UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

# Augmenting phenotype prediction models leveraging a genomic Large Language Model

*Author:*
Georgia ZAVOU

*Supervisor:*
Dr. Jordi ABANTE LLENAS

*A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science*

*in the*

Facultat de Matemàtiques i Informàtica

June 30, 2025

UNIVERSITAT DE BARCELONA

# *Abstract*

Facultat de Matemàtiques i Informàtica

MSc

**Augmenting phenotype prediction models leveraging a genomic Large Language Model**

by Georgia ZAVOU

UNIVERSITAT DE BARCELONA

# *Abstract*

Facultat de Matemàtiques i Informàtica

MSc

**Augmenting phenotype prediction models leveraging a genomic Large Language Model**

by Georgia ZAVOU

Huntington's disease (HD) is a progressive neurodegenerative disorder caused by CAG repeat expansion in the HTT gene. While the length of this expansion explains a large portion of the variability in age of onset (AO), additional genetic modifiers, including regulatory variants, contribute to the remaining variability. In this work, we investigate the utility of genomic language models (gLMs), specifically Borzoi, for predicting tissue-specific gene expression changes from individual genomic data. We applied Borzoi to whole-genome sequencing data and integrated RNA-seq coverage predictions for relevant brain regions, including putamen and caudate. After weighting logSED scores using enhancer proximity, we aggregated these expression predictions at the gene level. We then trained multiple machine learning models to classify AO residuals such as a baseline XGBoost model using coding SNPs, CAG repeat length, and sex, an expression-based model using Borzoi-derived features and a multimodal model combining both genomic and predicted expression features. Our results show that Borzoi expression predictions capture meaningful regulatory signals, with functional enrichment analysis highlighting genes involved in transcription regulation, DNA repair, and glutamate signaling. While genotype-based models achieved the highest predictive performance, the multimodal model demonstrated complementary information from expression-based features. This study illustrates the potential of incorporating gLM-based expression predictions into phenotype modeling, offering insights into HD molecular mechanisms and genetic modifiers. The corresponding notebooks and scripts for this thesis, can be found in the following GitHub FPDS Thesis GitHub Repository

...

# *Acknowledgements*

I would like to express my sincere gratitude to the Universitat de Barcelona for giving me the opportunity to work on this project. I am especially thankful to my supervisor, Jordi Abante Llenas, for his outstanding guidance, constant support, and patience throughout every stage of this thesis. I would also like to extend my appreciation to my colleague Caterina Fuses for her valuable collaboration, helpful discussions, and continuous assistance during the development of this work.. . .

# Contents

**Abstract** iii

**Acknowledgements** vii

**1 Introduction** 3
- 1.1 Motivation . . . 3
- 1.2 Objectives . . . 3
- 1.3 Contributions . . . 4
- 1.4 Layout . . . 4

**2 Background** 5
- 2.1 Huntington's Disease . . . 5
  - 2.1.1 Genetic Etiology . . . 5
  - 2.1.2 Genetic Modifiers . . . 6
- 2.2 Genome-wide Association Studies (GWAS) . . . 7
  - 2.2.1 Genotype Data . . . 7
  - 2.2.2 Phenotype Prediction . . . 8
  - 2.2.3 Variant Effect Prediction . . . 9
- 2.3 Genomic Language Models (gLMs) . . . 9
  - 2.3.1 State of the Art . . . 10
  - 2.3.2 Model Training & Inference . . . 11
  - 2.3.3 Limitations and Challenges . . . 12
  - 2.3.4 Transformers . . . 13
  - 2.3.5 Architectural Overview . . . 14
  - 2.3.6 Capabilities and Applications . . . 15

**3 Methodology** 17
- 3.1 Genotype Data . . . 17
  - 3.1.1 Protein-coding regions . . . 18
  - 3.1.2 Regulatory regions . . . 18
- 3.2 RNA-seq Data . . . 18
  - 3.2.1 Retrieval of RNA-seq Data . . . 18
- 3.3 Genomic Language Model Gene Expression Prediction . . . 19
  - 3.3.1 Liftover of VCF Files . . . 19
  - 3.3.2 Setup and Configuration of Borzoi . . . 20
- 3.4 Phenotype prediction . . . 25
  - 3.4.1 Baseline Prediction Model . . . 25
  - 3.4.2 A Phenotype Prediction Model Based on RNA-seq Coverage . 26
  - 3.4.3 Multimodal Phenotype Prediction Model . . . 26
  - 3.4.4 Evaluation Metrics . . . 27

**4 Results and Discussion** **29**
4.1 Tissue-specific Gene Expression Predictions . . . . . . . . . . . . . . . . . 29
4.2 Gene Expression Predictions Are Informative . . . . . . . . . . . . . . . . 29
4.3 Augmentation of an HD Phenotype Prediction Model . . . . . . . . . . . 29

**5 Discussion** **33**

**6 Conclusion and Future Work** **35**

**Bibliography** **37**

# Chapter 1

# Introduction

## 1.1 Motivation

Huntington's disease (HD) is a progressive neurodegenerative disorder caused by an abnormal expansion of CAG trinucleotide repeats in the HTT gene (Jurcau and Anamaria, 2022). Although the disease is monogenic, considerable variation in the age of onset and symptom severity has been observed among individuals with the same CAG repeat length. This variability is understood to be influenced by genetic modifiers and regulatory variation across the genome. Research has shown that intermediate alleles with 27–35 CAG repeats, even if it is traditionally considered non-pathogenic, can still have mild effects on motor, cognitive, or psychiatric traits. This underscores the importance of advancing research and deepening our understanding of gene regulation.

## 1.2 Objectives

The objective of this thesis is to explore the use of genomic language models to predict tissue-specific gene expression changes from genomic variants, and to evaluate their relevance for phenotype prediction in HD. Specifically, we aim to:

- Generate gene expression predictions using Borzoi for tissues relevant to HD, such as putamen and caudate, based on individual genotype data.
- Investigate whether Borzoi-derived expression predictions contain functional information that can help predict residual age of onset (AO), after accounting for the effect of CAG repeat length.
- Compare different phenotype prediction models using XGBoost: one using only genotype information, one using Borzoi-derived expression features, and a multimodal model combining both.
- Analyze which genes and regulatory pathways are prioritized by the models, and whether expression-based features reveal candidate genetic modifiers involved in HD progression.

By addressing these objectives, the thesis seeks to better understand how regulatory variation contributes to HD variability, and whether gene expression predictions can improve genotype-to-phenotype modeling.

## 1.3 Contributions

In this thesis, we applied genomic language models to improve phenotype prediction in HD. Using Borzoi, a transformer-based model trained to predict RNA-seq coverage directly from genomic sequence, we generated tissue-specific expression predictions for brain regions relevant to HD like putamen and caudate. These predictions were based on individual-level variant data from the Enroll-HD cohort, combining both protein-coding variants and variants located in regulatory regions such as enhancers and promoters. We integrated these Borzoi-derived expression features into phenotype prediction models for residual AO, alongside traditional genotype-based features, CAG repeat length, and sex. Multiple XGBoost classification models were trained and compared such as a genotype-only model, an expression-only model using Borzoi predictions, and a multimodal model combining both types of features. Finally, feature importance analysis was performed to identify key genes, regulatory elements, and biological pathways contributing to AO.

## 1.4 Layout

The remainder of this thesis is organized as follows. Chapter 2 provides background on HD, genome-wide association studies (GWAS), and genomic language models (gLMs), including Borzoi. Chapter 3 describes the methodology, covering data processing for genotype and RNA-seq data, generation of gene expression predictions using Borzoi, and the development of phenotype prediction models, including baseline, expression-based, and multimodal models. Chapter 4 presents the results and discussion, structured into gene expression predictions and augmentation of phenotype prediction models. Chapter 5 discusses the implications and limitations of the findings, while Chapter 6 summarizes the conclusions and outlines future research directions.

# Chapter 2

# Background

## 2.1 Huntington's Disease

Huntington's disease (HD) (Jurcau and Anamaria, 2022) is an incurable neurodegenerative disease (NDD) that is mainly inherited and results in the progressive breakdown of neurons in the brain. The cause is a genetic mutation in the *HTT* gene characterized by a gradual degeneration of neurons in the brain. This mutation can lead to motor dysfunction, cognitive decline, and psychiatric disturbances. Only one copy of the altered gene is enough for an individual to develop the disorder since the disease is classified as an autosomal dominant condition.

As there is currently no cure, it leads to premature death, often 15 to 20 years after initial diagnosis. The onset of symptoms usually occurs between the ages of 30 and 50, even if both juvenile and late-onset cases are observed. The disease is chronic and progressive, therefore symptoms are worsening over time and ultimately resulting in total dependency and death, often from secondary complications such as pneumonia, heart failure, or aspiration. HD affects approximately 4 to 15 individuals per 100,000 people of European descent  Network, 2024, making it one of the most common inherited NDD. It affects males and females equally.

### 2.1.1 Genetic Etiology

HD is caused by a well-defined genetic mutation in the *HTT* gene located on chromosome 4 (locus 4p16.3) (James F Gusella, 2021). The mutation involves an unstable CAG trinucleotide repeat expansion in the first exon of the gene. This exon encodes a polyglutamine tract near the amino terminus, an alpha-helical solenoid-like scaffold. In the general population repeats in this expansion are observed to be up to 35 but length polymorphisms exceeding this number can cause HD, affecting the structure, phosphorylation pattern and activities of the protein. There are cases of juvenile-onset HD that involve expansions greater than 60 repeats and are associated with earlier onset and a more aggressive disease course. The allele is classified into four categories, namely normal, intermediate, reduced penetrance and full penetrance based on the length of the expansion (Table 2.1).

A count of 36 or more CAG repeats leads to the production of a mutant HTT (mHTT) protein with an expanded polyglutamine tract that misfolds and forms toxic aggregates inside neurons. As a result, these aggregates interfere with essential cellular functions, including axonal transport, transcriptional regulation, mitochondrial activity, and protein degradation. Over time, this contributes to neuronal dysfunction and cell death. This affects critical regions for motor control and cognitive function such as striatum and cerebral cortex.

It is important to note that a feature of HD's inheritance is the phenomenon of genetic anticipation, in which the disease tends to appear at an earlier age. This is

TABLE 2.1: Classification of CAG Expansion Alleles in *HTT*.

| **Allele Classification** | **CAG Repeats** | **Expression** |
|---|---|---|
| Normal Allele | <27 | Not associated with a phenotype; inherited stably. |
| Intermediate Allele | 27–35 | Typically not linked to HD; may show germ line instability. |
| Reduced Penetrance HD Allele | 36–39 | Disease may or may not develop due to reduced penetrance. |
| Full Penetrance HD Allele | >39 | High likelihood of developing Huntington's Disease. |

most often observed when the mutation is inherited from the father, due to increased instability of the CAG repeat during spermatogenesis.

### 2.1.2 Genetic Modifiers

Genetic modifiers (GeMs) are the genes whose natural polymorphic variation contributes to modifying the development of disease symptoms (Gusella and MacDonald, 2009). In greater depth, a gene is considered a disease modifier if changes in its sequence or expression influence the onset of the symptoms caused by the primary disease mutation, in this case, the HD CAG expansion. Searching for these modifiers aims to determine the biochemical changes that occur many years before diagnosis in order to provide validated target proteins and pathways to guide the development of strategically designed therapeutic approaches. Additionally, identifying GeMs in human studies ensures that the associated pathways are already validated to modify the pathogenic process in HD patients. This helps to overcome a major obstacle early in the drug development process.

To highlight the importance of this study, it is worth noting that two individuals with identical HD CAG repeat lengths are unlikely to develop motor symptoms at the exact same age (Gusella and MacDonald, 2009). While the presence and length of the expanded CAG repeat are the primary factors in determining if and when an individual will develop HD, the specific symptoms and their timing can also be significantly influenced by other factors. This emphasizes the role of additional GeMs.

