Context-Dependent Genetic Modifiers of Huntington's Disease Revealed through Multimodal Machine Learning

C. Fuses[1-3], G. Zavou[1-3], B. Ros[1-3], J.M. Canals[1-3], and J. Abante[1-3*]

[1]Dept. of Biomedical Sciences, Universitat de Barcelona, Barcelona, Spain
[2]Creatio, Universitat de Barcelona, Barcelona, Spain
[3]Institut de Neurociències (UBNeuro), Universitat de Barcelona, Barcelona, Spain
[*]Corresponding author: jordi.abante@ub.edu

Abstract

Huntington's disease (HD) exhibits substantial variability in age of onset (AO), only partially explained by the length of the CAG repeat in the *HTT* gene. While most studies seeking additional genetic modifiers (GeMs) have relied on linear models, we investigate the potential of non-linear machine learning (ML) approaches, such as tree-based models and graph neural networks (GNNs), to capture complex, context-dependent genetic interactions influencing AO. To address the challenges posed by high-dimensional genotyping data, we introduce a strategy based on gene-specific variational autoencoders for genotype compression. This framework reveals novel modifiers with effects dependent on CAG repeat length, underscoring the importance of accounting for feature interactions. Additionally, we integrate predicted gene expression levels from Borzoi — a genomic language model — into a multimodal prediction architecture. This integration allows us to identify regulatory variants likely to affect AO through expression changes. To our knowledge, this is the first application of a gLM in multimodal genotype-to-phenotype prediction, offering a new paradigm for interpretable modeling of complex traits in HD and related polyglutamine disorders.

1 Introduction

Huntington's disease (HD) is an autosomal dominant hereditary neurodegenerative disorder with symptom onset varying widely across individuals. It is well established that the age of onset (AO) correlates with the length of a CAG trinucleotide expansion in the huntingtin gene (*HTT*) [1], encoding a long glutamine tract. This expansion is unstable, increasing in 80% of intergenerational transmissions [2], meaning intermediate alleles (27–35 repeats) can expand into the pathogenic range. Interestingly, somatic expansions also occur, particularly in brain regions vulnerable to degeneration [3], with striatal neurons showing especially high rates, producing toxic isoforms that drive neuronal death [4]. Somatic expansion is believed to accelerate disease progression, and critical protein interactions seem to underlie its molecular mechanisms [5, 6].

Currently, both genetic testing and AO prediction largely rely on this repeat length. However, the size of this expansion explains only part of the variability, especially for shorter expansions, with the CAG repeat length accounting for 40–70% of AO variability [7, 8]. Thus, researchers have leveraged genome-wide association studies (GWAS) to identify genetic modifiers (GeMs) that can explain the remaining variability, identifying many relevant genes involved in DNA maintenance pathways [9–12]. Nevertheless, these studies were based on simple linear models that generally fail to capture nonlinear effects and interactions.

Recent work has explored the potential of more sophisticated machine learning (ML) approaches for phenotype prediction. For example, a recent benchmark on UKBiobank data compared the

Proceedings of the 20th Machine Learning in Computational Biology meeting (MLCB 2025).

performance of various phenotype prediction ML algorithms, and found that tree-based models are especially effective [13]. HD phenotype prediction over clinical data has also shown the potential of tree-based models due to its explainability [14] and superior performance [15]. These models are particularly useful in this context, since they address the shortcomings of linear models while remaining interpretable. Prediction models based on artificial neural networks (ANN) have also been proposed as a powerful alternative for phenotype prediction [16, 17]. In particular, graph neural networks (GNNs) have caught some attention given their capacity to incorporate prior biological knowledge into the model, such as protein-to-protein interaction (PPI) networks, and their interpretability [18, 19]. Lastly, recent work also explores the usage of variational autoencoders (VAEs) to compress genotype information [20, 21], although the impact of this compression in phenotype prediction models has not been explored.

A common downstream task after identifying associations between genetic variants and a given phenotype is variant effect prediction (VEP), which allows researchers to probe the functional implications of the identified variants. Recently, genomic language models (gLMs), such as Borzoi [22], have emerged as powerful tools to perform VEP. These models are capable of performing tissue-specific RNA coverage predictions from sequence alone. To the best of our knowledge, however, their contribution to predictive genotype-to-phenotype modeling is just beginning to be explored [23]. However, it seems plausible that augmenting phenotype prediction models with RNA coverage predictions would (i) improve the predictive power of the models, (ii) facilitate to the discovery of important variants in regulatory regions, and (iii) allow researchers to identify genes whose expression is critical in the phenotype of study. For example, in HD it is known that the expression levels of certain genes can alter the somatic expansion rate of the CAG triplet [24]. Thus, HD serves as a good case study to evaluate the capacity of gLMs to produce multimodal phenotype prediction models.

It stands to reason that many of these recent advancements could have important implications for phenotype prediction models. Thus, we explore these questions using an HD GWAS dataset as a case study. First, we compare the performance of state-of-the-art phenotype prediction models, showing how tree-based approaches provide superior classification performance. In addition, we show how these models can be used to discover context-dependent GeMs in HD, leading to unique biological insights. To study the impact of genotype compression in phenotype prediction models, we produce embeddings of protein-coding regions using gene-specific VAEs, and evaluate the classification performance of the resulting predictive models, including GNNs encoding PPI information. Finally, we explore the potential of multimodal phenotype prediction models, leveraging gLM RNA coverage predictions. Through our predictions we identify several known GeMs and identify well established GeMs, and we identify novel candidates associated with DNA mismatch repair and transcription regulation activity, a mechanism that has not received as much attention in the HD GeM literature.

2 Methodology

2.1 Genotype Data

The data used for this study was assembled from different GWAS studies by Lee et al. [25], combining data from the GeM-HD Consortium, Enroll-HD and Registry, worldwide observational studies aimed towards the development of therapeutics for HD. The dataset contains whole-genome single nucleotide polymorphisms (SNPs) genotypes of 9,064 patients, their CAG trinucleotide uninterrupted expansion length in number of CAG triplet repeats, and their AO. The genomic information was obtained sequencing blood samples. The CAG expansion rate is much smaller in blood than in the striatum, where the main neurotoxicity events take place, but we use it as an approximation of the initial length of the mutation at birth as commonly done [26].

Since our goal is to identify genes that reduce the unexplained variability after accounting for CAG length, we trained our models on the residuals of a linear model that predicts the AO from CAG length alone. In addition, since machine learning algorithms tend to achieve better results in classification tasks than in regression [27], we decided to formulate the problem as a classification task. We first computed the first and second order linear effects of CAG alone on AO via a linear regression, i.e., $\text{AO} \sim \text{CAG} + \text{CAG}^2$ (Fig. 1A), and then stratified the residuals into 5 quantiles, grouping samples based on how much their AO deviates from the expected value by their blood CAG length (Fig. 1B), resulting in a balanced dataset with 5 classes.

To reduce the computational burden, we first created a list of protein-coding genes from gene ontology (GO) terms related to processes that could potentially be involved in HD pathogenesis [28, 29], motivated by studies that show how critical protein interactions underlie the molecular

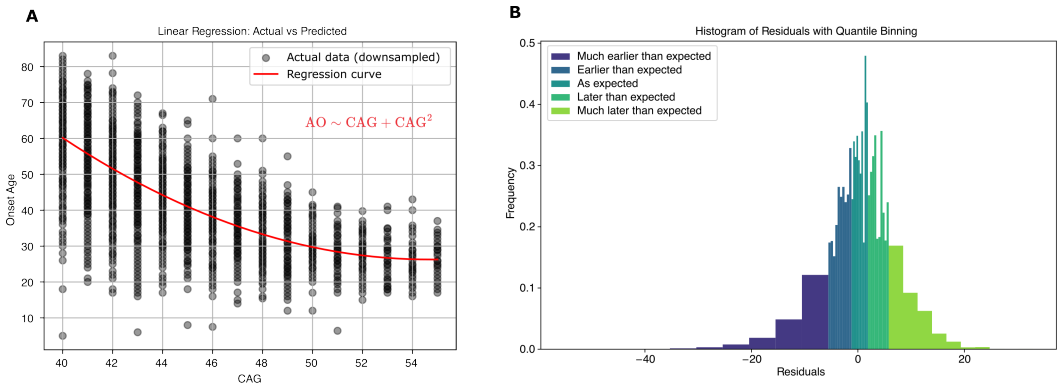

Figure 1: *Age of onset in Enroll-HD subjects. (A) AO linear regression represented over distribution of samples over their CAG length and AO (left), and residuals of the regression (right). (B) Histogram of AO residuals binned into 5 balanced classes.*

mechanism driving processes like somatic expansion [5]. We used this list to filter the SNPs, resulting in a total of 339,886 SNPs pertaining to 2,774 protein-coding genes.

2.2 Non-linear Dimensionality Reduction

Among our goals, we wanted to (i) study the impact of dimensionality reduction in predicting a phenotype, and (ii) develop models based on geometric deep learning. To that end, we took advantage of the correlative nature of genotyping data given linkage-disequilibrium (LD) blocks, clusters of genetically correlated variants (Fig. S3A). We assembled a second version of the dataset encoding the genotypes using a non-linear dimensionality reduction technique [20, 30]. In particular, we trained a variational autoencoder (VAE) for each gene, compressing the dimensionality to 30 dimensions for those genes with more than 30 SNPs, resulting in a second dataset with a total of 20,111 features, representing a compression of $\sim$95% with respect to the original dataset. The latent dimension of the autoencoder for each gene was chosen as an approximation of the number of LD blocks within that gene (Fig. S3C). To estimate this, we computed the correlation matrix of the gene's variants and computed the spectral decomposition. All eigenvalues greater than 20% of the largest eigenvalue were counted, and this count is what we used to estimate the number of blocks (Fig. S3B).

