Novel Machine Learning Algorithms for Analyzing Large-scale Genomic and Genetic Data

by

Ye Wang

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 12, 2020

Keywords: Machine learning, Human microbiome, Human genome

Approved by

Xiao Qin, Professor of Computer Science and Software Engineering
Li Chen, Assistant Professor of Health Outcomes Research and Policy
Cheryl Seals, Professor of Computer Science and Software Engineering
Richard Chapman, Associate Professor of Computer Science and Software Engineering
Debswapna Bhattacharya, Assistant Professor of Computer Science and Software Engineering

Abstract

With the advancement of next-generation sequencing technology, numerous disease/ phenotypic associations with the human microbiome and human genome are uncovered and revealed. In this dissertation, we take advantage of this and explore these association patterns using machine learning methods. We first design a deep learning method MDeep for microbiome-based prediction by considering both the taxon abundance and phylogenetic tree. MDeep models the taxonomic rank by the convolutional layers and captures the phylogenetic correlation on each taxonomic rank via the convolutional operation. Our simulations and real data analysis demonstrate that MDeep outperforms competing methods in both regression and binary classifications. In order to explore the diseases/ phenotypic associations with the human genome, we propose two machine learning frameworks. The first framework, WEVar, is a supervised learning framework by integrating the pre-computed scores from representative existing scoring methods, which will benefit from each individual method by automatically learning the relative contribution of each method and produce an ensemble score for the final prediction. Using simulation and real data studies, we show both context-free WEVar and context-dependent WEVar outperform the individual scoring methods on the state-of-the-art benchmark datasets. Furthermore, we find WEVar can prioritize experimentally validated non-coding variants in an LD block. The second framework, DeepMFIVar, is a deep multimodal learning framework for the functional interpretation of genetic variants. DeepMFIVar learns a predictive model linking DNA sequence context and clinical information to quantitative epigenetic signals. The mutation effect of the 210 million genetic variants is generated by the difference of the predicted epigenetic signal for the reference and for alternative alleles. The application to DNA methylation and histone modification demonstrate that DeepMFIVar can accurately predict locus-specific epigenetic signals using DNA sequence and clinical information, and it is also capable of prioritizing variants for downstream experiments.

Acknowledgments

I am grateful and want to express my deepest appreciation to all of my advisors, colleagues and friends who made my Ph.D. experience an unforgettable and unique journey of my life.

First and foremost, I give my most sincere and genuine gratitude to my advisor, Dr. Xiao Qin, and Dr. Li Chen for their countless research advice, inspiring mindset, and enthusiastic help and support. This dissertation wouldn't be here without their constructive ideas and continuous encouragement. I want to thank Dr. Li Chen, who led me into the magical area of Genome and Biology. Discussions with him were always enriching and rewarding.

I am also grateful for all of my committee members, Dr. Cheryl Seals, Dr. Richard Chapman and Dr. Debswapna Bhattacharya, for their invaluable advice during my research proposal presentation, by which lead to a substantially improvement of this dissertation. I also want to thank Dr. Jingjing Qian, for being my university reader.

Finally and most importantly, I want to give special appreciation to my family and friends for their unconditional love and unbreakable friendship. I want to thank my parents, although far away, they support me with their best, and consistent love. Without them, I will not be here writing my Ph.D. dissertation paper. I want to thank all of my friends, especially Hairuo Xu, Chengyu Tang, and Kenan Xiao for all the fun and depressing times we spent together. And I am very grateful to my friend, Qiang Chen, and Chuanpeng Dong, for their accompanies in Indiana. When I just moved to this brand new city, they helped and walked me through most struggled times.

To my parents

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Background

With the development of next-generation sequencing technologies and the advancement in machine learning, the computational models have been at the heart of many recent breakthroughs in the human microbiome and the human genome studies. Machine learning models complement biological experiments by providing the capacity to make accurate predictions on unseen samples and to summarize predictive features that elucidate the mechanisms of important biological or pathological processes. In this dissertation, we study machine learning methods for exploring disease / phenotypic association patterns with human microbiome and human genome, two research areas where traditional methods based on biological experiments fall short in power and efficiency.

The human microbiome, a collection of microbes that live in and on our bodies, including bacteria, archaea, viruses, and eukaryotes[1], plays a critical role in human health and disease. Numerous human microbiome studies revealed that the abnormal change of relative abundances of microbiota could lead to various diseases such as infection (e.g. Clostridium difficile, Helicobacter pylori, Bacterial vaginosis), liver diseases (e.g. acute-on-chronic liver failure), gastrointestinal malignancy (e.g. gastric cancer, colorectal cancer), metabolic disorders (e.g. obesity, type 2 diabetes), autoimmune diseases (e.g. Crohn's disease) and even mental or psychological diseases (e.g. autism spectrum disorder). Due to the decreasing cost of next-generation sequencing technologies, large-scale microbiome datasets with more than hundreds' samples are generated recently from modern high-throughput sequencing. Large-scale datasets allow researchers to employ more sophisticated modeling methods such as machine learning approaches to study the relationship between the microbiome and various phenotypes and diseases.