Figure 2.1 illustrates the inverse correlation between CAG repeat length and age at neurologic onset in HD. For each individual, the age where a person first shows motor symptoms of HD (x-axis) is plotted against the measured CAG repeats in the HTT gene (y-axis). Each dot represents a single individual (from a dataset of 1,200). There is clearly an inverse correlation since when CAG repeat length increases, age of onset decreases. The curve drawn through the dots is a logarithmic regression line that fits best to the data. The CAG repeat length accounts for approximately 67% of the overall variation in age at onset which makes it the main factor, but still not the only one. The remaining 33% variation is influenced by other heritable factors and the environment (Gusella and MacDonald, 2009). From that remaining variation, about 56% is heritable, suggesting other genes besides CAG length influence onset timing. These findings support the idea that, even though CAG length is the primary predictor, individuals with the same repeat length can show symptoms at different ages.

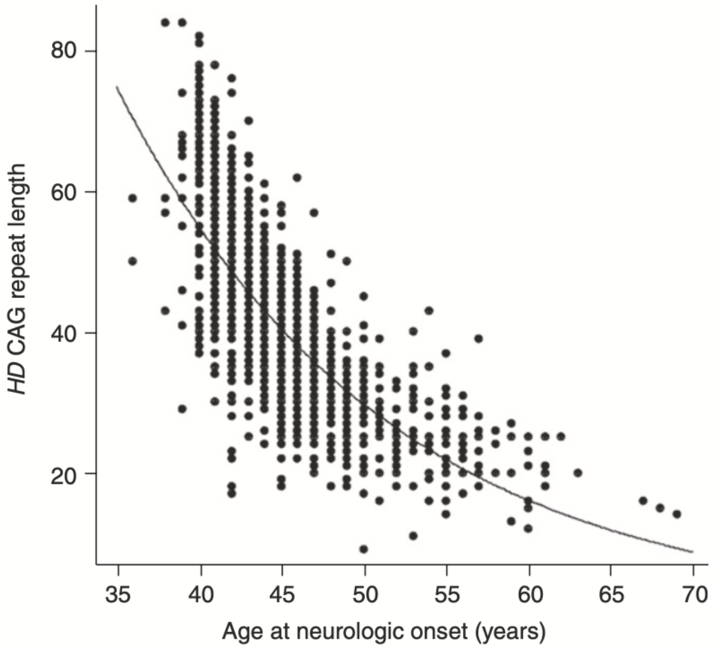

FIGURE 2.1: Neurologic Onset (adapted from (Gusella and MacDonald, 2009)).

## 2.2 Genome-wide Association Studies (GWAS)

A genome-wide association study (GWAS) is an observational study in genomics that contributes to identifying genetic variations associated with a particular disease (Cano-Gamez and Trynka, 2020). This method studies the entire genome of a large group of people, searching for small variations such as single nucleotide polymorphisms (SNPs). The goal is to estimate if any variant is associated with a trait by comparing them among people who have a particular trait or disease (referred to as phenotype) with people without it.

Due to the lack of variance explained in the phenotype by the CAG repeat length, researchers began searching for HD GeMs (Huntington's Disease (GeM-HD) Consortium, 2015). When advancing GWAS studies, many modifier loci (position on a genome) were identified. For instance, the GeM-HD Consortium in 2015 identified important loci on chromosomes 8 and 15 that accelerated or delayed onset with respect to the mean. These discoveries were later validated as HD modifiers through biological models confirming key genes involved in DNA repair such as *FAN1* in chromosome 15, and *RRM2B*, in chromosome 8, a ribonucleotide reductase. More recent research has continued to uncover additional modifier loci, including hits in chromosome 7. The potential of such studies is still being exploited.

### 2.2.1 Genotype Data

Every biological species is defined by a common set of genetic characteristics, but there still exists variation even between individuals of the same species. In humans, such variation is noted in physical features like eye color, hair texture, or disease susceptibility. This genetic variation is caused by inherited differences in DNA sequences.

Genotyping is the process of identifying differences in the genotype of an individual by examining their DNA sequence. It typically concentrates on specific genetic variants known as single nucleotide polymorphisms (SNPs), which consist in genetic variation in a particular genetic locus. As previously discussed, these loci are the subject of study, and their relationship with phenotypes, in GWAS (Gallagher and Chen-Plotkin, 2018). Scientists typically perform genotyping by analyzing particular genetic variations which occur at SNPs. These particular genetic variations show connections to both disease risk and gene regulation and the study of traits of interest (L. Chen et al., 2022; Gallagher and Chen-Plotkin, 2018).

Within this study scope, the genotyping data is analyzed to find what variants are carried by individuals in non-coding regulatory regions like enhancers and promoters. In coding regions, SNPs can lead to protein isoforms that disrupt the normal cellular functioning. Furthermore, in regulatory regions, such as enhancers or promoters, SNPs can disrupt the regulation of downstream target genes. In the presence of the latter, we can estimate how much an individual's gene expression profile might be altered due to their unique set of variants (R. v. d. Lee et al., 2022; Fulco et al., 2019).

### 2.2.2 Phenotype Prediction

In general, phenotype prediction through GWAS is the analysis of genetic data aiming to identify links between particular genetic variations and observable traits. This allows researchers to estimate traits based on an individual's genotype. Taking into account the strong associations between genotypes and disease phenotypes, especially in brain disorders, machine learning techniques can be utilized for phenotype prediction across different scales.

More specifically, phenotype prediction is applied to estimate observable traits or disease characteristics in individuals by using genetic and clinical data. In the case of HD, it is usually the prediction of traits such as age of onset, symptom severity, or progression speed based on a person's genetic profile. Knowing that the expanded CAG repeat in the *HTT* gene is the main cause of HD, individuals carrying intermediate CAG alleles with 27–35 repeats which are typically considered non-pathogenic have still shown evidence of subtle motor, cognitive, or psychiatric changes. This demonstrates that distinction of non-pathogenic and pathogenic CAG repeat lengths seems more complex than previously assumed (Meléndez et al., 2023).

Moreover, genetic modifiers outside the *HTT* locus have been shown to influence phenotype, especially age of onset. As previously discussed, recent work leveraging GWAS has led to the discovery of several modifier genes, such as *FAN1*, *MSH3*, and *RRM2B*, which are involved in DNA repair and regulatory pathways (Huntington's Disease (GeM-HD) Consortium, 2015). As a result, phenotype prediction models that integrate both CAG repeat length and genetic variants can give more accurate predictions compared to models based only in CAG.

Phenotype prediction is becoming more feasible because of the advances in machine learning and the availability of large-scale genetic datasets, such as the UK Biobank (Meléndez et al., 2023). When combining and applying these approaches, the goal is to uncover complex and potentially nonlinear relationships between genotype and phenotype in order to better understand the molecular basis of the disease. This will later on support the development of personalized prognostic tools or therapeutic strategies to delay symptom onset.

### 2.2.3 Variant Effect Prediction

Variant effect prediction (VEP) consists in the prediction of the biological impact of genetic variants. More specifically, VEP allow researcher to study how a change in the DNA sequence might affect genes, proteins, or traits in an organism. VEP in HD involves identifying modifier genes apart from *HTT* that can influence the onset of HD in individuals. It includes assessing regulatory variants since some can possibly affect gene expression or splicing, especially in brain-specific pathways. Additionally, it engages the prediction of protein impact as missense variants might affect protein function in pathways involved in neurodegeneration, inflammation, or DNA repair such as *FAN1*, *MLH1*, *MSH3* which are known modifiers in HD.

In addition to coding mutations, non-coding variants the ones found in regulatory regions such as enhancers, promoters, and introns also play a crucial role in modulating gene expression without altering protein sequence (Li et al., 2017). These variants apply strong regulatory effects even if they may be silent in most tissues. For example, a variant located within a brain-specific enhancer or promoter could influence the expression levels of *HTT* or modifier genes, and contribute to inter-individual differences in disease onset.

Recent studies have proved that rare non-coding variants are enriched near genes with extreme expression and show higher conservation (Chandrashekar et al., 2023). This is indicating that they are relevant in disease phenotypes. Moreover, models that integrate genomic and transcriptomic data rather than just relying on sequence annotations only, can improve the prediction of regulatory variant effects across tissues. These findings highlight the importance of incorporating tissue-specific expression and regulatory context into VEP, especially in the study of complex, neurodegenerative diseases like HD.

## 2.3 Genomic Language Models (gLMs)

Genomic Language Models (gLMs) are Large Language Models (LLMs) trained on genomic sequences like DNA and RNA, instead of natural language (Benegas et al., 2024). Analogous to the goal of Natural Language Processing (NLP) which is to analyze languages and understand large sequences of words, gLMs aims to understand biological sequences. Just like language models learn patterns in words, gLMs learn patterns in genomic sequences. This gives them the capability to comprehend genomes and how DNA elements at various scales interact.

What makes genomic gLMs particularly powerful is the ability to learn contextual representations of DNA, which allows them to capture functional regions and long-range dependencies in the genome and, therefore, identify transcription factor binding sites, splicing signals, and other regulatory motifs (Benegas et al., 2024). In addition they can provide strong transfer learning capabilities as after training, their learned representations can be fine-tuned for a wide range of tasks like gene expression prediction, enhancer/promoter detection, or genome annotation.

A major application of genomic LLMs is predicting the impact of genetic variants, such as SNPs. In the context of gene expression, gLMs can be fine-tuned to predict how specific variants or sequence changes influence transcriptional output across tissues. This facilitates the interpretation of genetic variants in complex traits and diseases, including Huntington's disease.

As gLMs continue to advance, they are becoming vital for interpreting genome function, evolution, and disease (Benegas et al., 2024). Their ability to generalize across cell types, and biological contexts makes them a key mechanism for personalized medicine, variant effect prediction, and innovative therapeutic discovery.

### 2.3.1 State of the Art

Models like DNABERT and the Nucleotide Transformer (Consens et al., 2025) are built upon the approach of gLMs by using self-supervised learning on large, unlabeled genomic datasets. Their aim is to learn biologically meaningful features such as transcription factor binding sites, splice junctions, and enhancer regions. These foundation models have proven effective for a range of genomics tasks, including regulatory element identification, variant effect prediction, and increasingly, gene expression modeling. The Nucleotide Transformer, for instance, was trained on hundreds of genomes from multiple species and demonstrated strong performance across various tasks related to the identification of chromatin features, DNA regulatory elements and splice sites in the human genome (Consens et al., 2025). Its embeddings generalize well to expression-related tasks via fine-tuning, even if it is not trained on expression data directly.

Given that current tools do not predict RNA-seq expression profiles due to the complexity of modeling regulatory processes, Borzoi (Linder et al., 2025) was introduced to overcome these challenges through an integrated modeling approach. It is a supervised transformer-based model, derived from Enformer. Just like Enformer, Borzoi is trained on RNA-seq data to predict RNA coverage across the genome. Borzoi distinguishes itself by modeling transcription, splicing, and polyadenylation from a single input, enabling more direct predictions of steady-state gene expression across tissues. In contrast to the methods mentioned above, instead of using self-supervised learning, models like Enformer and Borzoi use supervised learning, trained on labeled datasets such as RNA-seq coverage and epigenetic profiles, instead. Therefore, a key strength of gLMs, including Borzoi, lies in their potential to improve gene expression prediction directly from raw DNA sequence. Moreover, it is important to highlight that Borzoi also supports VEP, one of the major applications of gLMs. It scores variants by estimating their effect on predicted RNA-seq coverage and was shown to outperform the Enformer model in identifying functional regulatory variants.

Borzoi as a gLm trained on RNA-seq data, has the flexibility to be fine-tuned for specific biological contexts, including those relevant to Huntington's Disease (HD). Although HD is driven by a CAG repeat expansion in the HTT gene, variation in disease onset and progression is also influenced by non-coding variants and regulatory elements in modifier genes such as *FAN1*, *MSH3*, and *MLH1*. The effects of these modifiers act through gene expression regulation and DNA repair pathways, especially in brain tissues. As Borzoi learns from RNA-seq coverage, it can be adapted to model tissue-specific expression patterns and to assess variant effects even in non-coding regions. This is crucial for HD where pathogenicity may not be driven by protein changes alone.