The encoder first reduces the gene's genotype dimension by 50% through a linear projection. From this intermediate representation, it further compresses to the latent dimension (estimated from the number of LD blocks), producing the mean and standard deviation of the approximate posterior. The decoder reconstructs the input from latent samples using a symmetric upscaling path, yielding genotype probabilities modeled with a relaxed Bernoulli distribution over $[0, 1]$. These probabilities are then discretized by binning them into 3 categories corresponding to possible genotype (0, 1 or 2 for reference homozygous, heterozygous or alternative homozygous respectively). Being d the decoder, z the latent variable and x the observed data, the likelihood is a relaxed Bernoulli $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) = \text{RelBer}(x \mid \sigma(d_\theta(z)))$, where $\sigma(\cdot)$ is a sigmoid function. The minimized loss is the Evidence Lower Bound (ELBO), which is a combination of the reconstruction loss $L_{\boldsymbol{\theta}}(x)$ and the Kullback-Leibler divergence D_{KL}:

$$\text{ELBO} = -\left(L_{\boldsymbol{\theta}}(x) + \cdot D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p_{\boldsymbol{\theta}}(\boldsymbol{z})\right)\right) . \tag{1}$$

2.3 Prediction Models

The classification models are created to solve the task of predicting the residual age of onset bin $\mathbf{Y} \in \{0, 1, ..., C-1\}^n$, where $C = 5$ is the total number of classes (Fig. 1B), representing how much a sample deviates from the expected AO given solely by CAG expansion length. We let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the feature matrix, where n and d are the number of samples and features, respectively. The feature matrix $\mathbf{X}$ can consist of a table with all SNPs from the selected genes as individual features, or a tensor where we represent genes by a 30-dimensional vector. In each case we consider a linear model, given by a regularized multinomial logistic regression that we fit via penalized maximum likelihood similar to SNPnet [31], and XGBoost [32], a tree-based method that has shown great performance in similar settings [13]. In addition, we also consider deep geometric learning algorithms trained on the learned representations through non-linear dimensionality reduction.

2.3.1 Multinomial Logistic Regression

The linear model is a Multinomial Logistic Regression (MLR) with *L1* regularization to implicitly perform feature selection. Unnormalized scores for each prediction are computed as $Z = XW + b$, and from these we compute softmax probabilities:

$$P(y_i = c \mid \mathbf{x}_i) = \frac{\exp(z_{ic})}{\sum_{k=1}^{C} \exp(z_{ik})} \,. \tag{2}$$

The minimized loss is the negative log-likelihood plus the regularization term:

$$\mathcal{L}(\mathbf{W}, \mathbf{b}) = -\frac{1}{n} \sum_{i=1}^{n} \log(\hat{p}_i) + \lambda \|\mathbf{W}\|_1, \qquad \|\mathbf{W}\|_1 = \sum_{j=1}^{d} \sum_{c=1}^{C} |w_{jc}| \,. \tag{3}$$

We optimized hyper-parameter λ by performing a grid search with 5-fold cross validation, assessing the performance via a balanced accuracy score (BA), i.e., the mean of recall obtained in each class (Appendix A.1). To prevent overfitting, we implemented early stopping based on a validation set, and testing of the resulting model was performed with 20% of the samples left out from training. To interpret the model and identify candidate GeMs, we performed feature importance by adding the 5 optimized weights corresponding to each class. These values were ranked in descending order, and we considered the top ones as the most important features contributing to the model classification.

2.3.2 XGBoost

In XGBoost [32] the model makes predictions by summing K regression trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}, \quad \mathcal{F} = \{\text{regression trees}\} \,. \tag{4}$$

Each regression tree $f_k(x_i) \in \mathbb{R}^K$ output a vector of logits (one per class), which go through a softmax function. The training objective to minimize is regularized:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{\mathbf{y}}_i^{(t)}) + \sum_{k=1}^{t} \omega(f_k) \,, \tag{5}$$

where $\hat{\mathbf{y}}_i^{(t)}$ is the softmax prediction of the sum of the outputs of all trees $f_k(\mathbf{x}_i)$ for a given sample i at step t, $\ell(y_i, \hat{\mathbf{y}}_i)$ is the multi-class log loss, and $\omega(f_k)$ is the tree complexity regularization. The boosting strategy consists in adding a new function f_t at each round to minimize the regularized objective through additive training. This is done until validation loss does not improve for a certain amount of consecutive rounds to avoid overfitting. Tree depth and learning rate were optimized through a grid search with 5-fold cross validation, using BA as the performance metric. In this case, we used the gain value of each feature to assess feature importance. This value quantifies how much a split on a certain feature reduces the loss function $\mathcal{L}$. The higher the gain, the better the split, and hence the feature used to split is more informative. We rank SNPs and genes by their overall gain in the model, and inspect the structure of the individual trees in the booster, calculating at which level the most relevant features are used, and getting the distribution of their splitting values. This provides insight into the interactions explaining de deviation of AO from the expected mean.

2.3.3 Geometric Deep Learning Models

It has recently been hypothesized that critical protein interactions underlie the molecular mechanisms driving somatic expansion in Huntington's disease [5]. This suggests that incorporating protein-protein interaction (PPI) data could enhance the identification of affected pathways that either hasten or delay disease onset, a strategy already explored in other diseases [18] and in related modalities such as gene expression [33]. Geometric models are neural networks that use graph data as input. At a high level, Graph Neural Networks (GNNs) learn representations of graphs by aggregating information from neighboring nodes with its own, what is known as message passing. A layer l, node i updates its representation via:

$$\mathbf{h}_i^{(l)} = \text{Update}^{(l)} \left(\mathbf{h}_i^{(l-1)}, \text{Aggregate}^{(l)} \left(\{ \mathbf{h}_j^{(l-1)} \mid j \in \mathcal{N}(i) \} \right) \right) \,, \tag{6}$$

where $\mathcal{N}(i)$ is the set of neighbors of node i, the aggregate function is summation and the update is a nonlinear layer (ReLU). In particular, we considered the two most broadly used types of GNN

in this setting, namely Convolutional (ConvGNN) and Graph Attention Networks (GAT), using *PyTorch Geometric* [34]. The implemented architecture can be found in Appendix A.2. Both GNN methods produce a reduced graph, which is then summarized into a single vector using global mean pooling (by averaging node encodings component-wise). This graph-level representation is passed through a fully connected layer. The output of this layer is then concatenated with the CAG repeat length and sex features. Finally, the combined vector is fed into a last fully connected layer with a softmax activation to compute the class probabilities $\hat{p}_{ic}$. We trained these models optimizing the cross-entropy:

$$\mathcal{L}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} p_{ic}\log(\hat{p}_{ic})\ . \tag{7}$$

The graphs that were used to train these models are PPIs between the genes included in this study, as obtained from StringDB [35]. Each graph corresponds to a single sample: nodes represent genes, and weighted edges reflect experimentally derived probabilities of physical interactions between their protein products. Compressed embeddings for each gene are used as node features. GNNs are well-suited to this structure, enabling us to assess the value of incorporating PPI information into genotype-to-phenotype prediction.

2.3.4 Genomic Language Model

The genotype data used to assemble our genotype dataset comes from protein-coding regions of the curated set of genes, as it has been recently hypothesized that critical protein interactions underlie the molecular mechanism driving the somatic expansion [5]. Nevertheless, genetic variation in regulatory regions, such as enhancers, can directly affect the expression level of these proteins. As expression is highly tissue-specific, we hypothesized that predicting the differential expression in the most affected tissues in HD (putamen and caudate) would give us information closer to the real molecular context taking place in HD brains. To that end, we took advantage of recent developments in genomic language models (gLMs) and used Borzoi [22] to predict tissue-specific RNA sequencing coverage given a genotype, allowing us to evaluate how alternative variants affect expression. We used this tool to predict tissue-specific log sum of expression differences (logSED) in both the putamen and caudate for each subject for each variant. The resulting scores were weighted using a Gaussian kernel taking into consideration their relative position to the center of the corresponding regulatory element (Fig. S4A). Then, we studied how augmenting the set of features with these predictions affects the accuracy and interpretability of our predictive models.

3 Results

3.1 Novel context-dependent genetic modifiers identified by XGBoost

First, we compared the classification performance of the MLR and XGBoost models when trained with the SNP dataset, consisting of a total of 339,888 features (including CAG length and sex). To that end, we computed the BA for each random partition of the data and we found that XGBoost significantly outperforms MLR (Fig. 2A). This is consistent with previous findings, showing that tree-based approaches outperform traditional linear methods in phenotype prediction tasks [13], which we owe to the fact that these models can implicitly take into consideration nonlinear effects and interactions, while remaining interpretable. As a result, we inspected the best performing model to (i) identify previously reported GeMs as positive controls, and (ii) to identify novel GeM candidates. Among the top 100 SNPs, we found 19 genes that have been previously identified in HD, such as DNA repair genes *FAN1*, *PMS2* or *MLH1*. In addition, we identified several novel candidates that have not been previously described to the best of our knowledge, such as *NRG1*, *ZFHX3* and *RORA* (Table S1). Sex was not an important feature in any model. To investigate which biological processes can influence AO, we performed GO enrichment analysis (Fig. 2C). As expected, we found that SNPs belonging to DNA maintenance were significantly overrepresented among the top 100 SNPs. Furthermore, we found that SNPs mapping to genes involved in folic acid metabolism were also significantly overrepresented. This is consistent with recent findings that regulation of one-carbon metabolism could affect the progression of HD [36], although this has not been previously described in the context of GWAS.

