KU LEUVEN

FACULTY OF BIOSCIENCE ENGINEERING
FACULTY OF ENGINEERING SCIENCE
FACULTY OF MEDICINE
FACULTY OF SCIENCES

Deciphering organism-wide gene regulation with genomic deep learning in C. Elegans

Promotor:
Prof. Stein Aerts
Department: KU Leuven Department of Human Genetics
& VIB Center for Brain & Disease Research
Division: Laboratory of Computational Biology

Dissertation presented in
fulfillment of the requirements
for the degree of Master of Science:
Bioinformatics

Cristiano Cordì

August 2025

Copyright Information:
student paper as part of an academic education and examination.
No correction was made to the paper after examination.

Acknowledgements

First, I would like to thank Prof. Stein Aerts for giving me the opportunity to conduct my master's thesis in his lab. Next, I would like to thank my supervisor, Casper H. Blaauw, for his constant guidance over the past year, as well as for the inspiration and valuable lessons I have learned from him. I am also grateful to the members of the Aerts lab for creating such a welcoming environment.

I would like to thank Prof. Mark Fiers and Prof. Sofie Demeyer for agreeing to evaluate this thesis. I am also grateful to KU Leuven, the VIB institutes, and the VSC for providing the resources that allowed me to do what I enjoy most.

I am thankful to my fellow master's students in the lab: Bram Stuyven, August Winderickx, Wannes Vanoyenbrugge, Eva Bosch, Senne Van Eynde, and Ismail Tabash for their companionship. Thanks also to my bioinformatics classmates for fostering such a cheerful atmosphere.

I would like to thank my friends Edoardo Trinca, Camillo Colleluori, and Robert Forsyth for their support and for the fun we enjoyed along the way.

My heartfelt thanks go to Laura Beltrame for the incredible support she has given me throughout the year. Without her, completing the Master of Bioinformatics would have been impossible.

Finally, I wish to thank my family: Alice, Silvia, and Mario, without whom none of this would have been possible.

Abstract

Deciphering how the genome regulates gene expression to produce distinct cell types remains a fundamental question in biology. While the DNA sequence encodes regulatory instructions, the specific mechanisms by which these sequences drive gene expression are not yet fully understood. In this work, we use deep learning to decode the regulatory grammar of the genome and predict gene expression at the single-cell level. Specifically, we develop deep learning sequence-to-function models based on convolutional or attention-based architectures to analyze genomic sequences from Caenorhabditis elegans.

This model organism offers a uniquely well-characterized cellular landscape, allowing for a comprehensive investigation of gene expression across all cell types. Our models aim to predict gene expression directly from DNA sequence, marking a significant step forward in the field.

Furthermore, we will generate a fully integrated dataset of cells from different developmental stages of the organism and employ explainable machine learning techniques such as gradient-based feature importance methods to interpret the learned regulatory features. We seek to provide novel insights into the regulatory logic of the genome. Our findings contribute to a deeper understanding of gene regulation at the organism-wide level, with potential implications for diverse biological applications, including developmental biology and disease research.

List of Abbreviations

ATAC-seq Assay for Transposase-Accessible Chromatin sequencing.

ChIP-seq Chromatin Immunoprecipitation Sequencing.

CNEs Non-Coding Elements.

CNN Convolutional Neural Network.

CREs Cis-Regulatory Elements.

CREsted Cis-Regulatory Element prediction from DNA sequence.

DNA Deoxyribonucleic Acid.

DNase-seq DNase I hypersensitive sites sequencing.

DPE Downstream Promoter Element.

ELBO Evidence Lower Bound.

GEO Gene Expression Omnibus.

GPU Graphics processing unit.

HPC High-performance computing.

IGV ntegrative Genomics Viewer.

INR Initiator.

KL Kullback-Leibler.

LDA Latent Dirichlet Allocation.

PWMs Position Weight Matrices.

RNA Ribonucleic Acid.

scATAC Single cell Assay for Transposase-Accessible Chromatin using sequencing.

sci-ATAC Single cell Combinatorial Indexing ATAC-seq.

scRNA Single cell RNA sequencing.

scVI Single-cell Variational Inference.

t-SNE t-distributed Stochastic Neighbor Embedding.

TSS Transcription Start Site.

UMAP Uniform Manifold Approximation and Projection.

VAE Variational Auto-Encoder.

VSC Vlaams Supercomputer Center.

WandB Weights & Biases.

Contents

Chapter 1

Context and Aims

The genome in biology has long been considered the entity that encodes information and decodes it when needed. Traditionally viewed as a blueprint, this perspective is now shifting. With current knowledge, it becomes increasingly clear that the genome actually functions analogously to a Variational Auto-Encoder (VAE) rather than a simple recipe. A Variational Auto-Encoder compresses information to its essential components in a latent space and then reconstructs it from that latent representation. Moreover, these "genomic VAEs" are not artificially generated or created, they are self-assembling structures [1, 2].

Every cell in an organism carries the same genomic blueprint, yet the activation of different genes in distinct cell types results in the remarkable diversity of tissues and functions. Understanding how genomic sequences regulate gene expression across cell types is a central challenge in molecular biology.

Recent advances in single-cell sequencing have enabled high-resolution analysis of gene expression across individual cells, but the underlying regulatory logic encoded in DNA remains unclear. Deep learning models, particularly Convolutional Neural Networks (CNNs) and transformers, have demonstrated the ability to extract meaningful features from DNA sequences, offering a powerful approach to decipher regulatory mechanisms.

In this thesis, we focus on Caenorhabditis elegans, a model organism widely used in gene regulation research due to its fully mapped cellular lineage and its suitability for single-cell sequencing. Using Single cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC) and Single cell RNA sequencing (scRNA) data from C. elegans, we develop deep learning models to predict gene expression directly from DNA sequences.

To interpret the learned representations and uncover biologically meaningful features, we apply explainable machine learning techniques and gradient-based feature importance methods.

To achieve this, we will also assemble a dataset with cells from various developmental stages of C. elegans.

Chapter 2

Literature Review

2.1 The Genomic Regulatory Code

Gene expression is a highly regulated process that determines when and where genes are activated, managing cellular function and development. This regulation is achieved through a complex interplay of DNA sequence elements and protein factors that bind to these sequences.

2.1.1 Cis-Regulatory Elements (CREs)

CREs are non-coding DNA sequences that control the transcription of genes positioned on the same DNA molecule. The term "cis" refers to this physical linkage to their target genes, where they operate within the same chromosome and typically affect nearby genes. CREs serve as binding platforms for transcription factors and other regulatory proteins, integrating multiple inputs to orchestrate precise spatiotemporal gene expression patterns [3].

CREs are distinguished by several key characteristics. First, they contain clusters of transcription factor binding sites, often with a specific syntax and spacing that determines cooperative or competitive interactions between bound proteins. Second, CREs display varying degrees of evolutionary conservation, with functionally critical regions often showing increased sequence constraint across species. Third, active CREs typically exhibit characteristic epigenetic signatures, including DNase I hypersensitivity, specific histone modifications (H3K4me1, H3K27ac), and nucleosome depletion [4].

The repertoire of CREs includes promoters positioned at transcription start sites, enhancers that

boost transcription from variable distances, silencers that repress gene activity, and insulators that establish boundaries between regulatory domains. The human genome contains hundreds of thousands of CREs, with estimates suggesting that regulatory sequences occupy a substantially larger portion of the genome than protein-coding regions [4].

Cis vs. Trans Regulation

In contrast to cis-regulation, trans-regulation involves diffusible factors that can influence gene expression regardless of their genomic location. Trans-regulatory elements primarily consist of the genes encoding transcription factors, cofactors, chromatin modifiers, and non-coding RNAs that interact with CREs or act far away to control gene expression [5, 6].

The interplay between cis and trans regulatory components creates a complex regulatory network that enables precise control of gene expression. CREs integrate inputs from multiple trans-acting factors, serving as information processing units that compute cellular states and environmental conditions to determine appropriate transcriptional outputs. This integration capacity allows the same trans-regulatory environment to elicit different responses from genes with distinct cis-regulatory architectures, enabling cell-type specific gene expression within organisms [7, 8].

Modern genomic approaches have revolutionized our understanding of CREs through techniques such as Chromatin Immunoprecipitation Sequencing (ChIP-seq) for mapping transcription factor binding sites, Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) and DNase I hypersensitive sites sequencing (DNase-seq) for identifying accessible chromatin regions, and massively parallel reporter assays for functional characterization. The combination of these methodologies with computational approaches has revealed the vast landscape of CREs across diverse cell types and developmental stages [9].

Promoters

Promoters are regulatory regions typically located immediately upstream of the Transcription Start Site (TSS). Core promoters contain sequence motifs that recruit the basic transcriptional machinery, though the specific elements vary across species. While mammalian promoters often contain TATA boxes, Initiator (INR) elements, and Downstream Promoter Element (DPE), C. elegans promoters show distinct organizational features with less well-defined core elements [10].
The proximal promoter elements that extend further upstream contain binding sites for specific transcription factors that modulate the rate of transcription initiation [10].

Enhancers

Enhancers are distal regulatory elements that can activate transcription regardless of their orientation or distance from the target gene. These elements contain clusters of transcription factor binding sites and can be located upstream, downstream, or even within the introns of their target genes. Enhancers function through chromatin looping, which brings them into physical proximity with promoters to stimulate transcription [11].

Silencers and Insulators

Silencers are DNA sequences that repress gene expression by binding transcriptional repressors or inducing repressive chromatin states. Insulators serve as boundary elements that block the interaction between enhancers and promoters when positioned between them, thereby preventing inappropriate gene activation [12].

2.1.2 Sequence-Based Regulatory Elements

The DNA sequence itself encodes information critical for gene regulation through specific motifs recognized by regulatory proteins.

Transcription Factor Binding Motifs

Transcription factors recognize specific DNA sequence motifs, typically 6-12 base pairs in length. These motifs often display sequence degeneracy, represented as Position Weight Matrices (PWMs) that capture the probability of each nucleotide occurring at each position [13].

CpG Islands

CpG islands are regions with high CG dinucleotide content that frequently overlap with promoters of housekeeping genes and developmental regulators. Their methylation status correlates with gene expression, with unmethylated CpG islands generally associated with transcriptionally active genes [14].

Structural DNA Elements

Certain DNA sequences influence local DNA structure, affecting protein binding and transcriptional activity. For example, poly(dA:dT) tracts resist nucleosome formation, creating nucleosome-depleted regions often found in active promoters [15]. G-quadruplex structures formed by guanine-rich sequences can affect transcription factor binding and RNA polymerase progression [16].

Conserved Non-Coding Sequences

Evolutionary conservation analysis has revealed numerous non-coding sequences under selective pressure, suggesting functional importance in gene regulation. These conserved Non-Coding Elements (CNEs) often correspond to enhancers and other regulatory elements active in development [17].

2.2 Caenorhabditis elegans

Caenorhabditis elegans has emerged as one of biology's most powerful model organisms, offering unique advantages for studying fundamental biological processes, from development to neurobiology.

2.2.1 Advantages as a Model Organism

C. elegans possesses several characteristics that make it exceptionally suitable for laboratory research. This small, free-living nematode measures approximately 1mm in length and can be cultivated on agar plates with E. coli as a food source, requiring minimal laboratory resources [18]. Its rapid life cycle—progressing from egg to reproductive adult in just 3-4 days at 20°C—enables quick experimental turnaround and multi-generational studies. The predominant self-fertilizing hermaphrodite reproductive strategy, complemented by occasional males, facilitates both strain maintenance and genetic crosses [19].

The worm's transparent body allows direct observation of cellular processes in living animals using simple light microscopy, while its relatively simple anatomy (959 somatic cells in hermaphrodites) makes it tractable for comprehensive analysis. Additionally, C. elegans was the first multicellular organism with a completely sequenced genome, comprising approximately 100 megabases with about 20,000 protein-coding genes, many with direct human orthologs [20]. This genetic conservation, combined with facile genetic manipulation through techniques including RNA interference, CRISPR/Cas9 editing, and transgenic approaches, has established C. elegans as an invaluable system for modeling human disease mechanisms.

2.2.2 Invariant Lineage and Cellular Characterization

Perhaps the most distinctive feature of C. elegans is its invariant cell lineage, a property termed "eutely." The developmental trajectory of every somatic cell from zygote to adult has been meticulously mapped, creating a deterministic lineage tree where each cell's identity and fate are predictable across individuals [21]. This invariant development has facilitated unprecedented

understanding of cell fate decisions, developmental timing, and morphogenesis.

The complete connectome of C. elegans has also been mapped through electron microscopy reconstruction, detailing all 302 neurons and their approximately 7,000 synaptic connections [22]. This comprehensive neural map has positioned C. elegans as an unparalleled system for studying the relationship between neural circuits and behavior.

Recent technological advances have enabled scRNA across all cells in the organism at multiple developmental stages, creating comprehensive transcriptomic atlases that reveal the molecular signatures underlying cell fate decisions [23, 24]. These resources, combined with chromatin accessibility mapping, have begun to illuminate the gene regulatory networks governing development and cell identity maintenance.

2.3 Single Cell RNA Sequencing

The emergence of Single cell RNA sequencing (scRNA) represents one of the most significant technological breakthroughs in modern molecular biology, allowing researchers to profile gene expression in individual cells rather than in bulk tissue samples. This technology has fundamentally transformed our understanding of cellular heterogeneity, developmental trajectories, and the molecular basis of complex biological processes.

2.3.1 scRNA-seq Technology and Its Impact

scRNA encompasses a family of methods that isolate individual cells, capture their transcriptomes, and generate sequencing libraries that preserve cellular identity information. The typical workflow involves cell isolation (through microfluidics, droplet-based methods, or plate-based sorting), cell lysis, mRNA capture (usually via poly(A) selection), reverse transcription with cell-specific barcoding, amplification, library preparation, and next-generation sequencing [25].

Several technological platforms have emerged with distinct characteristics. Droplet-based methods like 10x Genomics Chromium and Drop-seq enable massively parallel processing of thousands to tens of thousands of cells with moderate sensitivity and have become the dominant approach in the field. Plate-based methods such as Smart-seq2 offer higher sensitivity and full-length transcript coverage but with lower throughput, making them suitable for specialized applications requiring detailed transcript analysis [25].

The impact of scRNA on biological research has been profound. It has revealed previously unrecognized cell types and cell states across tissues and organisms, including rare cell popu-

lations that were masked in bulk analyses [26]. In developmental biology, scRNA has allowed the reconstruction of cellular differentiation trajectories with unprecedented resolution, revealing the molecular changes that accompany fate decisions [27].

2.3.2 scRNA Expression Representations: Tracks vs Scalars

The representation of gene expression data from scRNA experiments presents unique challenges and opportunities compared to traditional bulk sequencing approaches.

Scalar Representations

The most common representation of scRNA data reduces each cell's transcriptome to a vector of scalar values, with each value representing the expression level of a gene (typically as normalized counts or log-transformed values). This approach facilitates standard multivariate analyses including clustering, differential expression testing, and dimension reduction. Scalar representations are efficient computationally and conceptually straightforward, enabling the application of numerous machine learning techniques to identify patterns across cells [28].

Track Representations

More recently, track-based representations have emerged that preserve the spatial or temporal dimensions of expression data. Rather than representing a gene's expression as a single value per cell, track representations maintain expression profiles across genomic coordinates, capturing features such as isoform usage, alternative splicing patterns, and allele-specific expression [29].

2.4 Deep Learning in Genomics

The application of deep learning to genomic data has transformed our ability to interpret the complex patterns embedded in DNA sequences and predict their functional consequences. These computational approaches offer powerful frameworks for modeling the relationship between genomic sequences and biological functions.

2.4.1 Deep Learning Foundations in Genomics

Deep learning models have emerged as particularly well-suited for genomic analysis due to their capacity to learn hierarchical representations from raw sequence data [30]. Unlike traditional machine learning approaches that rely on manually crafted features, deep neural networks can

directly process nucleotide sequences and capture complex patterns at multiple scales (from local motifs to broader regulatory grammars [31]).

Early applications focused on predicting transcription factor binding sites and DNase I hypersensitivity regions [32]. These successes demonstrated that neural networks could learn biologically meaningful sequence features directly from data. The field has since expanded to address diverse challenges including variant effect prediction, enhancer-promoter interactions, RNA splicing patterns, and chromatin state inference [33].

Deep learning's impact in genomics stems from several key advantages: the ability to model non-linear relationships, capacity to integrate heterogeneous data types, and potential for transfer learning across related tasks. However, these approaches also face challenges including interpretability limitations, requirements for large training datasets, and difficulty in incorporating evolutionary conservation and three-dimensional genomic structure [31].

2.4.2 Convolutional Neural Networks for Sequence Analysis

Convolutional Neural Networks (CNNs) have become a foundational architecture for genomic sequence analysis, inspired by their success in computer vision. In genomics, CNNs typically process DNA as one-hot encoded matrices, where convolutional filters act as position weight matrix-like detectors that scan sequences for recurring patterns [34].

The key advantage of CNNs lies in their parameter sharing and local connectivity, which efficiently capture position-invariant sequence motifs regardless of their location within the input sequence. Early CNN architectures like DeepBind [32] demonstrated superior performance in predicting protein-DNA binding affinities compared to traditional methods. This success spawned more sophisticated models including DeepSEA [35], which predicts chromatin features across multiple cell types.

More recent CNN architectures have incorporated dilated convolutions to expand receptive fields and residual connections to facilitate training of deeper networks. These advances have enabled models like Basenji [36] to predict cell-type-specific epigenetic marks and expression levels across long genomic distances by incorporating wider sequence context. Notably, CNN architectures have proven effective at capturing regulatory syntax—the grammar-like rules governing how combinations of transcription factor binding sites collectively influence function [37].

2.4.3 Transformer Models in Genomic Sequence Analysis

Transformer architectures, which revolutionized natural language processing through self-attention mechanisms, have recently been adapted to genomic sequence analysis with remarkable success. Unlike CNNs, transformers can directly model long-range dependencies between distant positions in a sequence without requiring recurrent connections [38].

The self-attention mechanism allows each position in a sequence to attend to all other positions, enabling transformers to capture complex interactions between motifs separated by large distances. This capability is particularly valuable for modeling enhancer-promoter interactions, alternative splicing, and three-dimensional chromatin structure effects on gene regulation [37].

Notable genomic transformer implementations include DNABERT [39], which applies BERT-style pre-training to DNA sequences, and Enformer [37], which combines transformer blocks with convolutional layers to predict gene expression from sequences spanning 200kb regions. These models have demonstrated state-of-the-art performance in predicting regulatory effects of distal elements and non-coding variants.

The pre-training and fine-tuning paradigm common in transformer applications has proven especially valuable in genomics, where labeled data for specific tasks may be limited. By pre-training on abundant unlabeled genomic sequences, these models develop representations that capture fundamental biological patterns transferable to downstream tasks [39].

2.4.4 Sequence-to-Function Models for Gene Expression Prediction

Sequence-to-function models aim to predict functional outputs, particularly gene expression levels, directly from DNA sequence. These models address the fundamental question in regulatory genomics: how does the non-coding genome control when, where, and to what extent genes are expressed?

Early sequence-to-expression models focused on promoter regions, demonstrating that sequence features immediately surrounding the transcription start site contain sufficient information to predict expression levels across cell types [40]. More recent approaches have expanded to incorporate broader genomic contexts, including distal enhancers and insulators.

State-of-the-art models like Enformer [37] integrate information across 100-200kb regions to predict expression at single-nucleotide resolution across diverse cell types. These models effectively capture the contributions of distal regulatory elements, demonstrating that sequence alone contains substantial information about tissue-specific expression patterns.