Furthermore, the model's ability to model transcription, splicing, and polyadenylation in an integrated framework allows for the analysis of alternative isoforms and untranslated regions (UTRs). This could also play a role in HD phenotypes through post-transcriptional regulation. This is highly significant for analyzing the effects of intermediate CAG alleles (27–35 repeats) and understanding gene expression changes in HD-affected tissues.

Borzoi represents a state-of-the-art tool for linking genomic variation to expression-level changes in the context of complex diseases like HD. While foundation models like DNABERT and Nucleotide Transformer excel in versatility and transferability, task-specific models like Borzoi provide greater precision in expression prediction. More detailed training and evaluation analysis is presented in the following sections.

### 2.3.2 Model Training & Inference

Borzoi is built based on the Enformer's architecture as illustrated in Figure 2.2. It combines convolutional layers, downsampling, and self-attention blocks, with U-net-style upsampling for high-resolution predictions. It uses 524 kb DNA input sequences and outputs predictions in 32 bp bins. Borzoi is trained on uniformly processed RNA-seq data from ENCODE which consists of datasets with 866 human and 279 mouse samples. It is also trained with GTEx RNA-seq across various tissues, and lastly on Enformer multi-omics datasets such as CAGE, DNase-seq, ATAC-seq, and ChIP–seq for multi-modal learning. During these training sessions, the tiling strategy was applied, meaning the genome was divided into 524 kb windows, creating training examples where genes appear in variable locations within the window.

Because RNA-seq coverage incorporates effects of transcription, splicing and polyadenylation, hence, Borzoi learns all three processes from a single data type at the same time. The model learns to predict RNA-seq coverage across introns, exons, transcription start sites (TSSs), and polyadenylation sites (PASs), and is benchmarked on its ability to predict exon/intron boundaries and splicing dynamics. To assess model performance variance and enable ensembling, training process involved four randomly initialized replicate models. This implies that training was conducted using four separate Borzoi models, each starting from a different random initialization, with their outputs combined. This technique is standard in machine learning as it often leads to more reliable performance, especially when models are trained on complex, noisy biological data like RNA-seq. In Borzoi's case, this helps ensure that the model generalizes well across tissues and gene structures.

A primary metric used to evaluate how well Borzoi predictions match observed RNA-seq data is the Pearson correlation coefficient (R), which measures the average correlation between predicted and actual values across multiple test sequences. Borzoi makes base-resolution predictions of RNA-seq coverage and demonstrates strong performance at both the gene-level and bin-level. Specifically, gene-level prediction on held-out genes yields a Pearson's R of 0.87 as shown in the Figure 2.3, indicating that the model effectively captures general gene expression patterns. When evaluating bin-level coverage across exons and introns, particularly in the top 20% most variable genes, Borzoi achieves a Pearson's R of 0.88, showing strong predictive accuracy at finer resolution. Additionally, tissue-specific expression is evaluated by comparing residual expression across tissues using quantile-normalized data, resulting in a Pearson's R of 0.58, which reflects a moderate but meaningful performance in capturing expression differences across biological contexts.

Beyond general expression prediction, Borzoi was assessed on tissue-specific gene regulation tasks across five GTEx tissues (blood, liver, brain, muscle, and esophagus), including differential expression fold changes, transcription start site (TSS) usage, and alternative polyadenylation (APA) site usage. For tissue-specific fold changes, the model achieved Spearman's R values ranging from 0.52 to 0.75 as illustrated in Figure 2.4, suggesting that it effectively captures regulatory shifts across tissues.

A key application of Borzoi is variant effect prediction, particularly for assessing the functional consequences of non-coding variants such as expression quantitative trait loci (eQTLs), splicing QTLs (sQTLs), and polyadenylation QTLs (paQTLs). As

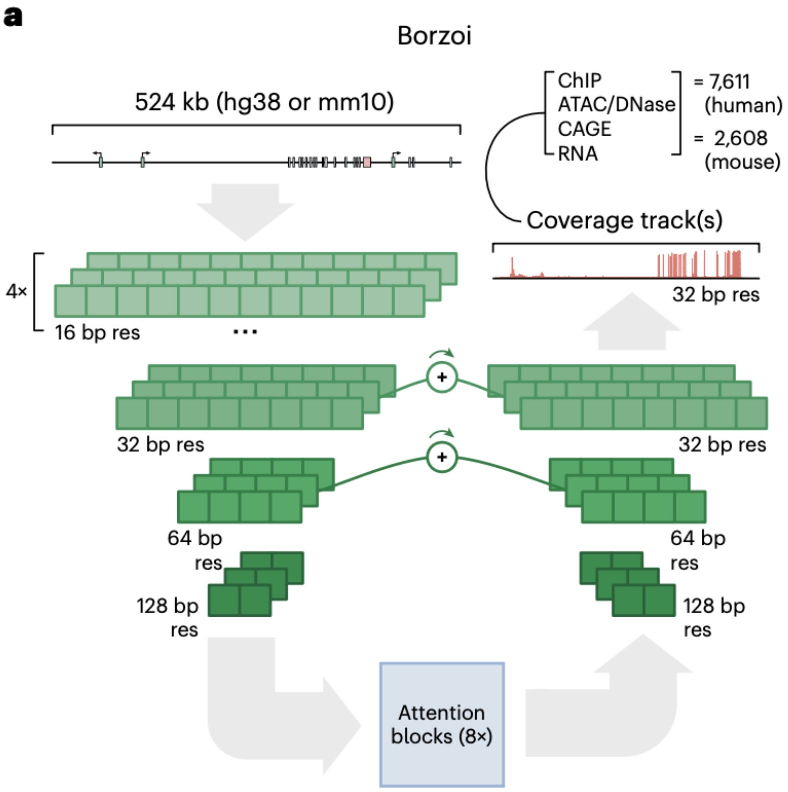

FIGURE 2.2: Borzoi neural network Architecture (adapted from (Linder et al., 2025)).

mentioned earlier, it is trained to model transcriptional, splicing, and polyadenylation dynamics directly from sequence, in order to evaluate variants by scoring their predicted impact on RNA-seq coverage. The model's performance in variant effect prediction is measured using metrics such as the Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and Spearman correlation with experimentally determined eQTL effect sizes. In comparisons with Enformer, Borzoi demonstrated superior performance, achieving a mean AUROC of 0.794 compared to 0.747 from Enformer. In addition, Borzoi's predicted variant scores showed a Spearman correlation of R = 0.334 with known eQTL coefficients, indicating a meaningful alignment between model predictions and biological data.

### 2.3.3 Limitations and Challenges

Borzoi shows strong predictive performance across gene expression and variant effect tasks, however, the model faces several important limitations that affect its interpretability, biological resolution, and generalizability.

One key challenge is its limited ability to model tissue-specific splicing events. Although Borzoi predicts RNA-seq coverage with high accuracy on average, it often fails to capture fine-grained, tissue-dependent transcript variants. This indicates a tendency to default to consensus transcript profiles rather than condition-specific isoforms, which reduces its effectiveness in studying alternative splicing patterns.

Furthermore, technical biases inherent in RNA-seq data, such as GC-content bias and 3' end bias, can impact the model's performance by introducing misleading

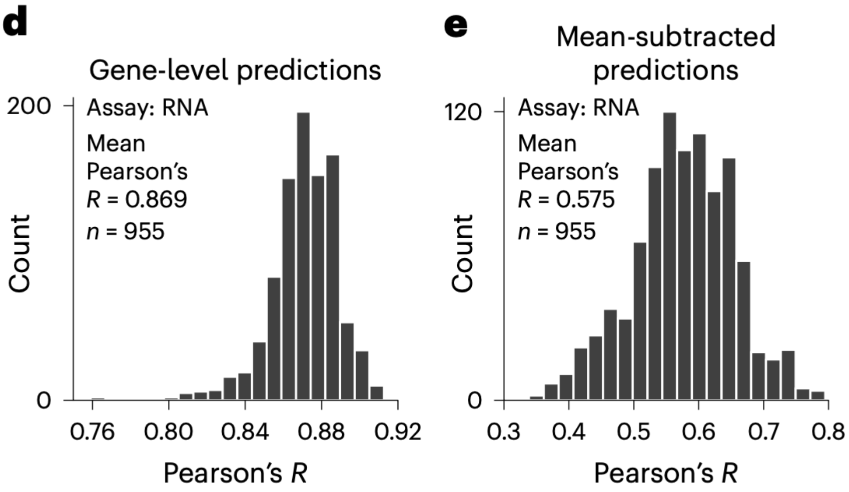

FIGURE 2.3: Pearson Correlation Coefficient (R) (adapted from (Linder et al., 2025)).

signals. These biases often lead to false positives, particularly in the prediction of alternative splice sites. This variation underscores a critical limitation of relying solely on RNA-seq data as a training signal, especially when interpreting regulatory elements.

A major challenge is disentangling overlapping regulatory signals, as RNA-seq data capture the combined output of multiple, interdependent regulatory programs. While Borzoi is trained to model transcription, splicing, and polyadenylation in parallel from sequence data, these layers of regulation are deeply interconnected, making it difficult to attribute signal components to individual processes. This complexity limits the precision with which Borzoi can isolate the effects of specific regulatory mechanisms or variants that act through only one layer.

Finally, the interpretability of Borzoi's predictions is highly dependent on the choice of attribution method. Different interpretation techniques, such as input gradients, *in silico* mutagenesis (ISM), and window-shuffled ISM, yield different results depending on the genomic context. For instance, while input gradients and *in silico* mutagenesis (ISM) produced high-quality attributions for splicing and enhancer–promoter communication, window-shuffled ISM performed better in 3′ UTR regions due to signal buffering effects. This variability introduces uncertainty in identifying causal regions and reduces confidence in variant interpretation.

Together, these limitations point to key areas for future improvement, including better modeling of splicing and isoform dynamics, incorporation of experimental data on mRNA stability, and development of more robust interpretability frameworks suited for diverse genomic contexts.

### 2.3.4 Transformers

In gLMs, transformers are used in the same way as in NLP, but in this case they are processing biological sequences like DNA, RNA, or proteins. A revolution has taken place after LLMs based on the transformer deep learning architecture were utilized for NLP. Hence, researchers started developing genome language models that are based on transformer architecture as soon as they observed the parallel between human language and the genome's biological code.

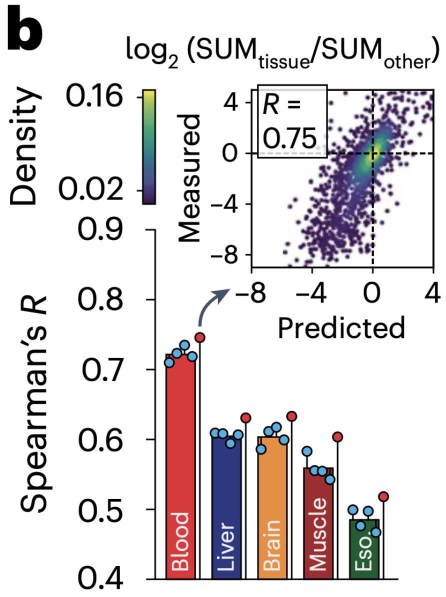

FIGURE 2.4: Spearman's R values for tissue-specific fold changes (adapted from (Linder et al., 2025)).

Transformers use a self-attention mechanism (Vaswani et al., 2017) that empowers the model to consider every position in a sequence and decide how important each other position is when making a prediction for a given token. This allows them to inspect the whole DNA sequence at once and decide which parts are relevant to each other, regardless of how far apart they are. This gives them the capability to capture relationships between distant nucleotides which is something that CNNs and RNNs ofter struggle with. This happens because RNNs read the sequences step-by-step and can forget context over long distances and because CNNs have visibility only of what's in their small window called the receptive field.