We then investigated the order in which the individual boosters used the most relevant features as well as their interactions to identify GeMs relevant at different CAG expansion lengths. We found that trees would first split subjects based on their CAG length in 35% of the cases, even when working

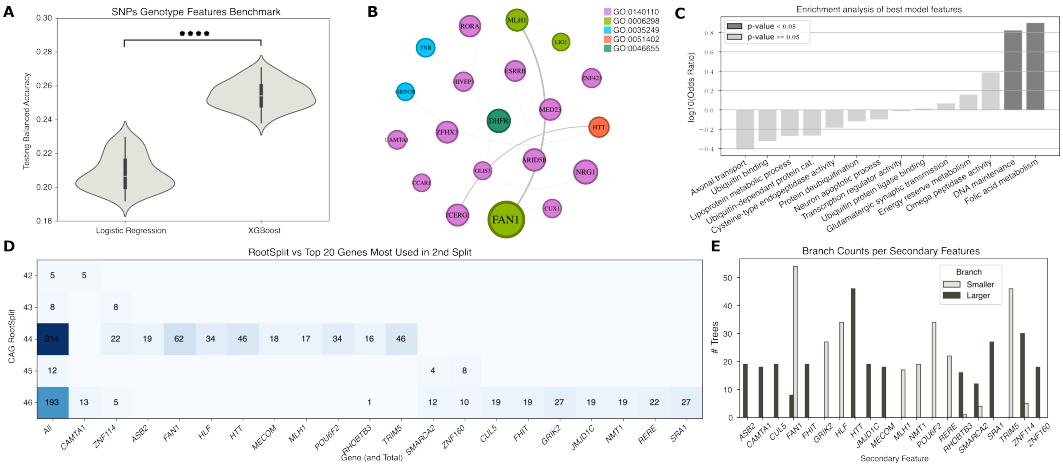

Figure 2: *State-of-the art benchmark. (A) Violin plot of balanced accuracy of the MLR and XGBoost models, showing the superior performance of the latter ($p_{Wilc} = 2e^{-15}$). (B) Protein-interaction network of the top 20 genes in the best performing XGBoost model colored by gene ontology (GO) term (node size reflects highest SNP gain). (C) Gene ontology results of top 100 SNPs for best performing XGBoost model. (D) Genes used by the best performing XGBoost model after splitting by CAG length showing that some of the most informative genes are a function of this length. (E) Frequency of how many trees with CAG at the root node use each feature for smaller or larger CAG expansions.*

with residuals of the CAG linear model, pointing at the existence of important interactions between this feature and genetic variants. In addition, looking at the SNPs that the model was using for subsequent splits, we found that several SNPs heavily depended on CAG length (Fig. 2D), suggesting that different biological processes might be relevant depending on this length. We also observed this effect when looking across all XGBoost models trained with the same dataset (Fig. S1B). We further looked into possible directionality of this effect, and found that features are consistently only used by the model when the repeat expansion is below a CAG threshold and another subset that is used above a CAG threshold (Fig. 2E). For example, we found DNA repair genes like *FAN1* and *MLH1* are consistently used by models to further split subjects with fewer than 44 CAG repeats. Indeed, genes associated with SNPs used to split subjects with shorter CAG repeat lengths were heavily enriched for DNA maintenance (Fig. S2A), whereas genes associated with larger CAG lengths were enriched for folic acid metabolism, neuron apoptotic process and ubiquitin-dependent protein catabolic process (Fig. S2B). Importantly, these observations would have not been possible when using standard linear models since such interactions are generally not included.

3.2 Efficient phenotype prediction models via nonlinear embeddings

To study the impact that nonlinear dimensionality reduction has in predicting the AO in HD, we produced a compressed dataset consisting of subject-specific gene embeddings that we later used to train our classifiers. To that end, we took advantage of the correlative nature of SNPs, defining the latent dimension on the basis of the eigenvalues of the correlation matrix (Fig. 3A). For each gene we trained various VAE models and, in each case, we kept the best performing one evaluating them on the test set, resulting in BA $> 95\%$, although the median accuracy decreased for non-reference genotypes (Fig. 3B).

We then trained classifiers using the resulting embeddings and compared the performance with classifiers trained on the SNP dataset. Interestingly, we found no differences in the median performance of the models, regardless of the classifier model used, suggesting that the learned representation encodes the essential information used by the previous models (Fig. 3C). Notably, we found that the GNN models outperformed the linear classifier. We owe this to the fact that, similar to tree-based method, GNN models can capture non-linearities and interactions between features. Nevertheless, the performance did not improve that of XGBoost, which remained the best performing model.

We next studied the set of genes deemed important among XGBoost models, resulting in a set of genes for the k-th model in each case $\mathcal{G}_{\text{SNP},k}$ and $\mathcal{G}_{\text{Emb},k}$, respectively. Comparing the set of genes obtained for the same feature set, we found that both classes of models were equally consistent, resulting in Jaccard indexes (JI) of $\text{med}(\text{JI}(\mathcal{G}_{\text{SNP},i}, \mathcal{G}_{\text{SNP},j})) = 0.053$ and $\text{med}(\text{JI}(\mathcal{G}_{\text{Emb},i}, \mathcal{G}_{\text{Emb},j})) = 0.053$, respectively. However, we did find smaller consistency when we compared the set of genes obtained

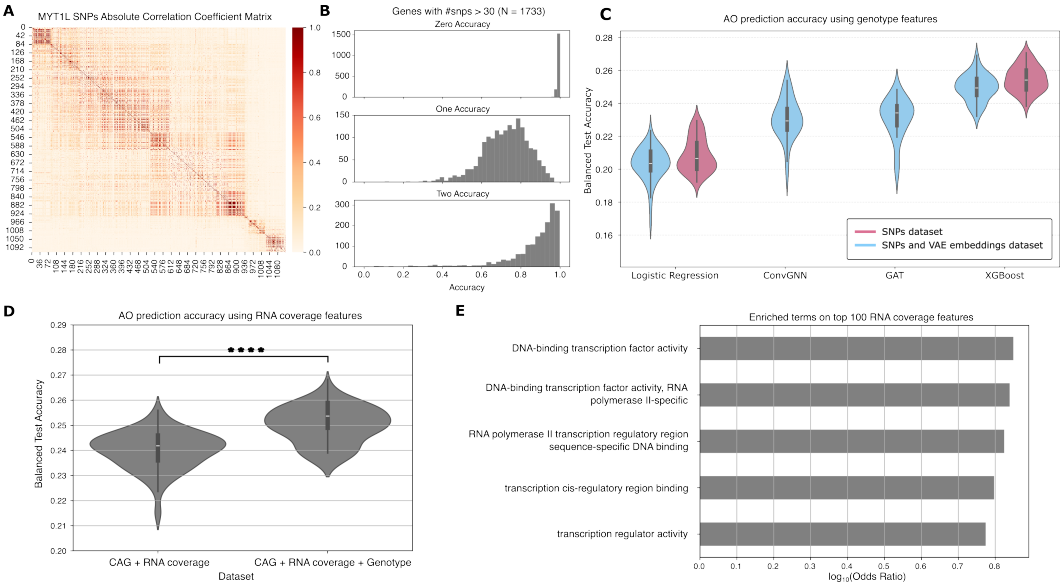

Figure 3: *Gene embedding and geometric deep learning. (A) SNP correlation matrix for gene MYT1L, showing a clearly defined block structured that we use to define the latent dimension. (B) Reconstruction accuracies for homozygous reference alleles, heterozygous alleles, and homozygous alternatives alleles. (C) Balanced accuracy for models with the original SNP dataset (pink) and the compressed version (blue). (D) Balanced accuracy for XGBoost models using RNA coverage prediction features without genotype features (left) and with genotype features (right), showing a better outcome with genotype information ($p_{Wilc} = 5e^{-12}$). (E) Top 5 enriched functionality terms among the expression features used by multimodal models.*

in the model trained on the compressed set of features with that of the model trained on the SNPs, producing a med(JI)$(\mathcal{G}_{\text{Emb},i}, \mathcal{G}_{\text{SNP},j}) = 0.023$, suggesting the models are finding slightly different sets of genes depending on the feature set used.

Lastly, we found that the XGBoost classifier trained on the compressed dataset was 5.5-fold faster and required 12.6% of the memory used by the model trained on the full dataset on average. Thus, although the set of identified genes slightly differs between models trained with SNP and the embeddings, training models with the latter results in similar prediction accuracy at a fraction of the time and computational cost.

3.3 Multimodal phenotype prediction model identified critical variants

Recent experiments showed the importance of GeMs by producing knockout experiments. These experiments showed that the expression of previously described GeMs, such as *FAN1* or *MLH1*, directly affect the somatic expansion rate [37, 10]. Thus, we hypothesized that including gene expression information the performance of the models could improve. To that end, we used Borzoi to score variants located in enhancers, extracted from GeneHancer [38], regulating the set of protein-coding genes being used. We also scored variants situated in the promoter region of each gene, defined as a 4 kbp window centered at the transcription start site (TSS). For each variant, Borzoi predicts the log fold change (LFC) that an alternative allele has on the expression with respect to the provided reference allele. Thus, for a given patient, we assembled an expression vector concatenating the aggregated logSED of both the putamen and caudate predictions.

First, we sought to evaluate whether these predictions alone had any predictive power over the classes of interest. To that end, we trained an XGBoost model using these predictions as input features, alongside the expansion length and sex. Interestingly, we found that the resulting models had a median BA of 0.242, significantly above what's expected by chance (p_{Binom}=0.2). In addition, we found that models with depth 2 (24% of trained all trained models) on average had 65% of boosters using CAG as the first splitting node feature, the same interaction we had observed in the genotype-only XGBoost models. Across the variants in regulatory regions affecting these genes, we identified 3 new variants producing important changes in terms of gene expression in brain tissue in enhancer loci, including 19_50651485_A_C (rs1809186999), 5_60241142_G_A, and 1_157069597_G_A that have not been previously reported.