A particularly valuable application of these models is predicting the effects of non-coding variants on gene expression. By performing in silico mutagenesis, systematically altering sequences and observing predicted expression changes, researchers can prioritize variants for experimental validation and gain insights into the regulatory code [41].

Despite their predictive power, current sequence-to-function models face limitations in modeling the influence of three-dimensional genome organization, epigenetic memory, and trans-acting factors that vary between cellular contexts.

2.4.5 Challenges and Future Directions

While deep learning has transformed genomic analysis, significant challenges remain. Interpretability continues to be a central concern, as understanding the biological mechanisms underlying predictions is crucial for scientific discovery. Approaches like attribution methods, feature visualization, and in silico mutagenesis offer partial solutions, but extracting biologically meaningful insights from complex.

Data limitations also constrain progress, particularly for rare cell types, developmental stages, and disease states. Transfer learning, data augmentation, and multi-task learning strategies are being developed to address these challenges. Additionally, integrating evolutionary conservation information and incorporating three-dimensional genome structure remain active areas of development.

Future directions include multimodal models that simultaneously process DNA sequence alongside epigenetic, expression, and imaging data; generative models that design sequences with desired functional properties; and self-supervised approaches that leverage massive unlabeled genomic datasets to learn general biological principles [42].

2.4.6 Explainable AI in Genomics

As deep learning models become increasingly powerful at predicting functional outcomes from DNA sequences, the need to interpret these complex models has emerged as a critical research priority. Explainable AI approaches seek to go beyond the "black box" nature of deep neural networks, providing insights into the patterns and features these models use to make predictions.

Explainable AI techniques have enabled researchers to extract biological knowledge from deep learning models trained on genomic data, effectively using these models as discovery tools rather than merely prediction engines [43].

Attribution Methods

Attribution methods, which identify input features most influential for model predictions, have proven particularly valuable for understanding regulatory sequences. Saliency maps, which compute the gradient of the output with respect to input nucleotides, have revealed binding motifs for transcription factors and regulatory elements without prior knowledge of their existence [44]. More sophisticated approaches like DeepLIFT and integrated gradients overcome limitations of simple gradient methods by incorporating reference sequences and integration paths, producing more accurate importance scores for each nucleotide position [43].

The application of these methods to models mentioned before in 2.4.2 like DeepSEA and Basenji, has uncovered known and novel transcription factor binding motifs, demonstrating that neural networks independently learn biologically meaningful sequence patterns [35, 36]. Studies like Enformer [37] and more recently Alphagenome [42] used attribution maps to identify regulatory elements controlling gene expression across 200kb (Enformer) and 1 Mb (AlphaGenome) regions, revealing long-range interactions that would be difficult to detect through traditional experimental approaches.

In Silico Mutagenesis

In silico mutagenesis systematically alters input sequences and observes changes in model predictions, mimicking experimental mutagenesis at scale. This approach has proven particularly effective for discovering regulatory syntax (the rules governing how motif combinations, spacing, and orientation affect function [45]).

DeepBind and related models subjected to in silico mutagenesis revealed that transcription factor binding depends not only on primary motif sequences but also on their flanking contexts [32]. More recent work has employed combinatorial mutagenesis to detect epistatic interactions between motifs, revealing cooperative and competitive relationships between transcription factors that collectively determine regulatory activity [37].

Feature Visualization

Feature visualization techniques generate synthetic sequences that maximally activate specific neurons or layers, revealing the patterns learned by the network. This approach has identified complex regulatory grammars, including gapped motifs and motif pairs with specific orientations and spacings [45].

Key Discoveries in Regulatory Logic

The development of sophisticated sequence-to-function models has revealed fundamental principles of gene regulation that were previously difficult to discern from experimental data alone. Models like Basenji established that convolutional neural networks could successfully predict cell-type-specific regulatory activity from DNA sequence, demonstrating that sequence contains sufficient information to explain substantial variance in epigenetic and transcriptional profiles across diverse cellular contexts [36].

Enformer represented a major advance in modeling long-range regulatory interactions through its expanded receptive field and transformer architecture. By capturing regulatory relationships across much larger genomic distances, the model revealed that distal enhancer-promoter interactions contribute significantly to gene expression patterns. The model's learned attention patterns indicated an understanding of canonical regulatory elements, including enhancers and insulators, suggesting that deep learning approaches can discover fundamental organizational principles of genome regulation without explicit programming of these concepts [37].

The extension to single-cell resolution through models like Decima demonstrated that regulatory rules of cell-type specificity and disease states can be learned directly from sequence data. This work showed that the regulatory code operates differently across cellular contexts, with the same genomic sequences producing distinct expression outputs depending on the cellular environment [46].

2.4.7 Challenges in Interpreting Deep Learning Models in Genomics

Despite significant progress, interpreting deep learning models in genomics presents substantial challenges that limit their utility for biological discovery.

2.4.8 Fine-Tuning

The application of transfer learning approaches, particularly fine-tuning pre-trained models, has emerged as a powerful paradigm in genomics machine learning. This approach utilizes large-scale self-supervised learning on vast genomic datasets to develop models with general sequence understanding that can be adapted to specific downstream tasks with limited labeled data. This approach contrasts with training models from scratch, which initializes parameters randomly and requires substantial task-specific data to learn meaningful representations.

The fine-tuning paradigm rests on the premise that pre-trained models learn generalizable rep-

resentations of data structure that transfer across related tasks. In natural language processing, models like BERT and GPT demonstrated that pre-training on massive text corpora enables adaptation to diverse language tasks with remarkable efficiency [47, 48]. This success has inspired analogous approaches in genomics.

Fine-tuning typically involves several key steps: first, a model with a suitable architecture is trained on a large dataset using a supervised objective; second, the pre-trained model is initialized with these learned parameters; and finally, the model is further trained on a smaller task-specific dataset, often with a modified output layer appropriate to the downstream task [49].

In genomics, foundation models like Enformer [37] and Borzoi [50] are typically trained in a supervised manner from the start on bulk genomic data, then fine-tuned to predict new cell types or assay types rather than relying on self-supervised pre-training approaches. During fine-tuning, parameters may be updated at different learning rates, with later layers typically receiving more substantial updates than earlier feature extraction layers.

Chapter 3

Materials and Methods

3.1 Data Collection

In this section we present datasets and the pre-processing applied to each of them.

3.1.1 scATAC Dataset

The scATAC dataset used in this thesis was obtained from Durham et al. [51] (GEO accession: **GSE157017**) and consists of 30,930 cells from C. elegans larvae at the middle L2 developmental stage. The original dataset was generated using Single cell Combinatorial Indexing ATAC-seq (sci-ATAC) protocol on wild-type worms, where nuclei were fixed, isolated, and frozen prior to library preparation.

The chromatin accessibility data underwent hierarchical topic modeling by the original authors. Initial processing involved bulk peak calling using MACS2 followed by Latent Dirichlet Allocation (LDA) modeling on the entire dataset to identify primary accessibility topics representing major cell types and tissues. Subsequently, iterative topic modeling was performed on cell type-specific subsets to refine cellular classifications and identify subtype-specific accessibility patterns.

The processed dataset includes multiple data formats: consensus peaks representing regions accessible across the entire dataset (filtered using interquartile range criteria to remove regions open in very few or very many cells), topic-specific peak sets corresponding to individual LDA topics from both whole-worm and subset analyses, and bigWig files containing normalized accessibility signals.

Our analyses started with primarily the consensus peak set containing 47,226 accessible chromatin regions of the C. elegans genome (WS235/ce11 assembly).

3.1.2 scRNA Datasets

The five scRNA datasets used in this thesis were taken from publicly available sources and are composed as follows.

Embryonic scRNA Dataset

The dataset was taken from Packer et al. [24] (GEO Accession number: GSE126954), and consists of 89,701 cells and 20,222 genes. These cells were taken from 2 different experiments conducted on C. elegans embryos, with a total of 7 different batches.

The cells are classified by lineage (36 classes), batch and timepoint (going from 0 to 850 minutes). As shown in **Figure 3.1**, most of the cells were collected during middle developmental stages. Quality control filtering was applied according to the criteria described in the original publication, with filter status indicated for each cell in the corresponding metadata column. After removing cells that failed quality control, the final dataset contained 86,024 cells.

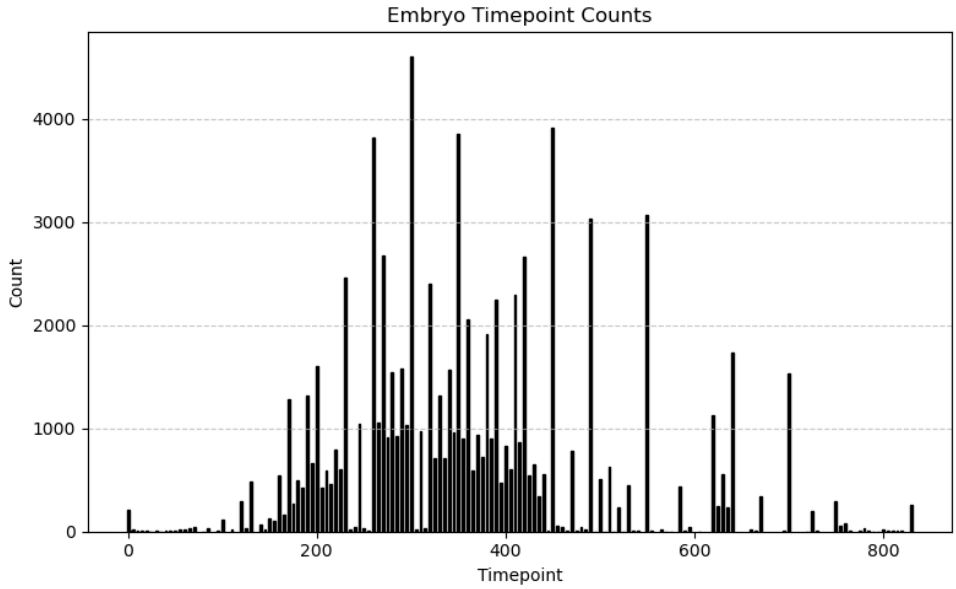


Figure 3.1: Distribution of cells by timepoints in minutes.

Stage L2 scRNA Dataset 1

This dataset contains 99,643 cells for 20,052 genes from C. elegans larvae at the L2 developmental stage, obtained from Toker et al. [52]. The dataset was originally comprised of multiple

species (C. elegans, C. briggsae, C. tropicalis). The dataset was used as provided by the original authors, with only the preprocessing steps described in the original publication applied. This dataset was also divided in 8 batches and 2 different tissue types annotations, "Broad Tissue" (20 classes) and "Fine Tissue" (107 classes).

Stage L2 scRNA Dataset 2

A second L2 stage dataset comprising 35,987 cells and 20,222 genes was obtained from Cao et al. [23] to complement L2 scRNA Dataset 1.

This dataset underwent pre-processing as described in the original publication. The cells were derived from 2 different experiments: Experiment 1 contributed 35,480 cells and Experiment 2 contributed 507 cells. Cell type annotations were provided for 117 distinct classes.

Stage L4 scRNA Dataset

The L4 larval stage dataset contains 100,955 cells for 22,469 from Taylor et al. [53]. Cells were annotated by cell type (169 classes) and tissue type (11 classes). Most of the cell types in this dataset are neurons (118 classes). This dataset underwent pre-processing as described in the original publication.

Adult scRNA Dataset

The Adult stage dataset contained 47,423 cells for 20,305 genes taken from Roux et al. [54]. The AnnData object containing the dataset was obtained from the Calico Labs website directly (http://c.elegans.aging.atlas.research.calicolabs.com/). No pre-processing was applied except for the steps already described in the original paper.

The cells here are annotated by timepoint (6 classes in days: "d1", "d3", "d5", "d8", "d11", "d15"), and by cell type (211 classes). In our study we decided to not use the timepoint annotation.

3.2 Computing Resources

3.2.1 High Performance Computing (HPC)

All computational analyses were performed on the Vlaams Supercomputer Center (VSC) HPC cluster. The cluster provided access to Nvdia H100 gpus for deep learning model training. Typical allocations for RAM were of 50 GB.

3.2.2 Keras and WandB

Deep learning model development and training were conducted using Keras version 3.4.0 as the primary framework for neural network construction and training. GPU acceleration through the underlying TensorFlow backend with CUDA 12.6 was utilized to optimize training performance on the HPC cluster.

Weights & Biases (WandB) version 0.18.1 was employed for experiment tracking, hyperparameter optimization and model performance monitoring. WandB facilitated systematic logging of training metrics, loss curves, and model artifacts across multiple experimental runs.

3.2.3 CREsted

Cis-Regulatory Element prediction from DNA sequence (CREsted) [55] framework version 1.5.0 was employed for chromatin accessibility prediction from DNA sequence. CREsted is a deep learning framework specifically designed for training convolutional neural networks on genomic sequences to predict chromatin accessibility patterns.

CREsted facilitated end-to-end workflows from raw DNA sequences to trained models, with modular design allowing for custom architectures while maintaining standardized preprocessing (sequence extraction, one-hot encoding, data augmentation). Built-in optimization strategies and performance monitoring ensured robust model development and validation.

3.2.4 pyCisTopic

pyCisTopic version 2.0 was employed for consensus peak calling and genomic region processing. The framework was used to merge topic-specific chromatin accessibility peaks from the Durham et al. dataset [51] into a unified consensus peak set.

Individual topic-specific peak files (subset refinement LDA summits) were imported and converted to PyRanges objects for genomic interval manipulation. Consensus peak calling was performed using the `iterative_peak_calling.get_consensus_peaks()` function with a peak half-width parameter of 350 base pairs, resulting in 700 bp consensus regions. Chromosome size information from the ce11 genome assembly was utilized to ensure proper genomic coordinate handling.

The final consensus peak set was exported as a BED format file for downstream analysis, providing a standardized set of chromatin accessibility regions suitable for integration with gene expression data and deep learning model training.

3.2.5 BEDTools

BEDTools suite version 2.30.0 was utilized for genomic interval operations.

Specific operations included filtering consensus peaks to remove exonic regions using `BEDTools intersect -wa` with C. elegans exon annotations taken from the C. elegans gtf file. This filtering step ensured that downstream analyses focused on intergenic and intronic regulatory regions by excluding protein-coding sequences from the chromatin accessibility dataset.

Command used to check on overlaps between regions:

```
BEDTools intersect -wa -f 0.5 -a consensus_peaks.bed -b ce11_exons.bed
```

3.2.6 IGV (Integrative Genomics Viewer)

ntegrative Genomics Viewer (IGV) version 2.17.3.03 was used for visual inspection and validation of genomic regions and chromatin accessibility patterns. The browser was used to manually examine consensus peaks, validate peak calling results, and assess the quality of chromatin accessibility signals in relation to genomic features such as gene annotations and regulatory elements.

This visual validation was essential for quality control of computational predictions and provided biological context for interpreting chromatin accessibility patterns across different genomic loci. **Figures 4.1 and 4.3**, were generated using IGV.

3.2.7 scVI

Single-cell Variational Inference (scVI) [56] version 1.3.0 was used for the integration of the datasets described in **section 3.1.2**. scVI is a deep generative model that trough variational autoencoders learns low-dimensional representations of single-cell gene expression data while accounting for technical confounders such as batch effects and sequencing depth.

The scVI model was trained on the combined scRNA datasets (Section 3.1.2) to generate a shared latent space representation that corrects for batch effects between different experiments and developmental stages. Model training utilized default hyperparameters with 65 training epochs. The learned latent representations were subsequently used for downstream analyses including dimensionality reduction, clustering and trajectory analysis across the integrated developmental stages.

3.2.8 Blender

Blender version 4.0 was used for 3D visualization of UMAP data calculated using the scanpy UMAP function `scanpy.tl.umap()`.

Custom Python scripts were developed to import UMAP embeddings and generate sphere instances for each UMAP coordinate while preserving their respective cell type labels, allowing for interactive 3D exploration of single-cell data structures.

3.3 Deep Learning Models

3.3.1 CREsted CNNs for scATAC Modeling

Convolutional Neural Networks within the CREsted framework were configured using the ChromBPNet architecture adapted for 700 base pair DNA sequences. The model utilized dilated convolutional layers (n=7) with standard activation functions and was trained to predict chromatin accessibility across all cell types in the scATAC dataset (**Section 3.1.1**) (35 output classes).

Training was performed using the Adam optimizer with a learning rate of 1e-3 and a batch size of 128. The loss function was set to "CosineMSELoss", combining cosine similarity and mean squared error. Model performance was monitored using multiple metrics including Mean Squared Error, Cosine Similarity, Pearson Correlation, and Concordance Correlation Coefficient.

Data augmentation strategies included stochastic sequence shifting (maximum 3 base pairs) and reverse complement augmentation to double the effective training dataset size. Training progress and hyperparameter tracking were managed through WandB logging system (**Section 3.2.2**).

3.3.2 Calico Pre-trained Model for scRNA Modeling

We used a pre-trained deep learning model developed by Calico as the foundation for scRNA expression prediction from DNA sequence. The base model architecture consists of a hybrid CNN-Transformer design with 131,072 base pair input sequences processed through convolutional tower layers with progressive pooling, followed by transformer blocks containing multi-head attention and feed-forward layers.

The original model was trained on large-scale genomic datasets for chromatin accessibility and gene expression prediction across different species. For adaptation to C. elegans scRNA data, transfer learning was applied by replacing the final prediction head with a global average pooling layer followed by a dense layer configured for tissue-specific expression prediction across all cell types in our dataset.

Model fine-tuning used the Adam optimizer with a reduced learning rate of 2e-5 to preserve pre-trained features while adapting to the target task. The loss function used was "CosineMSELoss". Training was performed for 25 epochs with a batch size of 2, incorporating data augmentation strategies including stochastic sequence shifting (maximum 500 bp) and reverse complement augmentation. Model performance was monitored using multiple metrics including Mean Squared Error, Cosine Similarity, Pearson Correlation, and Concordance Correlation Coefficient.

Chapter 4

Results

4.1 scATAC model and creation of regions

4.1.1 Training of the model

To develop an accurate accessibility prediction model, we initially targeted genomic regions with higher likelihood of chromatin accessibility. We computed consensus regions across different tissues using data and trained a peak regression model on these regions.
These consensus regions appeared to include most of the individual peaks and cover a great span of the genome. The initially covered regions were of uneven dimensions and appeared to capture non-significant peaks and regions where no peaks were present, so we proceeded to cut them into equal dimensions of 700bp using pyCisTopic 3.2.4, shown in **Figure 4.1**.