The fact that most genomic data do not have human-labeled annotations presents a significant challenge. However, transformers offer a powerful solution by pre-training on huge amounts of raw DNA sequence with unsupervised learning. In the same way that GPT is trained to predict the next word, genomic LLMs learn to predict masked nucleotides or relationships in sequences. This enables them to learn generalizable representations without the need for manual annotation, especially when data is sparse, which is typically the case in biological research.

### 2.3.5 Architectural Overview

The Transformer is a deep learning architecture originally introduced in 2017, designed to model relationships between elements in a sequence regardless of their position (Consens et al., 2025). In contrast to models like CNNs that focus on local patterns or RNNs that process sequences sequentially, transformers rely on a mechanism called self-attention. This mechanism allows every position in a sequence to attend to every other position to capture both local and global context.

The Transformer architecture consists of a variety of core components as shown in the Figure 2.5. One of them is the Self-Attention Mechanism that can take as an input a nucleotide or k-mer, which is then transformed into three vectors: Query (Q), Key (K), and Value (V). Next, the attention mechanism calculates how relevant each position in the sequence is to every other position based on these vectors. The output is a weighted sum of values, where the weights are the attention scores that reflect how much each token should "pay attention" to others.

Another key component is multi-head attention. Instead of computing one attention map, the transformer does it in parallel across multiple heads, allowing it to learn diverse patterns and dependencies from different perspectives. Transformers also include a Feed-Forward Neural Network (FFNN). After attention, each position is independently passed through this fully connected FFNN for transformation and interaction across layers. Since transformers do not inherently understand sequence order, positional embeddings are added to each input to provide information about the relative or absolute position in the sequence. This step is called Positional Encoding. After this, a normalization layer is necessary to stabilize and accelerate training. Each sub-layer includes layer normalization and skip connections that pass information forward more directly.

Lastly, the major core components are the Encoder and the Decoder. An encoder is the part of a transformer that processes the entire input sequence at once and transforms it into a rich, context-aware representation. It looks in both the left and right directions of the sequence, which means that it understands the full context of the input based on what comes before and after it. This makes it bidirectional, which is ideal for tasks like classifying DNA sequences, predicting whether a base belongs to a promoter, exon, enhancer,annotating genomic regions and finding motifs or splice sites. A decoder, on the other hand, is designed for generative tasks in which the model needs to predict the next token based only on the past. It is unidirectional, meaning that it only looks left-to-right from earlier positions in the sequence toward the current one. Decoders are typically used for generating sequences or filling in missing bases of DNA design or mutational simulations.

### 2.3.6 Capabilities and Applications

Transformer-based models have emerged as a powerful architecture in genomics due to their ability to learn deep, context-aware representations of DNA sequences. Unlike CNNs or RNNs which are restricted by local window sizes or sequential memory limits, they can capture global text and long-range dependencies across DNA sequences (Consens et al., 2025). This is due to the fact that they use a self-attention mechanism which allows them to learn contextual embeddings, considering upstream and downstream information and therefore have bidirectional and context representations. Moreover, they are pre-trained on unlabeled data using a self-supervised pre-training to learn from massive unannotated genomic datasets, which makes them extremely powerful. These pretrained transformers can be fine-tuned on specific genomic tasks with limited labeled data and can also perform zero-shot inference, meaning they make predictions on enhancer or splice site without task-specific fine-tuning.

Transformers have rapidly gained traction in genomics due to their versatility across a wide range of predictive tasks, many of which are central to understanding gene regulation and genome function. These models have been applied to various tasks that demonstrate strong generalization and interpretability across scales. An important application is the prediction of variant effects, which indicates that they can score genetic variants by predicting their functional effects on regulatory

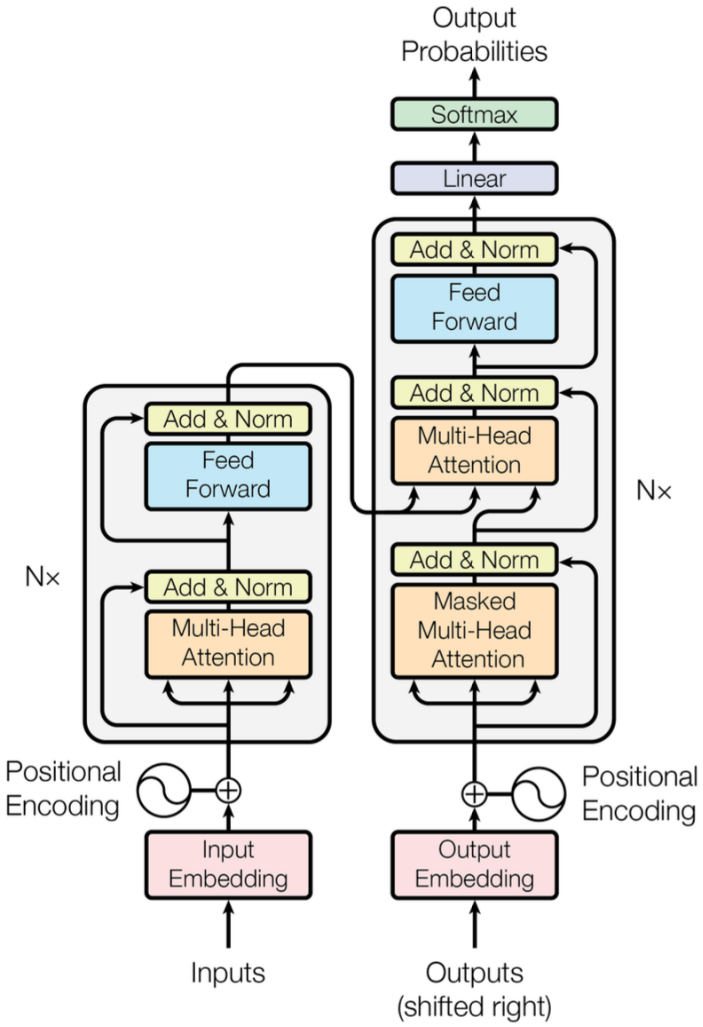

FIGURE 2.5: Transformer Architecture (adapted from (Consens et al., 2025)).

elements or gene expression. Secondly, transformers can model transcription, splicing, polyadenylation, and chromatin accessibility, Implying that they predict RNA-seq coverage and tissue-specific gene expression. In addition, transformers identify functional genomic elements like enhancers, promoters, and splice sites directly from sequence. Lastly, some hybrid transformer models like C.Origami are designed to predict 3D genome interactions and chromatin organization from sequence and epigenetic data which is crucial for understanding gene regulation beyond linear DNA.

# Chapter 3

# Methodology

## 3.1 Genotype Data

Studying rare diseases with machine learning (ML) is challenging because ML generally requires large-scale datasets. The data used for this study was assembled by Lee et al. (J. Lee et al., 2019), combining samples from several large HD observational studies, including the GeM-HD Consortium, Enroll-HD, and Registry. Enroll-HD, which is one of the largest contributors, is the world's largest observational HD study, with more than 20,000 participants enrolled globally (Sathe1 et al., 2021). However, for this study, the final combined dataset includes whole-genome SNP genotypes for 9,064 individuals, along with their CAG trinucleotide repeat lengths and recorded AO. These data were obtained by sequencing blood samples.

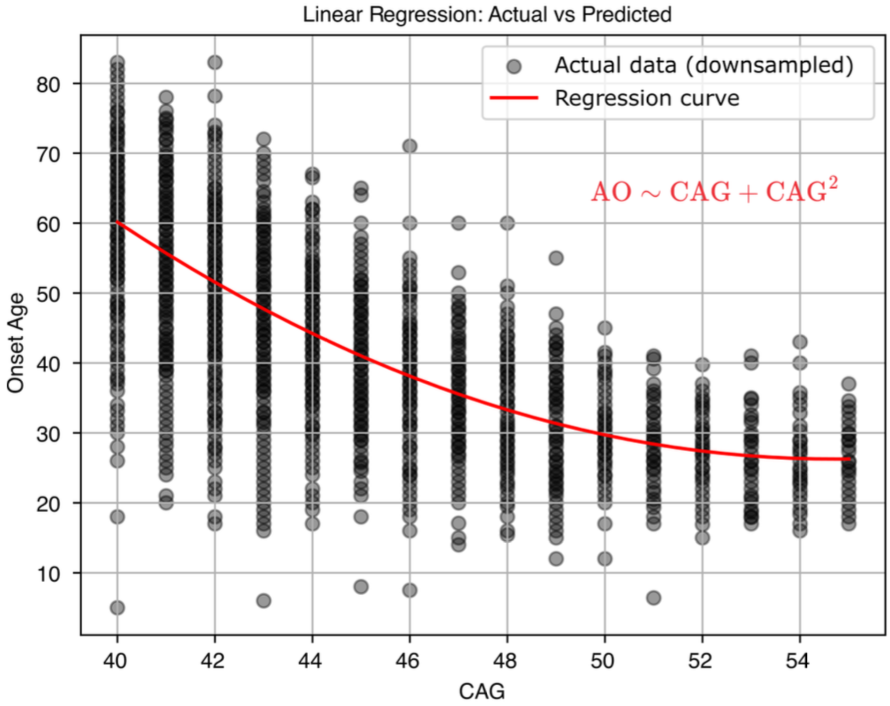

FIGURE 3.1

Since our goal is to identify genetic modifiers that contribute to the variability in AO beyond the known effect of CAG length, we first fitted a second-order linear regression model: $AO \sim CAG + CAG^2$. The residuals from this model represent the unexplained variability in AO after accounting for CAG length. To enable classification-based machine learning models, these residuals were divided into five quantiles, assigning individuals into five equally sized classes. Each class reflects how early or late an individual's onset is compared to what is expected given their CAG length as seen in figure 3.1.

Moreover, in this project, we are focusing on SNPs with a Minor Allele Frequency (MAF) greater than 1%. MAF refers to the frequency at which the less common allele occurs in a given population. This means that the alternative allele is present in at least 1% of the individuals. This threshold helps filter out rare or private mutations that are unlikely to contribute meaningfully to population-level gene regulation patterns. Combined with regulatory annotations, this forms the basis for assigning predicted expression changes per gene and per individual.

### 3.1.1 Protein-coding regions

To reduce computational burden and focus on biologically meaningful features, SNPs were filtered to include only those located in protein-coding genes that are functionally associated with HD pathogenesis. Specifically, we used Gene Ontology (GO) terms related to DNA repair, transcription regulation, and other relevant processes (Dickey and La Spada, 2018; Gatto et al., 2020). This filtering process resulted in a final dataset containing 339,886 SNPs across 2,774 protein-coding genes.

### 3.1.2 Regulatory regions

Regulatory regions were included by scoring variants located in enhancers and promoters. Enhancers were obtained from GeneHancer (Fishilevich et al., 2017), which provides enhancer-to-gene mappings. Promoters were defined as 4 kilobase (kbp) windows centered at the transcription start site (TSS) of each gene.

## 3.2 RNA-seq Data

### 3.2.1 Retrieval of RNA-seq Data

Since we hypothesized that what could give us information related to the real molecular contex taking place in HD brains is predicting the differential expression in the most affected tissues like putamen and caudate, we firstly obtained RNA-seq data from healthy individuals which served as the basis for our gLM predictions. To do so, transcriptomic reference data for downstream analysis, RNA-seq coverage files were downloaded for the putamen and caudate brain regions from the recount3 project. Specifically, we used data from the study SRP074904, available through the Sequence Read Archive (SRA). Recount3 provides uniformly processed RNA-seq data in BigWigformat. The relevant files were retrieved using `wget`, with download URLs obtained from the recount3 portal. After download, coverage tracks from individual samples were merged to create tissue-specific aggregate signal files. By doing this, we enable their use as reference inputs for modeling gene expression with Borzoi. Then, we had to convert raw RNA-seq data from the bigWig format into .w5 format that Borzoi expects for the targets.