In this thesis, we develop a novel deep learning prediction method MDeep (microbiome-based deep learning method) to predict disease and clinical outcomes. Conceptually, MDeep designs convolutional layers to mimic taxonomic ranks with multiple convolutional filters on each convolutional layer to capture the phylogenetic correlation among microbial species in a local receptive field and maintain the correlation structure across different convolutional layers via feature mapping. Taken together, the convolutional layers with its built-in convolutional filters capture microbial signals at different taxonomic levels while encouraging local smoothing and preserving local connectivity induced by the phylogenetic tree. We demonstrate that MDeep outperforms competing methods in both regression and binary classifications by simulation studies and real data applications.

A gene is a sequence of nucleotides. Genes are transcibed to RNA, and then translated to proteins which can be considered as "workhorses" of the cell, with all the functions necessary for life. However, over 98% of the human genome is not used to encode proteins, these non-coding elements, especially non-coding genetic variants play a crucial role in gene regulation as they might disrupt the promoter or enhancer regions and thus have an impact on the gene expression. The abnormal expressed genes may result in different kinds of diseases. To discover these functional genetic variants, numerous genome-wide association studies (GWAS) have been carried out to find the disease risk variants [2, 3]. However, these studies are limited by the sample size and linkage disequilibrium, which will mask true causal variants from the neural ones. Therefore, post-GWAS computational methods for fine tuning and prioritizing these regulatory variants is demanded. In the past ten years, multiple computational approaches have been established to accomplish such tasks [4]. Most of these methods can be roughly divided into three categories: (i) supervised learning methods that attempt to separate known disease variants from putative benign variants using a variety of genomic features [5, 6, 7]; (ii) unsupervised learning algorithms that try to integrate these different annotations into one measure of functional importance[8]; (iii) evolutionary methods that consider data on genetic variation together with functional genomic data and aim to predict the effects of non-coding variants on fitness[6, 9, 10, 11].

In this thesis, we develop two machine learning frameworks for genetic variants. The first framework, WEVar (Weighted Ensemble framework for predicting functional non-coding Variants), is a supervised learning framework by integrating the pre-computed scores from representative existing scoring methods. The second framework, DeepMFIVar (Deep Multimodal Learning framework for Functional Interpretation of genetic Variants), learns a predictive model linking DNA sequence and clinical information to quantitative epigenetic signals.

1.2 Dissertation Outline

The rest of this dissertation is organized as follows. The next chapter presents a new deep learning method for predicting disease and clinical outcomes by using human microbiome. In Chapter 3, we describe WEVar, a new weighted ensemble learning framework, to predict and prioritize functional relevant non-coding variations. In Chapter 4, we introduce a multimodal deep learning model, DeepMFIVar, that accurately predicts DNA methylation ratio and histone modification from the DNA sequence context and clinical outcome. Chapter 5 concludes the dissertation and points out future research.

Chapter 2

MDeep: a novel deep learning method for predictive modeling of microbiome data

2.1 Introduction

Human microbiota, including bacteria, archaea, viruses, and eukaryotes, colonizes the human body and affects host physiology to a great extent. The composition and function of microbiota vary across different body sites, ages, genders, races, and dietaries of the host [12]. The roles of human microbiota playing in human health are usually summarized in three aspects. First, the microbiota could potentially aid the digestive system by more efficiently extracting energy from food and harvesting nutrients [13, 14] as microbiota provides humans with enzymes and biochemical pathways [13] produced by versatile metabolic microbial genes that are far more than found in human genome. Second, the human microbiota protects its host against invasive pathogens by providing a physical barrier, producing antimicrobial substances or involving in competitive exclusion [15, 16]. Third, the microbiota is essential in the induction, training, and function of the host immune system [17, 18]. As a consequence, dysbiosis of human microbiota, that is the abnormal change of relative abundances of microbiota, could lead to various human diseases such as infection (e.g. *Clostridium difficile*, *Helicobacter pylori*, *Bacterial vaginosis*), liver diseases (e.g. Acute-on-chronic liver failure), gastrointestinal malignancy (e.g. Gastric cancer, Colorectal cancer), metabolic disorders (e.g. Obesity, Type 2 diabetes), autoimmune diseases (e.g. Crohn's disease) and even mental or psychological diseases (e.g. Autism spectrum disorder) [19].

16S rRNA gene-target sequencing is a cost effective metagenomic sequencing technology, which has been widely in microbiome studies for mainly uncovering bacteria and archaea by sequencing the structural components of the ribosome (V1-V9) that could be used as a molecular clock to identify phylogeny [20]. The raw sequencing reads could be processed using established bioinformatics pipelines such as Quantitative Insights Into Microbial Ecology (QIIME) [21], which clustered the reads into operational taxonomic units (OTUs) at different taxonomic levels. As a result, the processed microbiome data are generated, consisting of an OTU abundance matrix with rows as samples and columns as OTUs along with a phylogenetic tree based on which the phylogenetic information among OTUs could be inferred.

As human microbiota is closely associated with human health, it is natural to use the OTUs as "biomarkers" to predict host phenotypes or clinic outcomes. It should be noted that microbiome data is usually high-dimensional (more OTUs than samples), over-dispersed (large variability) and sparse (excessive zeros in the OTU abundance matrix). These data characteristics make robust machine learning models such as Random Forest [22] a desirable microbiome-based prediction model, which deals with high-dimensional features by randomly sampling a subset of features in each decision tree to reduce the possibility of overfitting and the final prediction is the aggregated predictions of all trees. Moreover, modern regression methods such as Lasso [23], MCP [24],and Elastic Net [25] are designed in nature for high-dimensional classification and regression and have been widely used in microbiome-based prediction [26, 27, 28]. These regression models usually incorporate a sparse penalty to select the most predictive taxa in the training set and thus improve the prediction performance in the testing set using the selected taxa.