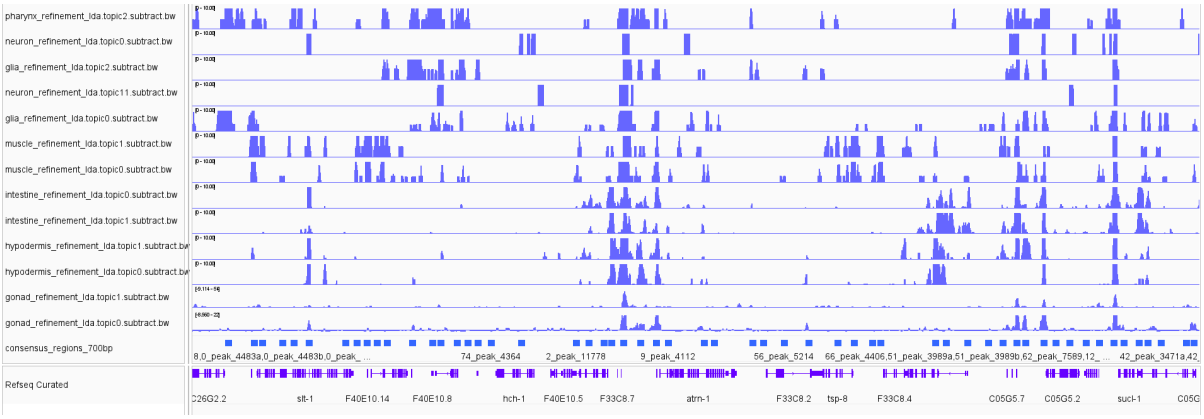
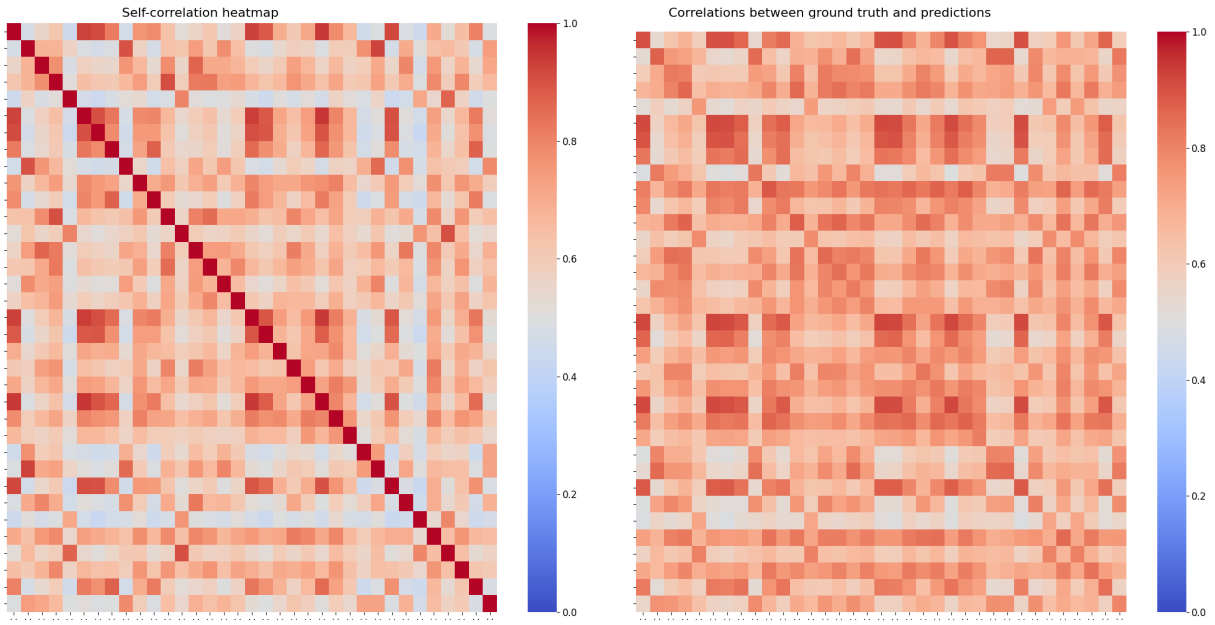


Figure 4.1: IGV Visualization of the initial 700bp consensus regions. Genomic Region: chrX:14,668,321-14,761,464

Using these initial 700bp consensus regions (amounting to a total of 47,226 regions), we proceeded in training a model using CREsted chrombpnet CNN model. This resulted in good performances overall (Validation Pearson Correlation of 0.84). To better evaluate the model we plotted correlation heatmaps, shown in **Figure 4.2**.



(a) Self-correlations of different topics

(b) Correlation between predicted values and ground truth

Figure 4.2: Heatmaps of self-correlations of different topics (left) and correlation between predicted values and ground truth (right).

4.1.2 Selection of regions

Initial region selection used existing BigWig files 3.1.1 to identify consensus accessible regions of 700 base pairs. Visual inspection using the ntegrative Genomics Viewer (IGV) revealed the genomic distribution and characteristics of these consensus regions (**Figure 4.1**). Based on these analyses, we proceeded to stratify the regions according to their overlap with exonic sequences.

Using BEDTools to stratify the regions, we found that of 47,226 regions, exactly 18,568 overlapped with exonic regions. Given the biological implausibility of this high overlap [57], we decided to analyze the differential impact of exonic versus non-exonic regions **Figure 4.3**.

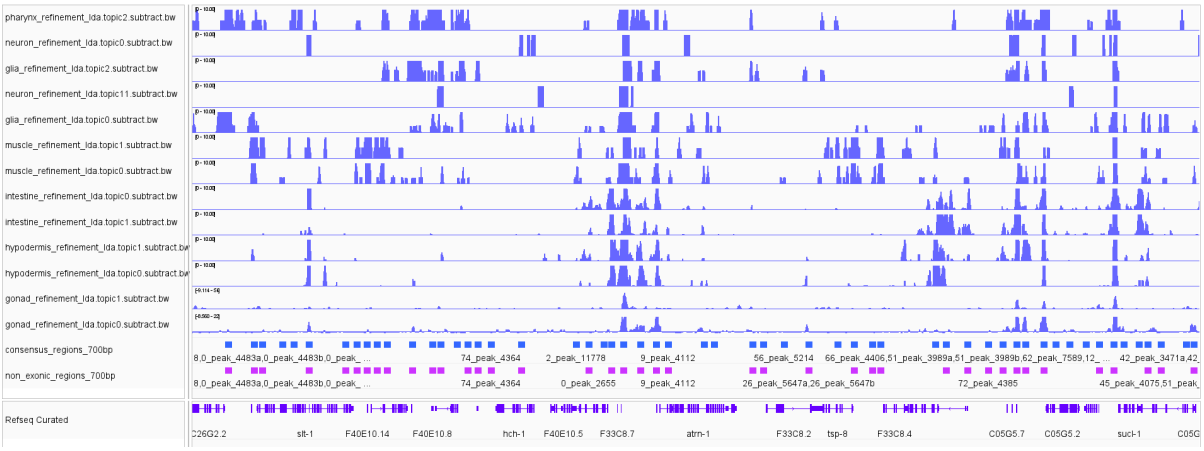


Figure 4.3: IGV Visualization of the track containing exonic consensus regions (upper) and the one without exonic consensus regions (bottom). Genomic Region: chrX:14,668,321-14,761,464

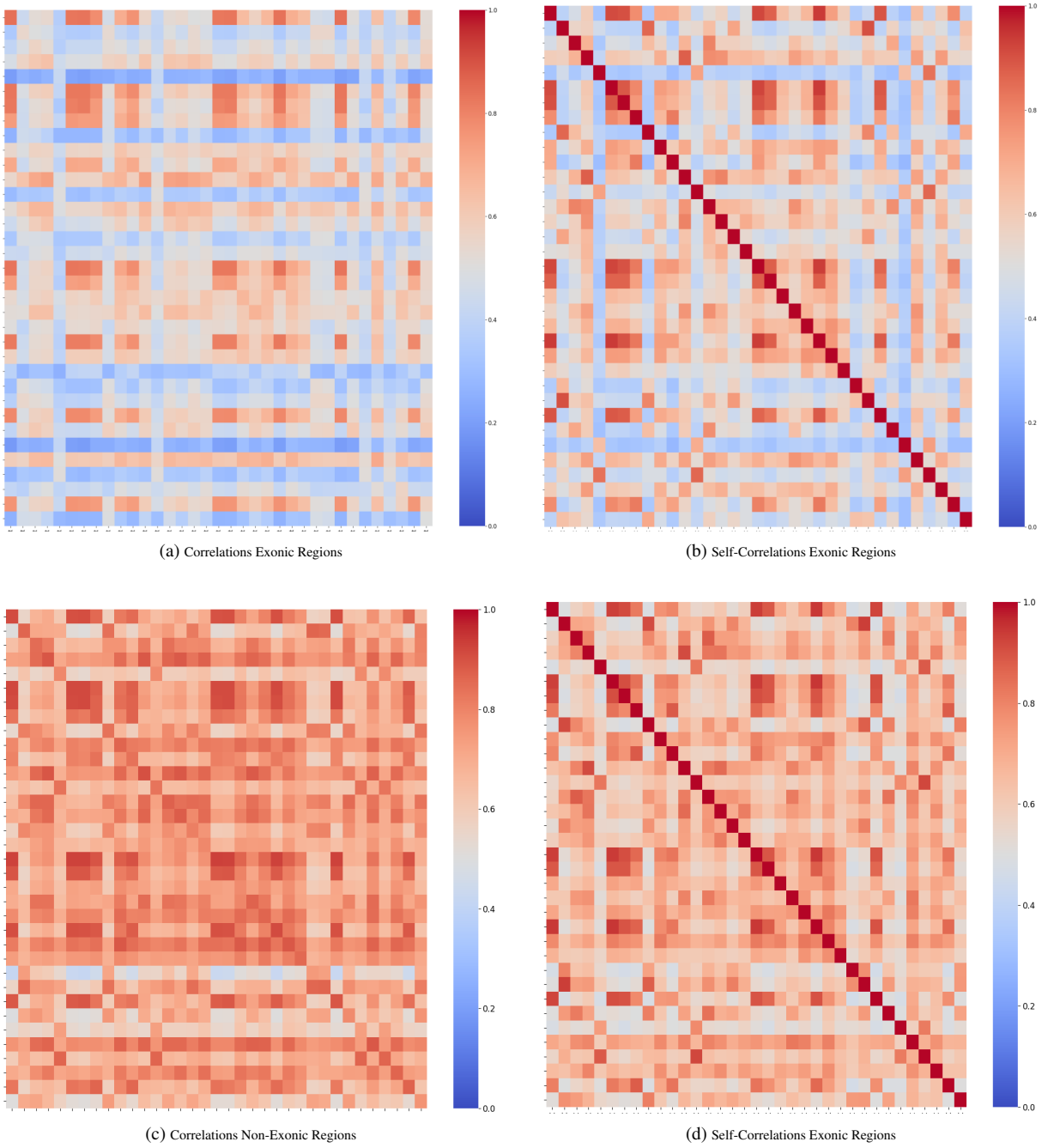
The remaining Non-Exonic regions amounted to 28,658. These regions are uniformly distributed across chromosomes, and were computed using BEDTools 3.2.5.

4.1.3 Models trained on Exonic and Non-Exonic regions

Using our originally trained model, we evaluated performance separately on Non-Exonic regions and Exonic regions. This stratification allowed us to assess whether accessibility prediction accuracy varied between coding and non-coding genomic contexts.

In **Figure 4.4**, we show the results of this analysis. Non-Exonic regions had higher correlations between predicted values and ground-truth values and contributed more on the model performances. Notably, the self-correlation patterns reveal fundamental differences in model performance quality between the two region types.

As shown by **Figures 4.4b, 4.4d**, the two sets of regions exhibit markedly different self-correlation structures between topics. In **Figures 4.4a, 4.4c**, the Non-Exonic regions maintain a clear diagonal pattern in their correlation heatmap, indicating that topics are most highly correlated with themselves, a fundamental indicator that the model has learned meaningful for each topic. In contrast, the Exonic regions show a completly absent diagonal structure, suggesting that the model struggles to maintain topic-specific predictions and may be producing less reliable or potentially artifactual outputs for these regions.



(a) Correlations Exonic Regions

(b) Self-Correlations Exonic Regions

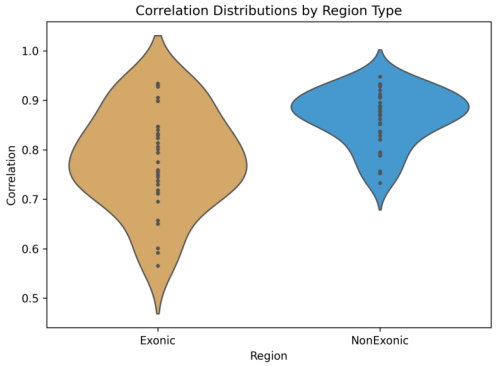
(c) Correlations Non-Exonic Regions

(d) Self-Correlations Exonic Regions

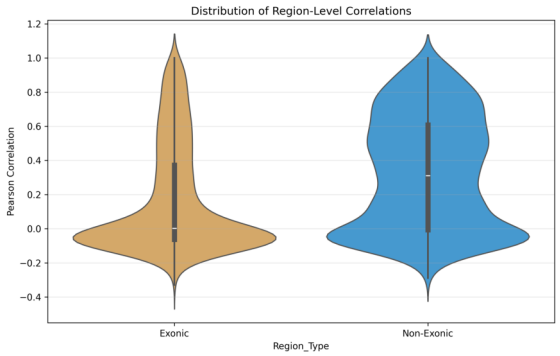
Figure 4.4: **A.** Heatmap of self-correlations of topics for the exonic consensus regions, **B.** Heatmap of self-correlations of topics for the Non-Exonic consensus regions, **C.** Heatmap of correlations between predictions and ground-truth of topics for the Exonic consensus regions, **D.** Heatmap of correlations between predictions and ground-truth of topics for the Non-Exonic consensus regions.

This loss of diagonal structure is a strong indicator that the model's learned representations for exonic regions lack the specificity necessary for accurate accessibility prediction.

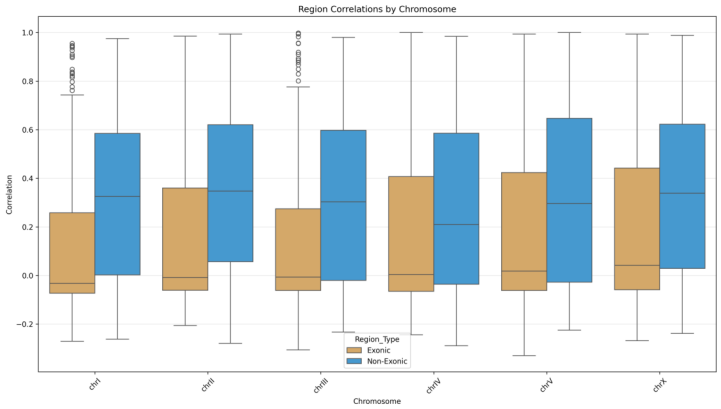
Given this fundamental difference in performance quality, we then proceeded to analyze the differences regarding the predictions correlations of these regions with respect to chromosomes and at both tissue and regional levels.



(a) Violin Plot Topics Correlations



(b) Violin Plot Region Level Correlations



(c) Box Plot Chromosome View Correlations

Figure 4.5: A. Violin Plot showing correlations in different topics, **B.** Violin Plot showing region-level correlations, **C.** Box Plot showing region-level correlations divided by chromosome.

As shown in **Figure 4.5a**, the difference in correlations between topics is substantial, with Non-Exonic regions achieving higher correlations across most topics compared to Exonic regions. For Exonic regions, the topics had a mean correlation between predictions and actual values of 0.7802 with a standard deviation of 0.0975. For Non-Exonic regions, this correlation was 0.8660 with a standard deviation of 0.0543.

The violin plot reveals that while the majority of topics show improved performance in Non-Exonic regions, the enhancement is not uniform across all topic classes, with some topics showing more pronounced improvements than others.

In **Figure 4.5b**, we show the correlation differences at the individual region level, comparing predictions versus ground truth for every region. The distributions reveal distinct patterns between the two region types. For Exonic regions, the mean correlation was 0.1644, with a median of 0.0028 and standard deviation of 0.3155. For Non-Exonic regions, the mean was 0.3259, with a median of 0.3103 and standard deviation of 0.3328.

The substantial difference between mean and median for Exonic regions (0.1644 vs 0.0028) indicates a highly skewed distribution, with most regions showing very poor prediction accuracy while a small subset achieves moderate performance. In contrast, Non-Exonic regions show nearly identical mean and median values, indicating a more symmetric distribution centered around moderate-to-good performance across most regions.

Finally, in **Figure 4.5c**, we show that even when examining region-level correlations from a chromosomal perspective, the difference between Exonic and Non-Exonic regions remains highly evident and consistent across all chromosomes. Notably, several chromosomes exhibit negative median correlations for Exonic regions.

4.1.4 Final Regions

Based on the previous comparative analysis of model performance across different genomic contexts, we established our final set of regions for downstream analysis.

In **Figure 4.6**, we show examples of how predictions and ground truth values correlate for all regions (independent of exonic status) across four different topics: two neuronal and two intestinal.

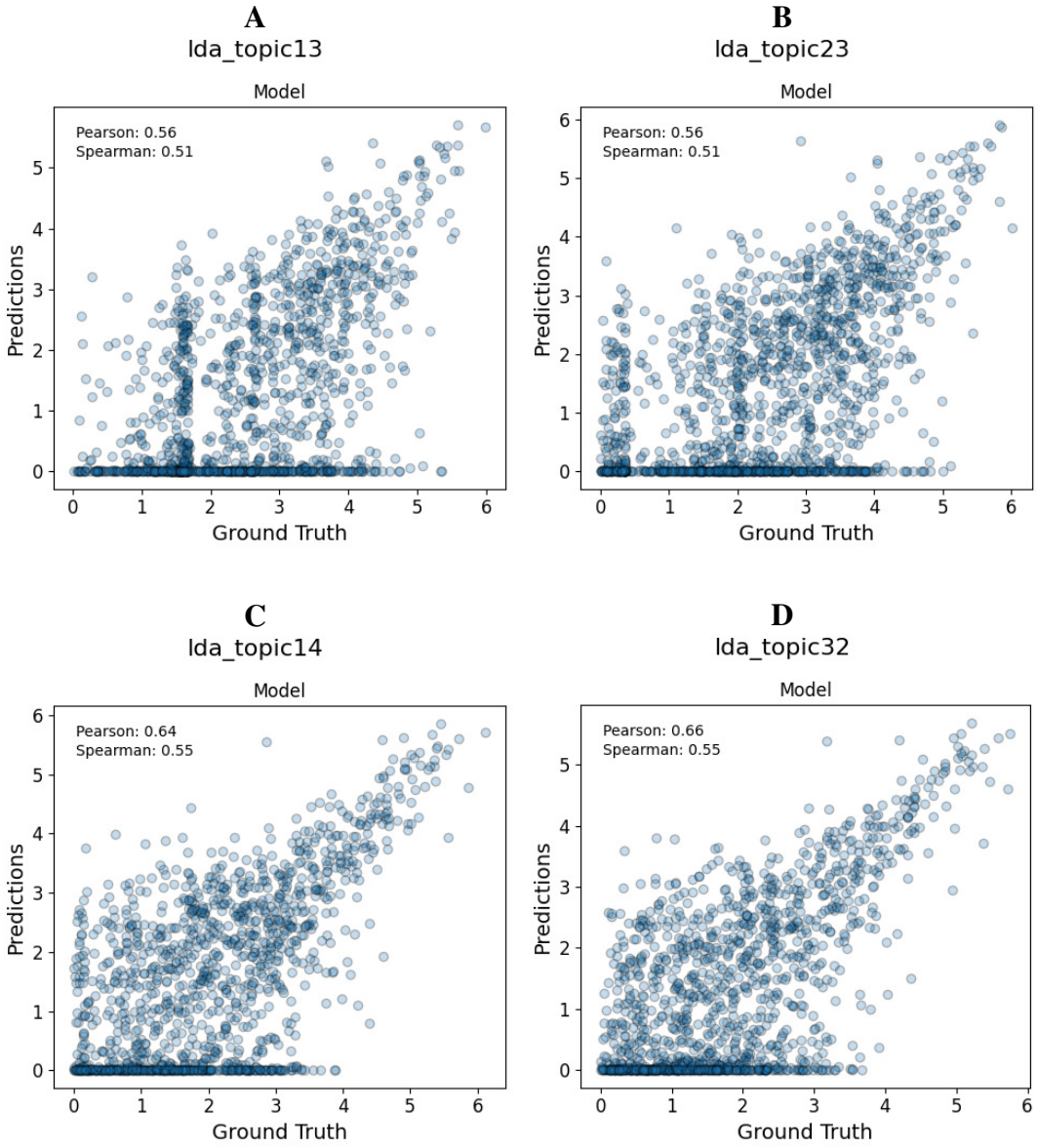


Figure 4.6: Scatter plots of regions and ground truth values for **Topic 13** (Intestine), **Topic 23** (Intestine), **Topic 14** (Neuron), **Topic 32** (Neuron).