Next, we generated the `targets\_human.txt` file that is required for inference and evaluation in Borzoi. This file is the configuration file that defines the target dataset, such as the tissues and expression bins to predict. Without this file, the model wouldn't know how to map internal predictions to interpretable biological outputs. As stated in the borzoi paper, it is used both during training to define the loss function and during evaluation or inference in order to interpret the predicted outputs. In this project, this file was customized to include only the tissues relevant to HD such as putamen and caudate to reduce complexity and focus the model on brain-relevant signals.

## 3.3 Genomic Language Model Gene Expression Prediction

Gene expression prediction is the estimation of mRNA presence for specific genes based on underlying genomic information. This includes predicting expression levels directly from DNA sequences, integrating additional biological features such as transcription factor binding motifs, chromatin accessibility, and regulatory element interactions. Accurate prediction of gene expression is essential to understand cellular function, tissue identity, and the regulatory architecture of the genome, especially in the context of disease (L. Chen et al., 2022).

In HD, predicting gene expression from genotypic data is very important because many disease-associated variants, especially those influencing the age of onset, are in non-coding regions of the genome, where they affect transcriptional regulation. These variants can alter the expression of the *HTT* gene or other modifier genes in DNA repair, neuronal signaling, and neuroinflammation like *FAN1*, *MSH3*, *MSH1*) (Huntington's Disease (GeM-HD) Consortium, 2015; Linder et al., 2025). Understanding how these variants affect gene expression can help us explain why symptoms differ between people with HD and can contribute to the search for new treatments.

As previosuly discussed, recent advances in deep learning have led to the development of gLMs, models that predict gene expression directly from DNA sequences. Transformer-based architectures such as Borzoi have demonstrated strong performance in modeling RNA-seq coverage from genomic sequence by capturing signals related to transcription, splicing, and polyadenylation (Linder et al., 2025). Unlike foundational gLMs like DNABERT or Nucleotide Transformer, Borzoi is trained in a supervised way using labeled RNA-seq data which allows it to produce tissue-specific predictions of gene expression with higher resolution.

Furthermore, these models can support VEP by simulating the presence of a variant and assessing how it alters predicted expression. For example, Borzoi's predictions can be used to compute log-fold changes in expression. This is possible because SNPs in enhancers or promoters can highlight regulatory variants that may not be apparent through GWAS alone (Linder et al., 2025; Gallagher and Chen-Plotkin, 2018). Such approaches are particularly useful in HD, where regulatory variation in brain-specific tissues is hypothesized to modulate disease onset and severity (Li et al., 2017).

### 3.3.1 Liftover of VCF Files

In order to replicate individual-specific regulatory effects in the genome, it was necessary to generate Variant Call Format (VCF) files which contain the genetic variants that are present in each individual. The VCF format is standard and widely accepted as a representation of genomic variation. These VCF files include information such as the genomic coordinates of variants, the reference and alternate alleles, genotype calls for each individual, and some metadata annotations. VCF files are critical for enabling personalized sequence modeling for the Borzoi prediction pipeline. They way Borzoi uses these files, is by predicting the impact of DNA sequence on RNA expression profiles. To personalize this, Borzoi modifies the reference genome sequence for each individual by embedding the specific genetic variants described in the VCF.

As previously mentioned, our analysis focuses on SNPs found in regulatory regions because these are the responsible for gene expression alterations. In addition, the genotype field in the VCF encodes whether an individual carries zero, one, or two copies of the alternate allele at a given position. Therefore, Borzoi can accurately model the effects of heterozygous or homozygous variants by taking advantage of this information.

Using VCF files in the way explained above, Borzoi is able to generate variant-aware, individualized predictions of transcriptional activity. This is necessary for understanding the functional consequences of genetic variation, especially in complex tissues like the brain. In this project, genomic variant data was originally available in the GRCh37 (hg19) reference genome coordinate system. However, the Borzoi model is trained and designed to operate on the more recent GRCh38 (hg38) reference genome. To ensure compatibility, a critical preprocessing step involved lifting over the variant coordinates from hg19 to hg38. This coordinate conversion is essential because even minor differences in genome assemblies can lead to incorrect positional mapping, which would affect prediction accuracy. The liftover process was performed using tools such as the UCSC LiftOver utility. These tools need precomputed files in order to map positions from one genome build to another.

This step was crucial in order to correctly use Borzoi because the model reads sequences directly from the hg38 reference and modifies them based on VCF input. Feeding it variants mapped to an older assembly would result in mismatched sequences and misaligned predictions. Therefore, accurate liftover was essential to ensure that each SNP is placed at the correct genomic position relative to regulatory elements.

### 3.3.2 Setup and Configuration of Borzoi

Before applying Borzoi to our custom large genomic input, it is essential to ensure that the model and its environment function in the way we expect. This is highly recommended by the model authors and was successfully completed in this project as a necessary check before proceeding to the more complex VEP workflow. Therefore, in this step we run Borzoi using the example inputs provided by the developers that were included in the GitHub. This process involved supplying preprocessed inputs, including a reference FASTA file, a sample VCF, and the `targets_human.txt` file provided by the authors, to ensure that the model could successfully generate RNA-seq coverage predictions. The resulting outputs are in the.h5 format, and a script was constructed in order to make them readable in the `.txt` format and check for correctness and consistency. This test served multiple purposes such as confirming that necessary software dependencies were correctly installed and verifying that the GPU environment and CUDA support were functional for efficient inference. Laslty, it provided a benchmark for comparing actual outputs with expected reference results.

After this first test, we proceeded with a second one, this time using a small subset of our own data. This involved replacing the example VCF with one derived from our dataset and using our own `targets_human.txt` file that was generated using brain-specific (Putamen) enhancer annotations. This extra test confirmed that the pipeline can process real, custom data correctly and that our inputs are properly formatted and compatible with Borzoi's requirements. In this test we used data only for the tissue putamen which allowed us to generate Putamen-specific regulatory effects (logSED) predictions, aligned with the biological focus of our study. To clarify whether using tissue-relevant data produces meaningful differences in predicted regulatory effects, we compared the logSED outputs from these two tests.

Firstly, we did a gene-level comparison by aggregating logSED values by gene and compared total predicted regulatory across the two tissues, putamen from our input and RNA-K562 from Borzoi's input. After observing the result in Figure 3.2, we can see that *FAN1* shows dramatic differences in predicted regulation between K562 and Putamen, indicating tissue-specific effects. Therefore, we will continue to analyze some more things based on these two tissues and the gene *FAN1*.

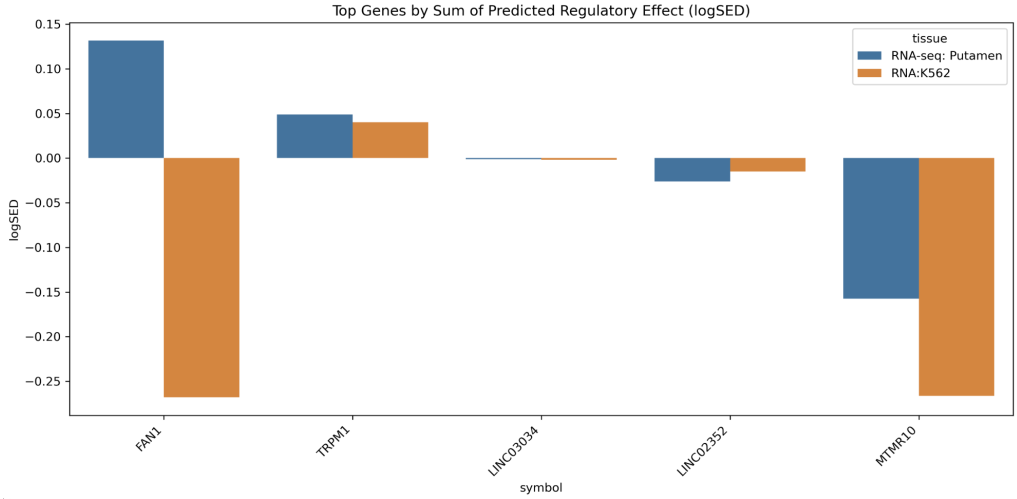

FIGURE 3.2: Top genes by sum of predicted regulatory effect.

Secondly, we did an SNP-level comparison for *FAN1* by aligning matching SNPs between the two runs and compared their individual logSED values. The resulting plots in Figure 3.3 show several variants with opposite or divergent effects. This means that there is need for tissue-specific analysis.

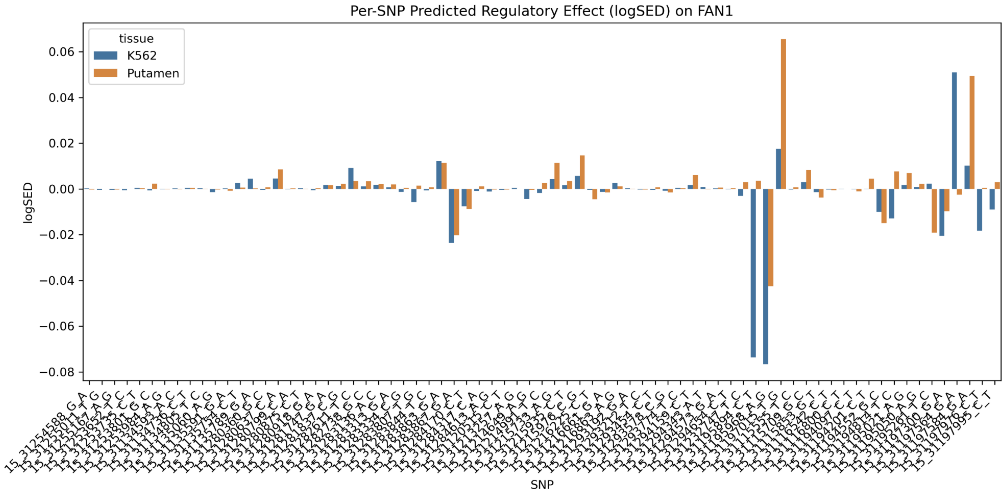

FIGURE 3.3: Per SNP predicted regulatory effect on FAN1.

In addition to gene- and SNP-level comparisons, we also examined the overall distribution of predicted regulatory effects across all genes for each tissue. Figure 3.4 illustrates the sum of logSED scores computed from all variant-gene pairs, grouped by tissue. The predictions based on Putamen-specific input showed an overall positive effect on gene regulation, while the predictions using K562 data showed a negative effect. This difference in both the direction and size of the predicted effects shows that the tissue used in the analysis has a big impact on the results. Because of this, it makes sense to use tissue-specific data for the rest of our analysis.

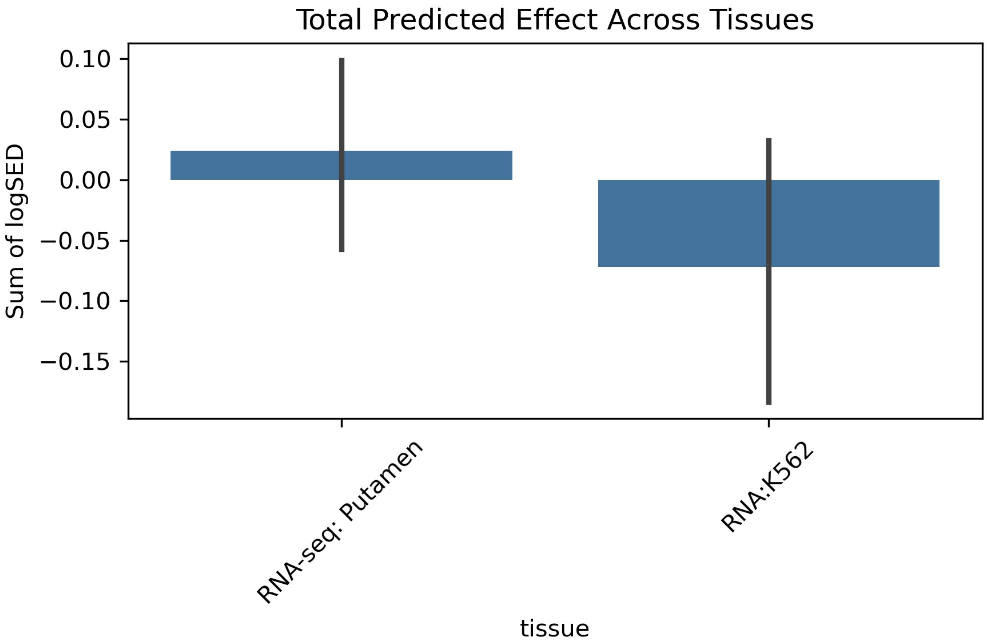

FIGURE 3.4: Total predicted effect across tissues

Finally we plotted one more graph Figure 3.5 that compares the predicted logSED of SNPs in the *FAN1* gene between two tissues. Each dot represents a single SNP, with its predicted effect in K562 plotted horizontally and in Putamen vertically. The red dashed line marks the line of equal effect (y = x), meaning SNPs that fall on this line have identical predicted effects in both tissues. Points above the line indicate a stronger effect in Putamen, while those below indicate a stronger effect in K562. Most SNPs do not lie on the red line which means that predicted regulatory effects differ between tissues. It is also notable that most of them are above the line, indicating a stronger regulatory influence in Putamen which once again validates our choice to run Borzoi with tissue-specific input.

