 1, are used as the foundation for threshold-based binary categorization as well as assessments of prediction confidence. By classifying enhancer-pathway relationships according to prediction strength, our dual-use method facilitates downstream functional annotation and allows for a thorough assessment of model performance across a range of thresholds.

Binary Classification Evaluation

The best threshold value, usually found via ROC curve analysis to balance sensitivity and specificity, was used to translate the continuous prediction probabilities into binary class labels for performance evaluation. Following binarization, several common measures were used to measure the performance of the model. By evenly weighting the true positive and true negative rates, balanced accuracy was used to lessen the consequences of class imbalance. Accuracy was calculated as the percentage of total right classifications. The percentage of real pathway-associated enhancers that the model correctly detected was measured by recall (sensitivity), and the percentage of non-associated enhancers that were accurately excluded was measured by specificity.

The F1 score, accuracy, and recall were utilized to further assess classification performance by summarizing the model's dependability in producing positive predictions as well as its capacity to identify genuine relationships. Furthermore, a more impartial and comprehensible indicator of model quality was offered by the Matthews Correlation Coefficient (MCC). This strong single-value statistic takes into consideration all four categories of the confusion matrix (TP, TN, FP and FN). An inverse measure of model performance was the error rate, which is the percentage of inaccurate predictions.

Using Gini important scores obtained from the Random Forest algorithm, feature importance analysis was carried out to obtain insights into the factors influencing model predictions. The biological interpretation of which histone modification profiles most strongly influence enhancer-pathway interactions was made possible by the identification and reporting of the top 20 most informative features (i.e., enhancer peak scores across various ChIP-seq datasets).

Aggregation and Output

Following evaluation, each pathway-specific model's outputs were methodically assembled into structured result files. TPR (True positive rate), FPR (false positive rate), and the associated decision thresholds were among the following: (i) ROC statistics; (ii) final predictions in binary and probabilistic form; (iii) feature importance scores that identified the most predictive histone

peaks; (iv) feature-to-target correlation values; and (v) the full suite of binary classification performance metrics. Because these files were kept apart for every pathway, meta-analysis, downstream integration, and focused review were made easier.

After prediction scores were generated, a post-processing phase was carried out to allow for pathway-level interpretation and contextualize the results within the genetic landscape. Each of the 60 biological pathways that were evaluated had its result file created by parsing the entire predicted output. The most strongly predicted enhancer-pathway associations were prioritized by sorting enhancer entries for each pathway in descending order based on their prediction scores. The foundation for subsequent analyses, including functional enrichment, visualization, and possible experimental confirmation of the top enhancer regions, was provided by these pathway-specific files.

3.1.6 Results

Machine Learning Prediction for Enhancer-Gene Association

The machine learning model consistently and reliably predicted gene-pathway associations. AUC values indicated good discriminative power, whereas evaluation metrics, including accuracy, balanced accuracy, recall (sensitivity), specificity, F1 score, and MCC score (Matthews Correlation Coefficient), demonstrated a robust prediction capacity across pathways. The model's capacity to sustain both precision and recall was validated by the F1 and MCC scores, while robustness against class imbalance was guaranteed by the balanced accuracy metric. The model's stability and generalizability were further demonstrated by the low error rates throughout the majority of paths. When taken as a whole, these metrics confirm that the approach is highly reliable in identifying regulatory enhancers.