In **Figure 4.7**, we provide an example of the contributions for the prediction of gene **ges-1**, which is reported by WormBase to be expressed mainly in intestinal cell types. This region was selected from the list of non-exonic regions.

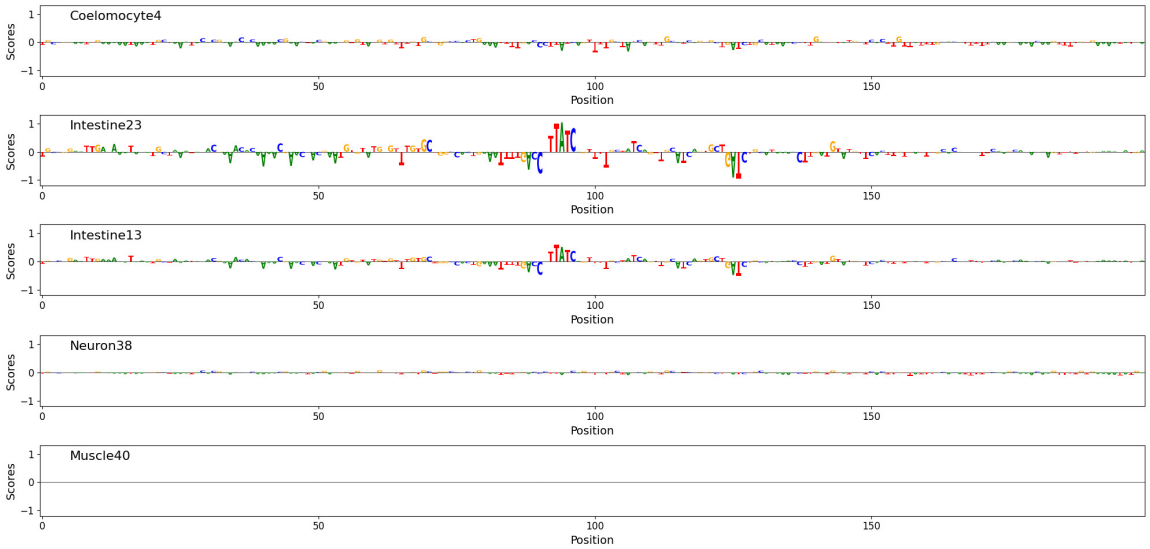


Figure 4.7: Visualization of contributions for region upstream of **ges-1** gene. Shown here are only 200 bp in the middle of genomic region: **chrV:1,432,914-1,433,613**.

The model correctly predicts peaks for intestinal topics and identifies **TTATC** as a motif, which closely matches the *elt-3* motif in JASPAR, which is known to be relevant in intestinal tissues.

Visualizing regions on model embeddings

To understand how our trained model internally represents different types of genomic regions, we extracted embeddings and visualized them using dimensionality reduction techniques.

In **Figure 4.8**, we show the embeddings extracted from the model plotted through UMAP and t-SNE, then colored by region status, whether it was a non-exonic or exonic region.

The visualization reveals substantial intermixing between exonic and non-exonic regions in the embedding space. While the two region types are not completely segregated, there are observable trends where certain areas of the embedding space show enrichment for one type over the other.

Furthermore, the embeddings do not indicate clear discrete clusters that would suggest strong cell-type specificity in the learned representations.

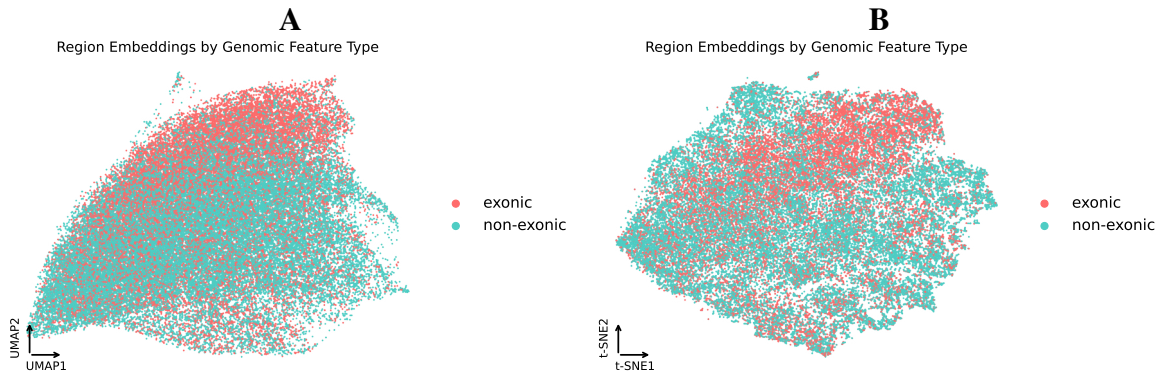


Figure 4.8: A. UMAP visualization of embeddings extracted from the scATAC model colored by region status, **B.** t-SNE visualization of embeddings extracted from the scATAC model colored by region status.

4.2 Datasets Integration

4.2.1 Individual Dataset Re-Labeling and UMAPs/t-SNEs

In order to create a unified dataset spanning different timepoints, we first prepared the original datasets by re-labeling and correcting for batch differences. Given the discrepancies in tissue and cell type labeling across the 5 datasets, we decided to unify labels referring to the same tissue or cell type to make the final dataset easily comparable.

In **Figures 4.9 to 4.13**, UMAPs of the 5 datasets originally used are displayed with the resulting new labeling. Original UMAPs can be found in methods section at figure (reference) for reference.

Embryonal Stage Dataset

For the embryonal stage dataset **((reference))**, standardizing the dataset resulted in 11 classes for Broad Tissue types (*Coelomocyte, Excretory, Glia, Hypodermis, Hypodermis_Seam, Intestine, Muscle, Neuron, Pharynx, Reproductive, Unannotated*) and 161 Fine Tissue types.

In **Figure 4.9**, we show the resulting UMAP of these assignments. After this initial labeling, 129 Fine Tissue classes remained marked as "*Unannotated*" in the Broad Tissue labels. This occurred because most of these cells were originally labeled using timepoints in minutes rather than tissue types, as explained in paragraph **((reference))** in the Methods section.

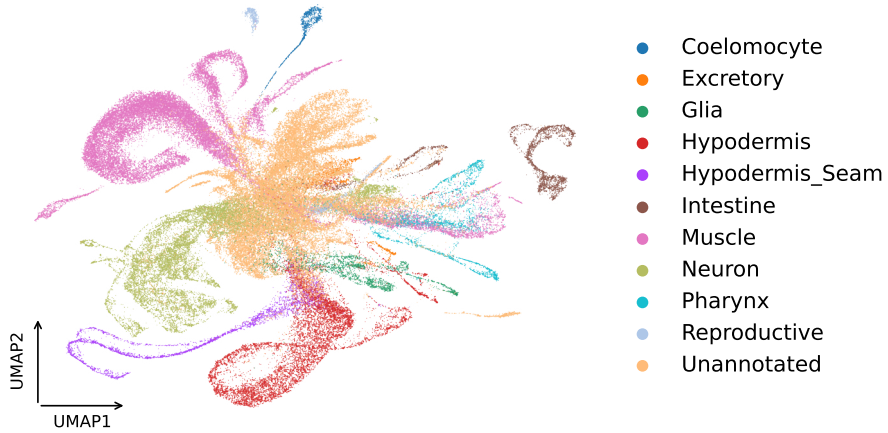


Figure 4.9: Embryonic stage dataset colored by initial broad tissue labels.

L2 Stage Hobert Dataset

For the L2 stage Hobert dataset (**(reference)**), standardizing the dataset resulted in 12 classes for Broad Tissue types (*Pharynx, Reproductive, Hypodermis_Seam, Muscle, Excretory, failed_Neuron, Glia, Intestine, Neuron, Coelomocyte, hmc, Failed_qc*) and 112 Fine Tissue types.

In **Figure 4.10**, we show the resulting UMAP of these assignments. Here, no cells were left unannotated.

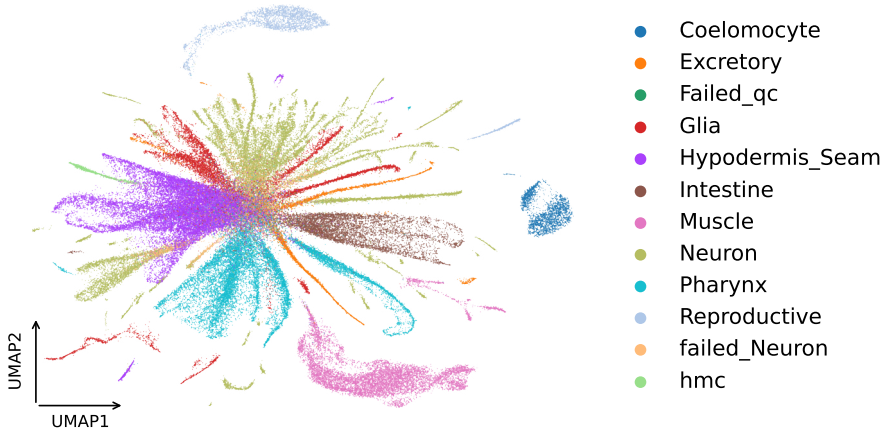


Figure 4.10: L2 stage Hobert dataset colored by initial broad tissue labels.

L2 Stage Cao Dataset

For the L2 stage Cao dataset (**(reference)**), standardizing the dataset resulted in 10 classes for Broad Tissue types (*Coelomocyte, Excretory, Glia, Hypodermis_Seam, Intestine, Muscle, Neuron, Pharynx, Reproductive, Unannotated*) and 117 Fine Tissue types.

In **Figure 4.11**, we show the resulting UMAP of these assignments. Here, 5962 cells were labeled as "*Unannotated*".

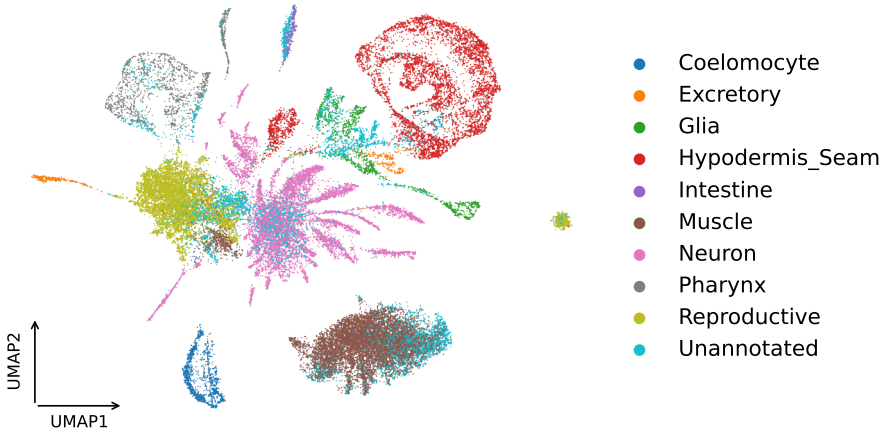


Figure 4.11: L2 stage Cao dataset colored by initial broad tissue labels.

L4 Stage Dataset

For the L4 Stage dataset (**(reference)**), standardizing the dataset resulted in 10 classes for Broad Tissue types (*Excretory, Glia, Hypodermis_Seam, Intestine, Muscle, Neuron, Pharynx, Rectal_cells, Reproductive, Unannotated*) and 155 Fine Tissue types.

In **Figure 4.12**, we show the resulting UMAP of these assignments. Here, 4195 cells were labeled as "*Unannotated*". This dataset is mainly populated by neurons.

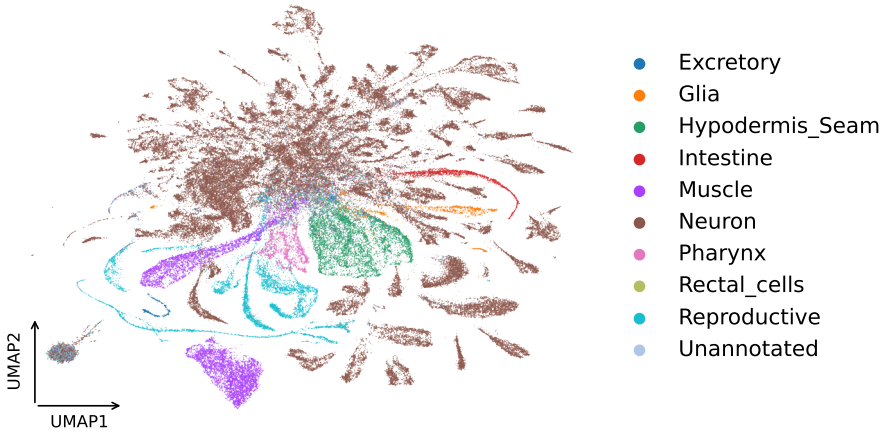


Figure 4.12: L4 stage dataset colored by initial broad tissue labels.

Adult Stage Dataset

For the Adult Stage dataset (**(reference)**), standardizing the dataset resulted in 12 classes for Broad Tissue types (*Coelomocyte, Excretory, Glia, Hypodermis_Seam, Intestine, Muscle, Neuron, Pharynx, Rectal_cells, Reproductive, Unannotated, hmc*) and 180 Fine Tissue types.

In **Figure 4.13**, we show the resulting UMAP of these assignments. Here, 912 cells were labeled as "*Unannotated*". This dataset is also mainly populated by neurons, like the L4 stage.

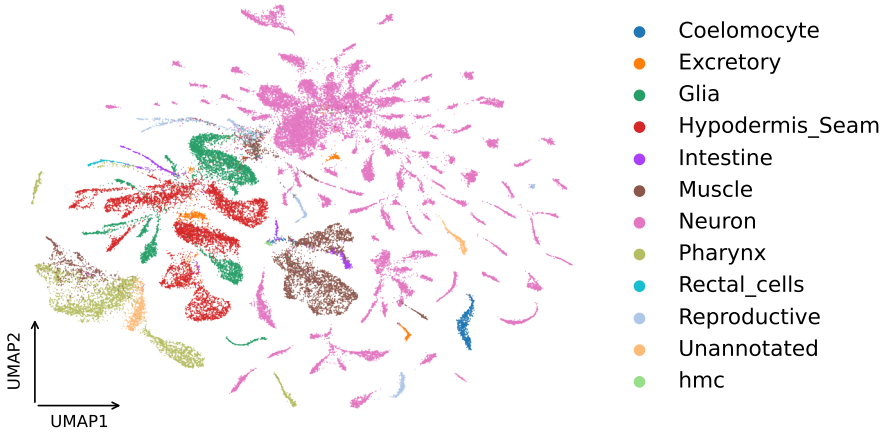
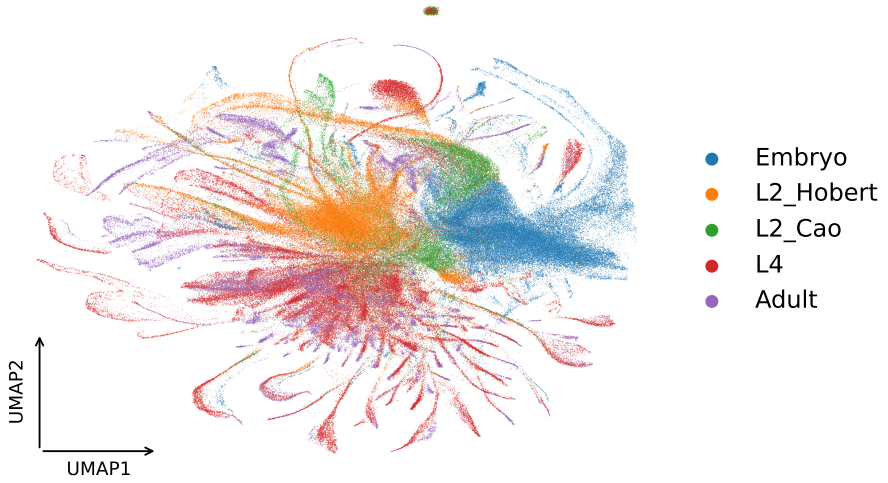


Figure 4.13: Adult stage dataset colored by initial broad tissue labels.

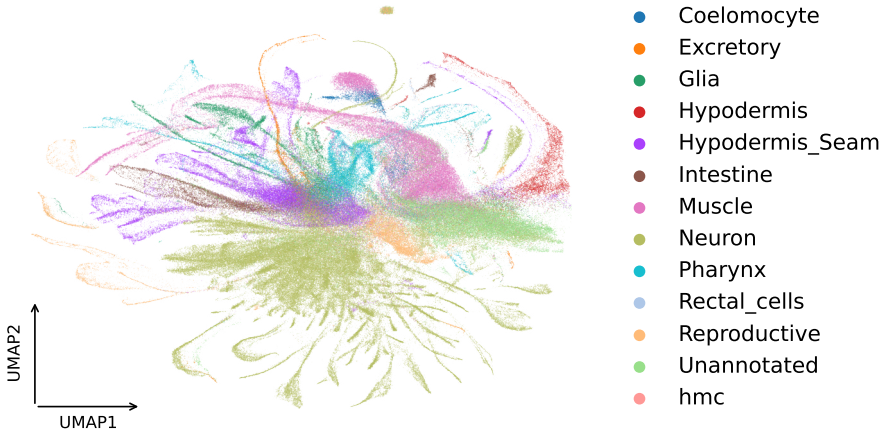
4.2.2 Pre-scVI unified dataset

Before training with scVI, the five datasets were joined together manually into one single AnnData object, without any other type of integration except the re-labeling done so far.

In **Figure 4.14**, we show the UMAP of the dataset resulting from this manual integration colored by stage and by tissue type.



(a) UMAP colored by developmental stage.



(b) UMAP colored by broad tissue type.

Figure 4.14: **UMAPs of the pre-scVI dataset. (A)** Colored by developmental stage. **(B)** Colored by broad tissue type.

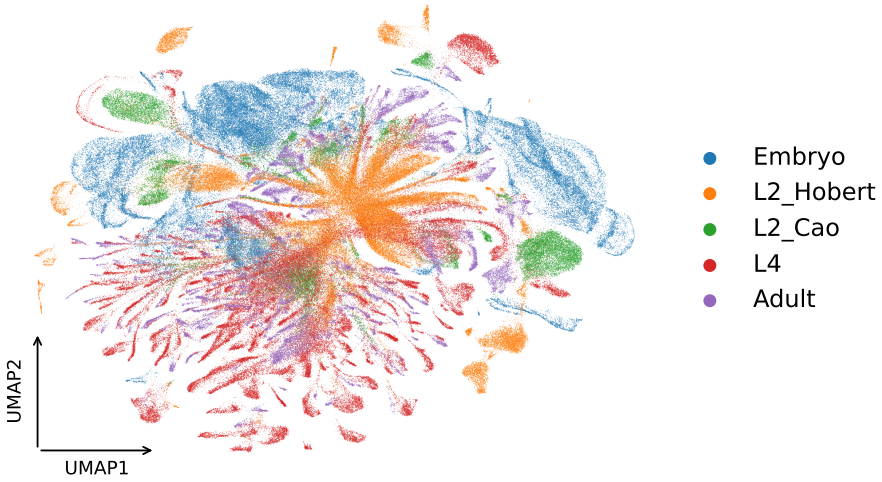
This AnnData object now has a total of 13 classes for the Broad Tissue type (*Coelomocyte, Excretory, Glia, Hypodermis, Hypodermis_Seam, Intestine, Muscle, Neuron, Pharynx, Rectal_cells, Reproductive, Unannotated, hmc*), 499 classes for the Fine Tissue type, and 724 Fine Tissue classes divided by stage (unique and non-unique). The dataset is composed of 357,627 cells and 16,804 genes.