In the Borzoi paper, authors explain that generating logSED predictions they apply a distance-based weighting step to the logSED values. This is suggested because SNPs located closer to the center of an enhancer are more likely to influence gene regulation. Therefore, after generating logSED predictions for all chromosomes using Borzoi, we applied to all of them the distance-based weighting step suggested in their publication. Specifically, we used enhancer annotations to map SNPs to nearby enhancer regions, and for each SNP-enhancer pair, we calculated a Gaussian weight based on the distance between the SNP position and the center of the enhancer. The standard deviation for the Gaussian kernel was set to 300 base pairs, consistent with the parameters described in the original Borzoi study. We then multiplied the original logSED value by this weight to obtain a weighted logSED score, which prioritizes SNPs located centrally within enhancer regions. This approach helps make the predicted effects more biologically realistic by giving more importance to SNPs that are

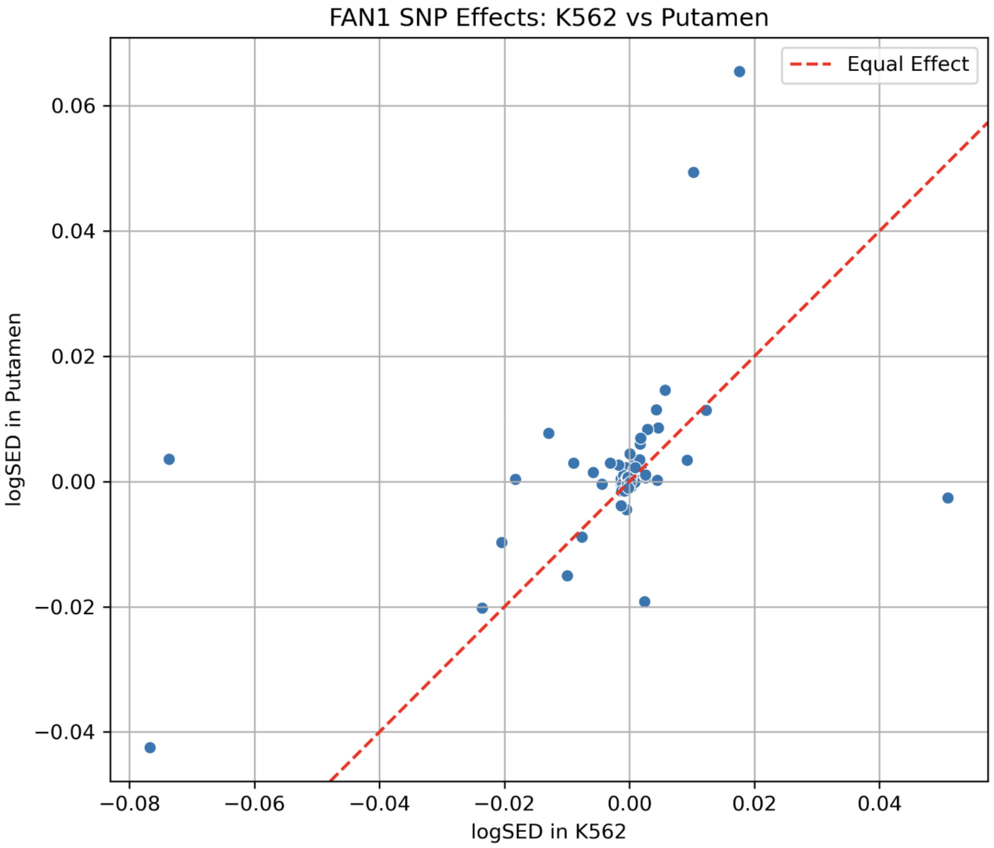

FIGURE 3.5: FAN1 SNP effect in both tissues

closer to the center of enhancer regions. It also reduces the impact of SNPs that are farther away or less likely to be relevant.

To better illustrate the Gaussian weighting applied to SNPs within enhancer regions, we include two example plots that show how the assigned weight depends on the SNP's position relative to the enhancer center. In both cases, the gray shaded region represents the enhancer's genomic span, the dashed vertical line marks the center of the enhancer, and the solid vertical line indicates the SNP position. The black curve is the Gaussian function used to calculate the weight as detailed before. In the first example Figure 3.6, the SNP lies within the enhancer but it is far from the center. As a result, the Gaussian curve assigns a relatively lower weight to this variant. This reflects the assumption that SNPs closer to the center of the enhancer are more likely to contribute to its regulatory activity. In contrast, the second example Figure 3.7 shows a SNP positioned almost exactly at the center of the enhancer. Here, the Gaussian function reaches its peak, assigning the maximum possible weight to the SNP's predicted regulatory effect. These two examples clearly illustrate how the weighting method emphasizes variants that are more centrally located within enhancer elements.

Flowing, to better visualize this effect, we created the histograms in Figure 3.8. These histograms demonstrate the distribution of logSEDs before and after applying Gaussian weighting for two brain tissues used in this project Putamen and Caudate. The histograms show the frequency of SNPs across different logSED values, plotted on a logarithmic y-axis to capture the wide range of frequencies. The gray bars represent the unweighted logSED values, while the black bars show the weighted values,

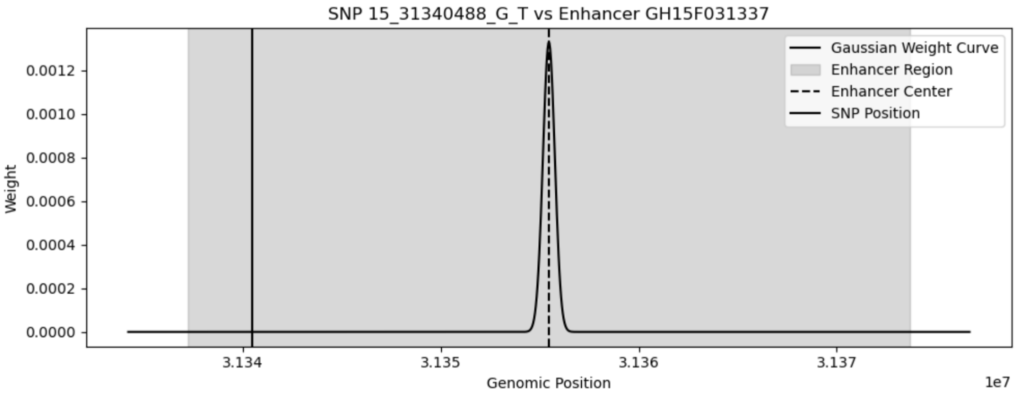

FIGURE 3.6: Gaussian Weighting of LogSED SNP1

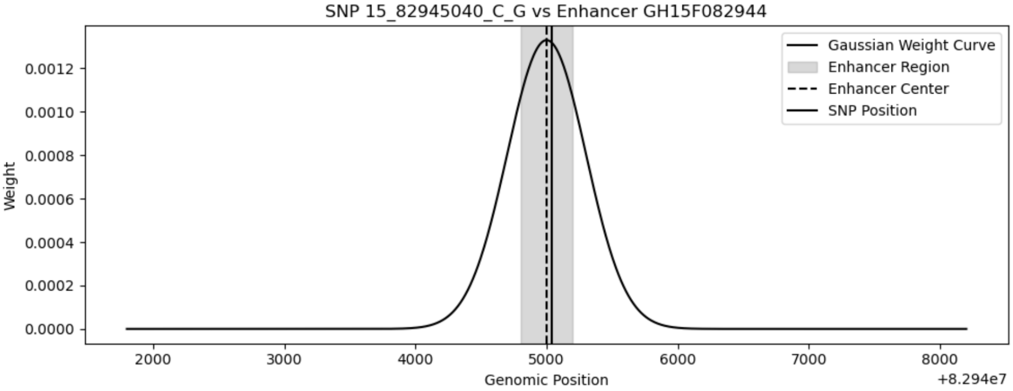

FIGURE 3.7: Gaussian Weighting of LogSED SNP2

which as explained above, are calculated based on the SNPs' distance from the center of enhancer regions. In both tissues, the weighted logSED values are more tightly centered around zero and show a reduced spread compared to the unweighted values. This verifies the necessity of the weighting step, which reduces the influence of more weakly associated SNPs. This helps to focus the analysis on those variants that are more likely to be biologically meaningful.

After applying the Gaussian weighting to adjust each SNP's logSED score based on how close it is to the center of an enhancer, we grouped the results by gene. This means that for each gene, we collected all the weighted logSED values from SNPs linked to that gene's enhancers. Then, for each individual (subject) we summed up the weighted logSED values for each gene. This gave us a single number per gene, per subject. It is a summary score showing how much that person's variants are predicted to affect expression of that gene in a specific tissue like Putamen or Caudate.

To make sure the aggregation and the overall logic coded for this, we created a notebook with toy examples in order to debug each step. Once we made sure that everything works as expected, we repeated this for every subject in the dataset and for all genes that had enhancer-linked SNPs. The result is a matrix where each row represents a subject, each column represents a gene, and each cell contains the summed, weighted logSED score for that gene in that subject. This matrix captures the predicted regulatory impact of variants on gene expression for each person.

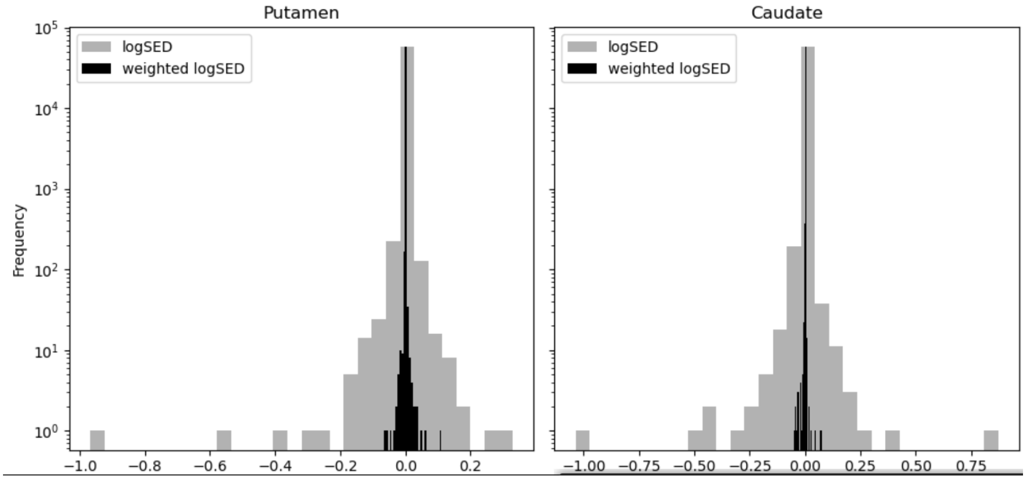

FIGURE 3.8: LogSED Distribution with Gaussian

## 3.4 Phenotype prediction

In the context of HD, this study explored whether predicted regulatory effects from gLMs like Borzoi's logSED scores could improve phenotype prediction models. More specifically for classifying residual AO. The residual AO is defined after regressing out CAG repeat length, and it captures variance in AO unexplained by CAG length alone. To evaluate this, several phenotype prediction models were developed using XGBoost. XGBoost (T. Chen and Guestrin, 2016) is a gradient-boosted decision tree algorithm that builds an ensemble of trees sequentially, where each new tree corrects errors made by the previous ones. The model's parameters, like tree depth and learning rate, were optimized via grid search using 5-fold CV. Feature importance was assessed using XGBoost's gain score, which reflects how much each feature contributes to reducing the loss function during training.