To assess the utility of gLM predictions in phenotype prediction models, we trained a series of XGBoost models using both genotype and expression information as features and evaluated how these predictions affected the classification performance (Fig. 3D). We found that the multimodal models generally performed significantly better than the models trained on RNA coverage predictions alone. The achieved accuracies by the former were comparable to those of the models trained on genotype data alone (p_{Wilc}=0.19). Nevertheless, feature importance analysis revealed that 44% of the 100 top-ranked features across all multimodal models were RNA coverage features, suggesting that these predictions hold meaningful information towards classifying subjects. Interestingly, when averaging the importance score for the same gene in both tissues we found that this proportion among the top ranking features was reduced to 35%, meaning that the importance of expression level is dependent on tissue. An ordered functional enrichment analysis over the expression features used by the multimodal models reveals the top 5 terms are related to transcription and DNA binding processes (Fig. 3E). Among the significantly enriched molecular functions we also find ubiquitin protein ligase binding and glutamate-gated receptor activity. Among the top 100 most important features was the expression of *GRIK1*, a glutamate ionotropic receptor, and *CUL2*, a gene involved in ubiquitin-dependent protein catabolic processes. Among the models of depth 2, we observe the same CAG length dependency effect for expression features as we saw for genotype features (Fig S5). For example, the expression levels of the transcription regulation mediator *MED23* is only considered for the classification in individuals with CAG less than 45-46 repeats.

4 Discussion

Here, we present a comprehensive analysis of age of onset prediction from genotype data. We first show how tree-based methods offer a valuable alternative to traditional linear methods, offering superior prediction accuracy, while remaining interpretable. The latter is particularly important in our setting, where we seek to identify GeMs. We identify known GeMs and discover new ones related to DNA transcription regulation, a biological process that has not gotten as much attention as DNA maintenance in the HD GeM literature. In addition, we show how GeMs can be context-dependent in HD, with different genes and mechanisms modifying the course of the disease depending on the CAG length measured in blood. We hypothesize this different context-dependency could potentially be related to the different regimes in spiny projection neuron phenotypes recently described [4], triggering different mechanisms. Nevertheless, to the best of our knowledge, this is the first time this context-dependency is described in the context of GeMs and could be potentially relevant in other polyglutamine diseases as well, such as spinocerebellar ataxia or spinal and bulbar muscular atrophy.

We also investigate the impact that non-linear dimensionality reduction of genotype data has on our phenotype models. Interestingly, we show that the classification performance of the models performance is not greatly impacted, even with a compression rate of about 95% which, in turn, allows for faster computations. Nevertheless, phenotype prediction models trained on such feature would require subsequent fine-mapping, similar to an admixture mapping setting. We also show how these embeddings can be used as node features in two different types of GNN architectures that we train to do classification on protein-interaction graphs. Although these models did outperform the linear phenotype prediction model, probably due to the possibility to model non-linear effects and interactions, they did not surpass XGBoost in this case. This suggests that while GNNs offer advantages in modeling complex interactions, XGBoost remains a strong baseline due to its robustness and computational efficiency.

Finally, we explored the possibility of augmenting the models with gene expression predictions produced by a state-of-the-art genomic language model. We show that these predictions alone have some predictive power. However, when combined with the genotype in protein-coding regions, we found the resulting models perform just as good as the best unimodal (genotype) XGBoost models. Nevertheless, we show that the multimodal model allows us to identify genes for which the expression level influences the outcome which, in turn, can be traced back to alternative variants in regulatory regions, such as enhancers, driving the change in expression. To the best of our knowledge, this is the first time that gLM predictions have been used to produce a multimodal phenotype prediction model. Interestingly, such a model could be used to perform *in silico* perturbations at scale to study possible gene therapies directed at affecting the expression of specific genes and directly assessing their effect on the phenotype of interest.

References

[1] M Duyao, A Lazzarini, A Falek, W Koroshetz, D Sax, E Bird, J Vonsattel, E Bonilla, J Alvir, J Bickham Conde', J.-H Cha, L Dure, F Gomez, M Ramos, J Sanchez-Ramos, S Snodgrass, M De Young, N Wexler, C Moscowitz, G Penchaszadeh, H Macfarlane, M Anderson, B Jenkins, J Srinidhi, G Bames, J Gusella, and & M Macdonald. Trinucleotide repeat length instability and age of onset in Huntington's disease, 1993. URL `http://www.nature.com/naturegenetics`.

[2] Vanessa C Wheeler, Wojtek Auerbach, Jacqueline K White, Jayalakshmi Srinidhi, Anna Auerbach, Angela Ryan, Mabel P Duyao, Vladimir Vrbanac, Meredith Weaver, James F Gusella, Alexandra L Joyner, and Marcy E Macdonald. Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse, 1999. URL `https://academic.oup.com/hmg/article/8/1/115/2356046`.

[3] Laura Kennedy, Elizabeth Evans, Chiung Mei Chen, Lyndsey Craven, Peter J. Detloff, Margaret Ennis, and Peggy F. Shelbourne. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Human Molecular Genetics*, 12:3359–3367, 12 2003. ISSN 09646906. doi: 10.1093/hmg/ddg352.

[4] Robert E. Handsaker, Seva Kashin, Nora M. Reed, Steven Tan, Won-Seok Lee, Tara M. McDonald, Kiely Morris, Nolan Kamitaki, Christopher D. Mullally, Neda R. Morakabati, Melissa Goldman, Gabriel Lind, Rhea Kohli, Elisabeth Lawton, Marina Hogan, Kiku Ichihara, Sabina Berretta, and Steven A. McCarroll. Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease. *Cell*, 188:623–639.e19, 2 2025. ISSN 00928674. doi: 10.1016/j.cell.2024.11.038. URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867424013795`.

[5] Ricardo Mouro Pinto, Ella Dragileva, Andrew Kirby, Alejandro Lloret, Edith Lopez, Jason St. Claire, Gagan B. Panigrahi, Caixia Hou, Kim Holloway, Tammy Gillis, Jolene R. Guide, Paula E. Cohen, Guo Min Li, Christopher E. Pearson, Mark J. Daly, and Vanessa C. Wheeler. Mismatch Repair Genes Mlh1 and Mlh3 Modify CAG Instability in Huntington's Disease Mice: Genome-Wide and Candidate Approaches. *PLoS Genetics*, 9, 10 2013. ISSN 15537390. doi: 10.1371/journal.pgen.1003930.

[6] Marc Ciosi, Alastair Maxwell, Sarah A. Cumming, Davina J. Hensman Moss, Asma M. Alshammari, Michael D. Flower, Alexandra Durr, Blair R. Leavitt, Raymund A.C. Roos, Peter Holmans, Lesley Jones, Douglas R. Langbehn, Seung Kwak, Sarah J. Tabrizi, and Darren G. Monckton. A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. *EBioMedicine*, 48:568–580, 10 2019. ISSN 23523964. doi: 10.1016/j.ebiom.2019.09.020.

[7] Jian-Liang Li, Michael R Hayden, Elisabeth W Almqvist, Ryan R Brinkman, Alexandra Durr, Catherine Dodé, Patrick J Morrison, Oksana Suchowersky, Christopher A Ross, Russell L Margolis, Adam Rosenblatt, Estrella Gómez-Tortosa, David Mayo Cabrero, Andrea Novelletto, Marina Frontali, Martha Nance, Ronald J A Trent, Elizabeth Mccusker, Randi Jones, Jane S Paulsen, Madeline Harrison, Andrea Zanko, Ruth K Abramson, Ana L Russ, Beth Knowlton, Luc Djoussé, Jayalakshmi S Mysore, Suzanne Tariot, Michael F Gusella, Vanessa C Wheeler, Larry D Atwood, L Adrienne Cupples, Marie Saint-Hilaire, Jang-Ho J Cha, Steven M Hersch, Walter J Koroshetz, James F Gusella, Marcy E Macdonald, and Richard H Myers. A Genome Scan for Modifiers of Age at Onset in Huntington Disease: The HD MAPS Study, 2003.

[8] L. Djoussé, B. Knowlton, M. Hayden, E. W. Almqvist, R. Brinkman, C. Ross, R. Margolis, A. Rosenblatt, A. Durr, C. Dode, P. J. Morrison, A. Novelletto, M. Frontali, R. J.A. Trent, E. McCusker, E. Gómez-Tortosa, D. Mayo, R. Jones, A. Zanko, M. Nance, R. Abramson, O. Suchowersky, J. Paulsen, M. Harrison, Q. Yang, L. A. Cupples, J. F. Gusella, M. E. MacDonald, and Richard H. Myers. Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. *American Journal of Medical Genetics*, 119 A:279–282, 6 2003. ISSN 15524825. doi: 10.1002/ajmg.a.20190.

[9] Jong Min Lee, Vanessa C. Wheeler, Michael J. Chao, Jean Paul G. Vonsattel, Ricardo Mouro Pinto, Diane Lucente, Kawther Abu-Elneel, Eliana Marisa Ramos, Jayalakshmi Srinidhi Mysore, Tammy Gillis, Marcy E. MacDonald, James F. Gusella, Denise Harold, Timothy C. Stone, Valentina Escott-Price, Jun Han, Alexey Vedernikov, Peter Holmans, Lesley Jones, Seung Kwak, Mithra Mahmoudi, Michael Orth, G. Bernhard Landwehrmeyer, Jane S. Paulsen, E. Ray Dorsey, Ira Shoulson, and Richard H. Myers. Identification of genetic factors that modify clinical onset of huntington's disease. *Cell*, 162:516–526, 8 2015. ISSN 10974172. doi: 10.1016/j.cell.2015.07.003.