ID	Pathway	Cutoff	Accuracy	Balanced-Accuracy	Recall/Sensitivity	Specificity	F1	MCC	Error-rate	AUC
1	GO_POSITIVE_REGULATION_OF_VIRAL_TRANSCRIPTION	0.414	0.95454545	0.95	0.9	1	0.947368	0.911465	0.045454545	1
2	GO_CARDIAC_CHAMBER_DEVELOPMENT	0.61025	0.9266055	0.927762526	0.936170213	0.919354839	0.916667	0.851734	0.073394495	0.957275
3	GO_DNA_DEPENDENT_DNA_REPLICATION_MAINTENANCE_OF_FIDELITY	0.83	0.58333333	0.583333333	0.5	0.666666667	0.545455	0.169031	0.416666667	0.638889
4	GO_CIRCADIAN_RHYTHM	0.568667	0.90217391	0.904285714	0.928571429	0.88	0.896552	0.805699	0.097826087	0.929048
5	GO_PHOSPHATIDYLSERINE_ACYL_CHAIN_REMODELING	0.7	0.66666667	0.666666667	0.666666667	0.666666667	0.571429	0.316228	0.333333333	0.666667
6	GO_SPINAL_CORD_DEVELOPMENT	0.709	0.92857143	0.929280397	0.923076923	0.935483871	0.935065	0.856086	0.071428571	0.974359
7	GO_PLATELET_DERIVED_GROWTH_FACTOR_RECEPTOR_SIGNALING_PATHWAY	0.69325	0.92857143	0.928571429	0.857142857	1	0.923077	0.866025	0.071428571	0.994898
8	GO_CELLULAR_RESPONSE_TO_LIPOPROTEIN_PARTICLE_STIMULUS	0.748333	0.75	0.728571429	0.857142857	0.6	0.8	0.478091	0.25	0.771429
9	GO_REGULATION_OF_NLRP3_INFLAMMASOME_COMPLEX_ASSEMBLY	0.93	0.71428571	0.65	0.5	0.8	0.5	0.3	0.285714286	0.9
10	GO_POSITIVE_REGULATION_OF_EPITHELIAL_CELL_DIFFERENTIATION	0.606	0.92	0.91626409	0.962962963	0.869565217	0.928571	0.84069	0.08	0.971014
11	GO_POSITIVE_REGULATION_OF_KINASE_ACTIVITY	0.476667	0.95392954	0.95489418	0.994444444	0.915343915	0.954667	0.910994	0.046070461	0.97134
12	GO_NEGATIVE_REGULATION_OF_TRANSCRIPTION_FACTOR_IMPORT_INTO_NUCLEUS	0.583333	0.875	0.874509804	0.866666667	0.882352941	0.866667	0.74902	0.125	0.941176
13	GO_POTASSIUM_ION_TRANSPORT	0.54075	0.9122807	0.912280702	0.964912281	0.859649123	0.916667	0.829168	0.087719298	0.95968
14	GO_REGULATION_OF_T_CELL_RECEPTOR_SIGNALING_PATHWAY	0.77375	0.83333333	0.811688312	0.714285714	0.909090909	0.769231	0.644658	0.166666667	0.935065
15	GO_CARDIAC_MUSCLE_ADAPTATION	0.735	0.75	0.75	0.75	0.75	0.8	0.478091	0.25	0.71875
16	GO_NEGATIVE_REGULATION_OF_EPITHELIAL_CELL_PROLIFERATION	0.49425	0.97	0.970493778	0.962264151	0.978723404	0.971429	0.940045	0.03	0.997591
17	GO_MOVEMENT_IN_ENVIRONMENT_OF_OTHER_ORGANISM_INVOLVED_IN_SYMBIC	0.698917	0.9245283	0.923850575	0.916666667	0.931034483	0.916667	0.847701	0.075471698	0.987069
18	GO_REGULATION_OF_PROTEIN_TARGETING_TO_MITOCHONDRION	0.307833	0.890625	0.895098039	0.966666667	0.823529412	0.892308	0.79214	0.109375	0.930392
19	GO_APICAL_PROTEIN_LOCALIZATION	0.71	0.66666667	0.625	0.5	0.75	0.5	0.25	0.333333333	0.75
20	GO_REGULATION_OF_ESTABLISHMENT_OF_PLANAR_POLARITY	0.523333	0.91428571	0.913151365	0.903225806	0.923076923	0.903226	0.826303	0.085714286	0.964433
21	GO_FOREBRAIN_NEURON_DEVELOPMENT	0.5285	0.93333333	0.9375	0.875	1	0.933333	0.875	0.066666667	0.977679
22	GO_POSITIVE_REGULATION_OF_PROTEIN_MATURATION	0.60125	0.90909091	0.916666667	0.833333333	1	0.909091	0.833333	0.090909091	1
23	GO_NEUROMUSCULAR_JUNCTION_DEVELOPMENT	0.635	0.91428571	0.904761905	0.857142857	0.952380952	0.888889	0.820768	0.085714286	0.979592
24	GO_MITOTIC_CYTOKINESIS	0.589	0.95454545	0.95	0.9	1	0.947368	0.911465	0.045454545	1
25	GO_NEGATIVE_REGULATION_OF_RESPONSE_TO_ENDOPLASMIC_RETICULUM_STRES	0.606667	0.78947368	0.803571429	0.857142857	0.75	0.75	0.586556	0.210526316	0.892857
26	GO_SMAD_PROTEIN_SIGNAL_TRANSDUCTION	0.54525	0.88888889	0.888888889	0.944444444	0.833333333	0.894737	0.782624	0.111111111	0.932099
27	GO_CYTOPLASMIC_TRANSLATION	0.65475	0.875	0.864285714	0.8	0.928571429	0.842105	0.741941	0.125	0.942857
28	GO_MEIOTIC_CHROMOSOME_SEGREGATION	0.416667	0.96875	0.966666667	0.933333333	1	0.965517	0.938872	0.03125	1
29	GO_POSITIVE_REGULATION_OF_CALCIUM_ION_TRANSPORT	0.564333	0.86842105	0.867636868	0.897435897	0.837837838	0.875	0.73732	0.131578947	0.957034
30	GO_REGULATION_OF_DOUBLE_STRAND_BREAK_REPAIR	0.79725	0.85	0.843434343	0.777777778	0.909090909	0.823529	0.697518	0.15	0.909091
31	GO_RNA_DEPENDENT_DNA_BIOSYNTHETIC_PROCESS	0.30275	0.91666667	0.928571429	0.857142857	1	0.923077	0.845154	0.083333333	1
32	GO_REGULATION_OF_B_CELL_RECEPTOR_SIGNALING_PATHWAY	0.86	0.75	0.666666667	0.333333333	1	0.5	0.48795	0.25	0.666667
33	GO_REGULATION_OF_G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATH	0.526417	0.91011236	0.910353535	0.931818182	0.888888889	0.911111	0.821122	0.08988764	0.972222
34	GO_DENDRITE_DEVELOPMENT	0.537417	0.91176471	0.907986111	0.972222222	0.84375	0.921053	0.827547	0.088235294	0.934028
35	GO_REGULATION_OF_RESPIRATORY_BURST	0.72	0.83333333	0.875	0.75	1	0.857143	0.707107	0.166666667	1
36	GO_G_PROTEIN_COUPLED_RECEPTOR_INTERNALIZATION	0.565	0.77777778	0.75	0.666666667	0.833333333	0.666667	0.5	0.222222222	0.888889
37	GO_MEMORY	0.569333	0.90789474	0.906098406	0.974358974	0.837837838	0.915663	0.82223	0.092105263	0.934858
38	GO_NEURON_DEVELOPMENT	0.64675	0.91170825	0.912540668	0.926530612	0.898550725	0.908	0.823803	0.088291747	0.960197
39	GO_REGULATION_OF_GOLGI_ORGANIZATION	0.425	0.83333333	0.75	0.5	1	0.666667	0.632456	0.166666667	1
40	GO_ENDOTHELIAL_CELL_DEVELOPMENT	0.39625	0.94736842	0.946778711	0.952380952	0.941176471	0.952381	0.893557	0.052631579	0.960784
41	GO_POSITIVE_REGULATION_OF_MYOTUBE_DIFFERENTIATION	0.3475	0.88888889	0.883116883	0.857142857	0.909090909	0.857143	0.766234	0.111111111	0.987013
42	GO_REGULATION_OF_CELL_ACTIVATION	0.57875	0.915625	0.914475489	0.934911243	0.894039735	0.921283	0.830825	0.084375	0.96003
43	GO_MULTICELLULAR_ORGANISM_AGING	0.52375	0.91666667	0.909090909	0.818181818	1	0.9	0.842075	0.083333333	0.986014
44	GO_ATP_DEPENDENT_CHROMATIN_REMODELING	0.567917	0.93333333	0.933794466	0.954545455	0.913043478	0.933333	0.867589	0.066666667	0.980237
45	GO_ANATOMICAL_STRUCTURE_FORMATION_INVOLVED_IN_MORPHOGENESIS	0.595917	0.9112782	0.911391718	0.936555891	0.886227545	0.913108	0.823671	0.088721805	0.958749
46	GO_ION_TRANSPORT	0.546417	0.91292135	0.91223335	0.961218837	0.863247863	0.917989	0.829377	0.087078652	0.947992
47	GO_LIPID_MODIFICATION	0.728667	0.94814815	0.948382809	0.950819672	0.945945946	0.943089	0.895576	0.051851852	0.967435
48	GO_NEGATIVE_REGULATION_OF_KIDNEY_DEVELOPMENT	0.885	0.83333333	0.857142857	0.714285714	1	0.833333	0.714286	0.166666667	0.971429
49	GO_REGULATION_OF_HEPATOCYTE_PROLIFERATION	0.75	0.75	0.75	0.666666667	0.833333333	0.727273	0.507093	0.25	0.847222
50	GO_ACYLGLYCEROL_HOMEOSTASIS	0.75	0.9	0.898989899	0.888888889	0.909090909	0.888889	0.79798	0.1	0.949495
51	GO_NEGATIVE_REGULATION_OF_PHOSPHOPROTEIN_PHOSPHATASE_ACTIVITY	0.775	0.77777778	0.75	0.666666667	0.833333333	0.666667	0.5	0.222222222	0.916667
52	GO_NEGATIVE_REGULATION_OF_PROTEIN_COMPLEX_ASSEMBLY	0.54375	0.9375	0.937773609	0.948717949	0.926829268	0.936709	0.875274	0.0625	0.989994
53	GO_HEMOGLOBIN_METABOLIC_PROCESS	0.855	0.85714286	0.833333333	0.666666667	1	0.8	0.730297	0.142857143	1
54	GO_SINGLE_FERTILIZATION	0.564583	0.88571429	0.885339885	0.891891892	0.878787879	0.891892	0.77068	0.114285714	0.955774
55	GO_LEUKOCYTE_ACTIVATION	0.618	0.93928571	0.939461197	0.964028777	0.914893617	0.940351	0.879708	0.060714286	0.969667
56	GO_CHROMOSOME_ORGANIZATION	0.5345	0.91627172	0.916503604	0.945859873	0.887147335	0.918083	0.834115	0.083728278	0.966201
57	GO_REGULATION_OF_MITOCHONDRIAL_DEPOLARIZATION	0.706	0.71428571	0.733333333	0.666666667	0.8	0.75	0.447214	0.285714286	0.888889
58	GO_NEGATIVE_REGULATION_OF_MAP_KINASE_ACTIVITY	0.691417	0.98039216	0.980769231	0.961538462	1	0.980392	0.961538	0.019607843	1
59	GO_SYNAPTIC_VESICLE_LOCALIZATION	0.587667	0.9625	0.9625	0.975	0.95	0.962963	0.925289	0.0375	0.990625
60	GO_PHENOL_CONTAINING_COMPOUND_METABOLIC_PROCESS	0.719167	0.91935484	0.91875	0.9375	0.9	0.923077	0.838812	0.080645161	0.947917