Although these methods adopt different approaches to address the high dimensional prediction task, they are limited in exploring the phylogenetic relationship among taxa. The phylogenetic tree is an informative prior as the microbial community changes are not randomly distributed but tend to occur in clades at varying phylogenetic depths corresponding to different taxonomic ranks. In other words, the tree could provide a phylogenetically correlated structure among taxa, based on which we can cluster and aggregate taxa abundance to achieve better predictive performance. Moreover, the cluster size and signal density also vary. For diseases such as colorectal cancer or arthritis [29, 30], few marker taxa are found to be associated to the disease state, whereas effects on the overall composition are very mild. In contrast, obesity and inflammatory bowel disease are associated with marked changes in the overall composition [31, 32]. Taken together, an optimized prediction model should have robust prediction performance across different cluster size and signal density.

Recently, deep learning methods especially convolutional neural network (CNN) have been widely used in bioinformatics for various prediction tasks: (i) predicting genomic regions such as TF-DNA binding sites [33, 34] and modification sites [35]; (ii) predicting genomic features such as non-coding RNAs [36] and enhancer [37]. (iii) predicting genomic signals such as gene expression [38] and DNA methylation [39] (iv) predicting regulatory variants [40, 33]. The advantages of CNN lie on its ability of capturing the local correlations of features, enhancing the local connectivity across different levels and reducing the parameters via convolutional operations, which improve the prediction performance. In these studies, CNN achieves an overall better prediction than other methods and the sample size of training set usually ranges from thousands [36] to millions [40]. However, the prediction performance of CNN is rarely exploited when the sample size of training set is as few as hundreds. Moreover, a recent CNN-based deep learning method Ph-CNN has been proposed to perform the task of binary classification with the consideration of phylogenetic tree [41]. However, the CNN architecture of Ph-CNN is not optimized. In addition, the utilization of phylogenetic tree based on finding neighbors of each OTU via MultiDimensional Scaling projection is computationally intensive, which limits the scalability of Ph-CNN to high-dimensional microbiome data. Importantly, Ph-CNN is only designed for binary outcome only. Considering these, a computational efficient deep learning-based prediction, which can utilize the phylogenetic tree, for the purpose of predicting both continuous and binary outcome, is in demand.

In this work, we develop MDeep, a novel Microbiome-based deep learning method, for predicting continuous and binary outcome by utilizing both the taxon abundance and phylogenetic tree. Based on an evolutionary model, MDeep is designed to model the taxonomic rank by the convolutional layer and capture the phylogenetic correlation on each taxonomic rank via the convolutional operation. The main contributions of MDeep can be summarized as: i) predicting both continuous and binary outcome; ii) utilizing the phylogenetic information besides taxon abundance for improving prediction accuracy; iii) obtaining an overall better prediction than other methods especially CNN without using the phylogenetic tree in a conducted comprehensive simulation study with considerations of signal density, cluster size and informativeness of phylogeny; iv) outperforming existing methods in real datasets including CNN and Ph-CNN. To the best of our knowledge, MDeep is the first deep learning approach that can efficiently utilize the phylogenetic tree in predicting both continuous and binary outcome. Importantly, this is the first study systematically exploring the prediction performance of deep learning approaches in a comprehensive simulation study, which will inform deep learning-based approaches' favorable scenario, that is, dense and large-clustered signals. We believe MDeep will be a valuable addition to the microbiome research community.

2.2 Methods

2.2.1 Encoding phylogenetic information by convolutional operation

Before we introduce the deep learning predictive model, we introduce a phylogeny-induced correlation structure $C_{p\times p}$ among taxa based on a evolutionary model [42], which is defined as,

$$C_{p\times p}(\rho) = e^{-2\rho D_{p\times p}} \tag{2.1}$$

where D denotes the pairwise patristic distance between taxa (e.g. the length of the shortest path between two taxa in the tree) that could be estimated by the function `cophenetic` in the R package `ape` that takes a phylogenetic tree as the input. $\rho \in (0, \infty)$ characterizes the evolutionary rate: fast evolution corresponds to a large ρ (a small C). Alternatively, ρ can be interpreted as a parameter that controls the phylogenetic depth at which the taxa are grouped: a larger cluster at a lower phylogenetic depth indicates a larger ρ (a smaller C). In other words, a large ρ represents a small phylogenetic correlation among taxa and large clusters, making the tree less informative. Because a phylogenetic depth corresponds to a taxonomic rank, ρ has a similar effect as taxonomic grouping conceptually, where taxa at different taxonomic ranks are grouped together according to their taxonomy. Without loss of generality, we fix ρ as 2 here without being tuned.

2.2.2 Microbiome-based deep learning architecture

MDeep is essential a Phylogeny-Regularized Convolutional Neural Network, which is composed of multiple convolutional layers followed by fully-connected layers. The taxa are clustered based on C before the taxon abundance is passed into the network. Following the input layer, convolutional layers are designed to include the phylogenetic correlation across different phylogenetic depths as much as possible (Fig 2.1). Since the phylogenetically correlated taxa are grouped, convolutional operation is more efficient to capture phylogenetic correlation in the local receptive field, and thus encouraging local smoothing. Notably, convolutional layers not only bring a solution to high-dimensional input variables by reducing the number of parameters but also allow to encode the phylogenetic information in the local receptive fields, which encourages spatially local input patterns. Additionally, convolutional layers encourage local connectivity in the way of making hidden nodes only receive input from only a restricted subarea of the previous convolutional layer. In contrast to convolutional layers, fully-connected layers exploit the nonlinear and high-order interactions among input features globally, further improving the feature representation. In sum, MDeep maximizes the feature representation of microbiome data to improve the prediction performance potentially.