[10] Nan Wang, Shasha Zhang, Peter Langfelder, Lalini Ramanathan, Fuying Gao, Mary Plascencia, Raymond Vaca, Xiaofeng Gu, Linna Deng, Leonardo E Dionisio, Ha Vu, Emily Maciejewski, Jason Ernst, Brinda C Prasad, Thomas F Vogt, Steve Horvath, Jeffrey S Aaronson, Jim Rosinski, and X William Yang. Distinct mismatch-repair complex genes set neuronal cag-repeat expansion rate to drive selective pathogenesis in hd mice. *Cell*, 2 2025. ISSN 1097-4172. doi: 10.1016/j.cell.2025.01.031. URL `http://www.ncbi.nlm.nih.gov/pubmed/39938516`.

[11] Peter A. Holmans, Thomas H. Massey, and Lesley Jones. Genetic modifiers of mendelian disease: Huntington's disease and the trinucleotide repeat disorders, 10 2017. ISSN 14602083.

[12] James F. Gusella, Jong Min Lee, and Marcy E. Macdonald. Huntington's disease: Nearly four decades of human molecular genetics, 10 2021. ISSN 14602083.

[13] Alex Meléndez, Cayetana López, David Bonet, Gerard Sant, Ferran Marqués, Manuel Rivas, Daniel Mas Montserrat, Jordi Abante, and Alexander G Ioannidis. Assessing tree-based phenotype prediction on the uk biobank. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023. doi: 10.1109/BIBM58861.2023.10385960.

[14] Jinnie Ko, Hannah Furby, Xiaoye Ma, Jeffrey D. Long, Xiao-Yu Lu, Diana Slowiejko, and Rita Ghandy. Clustering and prediction of disease progression trajectories in huntington's disease: An analysis of enroll-hd data using a machine learning approach. *Frontiers in Neurology*, 1 2023. doi: 10.3389/fneur.2022.1034269.

[15] Jasper Ouwerkerk, Stephanie Feleus, Kasper F. van der Zwaan, Yunlei Li, Marco Roos, Willeke M.C. van Roon-Mom, Susanne T. de Bot, Katherine J. Wolstencroft, and Eleni Mina. Machine learning in huntington's disease: exploring the enroll-hd dataset for prognosis and driving capability prediction. *Orphanet Journal of Rare Diseases*, 18, 12 2023. ISSN 17501172. doi: 10.1186/s13023-023-02785-4.

[16] Blaise Hanczar, Farida Zehraoui, Tina Issa, and Mathieu Arles. Biological interpretation of deep neural network for phenotype prediction based on gene expression. *BMC Bioinformatics*, 21, 12 2020. ISSN 14712105. doi: 10.1186/s12859-020-03836-4.

[17] Arno van Hilten, Jeroen van Rooij, Bastiaan T. Heijmans, Peter A.C. 't Hoen, Joyce van Meurs, Rick Jansen, Lude Franke, Dorret I. Boomsma, René Pool, Jenny van Dongen, Jouke J. Hottenga, Marleen M.J. van Greevenbroek, Coen D.A. Stehouwer, Carla J.H. van der Kallen, Casper G. Schalkwijk, Cisca Wijmenga, Sasha Zhernakova, Ettje F. Tigchelaar, P. Eline Slagboom, Marian Beekman, Joris Deelen, Diana van Heemst, Jan H. Veldink, Leonard H. van den Berg, Cornelia M. van Duijn, Bert A. Hofman, Aaron Isaacs, André G. Uitterlinden, P. Mila Jhamai, Michael Verbiest, H. Eka D. Suchiman, Marijn Verkerk, Ruud van der Breggen, Jeroen van Rooij, Nico Lakenberg, Hailiang Mei, Maarten van Iterson, Michiel van Galen, Jan Bot, Peter van 't Hof, Patrick Deelen, Irene Nooren, Matthijs Moed, Martijn Vermaat, René Luijk, Marc Jan Bonder, Freerk van Dijk, Wibowo Arindrarto, Szymon M. Kielbasa, Morris A. Swertz, Erik W. van Zwet, M. Arfan Ikram, Wiro J. Niessen, Joyce B.J. van Meurs, and Gennady V. Roshchupkin. Phenotype prediction using biologically interpretable neural networks on multi-cohort multi-omics data. *npj Systems Biology and Applications*, 10, 12 2024. ISSN 20567189. doi: 10.1038/s41540-024-00405-w.

[18] Riccardo Smeriglio, Joana Rosell-Mirmi, Petia Radeva, and Jordi Abante. Leveraging protein-protein interactions in phenotype prediction through graph neural networks. In *2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8. IEEE, 8 2024. ISBN 979-8-3503-5663-2. doi: 10.1109/CIBCB58642.2024.10702170. URL https://ieeexplore.ieee.org/document/10702170/.

[19] Jose L. Mellina Andreu, Luis Bernal, Antonio F. Skarmeta, Mina Ryten, Sara Álvarez, Alejandro Cisterna García, and Juan A. Botía. Phenolinker: Phenotype-gene link prediction and explanation using heterogeneous graph neural networks. *arXiv*, 2 2024. URL http://arxiv.org/abs/2402.01809.

[20] Margarita Geleta, Daniel Mas Montserrat, Xavier Giro-I-Nieto, and Alexander G Ioannidis. Deep Variational Autoencoders for Population Genetics. *arXiv*, 2023. doi: 10.1101/2023.09.27.558320. URL https://doi.org/10.1101/2023.09.27.558320.

[21] Gizem Taş, Timo Westerdijk, Eric Postma, Jan H. Veldink, Alexander Schonhuth, and Marleen Balvert. Computing linkage disequilibrium aware genome embeddings using autoencoders. *Bioinformatics*, 40, 6 2024. ISSN 13674811. doi: 10.1093/bioinformatics/btae326.

[22] 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*, pages 1–13, 1 2025. ISSN 1546-1718. doi: 10.1038/s41588-024-02053-6. URL https://www.nature.com/articles/s41588-024-02053-6.

[23] Zihao Wu, Zhengliang Liu, Minheng Chen, Yanjun Lyu, Lu Zhang, Jing Zhang, Yiwei Li, Wei Ruan, Xiaowei Yu, Chao Cao, Tong Chen, Yan Zhuang, Xiang Li, Rongjie Liu, Chao Huang, Wentao Li, Tianming Liu, and Dajiang Zhu. GP-GPT: Large Language Model for Gene-Phenotype Mapping. *arXiv*, 2024. doi: 10.48550/arXiv.2409.09825. URL https://www.researchgate.net/publication/384074589.

[24] Kert Mätlik, Matthew Baffuto, Laura Kus, Amit Laxmikant Deshmukh, David A. Davis, Matthew R. Paul, Thomas S. Carroll, Marie Christine Caron, Jean Yves Masson, Christopher E. Pearson, and Nathaniel Heintz. Cell-type-specific cag repeat expansions and toxicity of mutant huntingtin in human striatum and cerebellum. *Nature Genetics*, 56:383–394, 3 2024. ISSN 15461718. doi: 10.1038/s41588-024-01653-6.

[25] Jong Min Lee, K. Correia, J. Loupe, Kyung Hee Kim, Douglas Barker, Eun Pyo Hong, Michael J. Chao, Jeffrey D. Long, D. Lucente, Jean Paul G. Vonsattel, Ricardo Mouro Pinto, Kawther Abu Elneel, Eliana Marisa Ramos, Jayalakshmi Srinidhi Mysore, T. Gillis, Vanessa C. Wheeler, Marcy E. MacDonald, James F. Gusella, Branduff McAllister, Thomas Massey, Christopher Medway, Timothy C. Stone, Lynsey Hall, Lesley Jones, P. Holmans, S. Kwak, Anka G. Ehrhardt, Cristina Sampaio, Marc Ciosi, Alastair Maxwell, Afroditi Chatzi, Darren G. Monckton, Michael Orth, G. Bernhard Landwehrmeyer, Jane S. Paulsen, E. Ray Dorsey, Ira Shoulson, and Richard H. Myers. CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell*, 178:887–900.e14, 8 2019. ISSN 10974172. doi: 10.1016/j.cell.2019.06.036.

[26] Jong-Min Lee, Zachariah L McLean, Kevin Correia, Jun Wan Shin, Sujin Lee, Jae-Hyun Jang, Yukyeong Lee, Kyung-Hee Kim, Doo Eun Choi, Jeffrey D Long, Diane Lucente, Ihn Sik Seong, Ricardo Mouro Pinto, James V Giordano, Jayalakshmi S Mysore, Jacqueline Siciliano, Emanuela Elezi, Jayla Ruliera, Tammy Gillis, Vanessa C Wheeler, Marcy E MacDonald, James F Gusella, Anna Gatseva, Marc Ciosi, Vilija Lomeikaite, Hossameldin Loay, Darren G Monckton, Christopher Wills, Thomas H Massey, Lesley Jones, and Peter Holmans. Genetic modifiers of somatic expansion and clinical phenotypes in huntington's disease reveal shared and tissue-specific effects. *arXiv*, 2024. doi: 10.1101/2024.06.10.597797. URL `https://doi.org/10.1101/2024.06.10.597797`.

[27] Lawrence Stewart, Francis Bach, Quentin Berthet, and Jean-Philippe Vert. Regression as Classification: Influence of Task Formulation on Neural Network Features. *arXiv*, 11 2022. URL `http://arxiv.org/abs/2211.05641`.

[28] Audrey S. Dickey and Albert R. La Spada. Therapy development in huntington disease: From current strategies to emerging opportunities, 4 2018. ISSN 15524833.

[29] Emilia M. Gatto, Natalia González Rojas, Gabriel Persi, José Luis Etcheverry, Martín Emiliano Cesarini, and Claudia Perandones. Huntington disease: Advances in the understanding of its mechanisms. *Clinical Parkinsonism & Related Disorders*, 3:100056, 2020. ISSN 25901125. doi: 10.1016/j.prdoa.2020.100056.