Table 3.1 Evaluation metrics for each of the GO pathways - ML model

Evaluation and Visualization of the Enhancer-Based Classification Model

To evaluate the implemented machine learning classification model on enhancers, a comprehensive set of plots was generated for each biological pathway. The discriminative ability of the enhancer score features, class distribution, feature importance, and model performance are evaluated by these visualizations. The plots for two of the biological pathways, "GO Positive Regulation of Viral Transcription" and "GO Cardiac chamber Development" pathways, are shown here:

The number of enhancer samples categorized into the two classes, positive and negative, is depicted in the **class distribution** bar plots. Orange represents negative class samples (0), whereas blue represents positive class samples (1), demonstrating that both datasets preserved an equal distribution of classes.

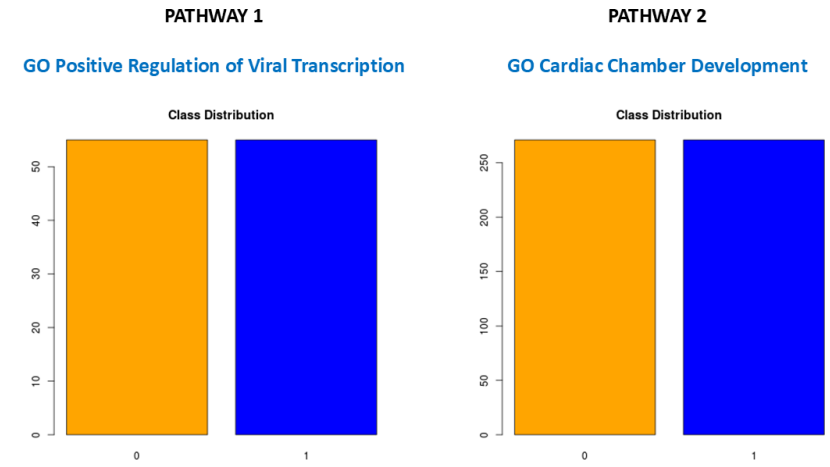


Figure 3.3: Class Distribution of Enhancer Samples

In the **confusion matrix**, true positives, true negatives, false positives, and false negatives for both classes are shown along with the model's prediction results. Most enhancer-gene correlations were successfully recognized by the model with few misclassifications, as indicated by a higher number of TP and TN cases.