To incorporate the phylogenetic tree information via convolutional operations more efficiently, Mdeep first clusters taxa on the first convolutional layer based on C, which will make the phylogenetically correlated taxa close to each other. MDeep then adopts multiple

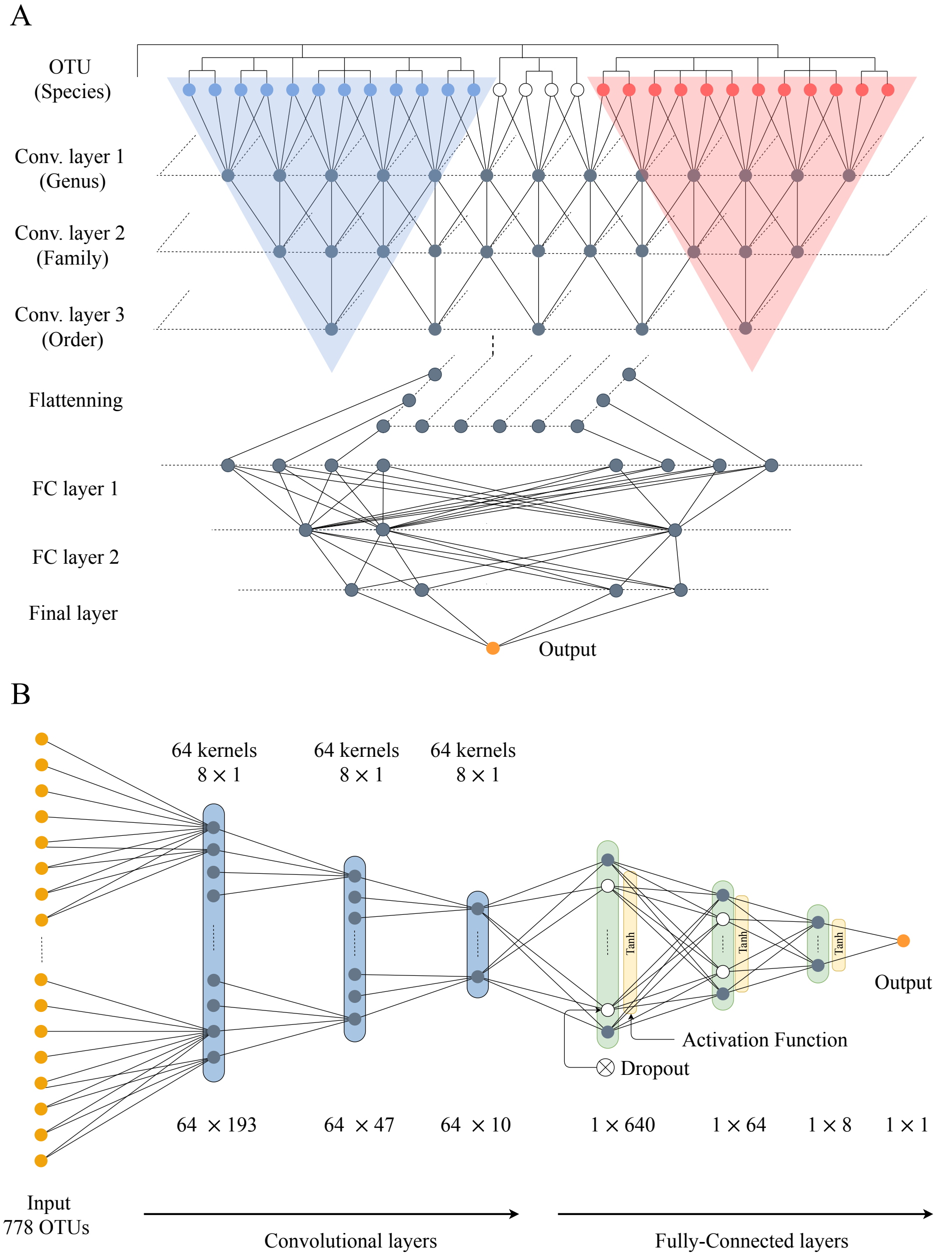

Figure 2.1: (A) The conceptual analogy between MDeep and taxonomic levels of the phylogenetic tree. OTUs on the species level are clustered based on the evolutionary model. This clustering step makes convolutional operation capture OTUs highly correlated in the phylogenetic tree. The number of hidden nodes decreases as the convolutional layer moves forward, reflecting the taxonomic grouping. (B) Input layer, convolutional layers, fully-connected layers and output layer of MDeep. There are three convolutional layers and three fully-connected layers in MDeep. The output layer, connected to the last fully-connected layer, contains a single node for continuous outcome and two nodes for binary outcome.