[30] Gizem Taş, Timo Westerdijk, Eric Postma, Jan H. Veldink, Alexander Schonhuth, and Marleen Balvert. Computing linkage disequilibrium aware genome embeddings using autoencoders. *Bioinformatics*, 40, 6 2024. ISSN 13674811. doi: 10.1093/bioinformatics/btae326.

[31] Ruilin Li, Christopher Chang, Johanne M. Justesen, Yosuke Tanigawa, Junyang Qian, Trevor Hastie, Manuel A. Rivas, and Robert Tibshirani. Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. *Biostatistics*, 23:522–540, 4 2022. ISSN 14684357. doi: 10.1093/biostatistics/kxaa038.

[32] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

[33] Ron Sheinin, Roded Sharan, and Asaf Madi. scNET: learning context-specific gene and cell embeddings by integrating single-cell gene expression data with protein–protein interactions. *Nature Methods 2025*, pages 1–9, 3 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02627-0. URL `https://www.nature.com/articles/s41592-025-02627-0`.

[34] Matthias Fey and Jan E Lenssen. Fast Graph Representation Learning with pytorch geometric. *ICLR*, 2019. URL `https://github.com/rusty1s/pytorch_geometric`.

[35] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian Von Mering. The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51:D638–D646, 1 2023. ISSN 13624962. doi: 10.1093/nar/gkac1000.

[36] Jiahua Xie, Farooqahmed S Kittur, Chiu-Yueh Hung, and Tomas T Ding. Regulation of one-carbon metabolism may open new avenues to slow down the initiation and progression of Huntington's disease. *Neural Regeneration Research*, 18(11):2401–2402, 2023.

[37] Branduff McAllister, Jasmine Donaldson, Caroline S. Binda, Sophie Powell, Uroosa Chughtai, Gareth Edwards, Joseph Stone, Sergey Lobanov, Linda Elliston, Laura Nadine Schuhmacher, Elliott Rees, Georgina Menzies, Marc Ciosi, Alastair Maxwell, Michael J. Chao, Eun Pyo Hong, Diane Lucente, Vanessa Wheeler, Jong Min Lee, Marcy E. MacDonald, Jeffrey D. Long, Elizabeth H. Aylward, G. Bernhard Landwehrmeyer, Anne E. Rosser, Jane S. Paulsen, Nigel M. Williams, James F. Gusella, Darren G. Monckton, Nicholas D. Allen, Peter Holmans, Lesley Jones, and Thomas H. Massey. Exome sequencing of individuals with huntington's disease implicates fan1 nuclease activity in slowing cag expansion and disease onset. *Nature Neuroscience*, 25:446–457, 4 2022. ISSN 15461726. doi: 10.1038/s41593-022-01033-5.

[38] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, Doron Lancet, and Dana Cohen. GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards. *Database*, 2017, 2017. ISSN 17580463. doi: 10.1093/database/bax028.

Data availability. Biosamples and data used in this work were generously provided by the participants in the Enroll-HD study and made available by CHDI Foundation, Inc. Enroll-HD is a clinical research platform and longitudinal observational study for Huntington's disease families intended to accelerate progress towards therapeutics; it is sponsored by CHDI Foundation, a nonprofit biomedical research organization exclusively dedicated to collaboratively developing therapeutics for HD. Enroll-HD would not be possible without the vital contribution of the research participants and their families. The RNA-seq data was obtained from the Sequence Read Archive (SRA) using accession number SRP074904.

Code availability. The code used in this paper is publicly available on Github at https://github.com/AbanteLab/MlcbFuses2025.

Author contributions statement. J.A. conceived the study; C.F. conducted the experiments with support from G.Z.; C.F. and J.A. analyzed the results; and C.F., B.R., J.M.C., and J.A. wrote and reviewed the manuscript.

Competing interests. No competing interest is declared.

Acknowledgements. This study was supported by "la Caixa" Foundation under the grant agreements LCF/BQ/PI24/12040007; and Red Española de Supercomputación (RES) under project BCV-2025-1-0010.

A Supplementary Material

A.1 Evaluation metric

For each model tested, we repeated the training process for 50 different seeds to study the model's susceptibility to training and testing sample sets splitting. Then, model performance was evaluated using the balanced accuracy (BA) score, an average of the recall obtained in each class:

$$\text{BA} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}$$

where 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 performance across all classes is equally weighted, making it particularly suitable for imbalanced datasets. By averaging the recall scores for each class, the balanced accuracy provides a more comprehensive evaluation of the model's ability to correctly classify instances from all classes, irrespective of class distribution.

A.2 GNN architecture

We implemented the two most broadly used types of GNN in this setting, namely Convolutional (ConvGNN) and Graph Attention Networks (GAT). In the former we used two layers of convolution to update node features:

$$\mathbf{h}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{1}{\sqrt{d_i d_j}} \mathbf{W}^{(l)} \mathbf{h}_j^{(l-1)}\right)$$

Each convolution is followed by a pooling stage that selects the most important nodes based on a learnable score vector $\mathbf{s} = \tanh(\mathbf{W}_p \mathbf{h}_i)$ that coarses the graph selecting the top $k = 80\%$ nodes to reduce the graph to its most informative structure.

In GAT we update node features using a single head of attention ($\mathbf{a}$), which scores the importance of the connections between nodes i and j. Each node learns to weight neighbors using:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^T \left[\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_j\right]\right)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})}$$

$$\mathbf{h}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}\mathbf{h}_j\right)$$

A.3 SNP and gene rankings

To identify important genetic features, including SNPs or embedding coordinates, in an XGBoost classifier we sorted the model features according to their gain using the booster's *get_score* attribute with *importance_type = 'gain'*. To compare the most important genes between a pair of classifiers trained on the full and the compressed dataset, we took the top 100 features and found the set of unique gens corresponding to these in each case. To evaluate the amount of overlap between two gene sets A and B we used the Jaccard index, given by

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where $|A \cap B|$ denotes the cardinality of the intersection of sets A and B, representing the number of elements common to both sets, and $|A \cup B|$ denotes the cardinality of the union of A and B, representing the total number of distinct elements across both sets. The Jaccard Index ranges from 0 to 1, with 0 indicating no overlap and 1 indicating that the sets are identical.

A.4 Functional Gene Ontology Analysis

To identify enriched gene ontology (GO) terms, we used Fisher's exact test to compute enrichment scores for a given subset of informative SNPs and a set of background SNPs. We assigned the GO term of the corresponding gene to each SNP in a protein-coding region and performed the enrichment analysis at the SNP level. To that end, we constructed the following 2x2 contingency table:

	In GO Term	Not in GO Term
In SNP Subset	a	b
In Background Set	c	d

where a represents the number of SNPs in the informative SNP subset that are annotated with the corresponding GO term, b is the number of SNPs in the same subset not annotated with the GO term, c denotes the number of background SNPs annotated with the GO term, and d is the number of background SNPs not annotated with the GO term. We computed the log odds ratio (OR) to quantify the strength of the association between gene subset membership and GO term annotation, given by

$$\log(\text{OR}) = \log\left(\frac{a \cdot d}{b \cdot c}\right)$$

and used Fisher's exact test to produce the corresponding *p*-value. This metric provides an interpretable measure of enrichment, where positive values indicate overrepresentation and negative values suggest underrepresentation of the GO term within the SNP subset.

B Supplementary Tables

Gene	Previously discussed in HD	Previously discussed in NDD
FAN1	McAllister 2022	
NRG1		AD (Ou 2021)
DHFR	Flower 2019	
ARID5B		AD (Guo 2024)
MLH1	Pinto 2013	
TCERG1	Lobanov 2022	
ZFHX3		Neural apoptosis (Yang 2021)
RORA		AD (Baker 2019)
MED23		
ESRRB		AD (Sato 2023)
HTT	Gusella 1983	
HIVEP3		AD (Wan 2014)
ZNF423		AD (Baker 2019)
CAMTA1		Neuronal degeneration (Long 2014)
TNR		Neurodevelopmental disorder (Wagner 2021)
CUX1		Developmental alterations (Oppermann 2023)
GLIS3		AD (Calderari 2018)
CCAR1	Reduces aggregates (Lugano 2023)	
LIG1	Ratcliffe 2025, Namuli 2025	
GRIN2B	Arning 2007	
TRERF1		AD (Guo 2024)
USP30		NDD (Harrigan 2018)
SLC1A1		Psychiatric Disorders (Underhill 2019)
WWOX		Neurogenesis and function (Steinberg 2024)
PMS2	Lee 2022	
FLYWCH1		Amyotrophic lateral sclerosis (Pappalardo 2024)
THRB		Association to PD (Mooradian 2025)
NPM1	Sonmez 2021	
WFS1		Psychiatric Disorders (Munshani 2021)
EXT1		Inhibits tau aggregates (Zhang 2018)

Table S1: *List of genes identified among the top 100 SNPs in the best performing XGBoost model, alongside previously reported associations with Huntington's Disease (HD) or other neurodegenerative disorders (NDD).*