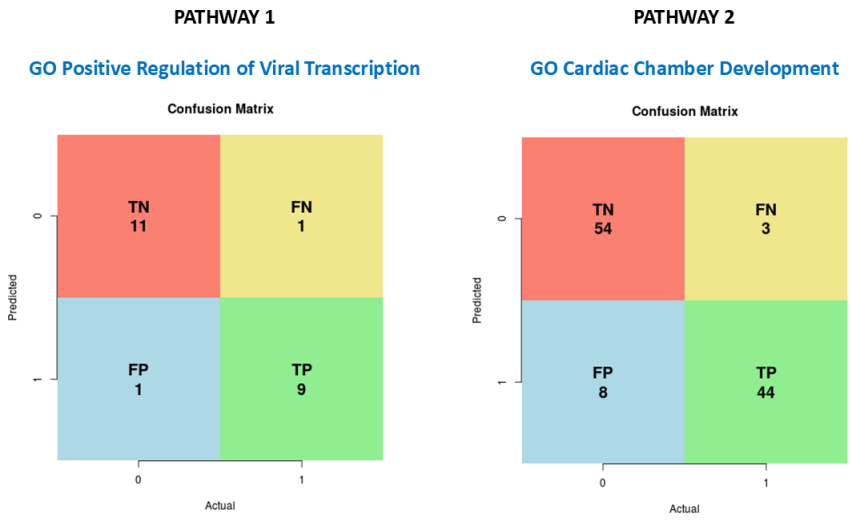


Figure 3.4: Confusion Matrix showing classification Outcomes

Plotting prediction score **histograms** allowed for the evaluation of the model's prediction confidence. The probability scores that the classifier awarded to the test samples are displayed in these histograms. With score peaks clustering close to 0 and 1, the bars, colored in sky blue, show a broadly bimodal distribution for both paths. This illustrates how confident the model is in its forecasts.

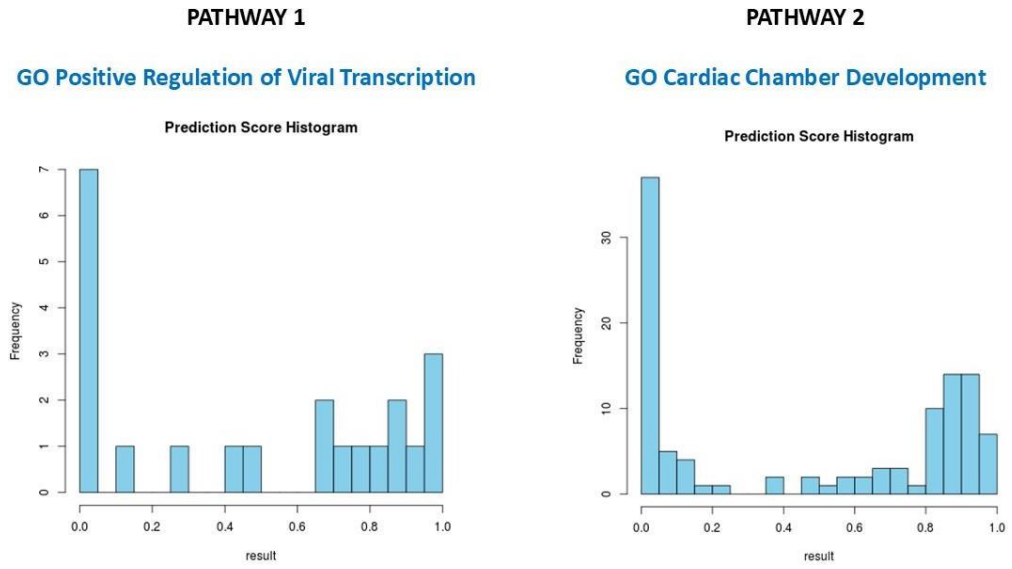


Figure 3.5: Distribution of Prediction Scores

The evaluation of the natural grouping of positive and negative samples is made possible by **Principal Component Analysis** (PCA) plots, which project the enhancer feature data into two primary components. Positive enhancer samples are displayed in blue in these plots, while negative samples are displayed in orange. Both pathways' PCA findings showed a modest but noticeable class separation, indicating that the enhancer scores reflect biologically significant variance pertinent to the classification task.