one-dimensional convolutional kernels to capture the phylogenetic correlations of taxa on each convolutional layer. Specifically, we let the training set in a batch consist of n labeled samples $(\mathbf{x}, y)_n$, where $\mathbf{x}$ is a taxon abundance vector of size p, $y \in \{0, 1\}$ for binary classification and $y \in \{-\infty, +\infty\}$ for regression. MDeep uses n_f filters, each of which has length l. Each filter will perform sliding window operations with stride s for consecutive movement from position 1 to $\left[\frac{p}{s}\right]$ across p taxa, resulting in a feature map $\mathbf{Z}$ of dimension $n_f \times \left[\frac{p}{s}\right]$. Specifically, $\mathbf{Z}$ could be derived from $\mathbf{x}$ and $\mathbf{W}$ as:

$$\mathbf{Z} = f_{conv}(\mathbf{x}) \tag{2.2}$$

$$z_{ij} = f(b_j + \sum_{k=1}^{l} \mathbf{W}_{kj}\mathbf{x}_{i+k-1}) \tag{2.3}$$

where z_{ij} is the feature map from ith position ($i \in \{1, ..., \left[\frac{p}{s}\right]\}$) of input and jth filter ($j \in \{1, ..., n_f\}$). $\mathbf{W}$ of dimension $n_f \times l$ is a matrix of weights for n_f filters with each row corresponding to an individual filter. b_j is a bias term specifically for filter j and f is a non-linear function such as the hyperbolic tangent.

The conceptual analogy between MDeep and taxonomy can be viewed in Fig 2.1A, where each convolutional layer mimics a taxonomic rank such that the input layer represents species level with each node corresponding to a species, followed by multiple convolutional layers starting from the representation of a lower phylogenetic/taxonomic level (e.g. genus) to a higher phylogenetic/taxonomic level (e.g. order). Multiple convolutional filters operate on each convolutional layer to encourage local smoothing, and feature maps between adjacent convolutional layers are restricted by the local connectivity, which resembles the scenarios that nearby taxa on the lower phylogenetic/taxonomic level are more likely close to each other on the higher phylogenetic/taxonomic level. In addition, the dimension of feature map decreases with a series of convolutional operations across multiple convolutional layers. The decreased dimension of feature map via convolutional operation mimics the structure of the phylogenetic tree, where the dimension of taxa decreases as the taxonomic rank increases, corresponding to phylogenetic depth from high to low (Fig 2.1B).

The feature map of the last convolutional layer is flattened as the input for the feed-forward neural network consisting of several fully-connected layers. The dimension of the fully-connected layer also decreases with the neural network moving forward in order to capture the high-order and nonlinear interactions of the features.

The output layer, connected to the last fully-connected layer, contains a single neuron for continuous outcome and two nodes for binary outcome. For binary outcome, the output is further passed through a sigmoid function to produce the prediction probability for $y = 1$, defined as $p(y = 1|\mathbf{x})$. The prediction probability for $y = 0$ can be obtained by $p(y = 0|\mathbf{x}) = 1 - p(y = 1|\mathbf{x})$ accordingly.

Prediction Mean Squared Error (PMSE) defined as $l_c = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, is used as the cost function for regression, whereas, cross entropy defined as $l_c = -\frac{1}{n}\sum_{i=1}^{n}(y_i log(p(y_i = 1|\mathbf{x}) + (1 - y_i) log(p(y_i = 0|\mathbf{x}))$, is used for binary classification. The objective function is $f_o = min_{(\mathbf{W})} l_c + \lambda \sum ||\mathbf{W}||^2$, where λ is the tuning parameter for the l_2 penalty of $\mathbf{W}$. f_o is used to minimize the training error and thus update the network parameters via a back-propagation algorithm. Specifically, MDeep adopts Adaptive Moment Estimation [43] as the optimizer in the process of back-propagation since it is an improved version of stochastic gradient descent method by computing adaptive learning rate for each weight.

An appropriate network architecture is essential to the model fitting: deep layers with many parameters will cause overfitting while shallow layers with few parameters will result in underfitting. Considering the sample size of a typical microbiome study is relatively smaller compared to a traditional CNN prediction task, it is feasible to manually select the optimal number of convolutional layers and fully-connected layers. To be specific, we start with a single convolutional layer and incrementally increase the number of convolutional layers with systematic exploration of the number of hidden nodes and dropout rates to find a proper architecture. We stop adding convolutional layers when increasing the number of layers does not improve prediction performance. We use the same procedure to select the number of fully-connected layers. The network architecture ends up with three convolutional layers followed by three fully-connected layers.

To improve generalization and to prevent overfitting, we leverage the dropout-layer technique and L_2 regularization for $\mathbf{W}$ in the cost function. Each fully-connected layer is followed by a dropout layer to avoid overfitting [44]. Specifically, 50% hidden neurons in the fully-connected layer are randomly dropped out.

Altogether, MDeep is a neural network consisting of one input layer, three convolutional layers, connected with three fully-connected layers and one output layer (Fig 2.1B). Each convolutional layer or fully-connected layer is activated by the hyperbolic tangent function. There are 64 kernels in each convolutional layer. The kernel size is 8 and the stride is 4. We run MDeep multiple epochs in the training. In an epoch, the whole training set is split into multiple batches and each batch pass forward and backward through the network. Here, we set batch size 16 and number of epochs 2000. When the objective function converges after multiple epochs, estimated parameters are stabilized and finalized. Given a testing sample, the trained MDeep $f(\mathbf{x}) = net^3(conv^3(\mathbf{x}))$ computes the prediction probability for binary classification and prediction values for regression.