C Supplementary Figures

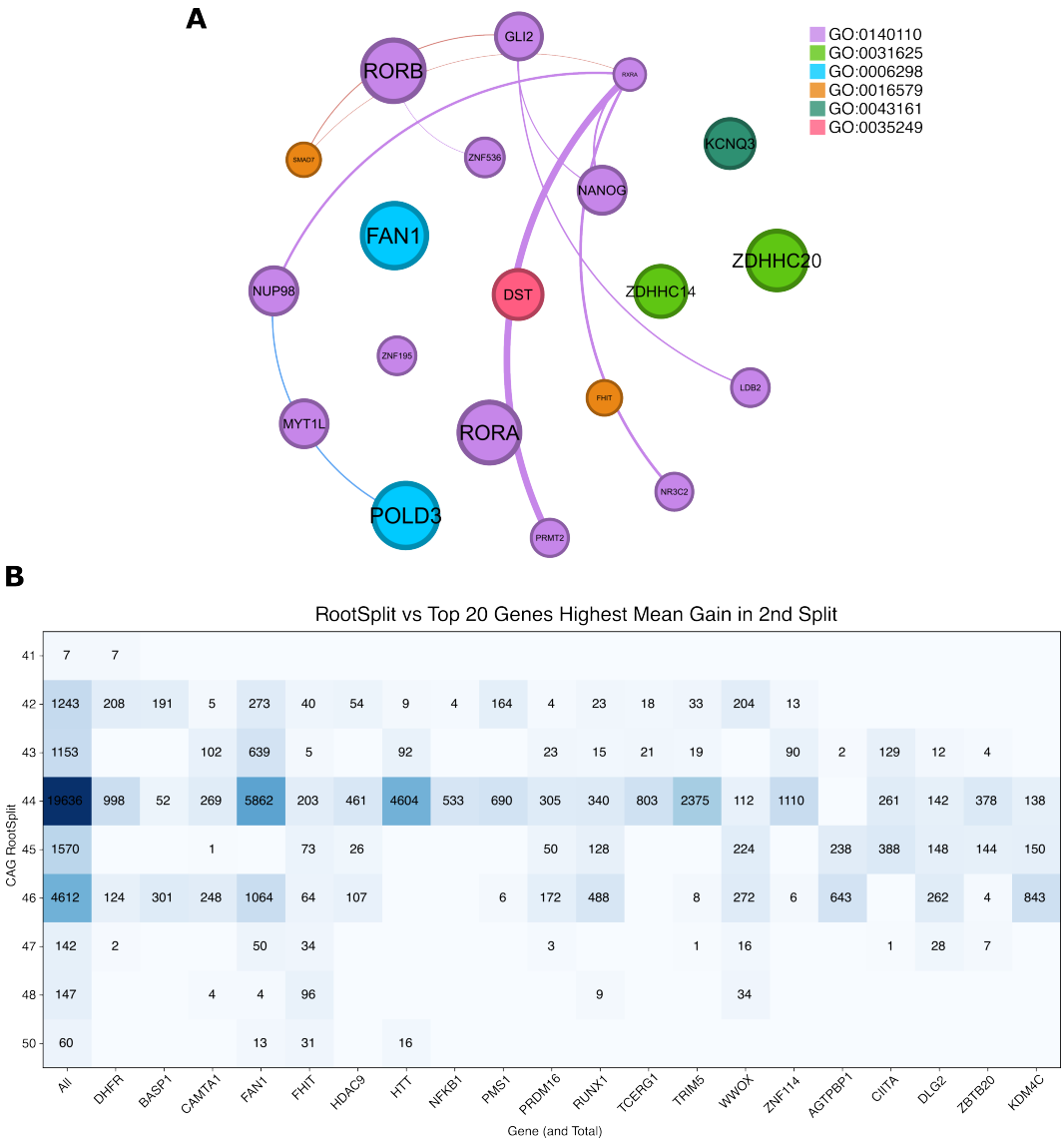

Figure S1: *Results of all XGBoost models trained with the SNPs dataset. (A) Protein-interaction network of the top 20 genes across XGBoost models colored by gene ontology. (B) Genes used in all models of depth 2 (50% of all models) by decision trees where the first node splits by CAG length (43.35% of trees on average), with what CAG splitting value was used before them.*

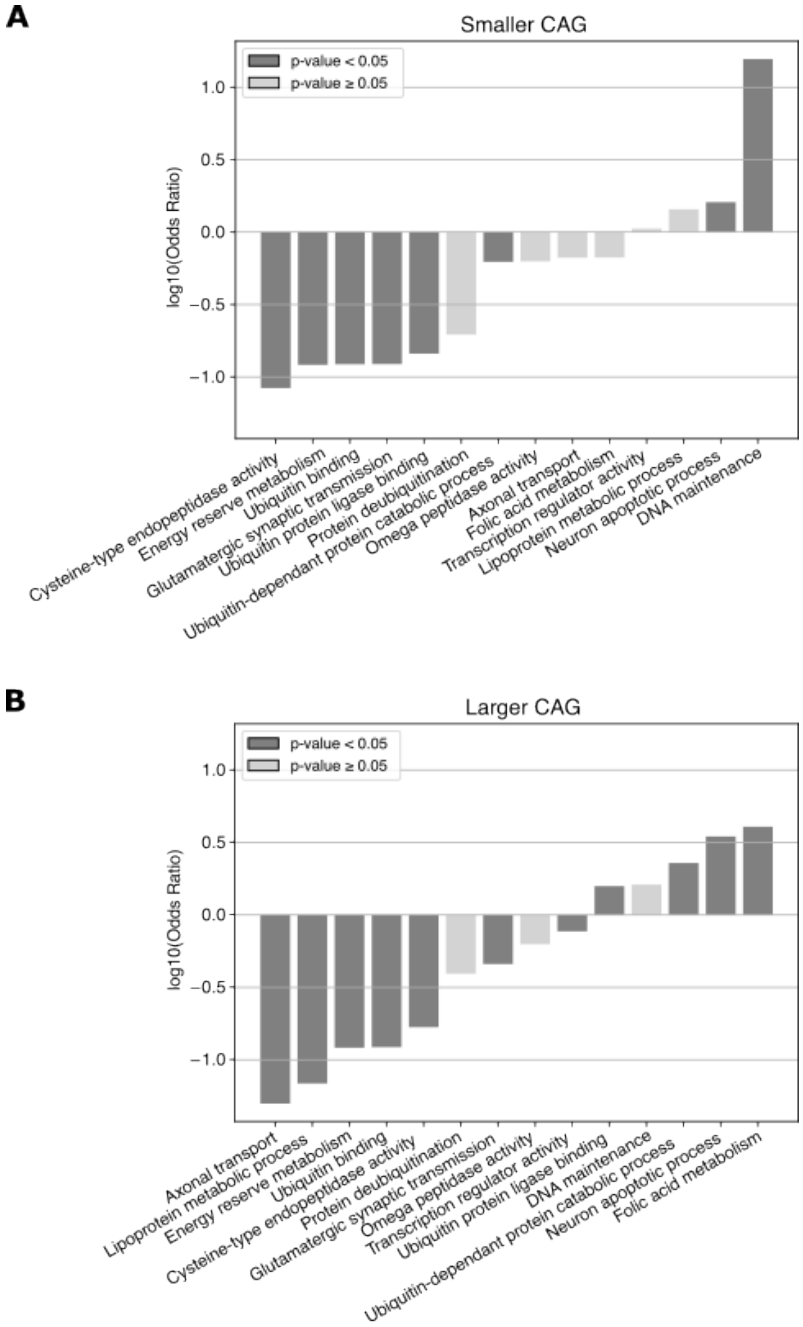

Figure S2: *Gene ontology of genes used for smaller and larger values of CAG expansions, showing that the directionality of the CAG interaction effect could be explained by the mechanisms affected. (A) Genes used in shorter CAG repeat lengths are heavily enriched for DNA maintenance. (B) Genes used in larger CAG lengths are enriched for folic acid metabolism, neuron apoptotic process and ubiquitin-dependent protein catabolic process.*