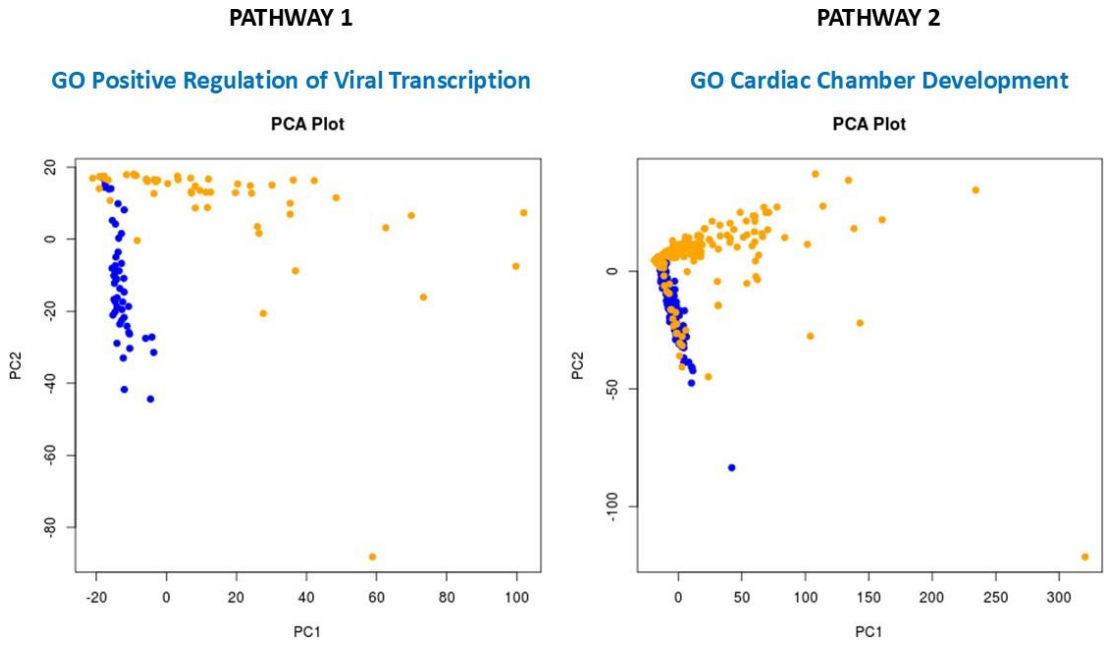


Figure 3.6: PCA Visualization of Enhancer Feature

Validation of Pathway-Gene Associations Using Enhancer Scores

We validated the top-ranked predicted enhancer areas at the route level to evaluate the machine learning model's predictive accuracy and biological relevance. Two well-known tools, GREAand Enrichr, which are often used for functional annotation of regulatory elements, were used to perform enrichment analysis. For validation, two complementary approaches were used:

The GREAT tool was used to determine whether the original pathway appeared enriched in the findings by ranking the top 100 enhancer regions for each biological pathway according to model-predicted scores. This method used enhancer-based genomic enrichment to evaluate direct route recall. For instance, GREAT enrichment analysis demonstrated the model's ability to recover functionally coherent enhancer sets in the case of "GO: Positive Regulation of Viral Transcription" by confirming a significant signal for this pathway among the top-ranked enhancer predictions [Figure 3.7]. Likewise, the model's capacity to detect enhancers important in epithelial lineage specification and differentiation was further supported by validation using Enrichr, which showed notable pathway enrichment for the "GO: Positive Regulation of Epithelial Cell Differentiation" route. The model captures significant biological regulatory relationships at the pathway level, as these data together show [Figure 3.8].

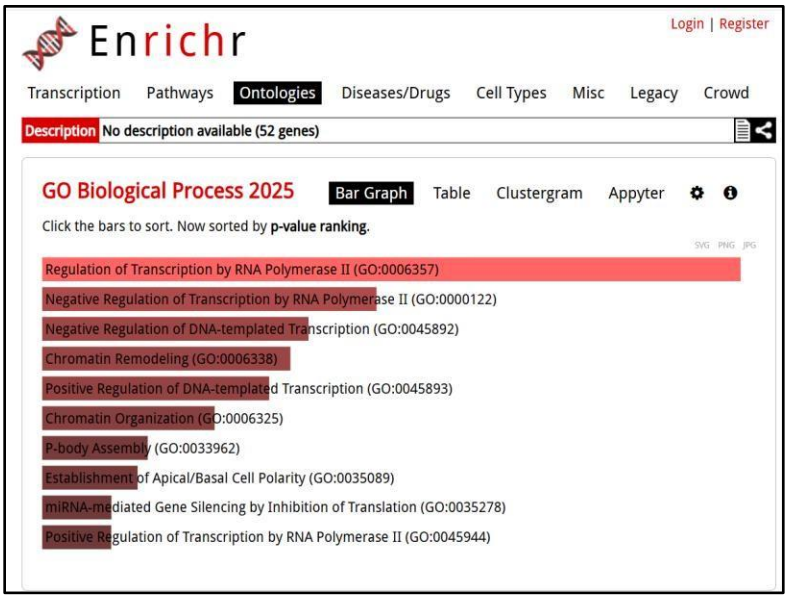


Figure 3.7: Enrichr output showing enrichment for the "GO: Positive Regulation of Viral Transcription" pathway.

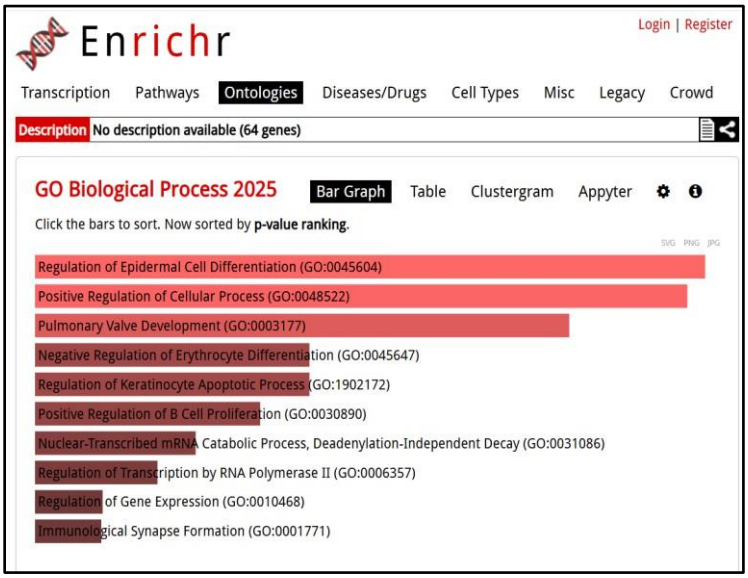


Figure 3.8: Enrichr output showing enrichment for the "GO: Positive Regulation of Epithelial Cell Differentiation" pathway

Validation of Pathway Genes

We tested our model on enhancer regions after training it mostly on promoter regions of known pathway-associated genes. Surprisingly, the most highly predicted enhancers were found adjacent to the same genes that were utilized in training, indicating that enhancers that control these genes are frequently found near their promoters. We used Venny to compare these gene lists with the initial training gene sets after mapping the closest genes to the top enhancer predictions using GREAT. Even when trained on promoter features, the model can reliably recover functionally relevant enhancer–gene connections, as evidenced by the observed overlap. This demonstrates the robustness of our methodology in finding biologically significant gene targets as well as the regulatory closeness of enhancers to promoters.