### 3.4.1 Baseline Prediction Model

As a baseline model, an XGBoost classifier was trained using genotype information derived from coding variants (Fuses et al., 2025). SNPs were first filtered to include only those located within protein-coding regions selected based on GO as explained previously in order to keep the ones related to HD, such as DNA repair and somatic expansion. This filtering resulted in a dataset containing 339,886 SNPs spanning 2,774 protein-coding genes as mentioned in the data preprocessing section. In addition to the genotype data, covariates including CAG repeat length and sex were included as features. The model aimed to classify subjects into five classes derived from the residuals of a second-order linear regression model predicting AO from CAG repeat length.

XGBoost performs classification by sequentially building an ensemble of $K$ regression trees, where the prediction for each individual is given by:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} = \{\text{regression trees}\}.$$

For multi-class classification, each tree outputs logits for each class. These are later transformed into probabilities through a softmax function. The model minimizes a regularized loss function of the form:

$$L(t) = \sum_i \ell(y_i, \hat{y}_i(t)) + \sum_j \omega(f_k),$$

where $\ell$ is the multiclass log loss and $\omega(f_k)$ penalizes tree complexity. The model's hyperparameters, such as tree depth and learning rate, were optimized via grid search using 5-fold cross-validation. Feature importance was assessed using the gain score, which reflects how much each feature contributes to reducing the loss function at each split.

### 3.4.2 A Phenotype Prediction Model Based on RNA-seq Coverage

In this model, only gene expression predictions derived from Borzoi were used, together with clinical covariates, to predict residual AO. Borzoi, a genomic language model trained to predict tissue-specific RNA-seq coverage, was applied to each subject's VCF file to generate logSED scores. Predictions were made using enhancer annotations for brain tissues such as putamen and caudate which are relevant to HD. Following Borzoi's recommended approach, the predicted logSED scores were weighted using a Gaussian kernel based on the distance between each variant and the center of its assigned enhancer as described earlier in this thesis. The weighted logSED scores were aggregated at the gene level to create matrices that represent the predicted regulatory impact for each individual.

These logSED features were combined with CAG repeat length and sex to form the input feature set. The classification target was defined by dividing residuals of a second-order linear regression model $AO \sim CAG + CAG^2$ into five quantiles, creating five equally sized AO classes. Model training was performed using XGBoost, with hyperparameters optimized via grid search and 5-fold cross-validation.

The purpose of this model was to check if Borzoi's predicted regulatory effects, even without using the actual genotype data, contain useful information for predicting AO. This allowed us to test how useful the tissue-specific regulatory predictions are on their own, before adding them together with the genotype data in the multimodal model.

### 3.4.3 Multimodal Phenotype Prediction Model

The multimodal prediction model was designed to combine both genotype and predicted expression features for phenotype prediction. This is aiming to test whether the integration of regulatory effect predictions with raw genotypes could enhance model performance.

Genotype data consisted of SNPs filtered to protein-coding regions selected based on GO processes related to HD, such as DNA repair and transcription regulation, resulting in 339,886 SNPs across 2,774 genes.

For predicted expression features, Borzoi was used to estimate tissue-specific gene expression effects based on subject-specific VCFs as mentioned earlier. The resulting logSED scores were aggregated into gene-level features by summing across variants, and combined with genotype features, CAG repeat length, and sex to form the multimodal feature set. The model was trained using XGBoost again.

### 3.4.4 Evaluation Metrics

The models were evaluated using Balanced Accuracy (BA) which is the average recall over all classes. For multi-class classification problems with potential class imbalance, BA provides a more reliable performance estimate than overall accuracy, as it equally weights the contribution of each class regardless of its prevalence. Specifically, BA is calculated as:

$$BA = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}$$

where $C$ represents the total number of classes, $TP_c$ denotes the true positives for class $c$, and $FN_c$ denotes the false negatives for class $c$. This metric ensures that the model's ability to correctly classify each class is fairly represented, making it particularly suitable for imbalanced datasets such as those encountered in residual AO prediction.

All models are evaluated using 10-fold cross-validation, ensuring that performance estimates generalize beyond a specific train-test split. To estimate the variability and confidence of model performance, we compute 90% confidence intervals through bootstrapping with 1,000 resamples of the test predictions. We also apply statistical significance tests, including the binomial test to assess whether model performance exceeds random chance and the Wilcoxon signed-rank test to compare the performance of alternative models. This multi-level evaluation strategy provides both quantitative accuracy and statistical confidence, ensuring a robust assessment of model effectiveness.

# Chapter 4

# Results and Discussion

## 4.1 Tissue-specific Gene Expression Predictions

## 4.2 Gene Expression Predictions Are Informative

To evaluate the predictive value of gene expression predictions generated by Borzoi, an phenotype prediction model was trained only on its gene expression predictions. In this setting, only Borzoi-predicted logSED scores, CAG repeat length, and sex were used as features. The logSED values were generated for putamen and caudate tissues, weighted according to their distance to enhancers using a Gaussian kernel with $\sigma = 300$, and aggregated at the gene level as previously explained.

The expression-only model achieved a median BA of 0.242. This is significantly better than random classifier, which would produce a BA of 0.242 ($p_{\text{Binom}} < 0.2$). Even though this is not as accurate as models trained on genotype data alone, the resulting BA suggests that the predictions produced by Borzoi contain some predictive value about the onset of the disease. This gives it the possibility to help identify new variants in regulatory regions that affect when HD symptoms start. In addition, feature importance analysis revealed that CAG repeat length remained an informative feature even when using only the expression features derived from Borzoi. Specifically, for models with tree depth 2, approximately 65% of decision trees selected CAG length as the first splitting feature in 24% of cases. This is indicating that even when Borzoi predictions are included, CAG repeat length often remains the most powerful predictor of AO.

This model also prioritized regulatory variants located in enhancer regions, including three variants not previously reported as HD genetic modifiers: 19_50651485_A_C (rs180918699), 5_60241142_G_A, and 1_157069597_G_A. These variants, identified through their predicted regulatory effects, suggest that Borzoi-based expression predictions can uncover novel candidate modifiers located outside protein-coding regions.

## 4.3 Augmentation of an HD Phenotype Prediction Model

To assess whether gene expression predictions from Borzoi improve phenotype prediction when combined with genotype information, a multimodal model was trained using both protein-coding SNP genotypes and gene-level logSED predictions, together with CAG repeat length and sex. To the best of our knowledge, this was the first time that a phenotype prediction model introduces predictions obtained from a gLM.

The multimodal model achieved significantly better performance than the model with only the expression as shown in Figure 4.1, demonstrating that combining

Borzoi-based expression predictions with genotype data improves classification accuracy. When comparing its performance to the genotype-only model, there is no statistically significant difference observed ($p_{\text{Wilc}}$ = 0.19). However, feature importance analysis indicated that Borzoi-based expression predictions contributed complementary information to the model. Among the top 100 most important features, 44% were expression-based features derived from Borzoi predictions. After averaging importance scores for the same gene across both tissues, the proportion of expression-derived features among the top features decreased to 35%. This suggests that the importance of expression level depends on the tissue.

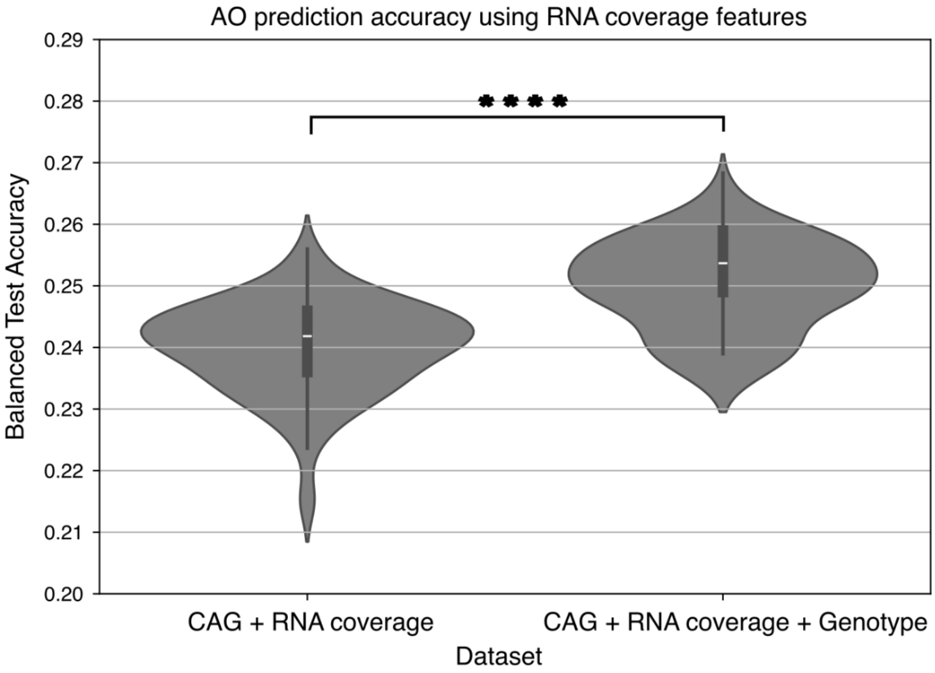

FIGURE 4.1: AO prediction accuracy improves when adding genotype to RNA coverage features.

Functional enrichment analysis of the top-ranked expression features revealed significant enrichment for biological processes such as transcription regulation, DNA binding, protein degradation via the ubiquitin-proteasome system, and brain signaling pathways like glutamate receptor activity. The model showed that several genes involved in these pathways contribute to predicting AO. Notably, two genes ranked highly among expression features. One gene is *GRIK1* which is a glutamate ionotropic receptor involved in neuronal signaling and the other one is *CUL2*, a gene involved in ubiquitin-dependent protein catabolic processes. Both of these processes have previously been linked to HD pathogenesis.

An additional analysis was performed to investigate how the importance of expression features depends on CAG repeat length. Specifically, trees from the multimodal XGBoost model with a maximum depth of 2 were examined. In these trees, CAG repeat length was often selected as the first splitting feature. The expression features that appeared in the second split were then analyzed separately for individuals with larger or smaller CAG expansions. In figure 4.2 for example, the expression of *MED23* was more frequently used for individuals with smaller CAG repeat lengths (less than approximately 45–46 repeats), while genes such as *MT1B* appeared more often for individuals with larger CAG expansions. Similar CAG-dependent effects were also observed for other genes like *DMBX1*, *EXOC3L1*, and *DMBX1*. This means that the relevance of different genes depends on the CAG repeat size.