The dataset appeared to exhibit separation by developmental stage or batch effects, which could confound downstream analyses. To address these technical artifacts and achieve proper integration across timepoints, we decided to apply scVI.

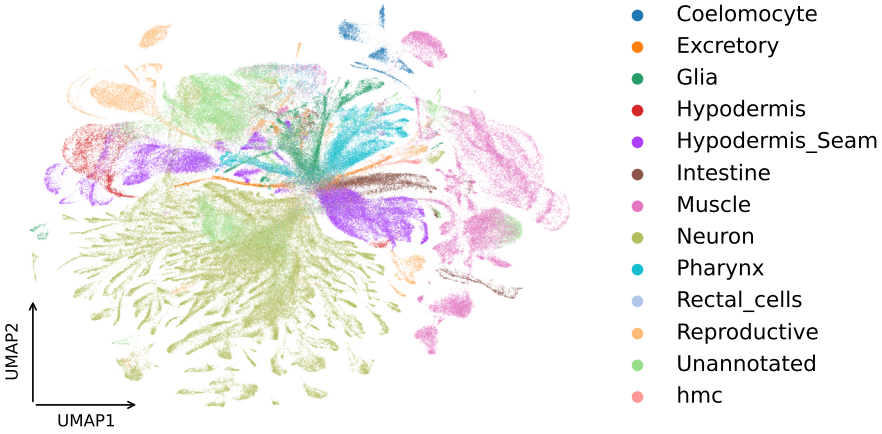
4.2.3 SCVI Integrated Dataset

Integration using scVI successfully converged with a final training loss of 1,811 and validation loss of 1,805, indicating stable model performance without overfitting. The Evidence Lower Bound (ELBO) reached 1,837 for training and 1,831 for validation. The reconstruction loss of 1,800 showed that the model successfully preserved gene expression patterns during integration. The local Kullback-Leibler (KL) divergence of approximately 31 for both training and validation datasets confirmed adequate batch correction across developmental stages, while the global KL divergence of 0.0 indicated that the model didn't require global constraints as the local corrections were sufficient.

The resulting UMAPs after this integration are shown in **Figure 4.15**.



(a) UMAP colored by developmental stage.



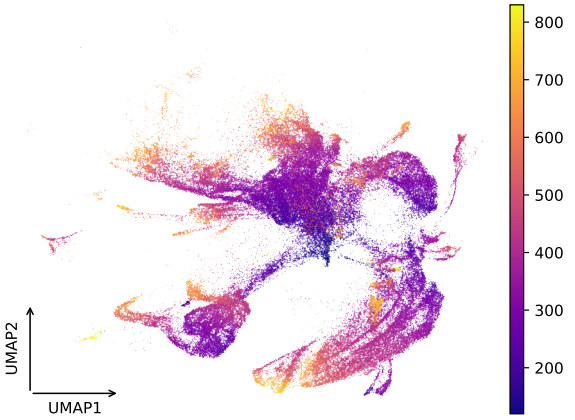
(b) UMAP colored by broad tissue type.

Figure 4.15: **A.** UMAP of integrated dataset colored by stage, **B.** UMAP of integrated dataset colored by Broad Tissue type.

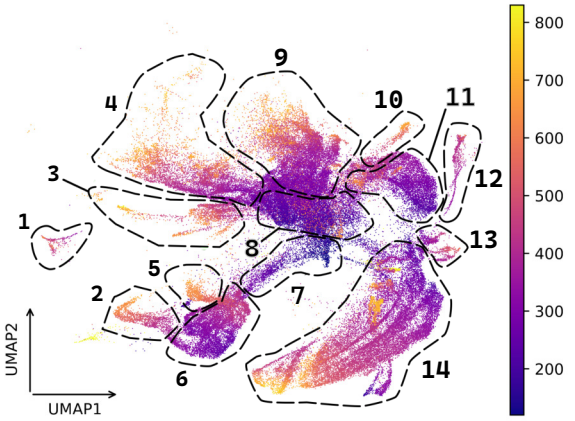
We then used the integrated dataset to further refine the labels for the cell types, particularly in embryonic cells where most of the unannotated cells were located.

After inspecting 3D UMAPs (**Figures 4.16c, 4.16d**) and 2D UMAPs (**Figures 4.16a, 4.16b**) by timepoints, we manually annotated the cells based on their developmental trajectory towards known clusters later in development (stages L2, L4, and Adult).

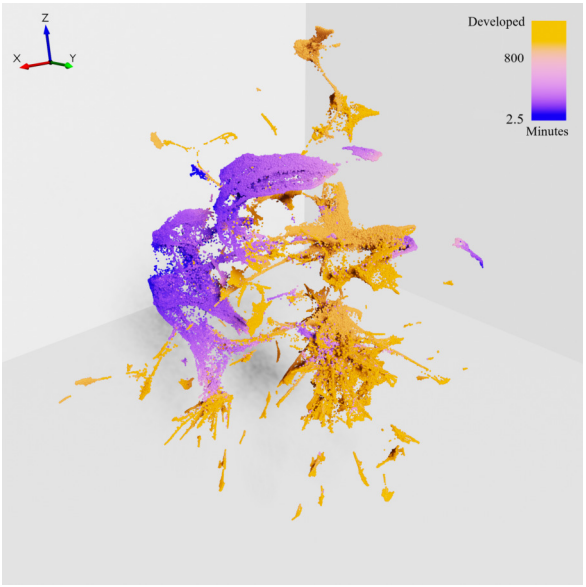
This approach allowed us to label most of the unannotated cells in the dataset and correct some misannotated ones.



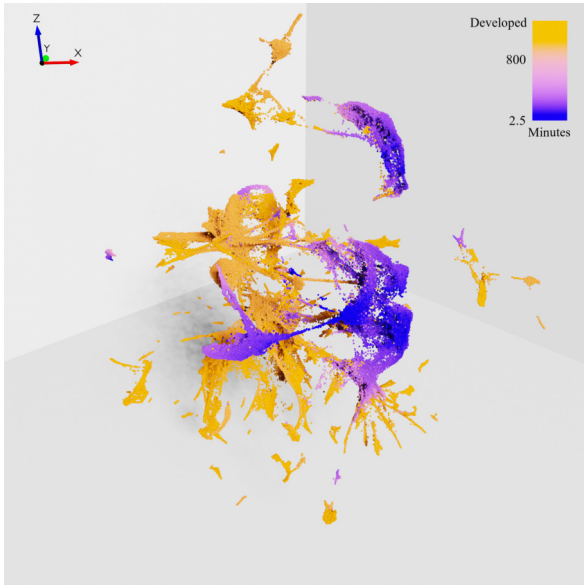
(a) Embryonic cells colored by timepoint.



(b) Labeled clusters of tissue types identified.



(c) 3D UMAP visualization of the embryonic cells labeled by timepoint (View 1).



(d) 3D UMAP visualization of the embryonic cells labeled by timepoint (View 2).

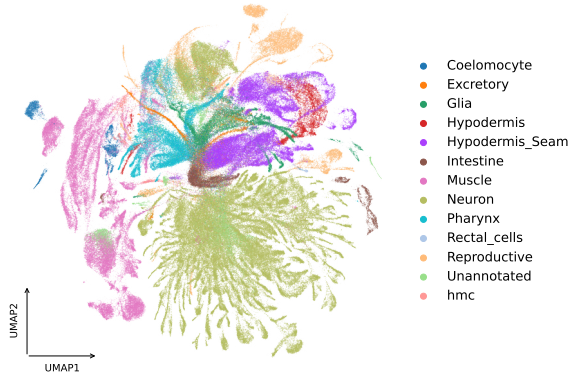
Figure 4.16: A. Embryonic cells colored by timepoints (2.5 minutes to 850 minutes) taken from integrated dataset without the L2 stage Hobert dataset, **B.** Embryonic cells colored by timepoints (2.5 minutes to 850 minutes) taken from integrated dataset without the L2 stage Hobert dataset, clustered by tissue type, **C.** 3D UMAP visualization of the embryonic cells labeled by timepoint (View 1), **D.** 3D UMAP visualization of the embryonic cells labeled by timepoint (View 2).

In particular, we successfully identified the 14 clusters shown in **Figure 4.16d**.

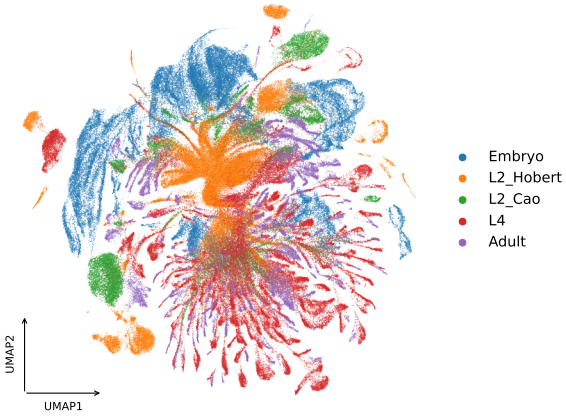
The clusters were labeled as follows: **1.** Reproductive, **2.** Seam Cells, **3.** Glia, **4.** Ciliated Amphid and Non-Amphid Neurons, **5.** Hypodermis, **6.** Precursor Seam Cells and Hypodermis, **7.** Gonadal Precursors, **8.** Neuronal Precursors, **9.** Motor Neurons and Interneurons, **10.** Arcade Cells, **11.** Pharyngeal Cells, **12.** Coelomocytes, **13.** Head Mesodermal Cells (HMC), **14.** Muscles.

Following the identification of these tissue types, we further subdivided the embryonic cells based on their developmental timepoints, organized into bins of approximately 100 minutes (e.g., 0-100 Muscle, 100-200 Muscle, 200-300 Muscle, 300-400 Muscle, 400-500 Muscle, 500-600 Muscle, 600-700 Muscle, 700-850 Muscle).

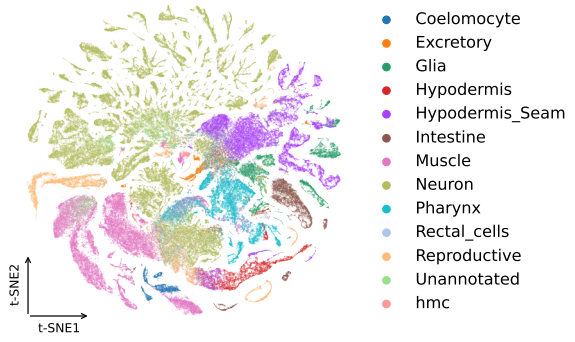
The final dataset was then composed of 13 Broad Tissue types, 449 Fine Tissue types, 611 Stage-Divided Fine Tissue types.



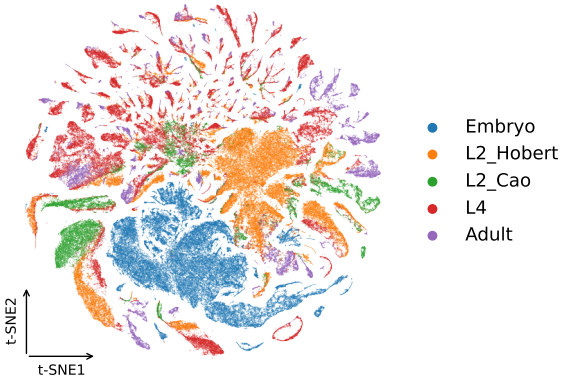
(a) UMAP plot of final integrated dataset colored by tissue.



(b) UMAP plot of final integrated dataset colored by stage.



(c) t-SNE plot of Final integrated dataset colored by tissue.



(d) t-SNE plot of Final integrated dataset colored by stage.

Figure 4.17: A. Final integrated dataset colored by Broad Tissue type (UMAP), **B.** Final integrated dataset colored by developmental stage (UMAP), **C.** Final integrated dataset colored by Broad Tissue type (t-SNE), **D.** Final integrated dataset colored by developmental stage (t-SNE).

4.3 Fine-Tuning of Pre-Trained Model

Using our integrated dataset containing most developmental stages of C. elegans and the consensus regions at L2 for reference, we proceeded to fine-tune the pre-trained model, referenced in the methods section (**(Calico model)**), to predict gene expression scalars across different tissues and developmental stages.

The model was trained using two distinct annotations: the base Fine Tissue type and the stage-divided tissue type annotation. The first approach utilized the combined power of all developmental stages for prediction, while the second allowed comparison of the same cell types across different developmental timepoints.

4.3.1 WandB training performances of different models

Training performance was monitored using WandB to track key metrics during the training of different models. In **Figure 4.18**, we show six training runs that summarize the findings of the training process.

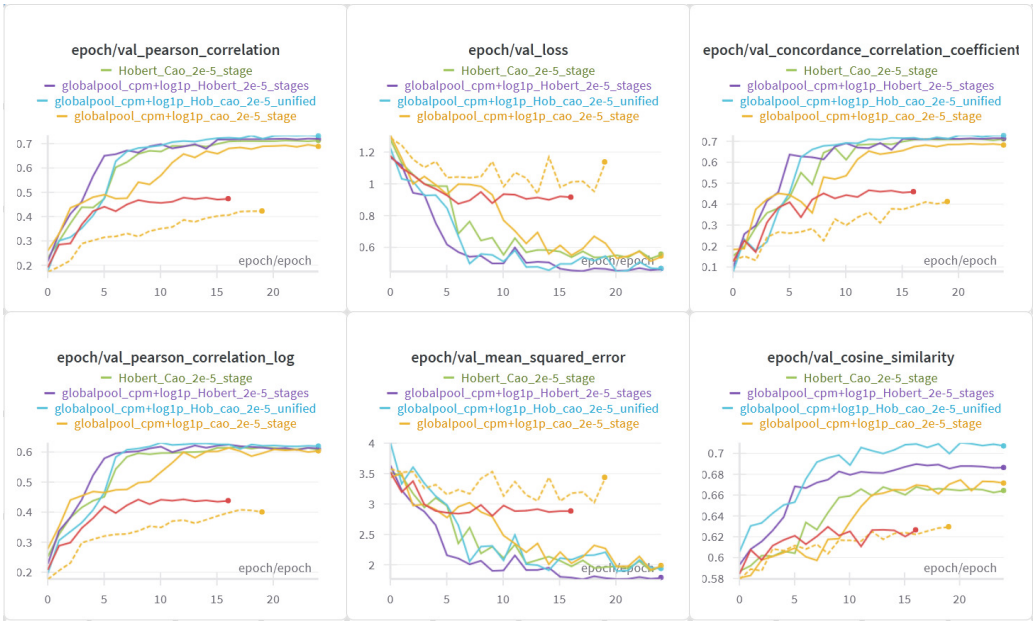


Figure 4.18: WandB metrics for the different model runs: **1.** Hobert_Cao_2e-5_stage, **2.** globalpool_cpm+log1p_Hobert_2e-5_stages, **3.** globalpool_cpm+log1p_Hobert_Cao_2e-5_unified, **4.** globalpool_cpm+log1p_Cao_2e-5_stage, **5.** globalpool_cpm+log1p_fine_tissue_5e-6, **6.** globalpool_cpm+log1p_fine_tissue_1e-5

Model 1 (Hobert_Cao_2e-5_stage) was trained using a learning rate of 2e-5, on the full integrated dataset using the stage-divided Fine Tissue annotation. This resulted in good perfor-

mance at epoch 24: Validation Pearson Correlation of 0.713, Validation Loss of 0.558, Validation Concordance Correlation Coefficient of 0.712, Validation Pearson Correlation Log of 0.608, Validation Mean Squared Error of 1.980, Validation Cosine Similarity of 0.664.

Model 2 (globalpool_cpm+log1p_Hobert_2e-5_stages) was trained using a learning rate of 2e-5, with global pooling and CPM+log1p normalization, on the Hobert dataset using the stage-divided annotation. This resulted in good performance at epoch 24: Validation Pearson Correlation of 0.720, Validation Loss of 0.466, Validation Concordance Correlation Coefficient of 0.714, Validation Pearson Correlation Log of 0.612, Validation Mean Squared Error of 1.790, Validation Cosine Similarity of 0.687.

Model 3 (globalpool_cpm+log1p_Hob_cao_2e-5_unified) was trained using a learning rate of 2e-5, with global pooling and CPM+log1p normalization, on the combined Hobert and Cao datasets using the unified Fine Tissue annotation. This resulted in good performance at epoch 24: Validation Pearson Correlation of 0.732, Validation Loss of 0.467, Validation Concordance Correlation Coefficient of 0.728, Validation Pearson Correlation Log of 0.620, Validation Mean Squared Error of 1.939, Validation Cosine Similarity of 0.707.

Model 4 (globalpool_cpm+log1p_cao_2e-5_stage) was trained using a learning rate of 2e-5, with global pooling and CPM+log1p normalization, on the Cao dataset using the stage-divided annotation. This resulted in good performance at epoch 24: Validation Pearson Correlation of 0.689, Validation Loss of 0.546, Validation Concordance Correlation Coefficient of 0.682, Validation Pearson Correlation Log of 0.604, Validation Mean Squared Error of 1.990, Validation Cosine Similarity of 0.672.

Model 5 (globalpool_cpm+log1p_fine_tissue_5e-6) was trained using a learning rate of 5e-6, with global pooling and CPM+log1p normalization, using the Fine Tissue annotation. This resulted in performance at epoch 19: Validation Pearson Correlation of 0.423, Validation Loss of 1.138, Validation Concordance Correlation Coefficient of 0.413, Validation Pearson Correlation Log of 0.401, Validation Mean Squared Error of 3.439, Validation Cosine Similarity of 0.630.

Model 6 (globalpool_cpm+log1p_fine_tissue_1e-5) was trained using a learning rate of 1e-5, with global pooling and CPM+log1p normalization, using the Fine Tissue annotation. This resulted in performance at epoch 16: Validation Pearson Correlation of 0.474, Validation Loss of 0.916, Validation Concordance Correlation Coefficient of 0.459, Validation Pearson Correlation Log of 0.438, Validation Mean Squared Error of 2.885, Validation Cosine Similarity of 0.627.

The six models can be categorized into three distinct experimental groups designed to test different hypotheses about model performance.

(Models 1 and 3) tested the impact of annotation strategies by training on the same full integrated dataset using two different labeling approaches: Model 1 used stage-divided Fine Tissue annotations, while Model 3 used unified Fine Tissue annotations.

(Models 2 and 4) were trained on subsets of the integrated data: Model 2 used only the Hobert dataset with stage-divided annotations, while Model 4 used only the Cao dataset with stage-divided annotations. This comparison was designed to assess whether individual dataset characteristics for the L2 Stage significantly influenced model performance.

(Models 5 and 6) represented early optimization experiments that tested learning rate sensitivity by training on the full integrated dataset with unified Fine Tissue annotations but different learning rates (5e-6 and 1e-5, respectively).