Training for the "GO: Regulation of Cell Activation" pathway involved the use of 472 genes. 86 closest genes were found from the GREAT analysis of the top enhancer predictions, 12 of which overlapped with the training gene set [Figure 3.9]. Similarly, out of the 469 genes that were initially linked to the "GO: Positive Regulation of Kinase Activity" route, 13 genes that overlapped with the training set were found in 85 predicted closest genes. The model's predictions' biological relevance is further supported by these regular overlaps between pathway training sets and projected enhancer-nearest genes. They show that the top-ranked enhancer regions have functional significance and probably play role in the control of genes specific to a given pathway (Figure 3.10).

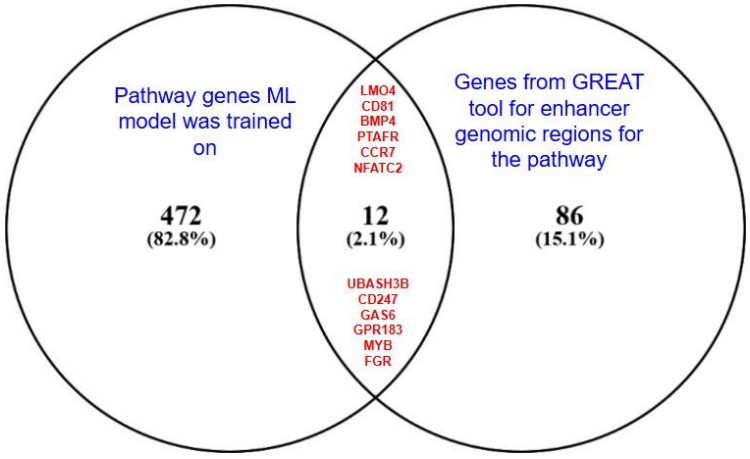


Figure 3.9: Validated Gene output for the pathway "GO: Regulation of Cell Activation"

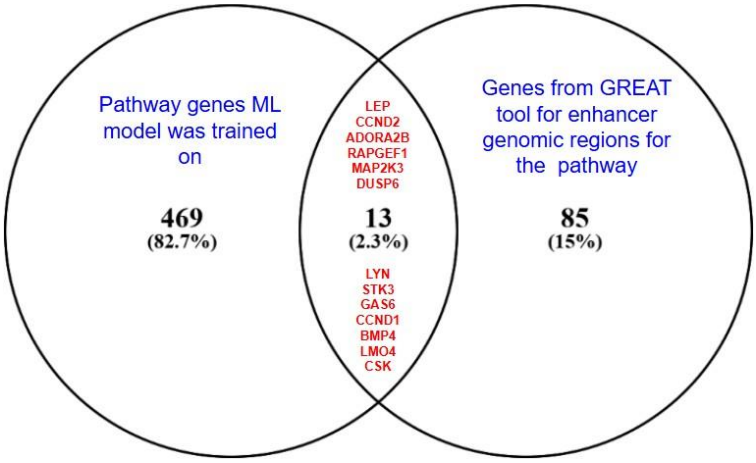


Figure 3.10: Validated Gene output for the pathway "GO: Positive Regulation of Kinase Activity"

3.2 Validation of Enhancer Perturbation Using Machine Learning Model

3.2.1 Objective and Study Selection

We used supervised machine learning to examine the prediction value of genomic characteristics and validate the curated perturbation dataset in ePerturbDB. Luo et al. [34], a CRISPR-based enhancer knockout study carried out in human embryonic stem cells (hESCs), was chosen as a high-quality dataset. Z-scores and $\log_2$ fold change ($\log_2$FC) values for 10,314 sgRNAs that the study provided target enhancer areas. Enhancer perturbations were classified as functionally significant if their z-score was greater than or equal to 2, and as non-significant if it was less than or equal to 2. Enhancer annotations were taken from the enhancerDB database, which contains peak scores for histone modifications like H3K27ac in various tissues, to prepare the input data. Bedtools intersect was utilized to overlap sgRNA-targeted regions with histone modification peaks to connect perturbation records with epigenomic characteristics. Annotated with genomic coordinates, binary classification labels (significant vs. non-significant), and quantitative histone mark scores from associated ChIP-seq experiments.

3.2.2 Model Building and Performance

We used the z-score of gene expression change as the target variable in a Random Forest regression model to evaluate the functional effects of enhancer perturbations. Using bedtools getfasta, the input feature set was obtained from the hg19 reference genome and comprised nucleotide sequences of enhancers, genomic locations, and histone modification peak scores (e.g., H3K27ac, H3K4me1). Only non-overlapping sgRNA locations were retained after we used de-duplication filtering to ensure data quality and minimize redundancy. However, the R2 score was 0.3, suggesting a comparatively low prediction accuracy when the model was trained and assessed using this filtered dataset. This implies that even while the model was able to capture some signal, it had trouble generalizing, most likely as a result of the small number of high-confidence perturbation events and the biological intricacy of enhancer control. The intrinsic heterogeneity in how enhancer perturbations impact gene expression—which is frequently non-linear and context-dependent—may also be reflected in the reported performance.

This investigation shows that epigenomic characteristics specifically histone modification scores, carry significant signals and may be able to predict enhancer activity when incorporated into more complex or context-sensitive models. In order to fully capture the intricate regulatory logic of enhancers, the model underscores the need for larger, higher-resolution datasets and potentially more enhanced modeling tools, while also highlighting the potential value of machine learning in enhancer functional annotation. The work lays the groundwork for future advancements by showing that machine learning offers a viable framework to quantitatively analyze enhancer perturbation data and direct experimental prioritization in regulatory genomics.

CHAPTER - 4

Conclusion

Enhancers and their transcribed end products, enhancer RNAs (eRNAs), represent regulatory components of the non-coding genome that are essential for regulating gene expression in a variety of cell types, developmental stages, and disease contexts. Despite increasing awareness of their significance, dynamic activity patterns of enhancers, tissue specificity, and the difficulty of manipulating and monitoring regulatory results in experiments have prevented thorough functional characterization. A comprehensive strategy that incorporates predictive modeling, integrative functional genomics, and carefully selected perturbation datasets was implemented to overcome these issues. To capture the regulatory impact of enhancer disruptions, a consolidated, searchable database was created by methodically gathering and classifying enhancer and eRNA perturbation data from several experimental platforms, such as CRISPR-based and RNA interference investigations.