2.2.3 Performance evaluation

We use R^2 and PMSE as the evaluation metric for regression, and R^2, AUC (Area Under the Curve) along with Sensitivity, Specificity, Accuracy, Precision, F1 score and Matthews

Correlation Coefficient (MCC) for binary classification. These metrics are defined as,

$$
\begin{aligned}
\text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}} \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\
\text{F1} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN}) \times (\text{TN} + \text{FP})}}
\end{aligned}
$$

In the above formula, TP and TN are the numbers of positive and negative samples that are correctly classified, FN and FP are the numbers of positive and negative samples that are mis-classified. Sensitivity or Recall ($\in [0, 1]$) is the proportion of true positives that are correctly predicted among all true positives. Specificity ($\in [0, 1]$) is the proportion of true negatives that are correctly predicted among all true negatives. Precision ($\in [0, 1]$) is the proportion of true positives that are correctly predicted among all predicted positives. Accuracy is the proportion of correctly predicted samples (TP and TN) among all samples. F1 ($\in [0, 1]$) considers both the Precision and Recall. MCC ($\in [-1, 1]$) indicates the correlation between predicted labels and true labels: $+1$ represents a perfect prediction, 0 means random guess and -1 indicates total disagreement between prediction and truth.

2.3 Simulation Studies

2.3.1 Simulation Strategy

We conduct comprehensive simulation studies to evaluate the prediction performance of MDeep along with other competing methods for both regression and binary classification. We synthetically generate 200 independent samples in the training set and an equal number of 200 independent samples in the testing set, respectively. The two classes are restricted to be balanced in both training and testing set for binary classification. By assuming that OTU abundance follows a Dirichlet-multinomial distribution (DM), we estimate parameters including dispersion and mean proportion from a real human upper respiratory tract microbiome data [45] consisting of 778 OTUs from 60 samples. Using these estimated parameters, we generate the read counts parametrically based on DM and the total read count is drawn from a negative binomial distribution with the mean of 5000 and the dispersion of 25. The OTU abundance is further normalized into proportion by dividing the total read counts. We then generate the outcome based on the abundance of outcome-associated OTUs (“aOTUs") that form in different number of outcome-associated OTUs (“aClusters") with different cluster size. We further evaluate the effect of number of clusters and cluster size on the prediction performance.

2.3.2 Constructing outcome-associated OTU clusters

For an informative tree, aOTUs are more likely clustered and have a similar magnitude and the same direction of effect to the outcome. In other words, aOTUs form outcome-associated clusters (“aClusters") to have an impact on the outcome. Since the density of aClusters and the size of aClusters may also vary, we therefore systematically investigate how cluster size and density of aClusters affect the prediction performance of MDeep along with other methods in the simulation study. In addition, we evaluate the effect of an non-informative tree on the prediction performance of MDeep, where clustered aOTUs have an opposite direction of effect to the outcome. To be specific, the parameters are set as follows,

- Cluster Size: aOTUs are clustered at different phylogenetic depths, resulting in different sizes of aClusters. We cluster OTUs into 50, 20, 10 clusters based on C, representing small, medium and large aClusters.
- Signal density: proportion of aClusters of all clusters, which is chosen, which is chosen from 10%, 20%, 40% to represent low, medium and high signal density.