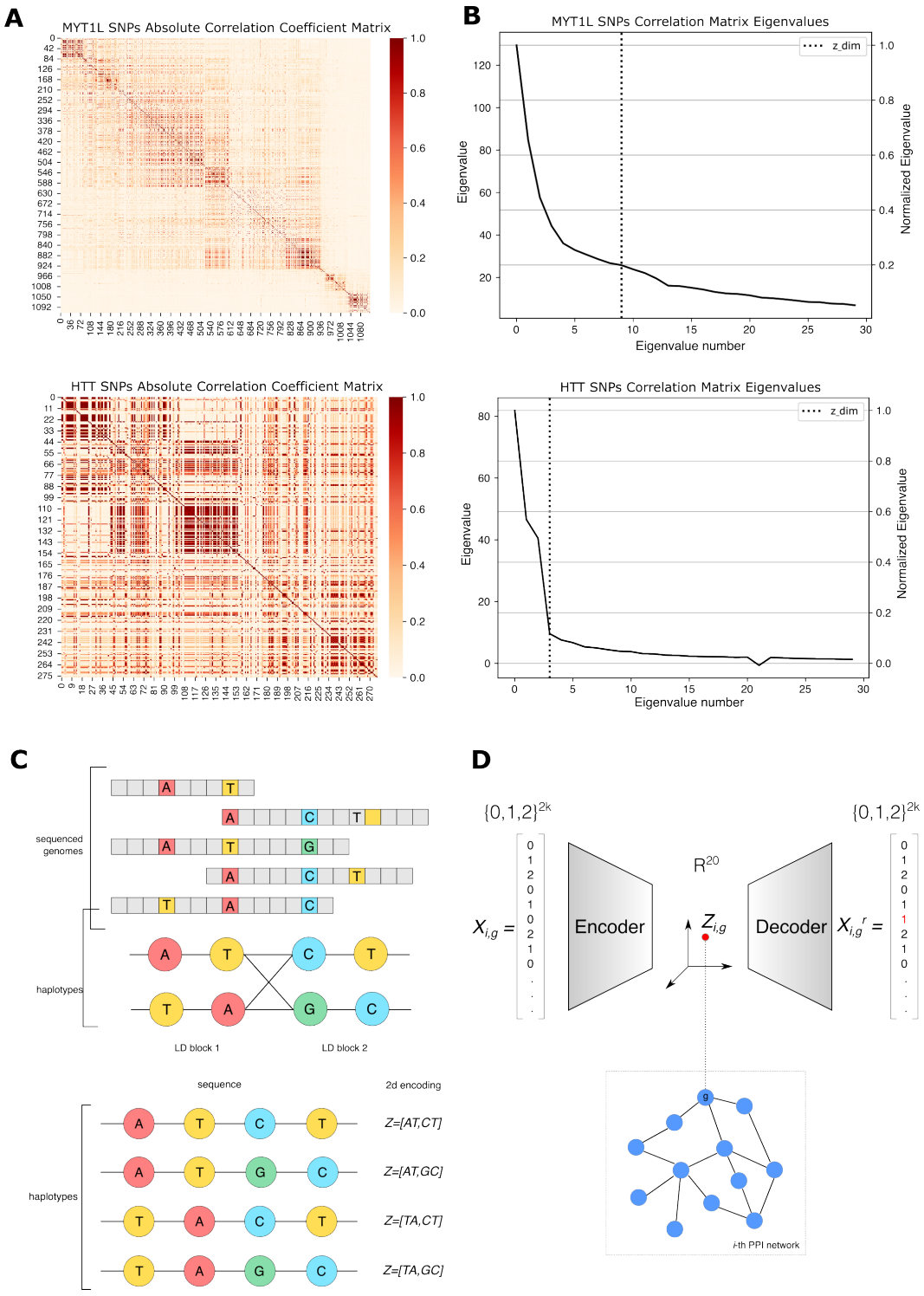

Figure S3: *Linkage disequilibrium blocks used to reduce dimensionality of dataset. (A) Correlation matrix of SNPs from two examples of genes (MYT1L and HTT). (B) Eigenvalues of the correlation matrices shown on the left. We select the number of LD blocks by taking the number of eigenvalues larger than 20% of the largest. (C) LD blocks recombination: during DNA recombination these groups tends to travel together, making haplotypes reducible to these blocks. (D) Autoencoders are trained for each gene with a latent dimension matching the number of LD blocks found with the method explained in B. The embeddings created in the resulting latent space are the reduced dimensionality features used to train GNNs, encoding each node of the PPI interaction nodes.*

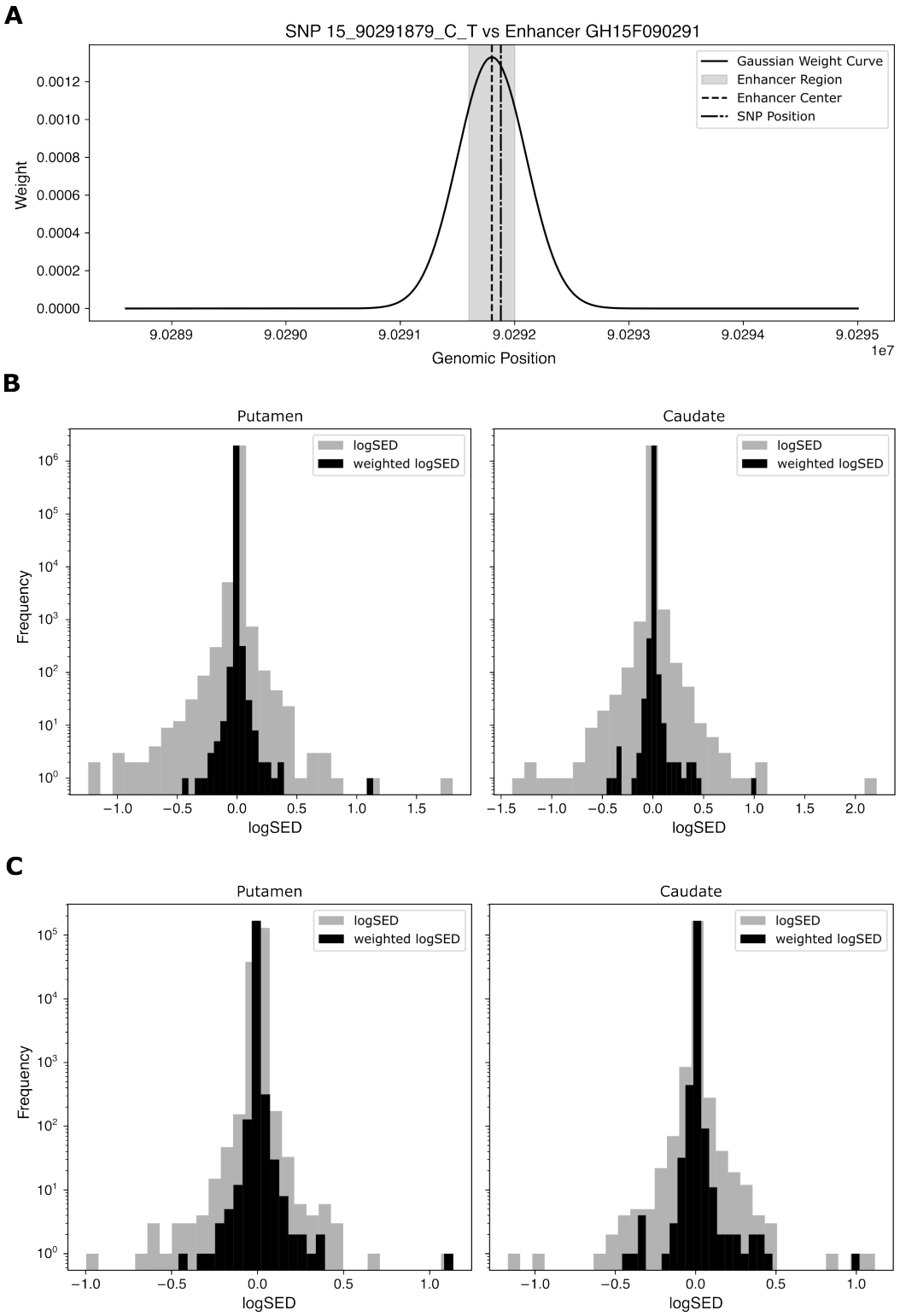

Figure S4: *Weights of Borzoi logSED scores. (A) Gaussian kernel centered at the center of enhancer GH15F090291 and weight assigned to SNP 15_90291879_C_T. (B-C) Histogram of raw and weighted logSED values for both tissues considered, putamen and caudate, including all predictions (B) and the subset of the 1,000 most variable genes in each case (C).*

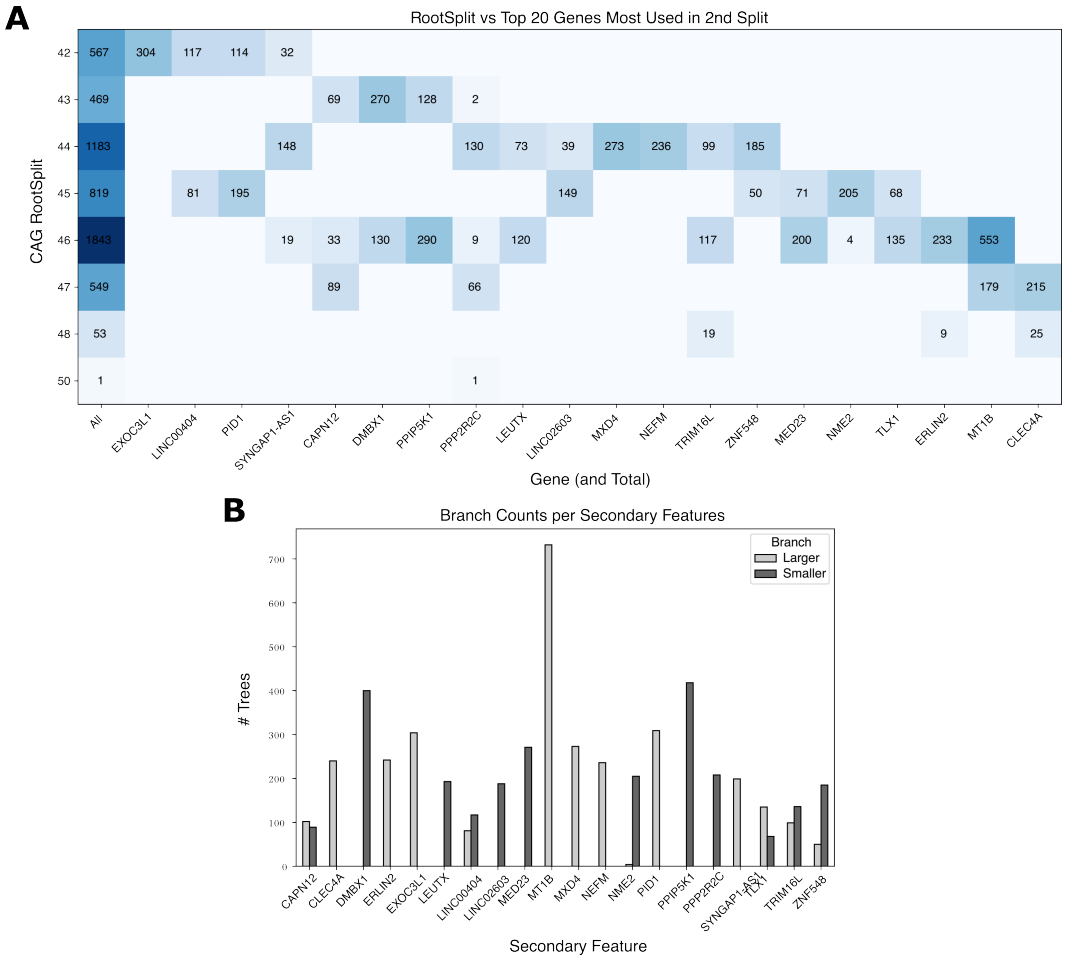

Figure S5: *Expression features of multimodal model. (A) Top genes whose predicted expression was used among the top 100 features in the multimodal models of depth 2 (45% of all models trained with the same data), used by decision trees where the first node splits by CAG length (45% of trees on average), with what CAG splitting value was used before them. (B) Frequency of how many trees with CAG at the root node use each feature for smaller or larger CAG expansions.*