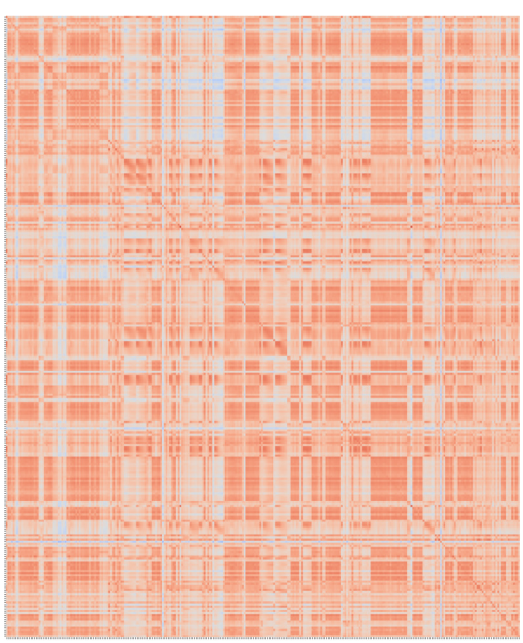
The results reveal several important findings. Models 1-4 achieved remarkably similar performance levels, with Pearson correlations ranging from 0.689 to 0.732, suggesting that neither dataset choice nor annotation strategy substantially affected model quality within this performance range. Specifically, the minimal difference between Models 1 and 3 (correlation difference of 0.019) indicates that both stage-divided and unified annotation approaches are equally effective for gene expression prediction. Similarly, the comparable performance between Models 2 and 4 demonstrates that individual dataset characteristics did not create significant performance differences, validating the robustness of our integration approach.

4.3.2 Performances of final models

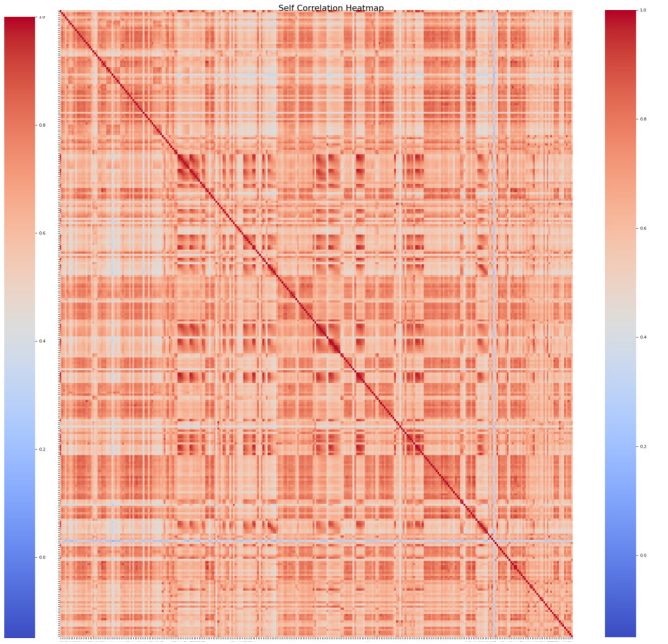
Our final models are then Models 1 and 3. From these models we started assessing the actual performances in prediction.

Regarding Model 3 (Fine Tissue type), as shown in **Figure 4.19**, the model successfully captured most of the data patterns and appears capable of correctly predicting the expression profiles of different tissues. Particularly, in **Figure 4.19c**, we can observe how the model, while not perfectly predicting the actual scalar values, reaches close approximations, frequently underestimating expression levels.

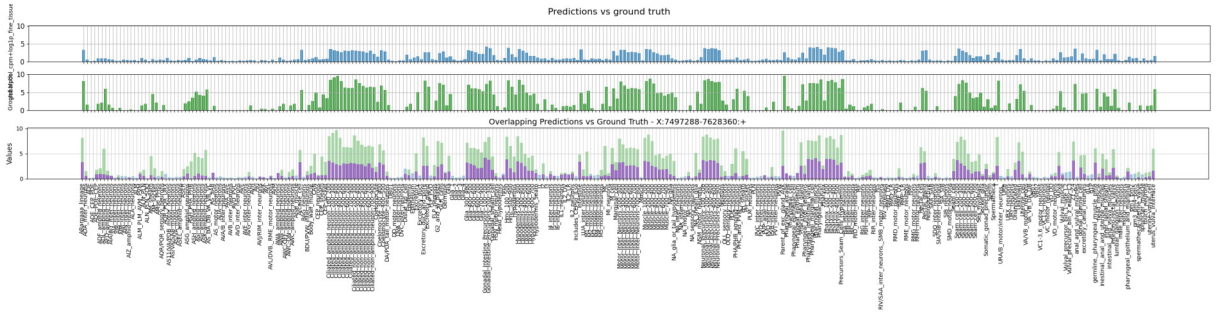
The model shows reduced accuracy primarily for highly specialized tissues such as Amphid Neurons and Interneurons.



(a) Correlations between predictions and ground-truth.



(b) Self-Correlations.



(c) Predictions for every tissue type and their relative ground-truth.

Figure 4.19: Performances relative to Model 3 (Fine Tissue annotation) **A.** Heatmap of correlations between predictions and ground-truths, **B.** Heatmap of Self-Correlations, **C.** (Top track) Predictions for every Fine Tissue annotation, (Middle track) Ground-Truths for every Fine Tissue annotation, (Bottom track) Overlay of the Top and Middle tracks (Blue: Predictions higher then Ground-Truths, Green: Ground-Truths higher than Predictions, Purple: Common area between the two tracks.). Region: X:7497288-7628360:+.

As shown in **Figure 4.19**, more general tissues exhibit higher correlations between predictions and ground truth, such as "Intestine" and "Body_Wall_Muscle". In contrast, highly specialized tissues like "AWC-OFF_amphid_neurons" show lower correlations.

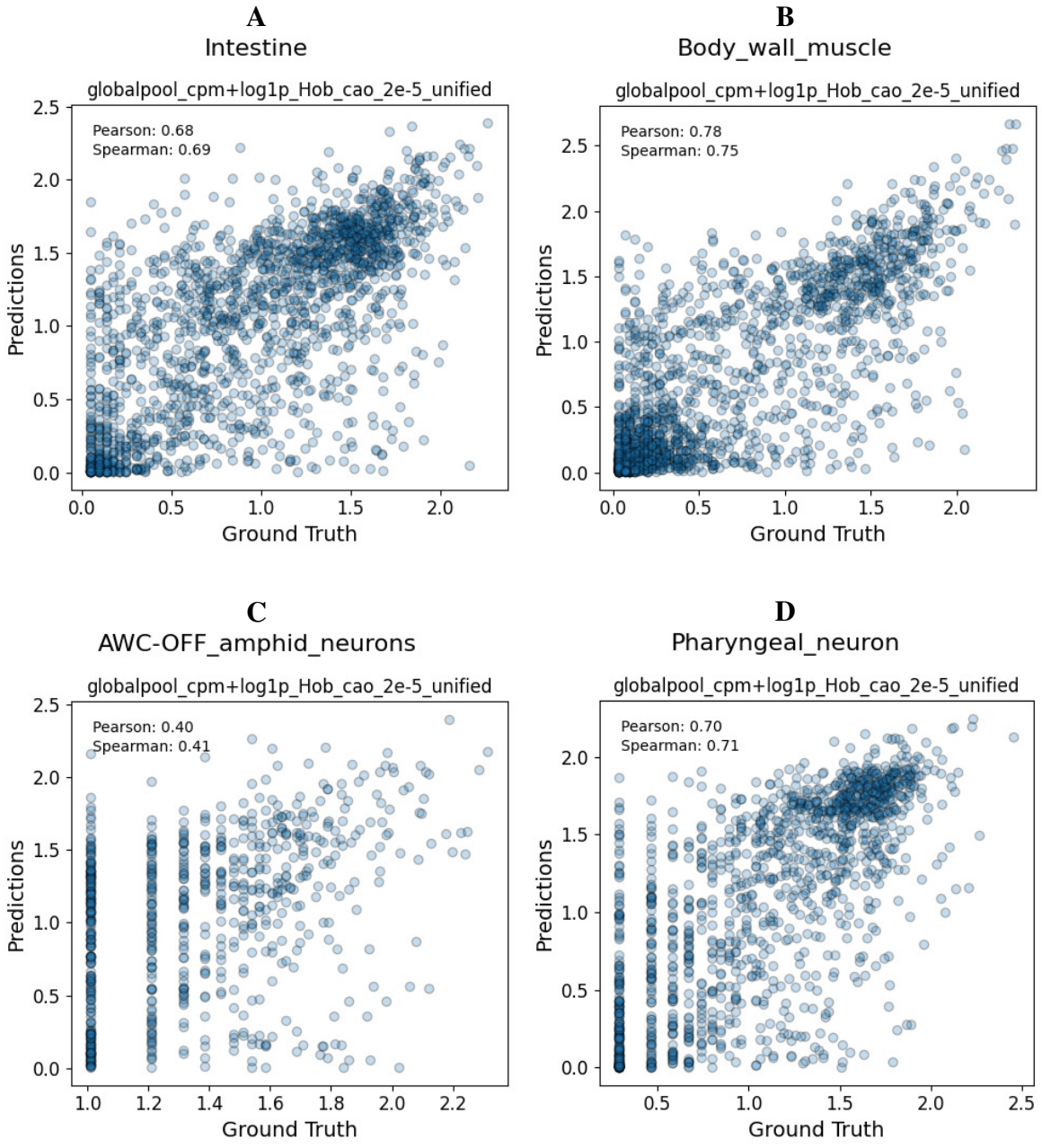
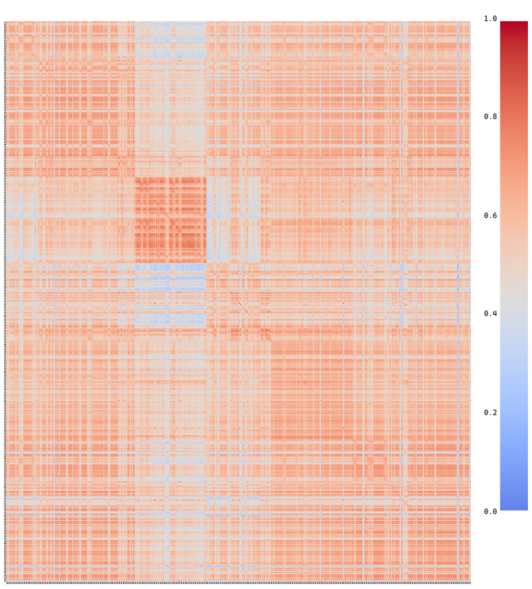


Figure 4.20: **A.** Scatter plot between predictions and ground-truths for every region in the Intestine class, **B.** Scatter plot between predictions and ground-truths for every region in the Body Wall Muscle class, **C.** Scatter plot between predictions and ground-truths for every region in the AWC-OFF amphid neurons class, **D.** Scatter plot between predictions and ground-truths for every region in the Pharyngeal Neuron class.

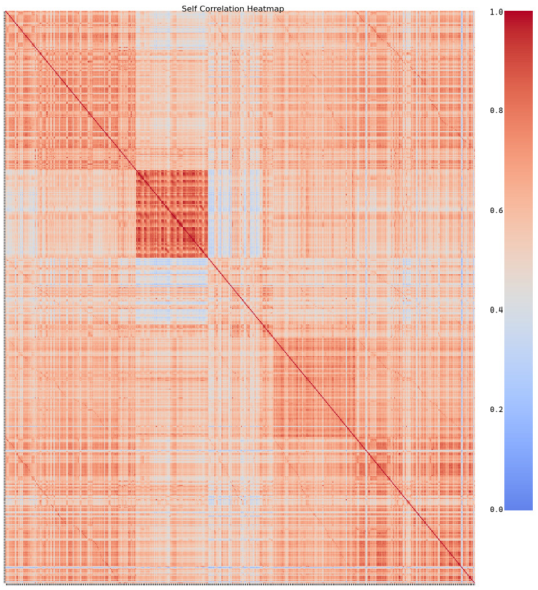
In **Subfigures 4.20c and 4.20d**, we can also observe how specialized cell types exhibit a striping pattern for lowly expressed genes.

For the Stage-Divided Fine Tissue annotation model (Model 1), as shown in **Figure 4.19**, the

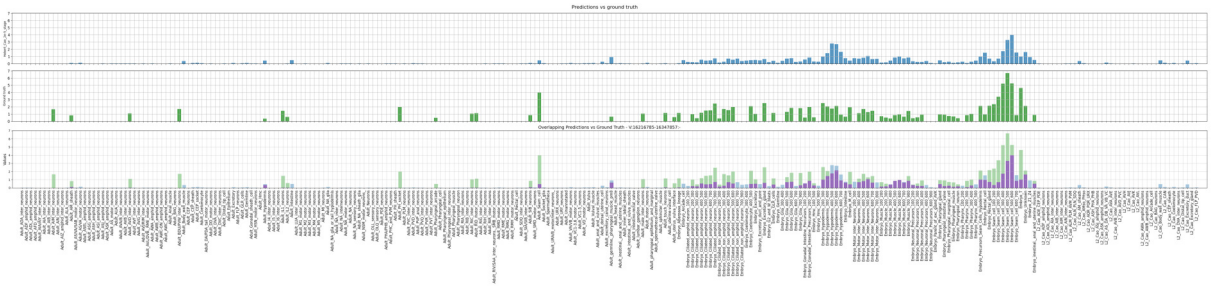
model successfully captured most of the data patterns and appears capable of correctly predicting the expression profiles of different tissues, similar to the Fine Tissue type model. Particularly, in **Figure 4.19c**, we can observe how the model, while not perfectly predicting the actual scalar values, reaches close approximations, frequently underestimating expression levels in this case as well.



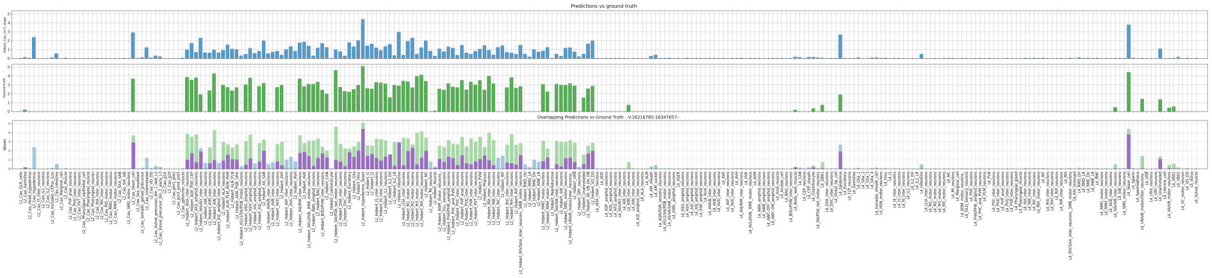
(a) Correlations between predictions and ground-truth.



(b) Self-Correlations.



(c) Predictions for the first 262 tissue types and their relative ground-truth.



(d) Predictions for the last 262 tissue types and their relative ground-truth.

Figure 4.21: Performances relative to Model 1 (Stage-Divided Fine Tissue annotation) **A.** Heatmap of correlations between predictions and ground-truths, **B.** Heatmap of Self-Correlations, **C-D.** (Top track) Predictions for every class, (Middle track) Ground-Truths for every class, (Bottom track) Overlay of the Top and Middle tracks (Blue: Predictions higher then Ground-Truths, Green: Ground-Truths higher than Predictions, Purple: Common area between the two tracks.). Region: V:19264027-19395099:+.

As shown in **Figure 4.22**, using this model we can, for istance, examine specific timeframes in embryonic cell development and compare them with later developmental stages. In these particular plots, we also notice how the correlation declines with cell specialization.

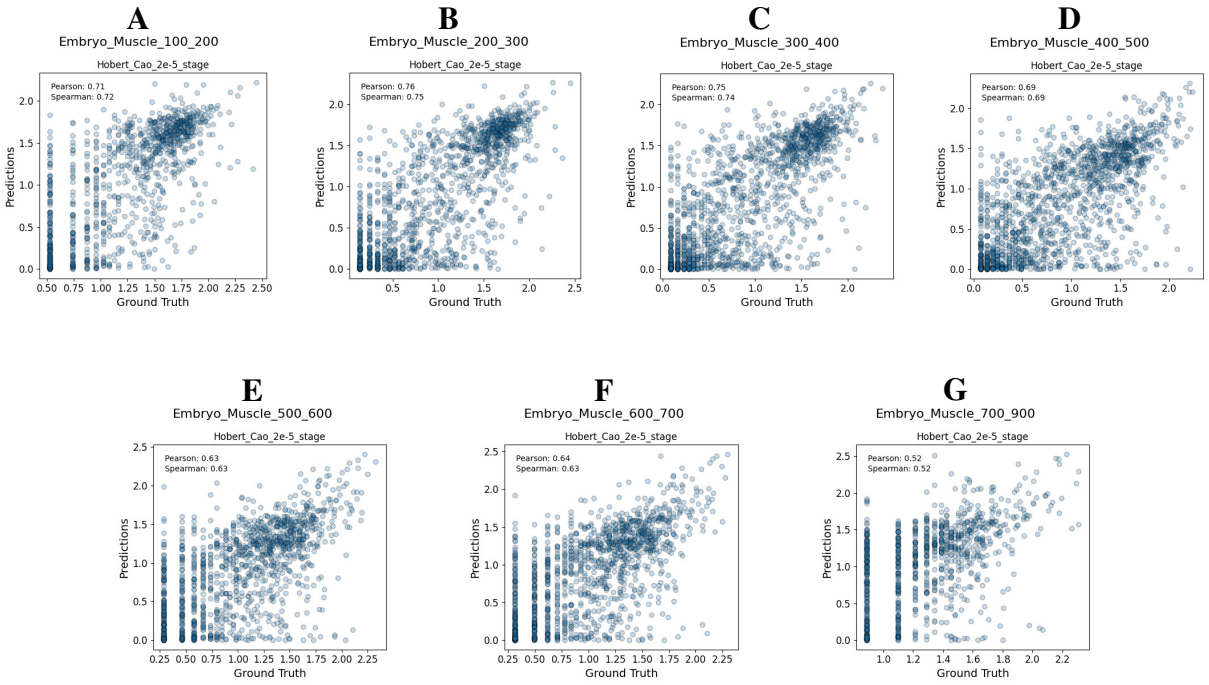


Figure 4.22: Scatter plots between predictions and ground-truths for each stage. **A.** Embryo Muscle (100–200), **B.** Embryo Muscle (200–300), **C.** Embryo Muscle (300–400), **D.** Embryo Muscle (400–500), **E.** Embryo Muscle (500–600), **F.** Embryo Muscle (600–700), **G.** Embryo Muscle (700–900).

When looking at regions individually, as shown in **Figure 4.23**, we can assess whether the models are predicting meaningful motifs. In this case we examined a region close to **vit-2**, a gene fundamental to vitellogenesis in intestinal tissues, and analyzed its contribution scores for the two models.

The models correctly identify meaningful motifs such as **TGATAA** and its reverse complement **TTATCA**, which are members of the elt family of transcription factors. The elt family is known to play crucial roles in intestinal development and function in C. elegans, making their identification in the vit-2 regulatory region biologically relevant for vitellogenesis regulation.