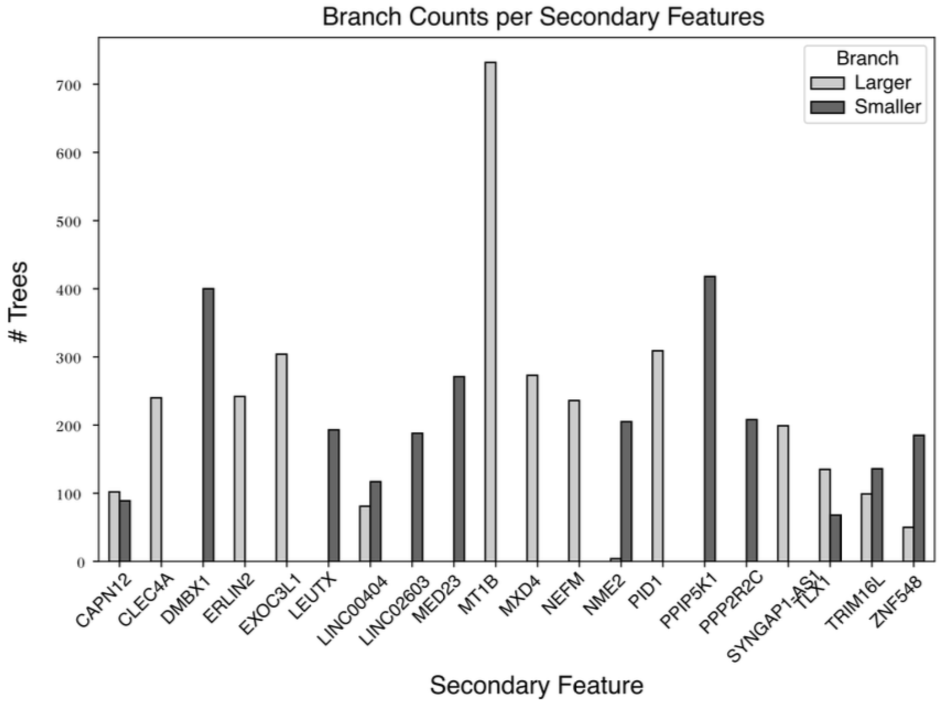

FIGURE 4.2: Number of trees using each gene as a secondary feature, split by branch direction (larger or smaller residual AO).

Overall, while Borzoi-based predictions did not significantly increase predictive accuracy beyond genotype-only models, they contributed biologically meaningful information, identifying candidate regulatory variants, CAG-dependent interactions, and tissue-relevant gene networks involved in HD.

# Chapter 5

# Discussion

In this thesis, we investigated the application of gLMs, specifically Borzoi, for predicting gene expression changes and their relevance for phenotype prediction in HD. Our main objective was to evaluate whether predicted gene expression changes based on a person's genetic variants could help improve predictions of the residual AO, beyond what is explained by CAG repeat length and protein-coding variants.

To address this, we generated tissue-specific gene expression predictions for brain regions relevant to HD, like putamen and caudate, by applying Borzoi to individual genotype data, including both protein-coding SNPs and variants located in regulatory regions like enhancers and promoters. These predictions were integrated into machine learning models built using XGBoost, where we compared three types of models, a baseline genotype-only model, an expression-only model based solely on Borzoi predictions, and a multimodal model combining both types of features.

The expression-only model achieved a balanced accuracy of approximately 0.24, significantly better than random classification. This demonstrates that predicted expression features from Borzoi, derived from individual variants, contain meaningful regulatory information relevant to AO. Although this model did not outperform the genotype-only model, it allowed us to capture functional variation outside of coding regions, particularly in enhancers. Several regulatory variants identified through Borzoi predictions have not been previously described as HD modifiers, suggesting the potential of this approach for highlighting novel non-coding genetic modifiers.

When combining both expression and genotype information in the multimodal model, we observed comparable performance to the genotype-only model ($p_{\text{Wilc}} = 0.19$), consistent with findings from previous work. While no significant improvement in predictive accuracy was achieved, feature importance analysis revealed that Borzoi-derived expression features contributed complementary information, enabling the model to prioritize additional regulatory regions not captured by genotype data alone.

Furthermore, feature analysis indicated that the relevance of specific expression features may depend on CAG repeat length. In particular, certain expression features, such as MED23, were more informative for individuals with smaller CAG expansions, while others, like MT1B, were more relevant for larger CAG repeat lengths. This CAG-dependent effect suggests that genetic modifiers may act through different regulatory pathways depending on the CAG expansion size. This is an observation that has not been extensively described in the HD modifier literature.

Functional enrichment analysis of the top expression features revealed significant enrichment for transcription regulation, DNA binding, ubiquitin ligase binding, and glutamate receptor activity. These are the all pathways that have been implicated in HD pathogenesis. Notably, genes such as *GRIK1*, involved in glutamate signaling, and *CUL2*, involved in ubiquitin-mediated protein degradation, were highlighted as important features contributing to AO variability.

To the best of our knowledge, this represents one of the first applications of genomic language models to produce multimodal phenotype prediction models in HD. This approach provides a framework not only for integrating genotype and regulatory information but also for identifying candidate regulatory variants that may affect gene expression and modify disease onset.

Overall, this work highlights the potential of integrating predicted gene expression features derived from genomic language models with genotype data to improve the understanding of regulatory mechanisms involved in HD, and opens new avenues for exploring the role of non-coding variants in complex neurodegenerative diseases.

# Chapter 6

# Conclusion and Future Work

In this thesis, we explored the application of genomic language models, specifically Borzoi, to study the regulatory impact of genetic variation in HD. By generating tissue-specific gene expression predictions for brain regions relevant to HD, we investigated whether these expression-based features could contribute to predicting residual AO, in addition to traditional genotype information.

Our results show that Borzoi-based expression predictions contain useful regulatory information that can help with phenotype prediction. The expression-only model performed better than random classification, suggesting that non-coding genetic variation carries important information. When combining expression features with genotype data in multimodal models, performance was similar to the genotype-only models. However, adding expression predictions allowed us to identify additional regulatory regions and genes that may influence age of onset. The feature importance and enrichment analyses highlighted biological pathways involved in transcription regulation, DNA binding, protein degradation, and glutamate signaling which are processes that are known to play a role in HD.

We also observed that some expression features were more important depending on CAG repeat length, suggesting that certain modifier effects may depend on the size of the CAG expansion, which should be further studied.

Although the results are promising, several limitations remain. Borzoi was trained on RNA-seq data from healthy individuals like GTEx and ENCODE (Linder et al., 2025), which may not fully capture the transcriptomic dysregulation that occurs in Huntington's Disease. The enhancer-gene assignments used for weighting regulatory variants are based on existing databases such as GeneHancer, which may not be fully complete or fully accurate for all regulatory elements. In addition, while the genotype dataset used in this study is one of the largest available for HD, even larger cohorts may be needed to detect more subtle modifier effects that were not captured here.

In future work, such models could enable large-scale *in silico* perturbations to explore potential gene therapy targets aimed at modulating gene expression to delay symptom onset.

# Bibliography

Benegas, Gonzalo et al. (Sept. 2024). “Genomic Language Models: Opportunities and Challenge”. In: *Trends in Genetics* 24.2, pp. 0168–9525. URL: https://www.cell.com/trends/genetics/abstract/S0168-9525(24)00295-6 (cit. on p. 9).

Cano-Gamez, Eddie and Gosia Trynka (May 2020). “From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases”. In: *Frontiers in Genetics* 11.7675, p. 424. URL: https://www.frontiersin.org/articles/10.3389/fgene.2020.00424/full (cit. on p. 7).

Chandrashekar, Pramod Bharadwaj et al. (2023). “DeepGAMI: deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype–phenotype prediction”. In: *Genome Medicine volume* 15, p. 88. URL: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-023-01248-6 (cit. on p. 9).

Chen, Lihua et al. (2022). “Predicting genotype-specific gene regulatory networks”. In: *Nature Communications* 13, p. 1234. URL: https://www.nature.com/articles/s41467-022-28990-4 (cit. on pp. 8, 19).

Chen, T. and C. Guestrin (2016). “XGBoost: A Scalable Tree Boosting System”. In: *ACM*, pp. 785–794. URL: http://dblp.uni-trier.de/db/conf/kdd/kdd2016.html#ChenG16 (cit. on p. 25).

Consens, Micaela E. et al. (Mar. 2025). “Transformers and genome language models”. In: *Nature Machine Intelligence* 7.4, pp. 346–362. URL: https://www.nature.com/articles/s42256-025-01007-9 (cit. on pp. 10, 14–16).

Dickey, Audrey S. and Albert R. La Spada (2018). “Therapy development in Huntington disease: From current strategies to emerging opportunities”. In: *American Journal of Medical Genetics Part A* 176.4, pp. 842–861. URL: https://pubmed.ncbi.nlm.nih.gov/29218782/ (cit. on p. 18).

Fishilevich, Simon et al. (2017). “GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards”. In: *Database* 2017 (cit. on p. 18).

Fulco, Charles P. et al. (2019). “Systematic mapping of functional enhancer–promoter connections with CRISPR interference”. In: *Science* 354.6313, pp. 769–773. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10520073/ (cit. on p. 8).

Fuses, C. et al. (2025). “Context-Dependent Genetic Modifiers of Huntington’s Disease Revealed through Multimodal Machine Learning”. In: (cit. on p. 25).

Gallagher, Melissa D. and Alice S. Chen-Plotkin (2018). “From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases”. In: *Frontiers in Genetics* 9, p. 424. URL: https://www.frontiersin.org/articles/10.3389/fgene.2018.00424/full (cit. on pp. 8, 19).

Gatto, Emilia M et al. (2020). “Huntington disease: Advances in the understanding of its mechanisms”. In: *PubMed Disclaimer* 158, p. 105482. URL: https://pubmed.ncbi.nlm.nih.gov/34316639/ (cit. on p. 18).

Gusella, James F and Marcy E MacDonald (Aug. 2009). “Huntington’s disease: the case for genetic modifiers”. In: *Genome Medicine* 1.1, p. 80. URL: https://pubmed.ncbi.nlm.nih.gov/19725930/ (cit. on pp. 6, 7).

Huntington's Disease (GeM-HD) Consortium, Genetic Modifiers of (July 2015). "Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease". In: *Cell* 162.3, pp. 516–526. URL: https://www.cell.com/cell/fulltext/S0092-8674(15)00840-5?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867415008405%3Fshowall%3Dtrue (cit. on pp. 7, 8, 19).

James F Gusella Jong-Min Lee, Marcy E MacDonald (Aug. 2021). "Huntington's disease: nearly four decades of human molecular genetics". In: *Human Molecular Genetics* 30.R2, R254–R263. URL: https://pubmed.ncbi.nlm.nih.gov/34169318/ (cit. on p. 5).

Jurcau and Anamaria (June 2022). "Molecular Pathophysiological Mechanisms in Huntington's Disease". In: *Biomedicines* 10.6, https://doi.org/10.3390/biomedicines10061432. URL: https://doi.org/10.3390/biomedicines10061432 (cit. on pp. 3, 5).

Lee, Jieun et al. (2019). "CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset". In: *Neurology* 93.11, e1040–e1047. URL: https://pubmed.ncbi.nlm.nih.gov/31398342/ (cit. on p. 17).

Lee, Renske van der et al. (2022). "Demystifying non-coding GWAS variants". In: *Genome Biology* 23, p. 111. DOI: 10.1186/s13059-022-02661-y. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02661-y (cit. on p. 8).

Li, Xin et al. (Oct. 2017). "The impact of rare variation on gene expression across tissues". In: *Nature* 550.7675, pp. 239–243. URL: https://www.nature.com/articles/nature24267 (cit. on pp. 9, 19).

Linder, Johannes et al. (Aug. 2025). "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation". In: *Nature Genetics* 57.1, pp. 949–961. URL: https://www.nature.com/articles/s41588-024-02053-6 (cit. on pp. 10, 12–14, 19, 35).

Meléndez, Alex et al. (Dec. 2023). "Assessing Tree-Based Phenotype Prediction on the UK Biobank". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 21.4, pp. 1234–1245. URL: https://ieeexplore.ieee.org/document/10385960 (cit. on p. 8).

Network, European Huntington's Disease (2024). *About HD*. Accessed: 2025-06-15. URL: https://ehdn.org/about-hd/ (cit. on p. 5).

Sathe1, Swati et al. (Aug. 2021). "Enroll-HD: An Integrated Clinical Research Platform and Worldwide Observational Study for Huntington's Disease". In: *Networks in Movement Disorders* 12.15. URL: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2021.667420/full (cit. on p. 17).

Vaswani, Ashish et al. (Mar. 2017). "Attention Is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* 7.4, pp. 5998–6008. URL: https://arxiv.org/abs/1706.03762 (cit. on p. 14).