Data-driven investigation of enhancer function in both physiological and pathological conditions is made possible by this resource, which makes it easier to retrieve genomic coordinates, perturbation techniques, impacted target genes, and related phenotypic consequences. Using epigenomic characteristics such histone modification patterns, chromatin accessibility, and genomic proximity, machine learning models were used in parallel to identify enhancer–gene–pathway relationships. These models were validated by enrichment studies and gene-level overlap with established biological pathways, and they showed a high predictive potential in identifying functionally relevant enhancers. Additionally, the therapeutic significance of functionally annotated enhancers was highlighted by the integration of enhancer annotations with transcriptional and survival data.

A scalable basis for comprehending enhancer biology has been established by the integration of high-throughput perturbation data, computational modeling, and curated knowledge. This integrated framework offers a useful toolkit for breaking down gene regulation at the genome-wide level, facilitates the creation of hypotheses for experimental follow-up, and improves the annotation of regulatory elements. Such methods will continue to be essential for converting non-coding genomic variation into mechanistic understandings and therapeutic prospects as the field of functional genomics develops.

References

[1] “Transcription regulation and animal diversity | Nature.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nature01763

[2] “Enhancer function: new insights into the regulation of tissue-specific gene expression | Nature Reviews Genetics.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nrg2957

[3] “Long-range enhancer–promoter contacts in gene expression control | Nature Reviews Genetics.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41576-019-0128-0

[4] “An atlas of active enhancers across human cell types and tissues | Nature.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nature12787

[5] “Widespread transcription at neuronal activity-regulated enhancers | Nature.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nature09033

[6] “Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation | Nature.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nature12210

[7] K. Mousavi *et al.*, “eRNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci,” *Molecular Cell*, vol. 51, no. 5, pp. 606–617, Sep. 2013, doi: 10.1016/j.molcel.2013.07.022.

[8] D. Hnisz *et al.*, “Super-Enhancers in the Control of Cell Identity and Disease,” *Cell*, vol. 155, no. 4, pp. 934–947, Nov. 2013, doi: 10.1016/j.cell.2013.09.053.

[9] “Enhancer loops appear stable during development and are associated with paused polymerase | Nature.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nature13417

[10] “Histone modifications at human enhancers reflect global cell-type-specific gene expression | Nature.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nature07829

[11] “The super-enhancer-driven lncRNA LINC00880 acts as a scaffold between CDK1 and PRDX1 to sustain the malignance of lung adenocarcinoma | Cell Death & Disease.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41419-023-06047-w

[12] H. Rahnamoun, P. Orozco, and S. M. Lauberth, “The role of enhancer RNAs in epigenetic regulation of gene expression,” *Transcription*, vol. 11, no. 1, pp. 19–25, Dec. 2019, doi: 10.1080/21541264.2019.1698934.

[13] S. H. Tan *et al.*, “The enhancer RNA ARIEL activates the oncogenic transcriptional program in T-cell acute lymphoblastic leukemia,” *Blood*, vol. 134, no. 3, pp. 239–251, Jul. 2019, doi: 10.1182/blood.2018874503.

[14] “Systematic mapping of functional enhancer–promoter connections with CRISPR interference | Science.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.science.org/doi/10.1126/science.aag2445

[15] “Integrative analysis of 111 reference human epigenomes | Nature.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nature14248

[16] J. Ernst and M. Kellis, “ChromHMM: automating chromatin-state discovery and characterization,” *Nat Methods*, vol. 9, no. 3, pp. 215–216, Mar. 2012, doi: 10.1038/nmeth.1906.

[17] A. S. Nord *et al.*, “Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development,” *Cell*, vol. 155, no. 7, pp. 1521–1531, Dec. 2013, doi: 10.1016/j.cell.2013.11.033.

[18] “The chromatin accessibility landscape of primary human cancers | Science.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.science.org/doi/10.1126/science.aav1898

[19] E. Geller *et al.*, “Massively parallel disruption of enhancers active in human neural stem cells,” *Cell Reports*, vol. 43, no. 2, Feb. 2024, doi: 10.1016/j.celrep.2024.113693.

[20] “Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations | Nature Genetics.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41588-019-0538-0

[21] M. Gasperini *et al.*, “A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens,” *Cell*, vol. 176, no. 1, pp. 377-390.e19, Jan. 2019, doi: 10.1016/j.cell.2018.11.029.

[22] M. Ding, Y. Liu, X. Liao, H. Zhan, Y. Liu, and W. Huang, "Enhancer RNAs (eRNAs): New Insights into Gene Transcription and Disease Treatment," *J Cancer*, vol. 9, no. 13, pp. 2334–2340, Jun. 2018, doi: 10.7150/jca.25829.

[23] "Enhancer redundancy provides phenotypic robustness in mammalian development | Nature." Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nature25461

[24] "Function-based identification of mammalian enhancers using site-specific integration | Nature Methods." Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/nmeth.2886

[25] R. Kang *et al.*, "EnhancerDB: a resource of transcriptional regulation in the context of enhancers," *Database*, vol. 2019, p. bay141, Jan. 2019, doi: 10.1093/database/bay141.

[26] T. Gao, B. He, S. Liu, H. Zhu, K. Tan, and J. Qian, "EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types," *Bioinformatics*, vol. 32, no. 23, pp. 3543–3551, Dec. 2016, doi: 10.1093/bioinformatics/btw495.

[27] "A multi-omic dissection of super-enhancer driven oncogenic gene expression programs in ovarian cancer | Nature Communications." Accessed: Jun. 06, 2025. [Online]. Available: https://www.nature.com/articles/s41467-022-31919-8

[28] C. Feng *et al.*, "KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors," *Nucleic Acids Res*, vol. 48, no. D1, pp. D93–D100, Jan. 2020, doi: 10.1093/nar/gkz881.