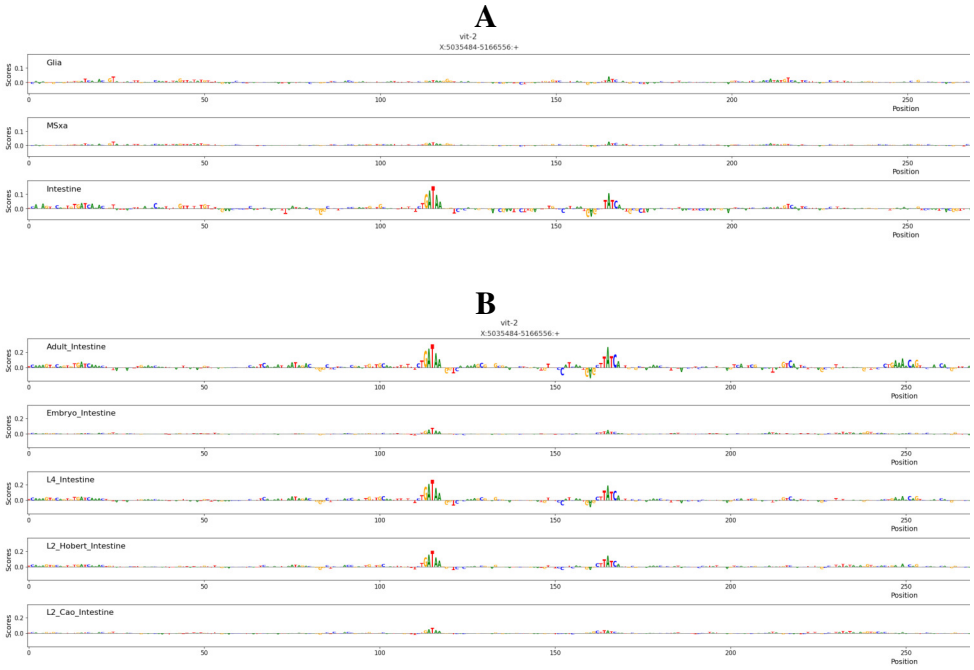


Figure 4.23: Visualization of contribution scores for a window of 260 bp. **A.** Contribution scores for **vit-2** given by the unified annotations model, B. Contribution scores for **vit-2** given by the stage-divided annotations model.

We also observe that not all developmental stages show strong contributions from these motifs, particularly the embryonic and L2 stages from the Cao dataset. This stage-specific pattern suggests that vit-2 regulation may be developmentally controlled, with elt-mediated transcription becoming more prominent in later larval and adult stages when vitellogenesis is most active.

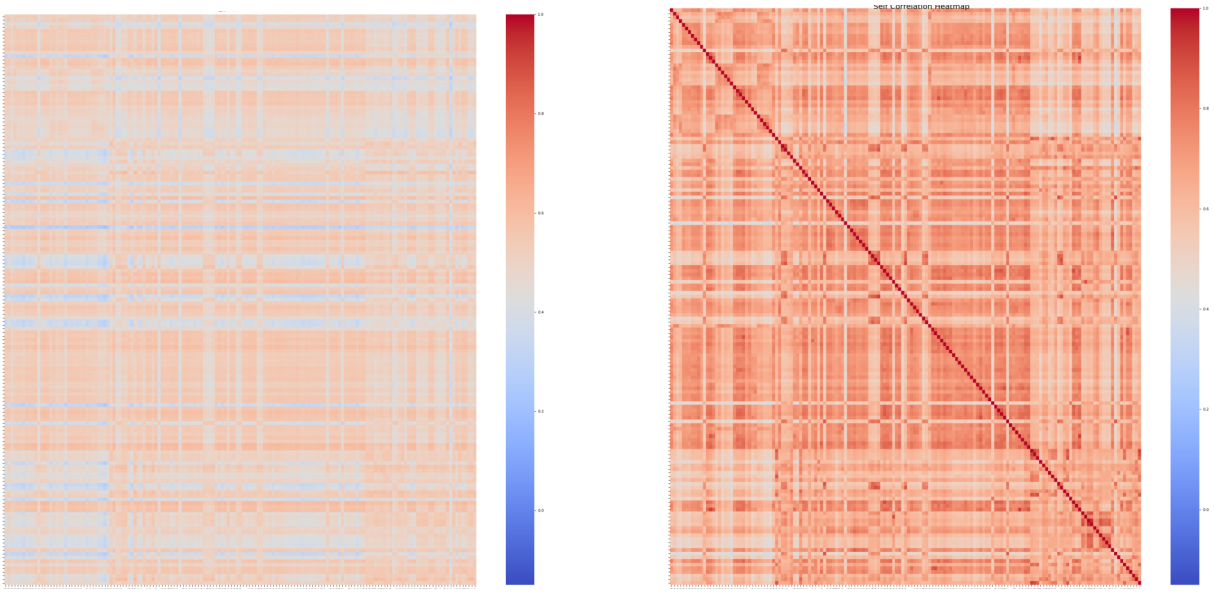
4.3.3 Fine-Tuning of single datasets

We also performed training on individual datasets to assess whether we would achieve the same performance as the integrated dataset. In **Figure 4.24**, we show the Weights & Biases metrics of these trained models.

We observe that datasets like L2 (Hobert) and Adult had lower validation Pearson correlations of **0.45** and **0.54**, respectively. Meanwhile, training on the embryo dataset stopped early at 6 epochs with a correlation of **0.58**.

Figure 4.24: Training of single datasets compared to the 2 final models (light green and cyan). Adult (purple), L2 (red), Embryo (teal).

As an example we show in **Figure 4.25**, the Correlation Heatmaps of the model trained on the single Adult dataset. Here the diagonal pattern from the self-correlation heatmap is almost completly absent in the correlation heatmap, indicating that the model learned little meaningful representations.



(a) Correlations between predictions and ground-truth.

(b) Self-Correlations.

Figure 4.25: Correlation heatmaps of a model fine-tuned on the single adult dataset.

4.4 Analysis of predictions

Given the two trained models, we can now inspect the regulatory patterns within different tissues and identify the motifs and transcriptional elements that contribute most to gene expression. We present here 2 broad tissues divided by stage and the analysis of embryonic muscle development.

4.4.1 Tissue-Specific Motif Contributions

To understand the regulatory mechanisms underlying our model's predictions, we performed motif contribution analysis across different tissue types mainly with the Stage-Divided annotations model.

Intestine Cell Type

In **Figure 4.26**, we present the motif analysis results for intestinal tissue, which reveals distinct patterns of transcriptional factor binding sites that contribute to tissue-specific gene expression.

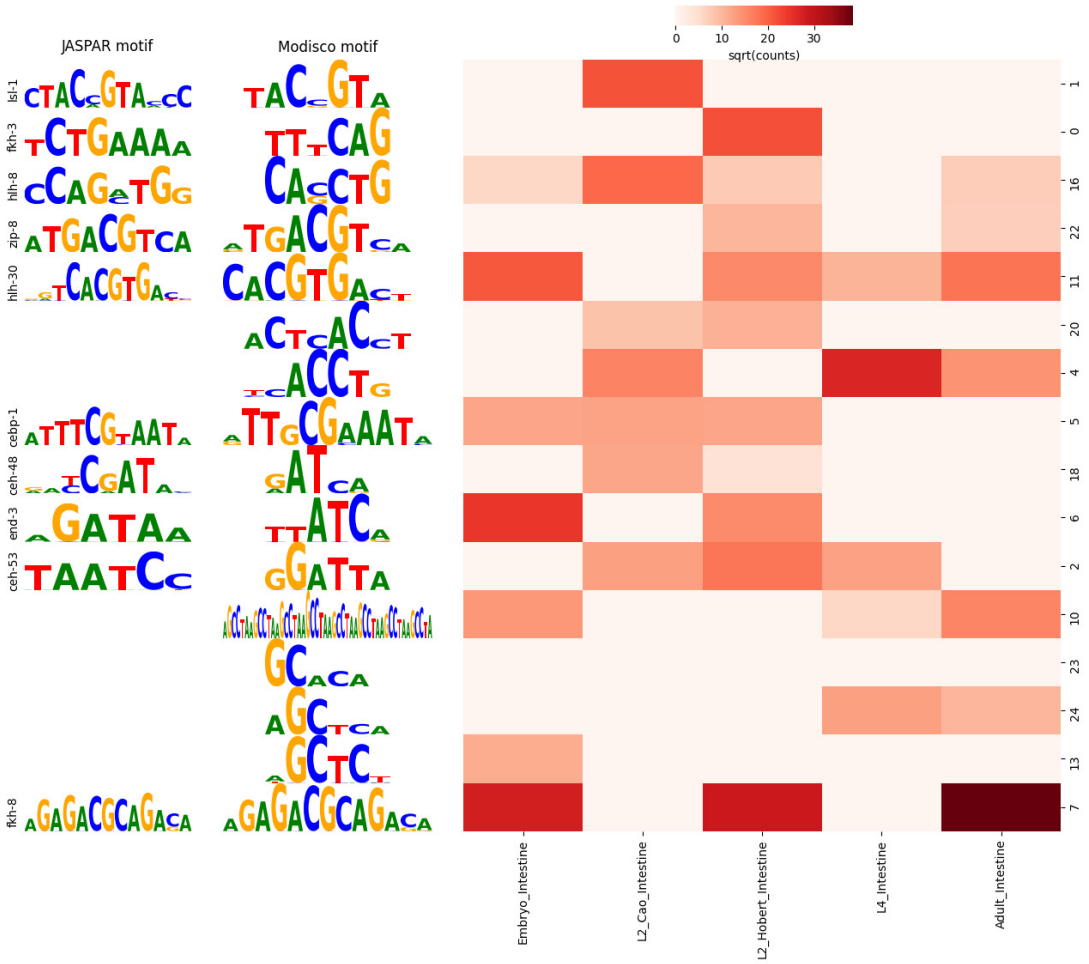


Figure 4.26: Top motifs contributing to intestine-specific gene expression matched on the JASPAR database. Columns from left to right: Embryonic stage, L2 stage (Cao dataset), L2 stage (Hobert dataset), L4 stage, and Adult stage.

The analysis identified several highly significant motifs with strong contributions to intestinal gene expression prediction. The motif logo representations display the sequence conservation and binding preferences of these regulatory elements, with larger letters indicating higher information content at specific positions. The accompanying contribution plots demonstrate the quantitative impact of each motif on the model's predictions, providing insights into which transcriptional regulators are most influential for tissue-specific expression patterns.

Notable motifs include **GATA** like motifs greatly expressed in intestines [58], **CACCTG/-CACGTG** which is *hlh-8*, a member of the bHLH family [59], **TTTCAG/CTGAAA** which is reported by JASPAR to be *fkh-3*. These motifs showed consistent enrichment patterns and high contribution scores, suggesting an important role in intestinal development and function. We also see other morifs related to homeodomain domains of the ceh family.

We also notice how some motifs seem to be present only in early stages (until L2 stage).

Hypodermis Cell Type

In **Figure 4.27**, we present the motif analysis results for hypodermis tissue.

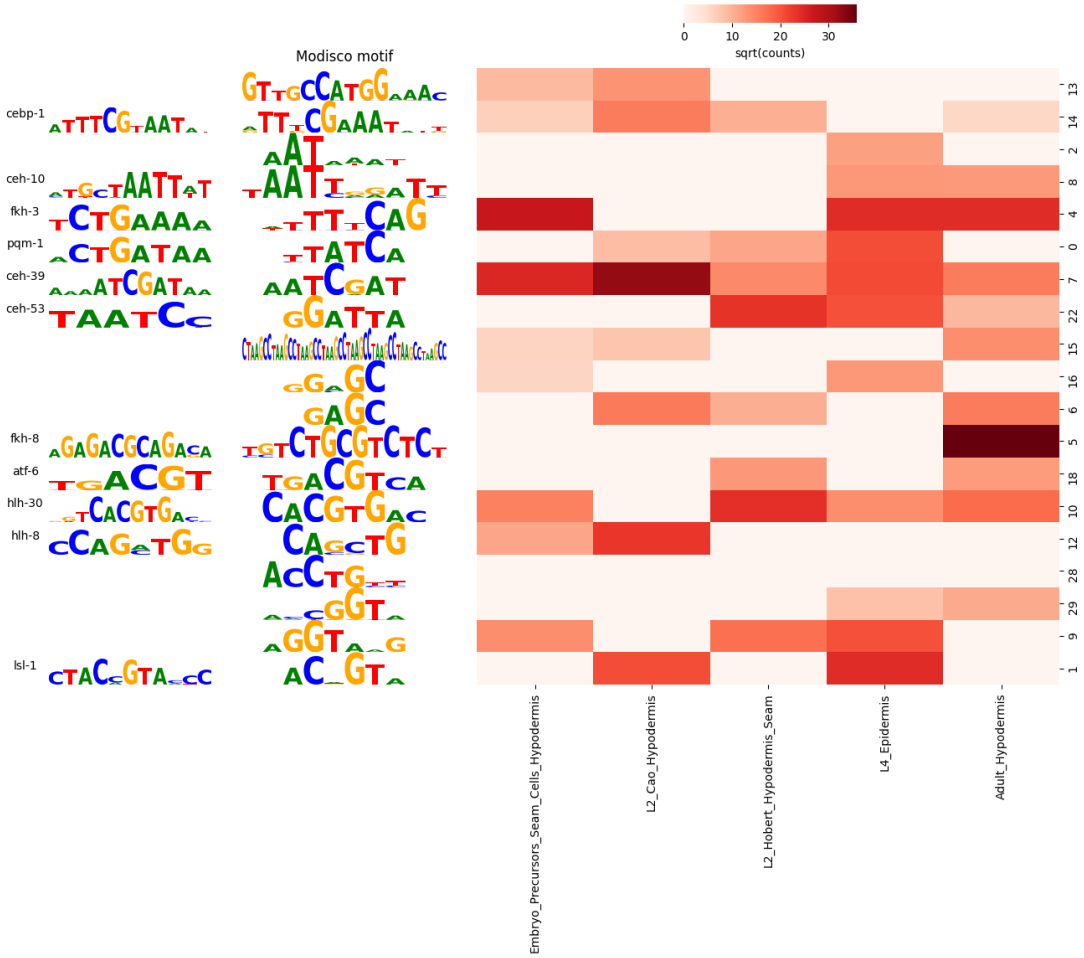


Figure 4.27: Top motifs contributing to hypodermis-specific gene expression, matched to the JASPAR database. Columns from left to right: Embryonic stage, L2 stage (Cao dataset), L2 stage (Hobert dataset), L4 stage, and Adult stage.

The analysis identified several highly significant motifs with strong positive contributions to hypodermal gene expression prediction. Notable motifs found by the model and matching JASPAR database include *pqm-1*, a transcription factor from the GATA family that plays a key role in C. elegans intestinal mTORC2 signaling and fat transport regulation in the hypodermis, that represses vitellogenesis [60]. Other motifs where from the hlh family [61] and homeodomain motifs.

4.4.2 Embryonic Muscle Cell Development Motif Contributions

In this section we present how different motifs behave during the embryonal development of muscle cells trough different timepoints.

In **Figure 4.28**, we show motifs that where found to be most contributing positively to muscle-specific gene expression.

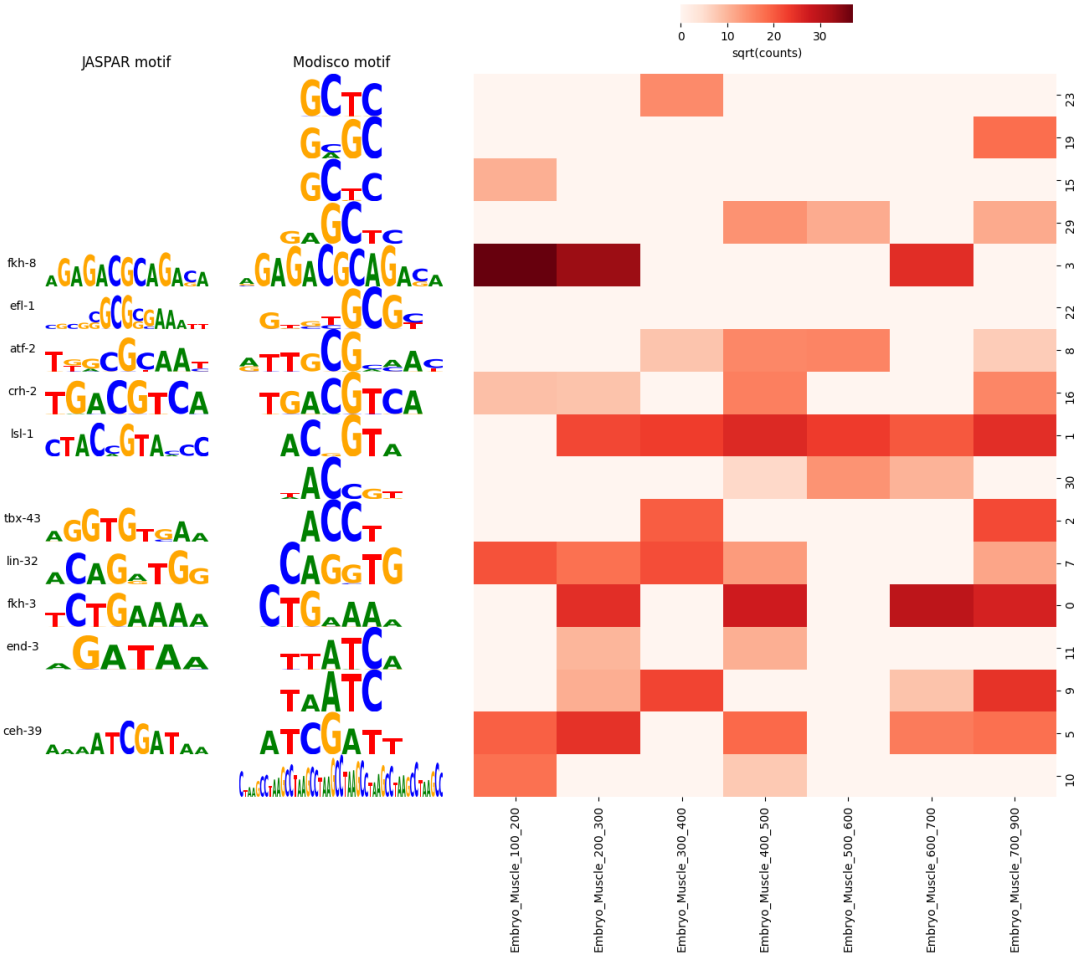


Figure 4.28: Top motifs contributing positively to embryonic muscle-specific gene expression, matched to the JASPAR database. Columns from left to right: Embryonic stage, L2 stage (Cao dataset), L2 stage (Hobert dataset), L4 stage, and Adult stage.

The analysis identified several highly significant motifs with strong positive contributions to embryonic muscle gene expression prediction. Notable motifs found by the model and matching entries in the JASPAR database included a *tbx-43*–like T-box motif. Members of the T-box transcription factor family are well established as key regulators of mesoderm specification and early muscle lineage commitment in C. elegans [62]. However the *tbx-43* motif specifically

remains poorly documented.

We also detected motifs from the basic helix–loop–helix (bHLH) family, which in C. elegans includes the master myogenic regulator *hlh-1* (MyoD homolog) [61]. We detect also forkhead motifs and homeodomain motifs.

Chapter 5

Discussion

This work represents the first comprehensive attempt to predict gene expression directly from DNA sequence across an entire multicellular organism's life cycle. By integrating single-cell RNA sequencing data from four major developmental stages of C. elegans, we have demonstrated that it is possible to build predictive models that capture the fundamental principles of gene regulation at organism scale.

5.1 scATAC Modeling and Regions analysis

We first began by analyzing regions that were most important for predicting accessibility of CREs in the L2 larval stage. Here, we trained a model using CREsted [55] to predict scATAC peaks and analyze the relative performance. What we found was that the original regions in the dataset overlapped with many exonic regions, and that these regions contributed very little to the model (as shown in **Figure 4.4**), indicating that most of the ATAC peaks are not located in coding sequences. This is consistent with findings that C. elegans has minimal regulatory activity in exonic regions, as described in papers such as Niu et al. [57].

However, even when focusing on non-exonic regions only, the model did not improve significantly in predictions, indicating that the presence of exonic regions did not substantially impair overall model performance. For instance, the validation Pearson correlation of the full model was **0.84**, for exonic regions it was **0.78**, and for non-exonic regions it was **0.86**.