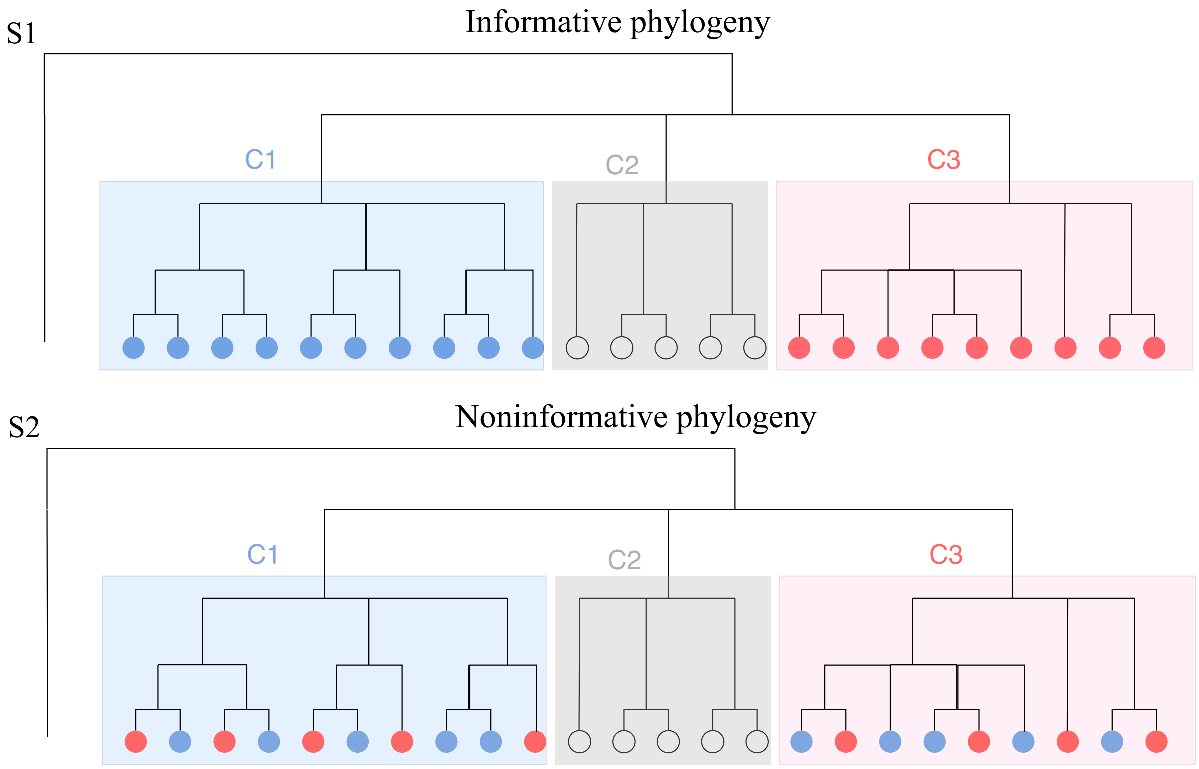

Figure 2.2: Illustration for the simulation strategy. S1 (informative phylogeny): all OTUs in C1 or C3 have the same effect size in same effect direction to the outcome. S2 (non-informative phylogeny), the adjacent OTUs in C1 or C3 have opposite effects to the outcome. C1 and C3 are two aClusters. C2 is a non-aCluster. Red circles represent aOTUs having positive effects to the outcome while blue circles represent aOTUs having negative effects to the outcome.

- Informativeness of phylogeny: for an informative tree (Scenario 1 or S1), we allow aOTUs in each aCluster have the same effect in the same direction to the outcome. For a non-informative tree (Scenario 2 or S2), we let adjacent aOTUs in an aCluster have opposite effects to the outcome, which violates the assumption that closely related aOTUs have similar biological functions and thus have similar effects to the outcome (Fig 2.2).

2.3.3 Generating outcomes based on aClusters

We denote $\mathcal{A}_l$ as the set containing indices of lth aCluster among m aClusters ($l \in 1, ..., m$), x_{ij} represents the jth OTU abundance in sample i, and η_i is the expected outcome value of sample i, which can be generated based on the following linear relationship,

$$\eta_i = \sum_{l=1}^{m} (\sum_{k \in \mathcal{A}_l} x_{ik}) \beta_l \tag{2.4}$$

$$\beta_l \sim N(0, \sigma_b^2) \tag{2.5}$$

Notably, β_l is sampled from a centered normal distribution and thus the effect of an aCluster can be either positive or negative.

We add random error from another centered normal distribution to obtain the continuous outcome y_i for ith sample,

$$y_i = \eta_i + \epsilon_i, \epsilon \sim N(0, \sigma_\epsilon^2) \tag{2.6}$$

We perform inverse logit transformation to η_i to obtain probability π_i, based on which the binary outcome y_i for ith sample is generated from a Bernoulli distribution,

$$y_i \sim \text{Bern}(\pi_i), \text{where } \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \tag{2.7}$$

We set σ_β^2 2 for continuous outcome and 4 for binary outcome. For continuous outcome, we adjust σ_β^2 and σ_ϵ^2 jointly to control the signal-to-noise ratio that fixed the percentage of variability explained by OTUs.

2.3.4 Competing Methods, Model Selection and Evaluation

We compare MDeep to sparse regression models including Lasso and Elastic Net (Enet), and machine learning models such as Random Forest (RF), feed-forward neural network (NN) and CNN. NN, CNN and MDeep are implemented in *TensorFlow* that is an open source artificial intelligence library developed by Google. The details of selecting number of hidden layers, fully-connected layers and other parameters of MDeep are described in the Method section. Sparse regression models such as Lasso and Enet are implemented in *glmnet R* package. RF is implemented in *randomForest R* package with the default parameter setting.

Particularly, NN, designed with the same number of fully-connected layers as in MDeep, is used as a baseline comparison to MDeep for demonstrating the advantages of convolutional operation to capture local correlations of OTUs at different phylogenetic depths and at the same time reduce possible overfitting. Since a typical CNN does not automatically exploit the phylogenetic information, we also include a typical CNN with the same architecture as MDeep as a comparison to demonstrate the importance of utilizing the phylogenetic information. Specifically, we randomly shuffle OTUs in the input layer of CNN instead of forcing phylogenetically correlated OTUs clustered as in MDeep. In this way, each convolutional operation will likely include both aOTUs and non-aOTUs, making the convolutional operation less effective.

Optimal tuning parameter of Lasso and Enet is selected based on five-fold cross-validation (5-CV). In 5-CV, the whole training set is divided into 5-folds, where 4-folds is used to train the model and 1-fold is used to obtain the metric for evaluating prediction performance of the trained model. Specifically, we used PMSE (Predicted Mean Square Error) as the metric for regression and AUC (Area Under the Curve) for binary classification. The optimal tuning parameter is chosen based on the average metric in 5-CV and a final model is fitted using the whole training set with the optimal tuning parameter. To evaluate an independent testing set, we use R^2 and PMSE for regression; R^2 and metrics including Sensitivity, Specificity, Accuracy, Precision, F1 score, and MCC for binary classification. In each simulation setting, 50 training and testing sets are generated and the average metric is reported.