[29] "The long noncoding RNA SNHG1 regulates colorectal cancer cell growth through interactions with EZH2 and miR-154-5p | Molecular Cancer | Full Text." Accessed: Jun. 17, 2025. [Online]. Available: https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-018-0894-x

[30] "Complementary Alu sequences mediate enhancer–promoter selectivity | Nature." Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41586-023-06323-x

[31] "RNAs interact with BRD4 to promote enhanced chromatin engagement and transcription activation | Nature Structural & Molecular Biology." Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41594-018-0102-0

[32] M. W. Lewis *et al.*, "CRISPR Screening of Transcribed Super-Enhancers Identifies Drivers of Triple-Negative Breast Cancer Progression," *Cancer Res*, vol. 84, no. 21, pp. 3684–3700, Nov. 2024, doi: 10.1158/0008-5472.CAN-23-3995.

[33] "Landscape of enhancer disruption and functional screen in melanoma cells | Genome Biology | Full Text." Accessed: Jun. 17, 2025. [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-03087-5

[34] "Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions | Nature Genetics." Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41588-023-01450-7

[35] "Deciphering essential cistromes using genome-wide CRISPR screens | PNAS." Accessed: Jun. 17, 2025. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.1908155116

[36] "Parallel characterization of cis-regulatory elements for multiple genes using CRISPRpath | Science Advances." Accessed: Jun. 17, 2025. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.abi4360

[37] X. Ren *et al.*, "CRISPR tiling deletion screens reveal functional enhancers of neuropsychiatric risk genes and allelic compensation effects (ACE) on transcription," Oct. 10, 2024, *bioRxiv*. doi: 10.1101/2024.10.08.616922.

[38] S. J. Kaplan *et al.*, "CRISPR screening uncovers a long-range enhancer for ONECUT1 in pancreatic differentiation and links a diabetes risk variant," *Cell Rep*, vol. 43, no. 8, p. 114640, Aug. 2024, doi: 10.1016/j.celrep.2024.114640.

[39] "Functional CRISPR screen identifies AP1-associated enhancer regulating FOXF1 to modulate oncogene-induced senescence | Genome Biology | Full Text." Accessed: Jun. 17, 2025. [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1494-1

[40] "JUN-induced super-enhancer RNA forms R-loop to promote nasopharyngeal carcinoma metastasis | Cell Death & Disease." Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41419-023-05985-9

[41] “Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA | Nature Communications.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/ncomms7520

[42] “Co-activation of super-enhancer-driven CCAT1 by TP63 and SOX2 promotes squamous cancer progression | Nature Communications.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41467-018-06081-9

[43] “Super-enhancer-driven lncRNA LIMD1-AS1 activated by CDK7 promotes glioma progression | Cell Death & Disease.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41419-023-05892-z

[44] “Enhancer RNA-driven looping enhances the transcription of the long noncoding RNA DHRS4-AS1, a controller of the DHRS4 gene cluster | Scientific Reports.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/srep20961

[45] “Super-enhancer-associated long noncoding RNA AC005592.2 promotes tumor progression by regulating OLFM4 in colorectal cancer | BMC Cancer | Full Text.” Accessed: Jun. 17, 2025. [Online]. Available: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-021-07900-x

[46] M. Ding *et al.*, “Enhancer RNA - P2RY2e induced by estrogen promotes malignant behaviors of bladder cancer,” *Int J Biol Sci*, vol. 14, no. 10, pp. 1268–1276, Jul. 2018, doi: 10.7150/ijbs.27151.

[47] “eNEMAL, an enhancer RNA transcribed from a distal MALAT1 enhancer, promotes NEAT1 long isoform expression - PMC.” Accessed: Jun. 17, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8139514/

[48] “Downregulation of enhancer RNA AC003092.1 is associated with poor prognosis in kidney renal clear cell carcinoma | Scientific Reports.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/s41598-024-64431-8

[49] “TALEN-mediated enhancer knockout influences TNFAIP3 gene expression and mimics a molecular phenotype associated with systemic lupus erythematosus | Genes & Immunity.” Accessed: Jun. 17, 2025. [Online]. Available: https://www.nature.com/articles/gene20164

[50] “Unveiling chromatin dynamics with virtual epigenome | Nature Communications.” Accessed: Jun. 22, 2025. [Online]. Available: https://www.nature.com/articles/s41467-025-58481-3

[51] “CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells | Science.” Accessed: Jun. 22, 2025. [Online]. Available: https://www.science.org/doi/10.1126/science.aah7111

[52] “TCeA.” Accessed: Jul. 16, 2025. [Online]. Available: https://bioinformatics.mdanderson.org/public-software/tcea/

[53] “VISTA Enhancer browser: an updated database of tissue-specific developmental enhancers | Nucleic Acids Research | Oxford Academic.” Accessed: Jul. 16, 2025. [Online]. Available: https://academic.oup.com/nar/article/53/D1/D324/7848837

[54] “ENdb-Home.” Accessed: Jul. 17, 2025. [Online]. Available: https://bio.liclab.net/ENdb/

[55] F. Spitz and E. E. M. Furlong, “Transcription factors: from enhancer binding to developmental control,” *Nat Rev Genet*, vol. 13, no. 9, pp. 613–626, Sep. 2012, doi: 10.1038/nrg3207.

[56] H. K. Long, S. L. Prescott, and J. Wysocka, “Ever-changing landscapes: transcriptional enhancers in development and evolution,” *Cell*, vol. 167, no. 5, pp. 1170–1187, Nov. 2016, doi: 10.1016/j.cell.2016.09.018.

[57] M. Osterwalder *et al.*, “Enhancer Redundancy Allows for Phenotypic Robustness in Mammalian Development,” *Nature*, vol. 554, no. 7691, pp. 239–243, Feb. 2018, doi: 10.1038/nature25461.