We then further visualized how the model captured this difference by extracting and plotting the region embeddings and coloring them by whether the region was exonic or non-exonic. This analysis revealed that the model successfully learned to distinguish between these two region

types at the embedding level.

All of this provided us with a validated set of regions to serve as reference for the other analyses we performed during scRNA fine-tuning and integration of the dataset.

5.2 Developmental Stages Integration

We then moved on to create an integrated dataset containing four developmental time points of C. elegans (Embryonic, L2, L4, Adult) so that we could use the temporal dynamics of the dataset to improve model performance. In the process, we ensured that the dataset was maximally compatible and easily accessible, with annotations that accurately described and grouped the cell types. The dataset achieved good levels of integration, as demonstrated by the clustering patterns observed in **Figure 4.17d** and **Figure 4.17c**, which show how the five datasets are well integrated with separations likely given by temporal differences and clear tissue type distinctions, respectively.

During the process of creating clear labels, we also discovered how embryonic cells predominantly followed developmental trajectories to clusters of later developmental stages (**Figure 4.16c** and **4.16d**), allowing us to accurately annotate the different cell populations and organize them into different developmental time bins. This trajectory analysis not only validated our integration approach but also provided biological insight into the development progression of C. elegans, confirming that our computational framework captured meaningful developmental relationships.

The successful integration of datasets from different studies and developmental time points represents a significant technical achievement, as batch effects and methodological differences between studies often pose substantial challenges for multi-dataset analysis. Our approach demonstrates the feasibility of using temporal information in developmental stages to enhance predictive modeling in C. elegans genomics.

5.3 scRNA Fine-Tuning

We then used this dataset to fine-tune a hybrid CNN-Transformer model to predict gene expression scalars. The training resulted in two successful models: one with annotations not divided by developmental stage and another with stage-divided annotations. These two models helped us understand which cell types were aided more by the presence of other developmental time points and which were not. In fact, we observed generally higher correlations in the first model

versus the second for the same tissues.

The predictions of both models proved to be reasonably effective at predicting gene expression values for the tissues, though they consistently underestimated expression values by a certain margin.

What we also found for these models is that comparing them with models fine-tuned only on single stages revealed that the integrated approach achieved better performances, as the single-stage models showed poor performance. While this result doesn't prove by itself that including multiple developmental stages in the datasets contributed to improved performance, we did not have time to further analyze the underlying reasons for this improvement.

5.4 Analyzing CREs

We were then able to identify through TF-MoDISco most of the motifs discovered in JASPAR for several tissue types. However, we did not observe strong differences in motif representations across the different developmental stages.

Our motif contribution analysis successfully identified key transcriptional regulators known to control tissue-specific gene expression in C. elegans. For intestinal tissues, we detected GATA-like motifs, which align with the well-established role of GATA factors in intestinal development and function [58]. The identification of *hlh-8* motifs (**CACCTG/CACGTG**) and *fkh-3* motifs (**TTTCAG/CTGAAA**) further validates our model's ability to capture biologically relevant regulatory patterns, as these factors are known intestinal regulators [59]. Similarly, in hypodermis tissue, the detection of *pqm-1* motifs confirms the model's sensitivity to tissue-appropriate transcriptional programs, given *pqm-1*'s established role in hypodermal metabolism and vitellogenesis regulation.

The muscle-specific analysis revealed T-box and bHLH motifs, consistent with the fundamental roles of these transcription factor families in mesoderm specification and myogenesis. The detection of tbx-43-like motifs, despite limited functional characterization of this specific factor, suggests our approach may identify potentially important but understudied regulators. The presence of *hlh-1*-related motifs aligns with the central role of this **MyoD** homolog in C. elegans muscle development [61].

Despite these tissue-specific successes, the lack of substantial motif representation changes across developmental stages was notable. This finding could reflect several factors: the core transcriptional machinery governing tissue identity may remain relatively stable across devel-

opment, with temporal regulation occurring through quantitative rather than qualitative changes in transcription factor activity. Alternatively, our analysis may have been limited by the resolution of our developmental time points or by integrating multiple stages in our training data, which could have obscured stage-specific regulatory signatures. Further investigation with stage-specific models would be needed to determine whether more subtle temporal regulatory dynamics exist.

5.5 Future Directions

In this thesis we managed to provide a comprehensive analysis of C. elegans gene regulation during most of its life cycle. However, the analysis could be extended in several meaningful directions.

Integration of C. briggsae, C. tropicalis

The regulatory principles learned from C. elegans could be validated and extended to other model organisms. Integration of other species like C. briggsae and C. tropicalis could help improve the model as it would give different temporal and species variation to learn meaningful grammar regulating the gene expression of these related species. Cross-species analysis would also enable identification of conserved regulatory motifs and transcriptional programs across nematodes, while highlighting species-specific adaptations in gene regulation.

Integration of L1 and L3 stages

Including the missing L1 and L3 developmental stages would provide a more complete temporal resolution of C. elegans development. Adding these time points could reveal stage-specific regulatory transitions that were missed in our current analysis and improve model performance through more comprehensive temporal coverage.

Adult day-by-day analysis

Adult C. elegans undergo significant physiological changes during aging, including reproductive senescence, metabolic shifts, and cellular deterioration. In this dataset we chose not to divide the adult cell types by timepoint and instead analyzed them in bulk.

A future direction could be to further subdivide these adult cell types by age (similar to our approach with embryonic stages), which would enable capture of fine-grained temporal dynamics during aging and identification of age-specific regulatory transitions within individual tissues.

Bibliography

[1] Kevin J. Mitchell and Nick Cheney. The Genomic Code: the genome instantiates a generative model of the organism. *Trends in Genetics*, 41(6):462–479, June 2025. Publisher: Elsevier.

[2] Peter Robin Hiesinger. *The self-assembling brain: How neural networks grow smarter*. Princeton University Press, Princeton, NJ, 2021.

[3] Hannah K. Long, Sara L. Prescott, and Joanna Wysocka. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell*, 167(5):1170–1187, November 2016.

[4] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C. J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A. L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Eric D. Green, Peter J. Good, Elise A. Feingold, Bradley E. Bernstein, Ewan Birney, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Mark Gerstein, Morgan C. Giddings, Thomas R. Gingeras, Eric D. Green, Roderic Guigó, Ross C. Hardison, Timothy J. Hubbard, Manolis

Kellis, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, Michael Snyder, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Morgan C. Giddings, Bradley E. Bernstein, Charles B. Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D. Ward, Robert C. Altshuler, Matthew L. Eaton, Manolis Kellis, Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P. Gunawardena, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Brian A. Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J. Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, Thomas R. Gingeras, Kate R. Rosenbloom, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, W. James Kent, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Terrence S. Furey, Lingyun Song, Linda L. Grasfeder, Paul G. Giresi, Bum-Kyu Lee, Anna Battenhouse, Nathan C. Sheffield, Jeremy M. Simon, Kimberly A. Showers, Alexias Safi, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Ewan Birney, Vishwanath R. Iyer, Jason D. Lieb, Gregory E. Crawford, Guoliang Li, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Oscar J. Luo, Atif Shahab, Melissa J. Fullwood, Xiaoan Ruan, Yijun Ruan, Richard M. Myers, Florencia Pauli, Brian A. Williams, Jason Gertz, Georgi K. Mari-

nov, Timothy E. Reddy, Jost Vielmetter, E. Partridge, Diane Trout, Katherine E. Varley, Clarke Gasper, The ENCODE Project Consortium, Overall coordination (data analysis coordination), Data production leads (data production), Lead analysts (data analysis), Writing group, NHGRI project management (scientific management), Principal investigators (steering committee), Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis), Broad Institute Group (data production and analysis), Barcelona RIKEN Sanger Institute University of Lausanne Genome Institute of Singapore group (data production and analysis) University of Geneva Cold Spring Harbor, Center for Genomic Regulation, Data coordination center at UC Santa Cruz (production data coordination), Austin University of North Carolina-Chapel Hill group (data production and analysis) EBI Duke University, University of Texas, Genome Institute of Singapore group (data production and analysis), and Stanford group (data production and analysis)-Caltech HudsonAlpha Institute, UC Irvine. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. Publisher: Nature Publishing Group.

[5] Boel Brynedal, JinMyung Choi, Towfique Raj, Robert Bjornson, Barbara E. Stranger, Benjamin M. Neale, Benjamin F. Voight, and Chris Cotsapas. Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *The American Journal of Human Genetics*, 100(4):581–591, April 2017. Publisher: Elsevier.

[6] Matthew V. Rockman and Leonid Kruglyak. Genetics of global gene expression. *Nature Reviews Genetics*, 7(11):862–872, November 2006. Publisher: Nature Publishing Group.

[7] P. J. Wittkopp. Genomic sources of regulatory variation in cis and in trans. *Cellular and Molecular Life Sciences: CMLS*, 62(16):1779–1783, June 2005.

[8] François Spitz and Eileen E. M. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, September 2012. Publisher: Nature Publishing Group.

[9] Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, April 2019. Publisher: Nature Publishing Group.

[10] Tamar Juven-Gershon and James T. Kadonaga. Regulation of Gene Expression via the Core Promoter and the Basal Transcriptional Machinery. *Developmental biology*, 339(2):225–229, March 2010.

[11] Job Dekker and Tom Misteli. Long-Range Chromatin Interactions. *Cold Spring Harbor Perspectives in Biology*, 7(10):a019356, October 2015.

[12] Miklos Gaszner and Gary Felsenfeld. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics*, 7(9):703–713, September 2006. Publisher: Nature Publishing Group.

[13] Gary D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, January 2000.

[14] Aimée M. Deaton and Adrian Bird. CpG islands and the regulation of transcription. *Genes & Development*, 25(10):1010–1022, May 2011.

[15] Eran Segal and Jonathan Widom. Poly(dA:dT) Tracts: Major Determinants of Nucleosome Organization. *Current opinion in structural biology*, 19(1):65–71, February 2009.

[16] Robert Hänsel-Hertsch, Marco Di Antonio, and Shankar Balasubramanian. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nature Reviews Molecular Cell Biology*, 18(5):279–284, May 2017. Publisher: Nature Publishing Group.

[17] Len A. Pennacchio, Nadav Ahituv, Alan M. Moses, Shyam Prabhakar, Marcelo A. Nobrega, Malak Shoukry, Simon Minovitsky, Inna Dubchak, Amy Holt, Keith D. Lewis, Ingrid Plajzer-Frick, Jennifer Akiyama, Sarah De Val, Veena Afzal, Brian L. Black, Olivier Couronne, Michael B. Eisen, Axel Visel, and Edward M. Rubin. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, November 2006. Publisher: Nature Publishing Group.

[18] S. Brenner. The Genetics of CAENORHABDITIS ELEGANS. *Genetics*, 77(1):71–94, May 1974.

[19] Ann K. Corsi, Bruce Wightman, and Martin Chalfie. A Transparent Window into Biology: A Primer on Caenorhabditis elegans. *Genetics*, 200(2):387–407, June 2015.

[20] Martin Chalfie. The worm revealed. *Nature*, 396(6712):620–621, December 1998. Publisher: Nature Publishing Group.

[21] Kaling Danggen and Varsha Singh. Sydney Brenner: The Tamer of an Elegant Worm. *Resonance*, 24(10):1061–1069, October 2019.

[22] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 314(1165):1–340, November 1986.

[23] Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, August 2017. Publisher: American Association for the Advancement of Science.

[24] Jonathan S. Packer, Qin Zhu, Chau Huynh, Priya Sivaramakrishnan, Elicia Preston, Hannah Dueck, Derek Stefanik, Kai Tan, Cole Trapnell, Junhyong Kim, Robert H. Waterston, and John I. Murray. A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science (New York, N.Y.)*, 365(6459):eaax1971, September 2019.

[25] Valentine Svensson, Roser Vento-Tormo, and Sarah A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, April 2018. Publisher: Nature Publishing Group.

[26] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, and Nir Yosef. The Human Cell Atlas. *eLife*, 6:e27041, 2017.

[27] Jonathan A Griffiths, Antonio Scialdone, and John C Marioni. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*, 14(4):e8046, April 2018.

[28] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019.

[29] Páll Melsted, Vasilis Ntranos, and Lior Pachter. The barcode, UMI, set format and BUStools. *Bioinformatics*, 35(21):4472–4473, November 2019.

[30] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature Genetics*, 51(1):12–18, January 2019. Publisher: Nature Publishing Group.

[31] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, July 2019. Publisher: Nature Publishing Group.

[32] Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, August 2015. Publisher: Nature Publishing Group.

[33] Amlan Talukder, Clayton Barham, Xiaoman Li, and Haiyan Hu. Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, 22(3):bbaa177, August 2020.

[34] David R. Kelley, Jasper Snoek, and John L. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, January 2016. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[35] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, October 2015. Publisher: Nature Publishing Group.

[36] David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, May 2018.

[37] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, October 2021. Publisher: Nature Publishing Group.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs].

[39] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, February 2021.

[40] Michael A. Beer and Saeed Tavazoie. Predicting Gene Expression from Sequence. *Cell*, 117(2):185–198, April 2004. Publisher: Elsevier.

[41] Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, August 2018. Publisher: Nature Publishing Group.

[42] Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model, July 2025. ISSN: 2692-8205 Pages: 2025.06.25.661532 Section: New Results.

[43] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences, October 2019. arXiv:1704.02685 [cs].

[44] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. arXiv:1312.6034 [cs].

[45] Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5, April 2020. arXiv:1811.00416 [cs].

[46] Avantika Lal, Alexander Karollus, Laura Gunsalus, David Garfield, Surag Nair, Alex M. Tseng, M. Grace Gordon, John Blischak, Bryce van de Geijn, Tushar Bhangale, Jenna L. Collier, Nathaniel Diamant, Tommaso Biancalani, Hector Corrada Bravo, Gabriele Scalia, and Gokcen Eraslan. Decoding sequence determinants of gene expression in diverse cellular and disease states, April 2025. Pages: 2024.10.09.617507 Section: New Results.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs].

[48] Christoph Alt, Marc Hübner, and Leonhard Hennig. Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction, June 2019. arXiv:1906.08646 [cs].

[49] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training.

[50] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics*, 57(4):949–961, April 2025. Publisher: Nature Publishing Group.

[51] Timothy J. Durham, Riza M. Daza, Louis Gevirtzman, Darren A. Cusanovich, Olubusayo Bolonduro, William Stafford Noble, Jay Shendure, and Robert H. Waterston. Comprehensive characterization of tissue-specific chromatin accessibility in L2 Caenorhabditis elegans nematodes. *Genome Research*, 31(10):1952–1969, October 2021.

[52] Itai Antoine Toker, Lidia Ripoll-Sánchez, Luke T. Geiger, Karan S. Saini, Isabel Beets, Petra E. Vértes, William R. Schafer, Eyal Ben-David, and Oliver Hobert. Molecular patterns of evolutionary changes throughout the whole nervous system of multiple nematode species, November 2024. Pages: 2024.11.23.624988 Section: New Results.

[53] Seth R. Taylor, Gabriel Santpere, Alexis Weinreb, Alec Barrett, Molly B. Reilly, Chuan Xu, Erdem Varol, Panos Oikonomou, Lori Glenwinkel, Rebecca McWhirter, Abigail Poff, Manasa Basavaraju, Ibnul Rafi, Eviatar Yemini, Steven J. Cook, Alexander Abrams, Berta Vidal, Cyril Cros, Saeed Tavazoie, Nenad Sestan, Marc Hammarlund, Oliver Hobert, and David M. Miller. Molecular topography of an entire nervous system. *Cell*, 184(16):4329–4347.e23, August 2021. Publisher: Elsevier.

[54] Antoine Emile Roux, Han Yuan, Katie Podshivalova, David Hendrickson, Rex Kerr, Cynthia Kenyon, and David Kelley. Individual cell types in C. elegans age differently and activate distinct cell-protective responses. *Cell Reports*, 42(8), August 2023. Publisher: Elsevier.

[55] Niklas Kempynck, Seppe De Winter, Casper H. Blaauw, Vasileios Konstantakos, Sam Dieltiens, Eren Can Ekşi, Valérie Bercier, Ibrahim I. Taskiran, Gert Hulselmans, Katina Spanier, Valerie Christiaens, Ludo Van Den Bosch, Lukas Mahieu, and Stein Aerts. CREsted: modeling genomic and synthetic cell type-specific enhancers across tissues and species, April 2025. Pages: 2025.04.02.646812 Section: New Results.

[56] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. Publisher: Nature Publishing Group.

[57] Wei Niu, Zhi John Lu, Mei Zhong, Mihail Sarov, John I. Murray, Cathleen M. Brdlik, Judith Janette, Chao Chen, Pedro Alves, Elicia Preston, Cindie Slightham, Lixia Jiang, Anthony A. Hyman, Stuart K. Kim, Robert H. Waterston, Mark Gerstein, Michael Snyder, and Valerie Reinke. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. *Genome Research*, 21(2):245–254, February 2011.

[58] Tetsunari Fukushige, Mark G. Hawkins, and James D. McGhee. The GATA-factor *elt-2* is essential for formation of the *Caenorhabditis elegans* intestine. *Developmental Biology*, 198(2):286–302, June 1998.

[59] Mary C. Philogene, Stephany G. Meyers Small, Peng Wang, and Ann K. Corsi. Distinct Caenorhabditis elegans HLH-8/twist-containing dimers function in the mesoderm. *Developmental Dynamics*, 241(3):481–492, 2012. _eprint: https://anatomypubs.onlinelibrary.wiley.com/doi/pdf/10.1002/dvdy.23734.

[60] Robert H. Dowen, Peter C. Breen, Thomas Tullius, Annie L. Conery, and Gary Ruvkun. A microRNA program in the C. elegans hypodermis couples to intestinal mTORC2/PQM-1 signaling to modulate fat transport. *Genes & Development*, 30(13):1515–1528, July 2016.

[61] Lihsia Chen, Michael Krause, Michael Sepanski, and Andrew Fire. The Caenorhabditis elegans MYOD homologue HLH-1 is essential for proper muscle function and complete morphogenesis. *Development*, 120(6):1631–1641, June 1994.

[62] Yoshiki Andachi. Caenorhabditis elegans T-box genes tbx-9 and tbx-8 are required for formation of hypodermis and body-wall muscle in embryogenesis. *Genes to Cells*, 9(4):331–344, 2004. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1356-9597.2004.00725.x.

[63] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

[64] H. Zhang and T. Blumenthal. Functional analysis of an intron 3' splice site in Caenorhabditis elegans. *RNA*, 2(4):380–388, January 1996. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[65] Warren A. Whyte, David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell*, 153(2):307–319, April 2013.

[66] Heidi A. Tissenbaum. Using C. elegans for aging research. *Invertebrate Reproduction & Development*, 59(sup1):59–63, January 2015. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07924259.2014.940470.

[67] Cynthia Kenyon, Jean Chang, Erin Gensch, Adam Rudner, and Ramon Tabtiang. A C. elegans mutant that lives twice as long as wild type. *Nature*, 366(6454):461–464, December 1993. Publisher: Nature Publishing Group.

[68] Nicholas Stroustrup, Winston E. Anthony, Zachary M. Nash, Vivek Gowda, Adam Gomez, Isaac F. López-Moyado, Javier Apfeld, and Walter Fontana. The temporal scaling of Caenorhabditis elegans ageing. *Nature*, 530(7588):103–107, February 2016. Publisher: Nature Publishing Group.

[69] Cynthia J. Kenyon. The genetics of ageing. *Nature*, 464(7288):504–512, March 2010. Publisher: Nature Publishing Group.

[70] Nicholas Frosst and Geoffrey Hinton. Distilling a Neural Network Into a Soft Decision Tree, November 2017. arXiv:1711.09784 [cs].