2.3.5 Additional data transformation and data generation strategy

We also apply CLR (centered log ratio transformation) to relative OTU abundance for evaluating the impact of data transformation on compositional data[46]. Additionally, we use negative binomial distribution with considering the microbial correlations in the data

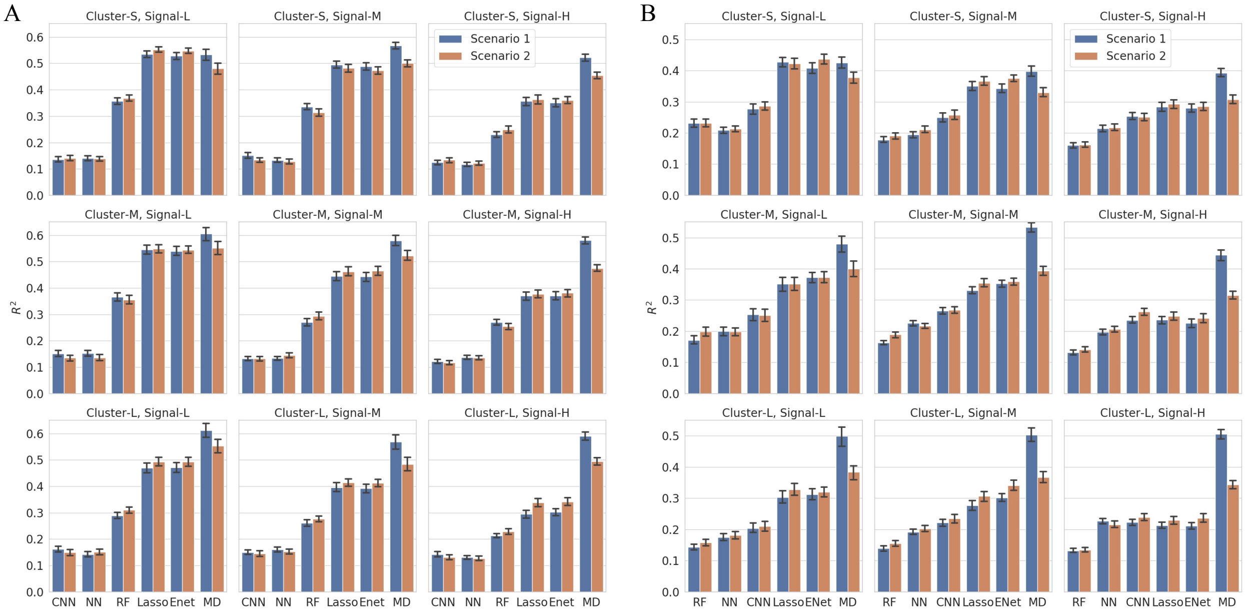

Figure 2.3: Prediction performance measured by R^2 for (**A**) continuous outcome and (**B**) binary outcome in the simulation study. Scenario1 (S1) represents informative phylogeny and Scenario2 (S2) represents non-informative phylogeny. Cluster-S, -M, and -L represent small, medium and large clusters. Signal-L, -M, and -H represent low, medium and high signal density, respectively. The whisker shows the standard deviation of R^2 for each method.

generation [47]. First, we estimate the microbial correlation from the real dataset using MAGMA[48]. Second, we generated correlated multivariate normal data with mean as 0 and correlation matrix as the estimated microbial correlations. Third, correlated multivariate normal data is transformed to the copula space using the cumulative distribution function (CDF) of the standard normal distribution. Last, correlated OTU counts following negative binomial distribution is generated by the inverse CDF of negative binomial distribution performed on the values obtained in the copula space. We then perform CLR on the relative OTU abundance.

2.3.6 Results of prediction for continuous outcome

Overall Comparison: When the phylogeny is informative, regardless of signal density and cluster size, MDeep outperforms other methods overall by obtaining a higher R^2

(Fig 2.3A) and lower PMSE (Fig 2.6A) across different scenarios and signal densities. Particularly, when the cluster size is large or the signal density is high, MDeep has an enormous advantage over other methods due to its ability for exploiting the phylogenetic structure. The improved prediction by MDeep confirms the benefits of encoding the phylogenetic information via the convolutional operation. Especially, NN does not perform well either due to overfitting or inability to utilize the phylogenetic information. MDeep has a significant performance gain than a typical CNN without considering the phylogenetic information, indicating the importance to include phylogeny in the predictive modeling. Moreover, it is expected and observed that other methods achieve similar performance between S1 & S2 as they are not able to utilize phylogenetic information. In contrast, the prediction performance of MDeep deteriorates in S2 compared to S1. Nevertheless, MDeep still outperforms other methods in general and has a clear advantage over other methods even when the tree is non-informative. These observations indicate the fact that utilization of phylogenetic information improves MDeep's prediction performance, while has little effect on methods without considering the tree.

Signal Density: For each cluster size, we observe an overall decrease in R^2 when the signal density increases in both S1 and S2, which can be explained by the result of decreasing individual effects with an increasing number of aOTUs while the percentage of variability explained by aOTUs is fixed. This reduction in individual effects is unfavorable for both sparse regression methods and tree-based methods. The decreasing R^2 of sparse regression methods is attributed to weaker individual effects of more aOTUs, a scenario where sparse regression methods tend to have a low sensitivity and specificity to identify aOTUs. As a result, the identified OTUs cannot explain an enough percentage of variability of the outcome, leading to decreasing prediction performance. The deteriorating performance of RF can be explained by an insufficient tree depth to accommodate an expansion of aOTUs when the signal density increases, resulting in a potential underfitting. However, the increase of signal density imposes little adverse effect on either NN, CNN or MDeep, which indicates the neural network architecture is robust across different levels of signal density. This robustness may be attributable to two reasons. First, neural architecture does not assume sparsity in the model while utilizes the information of all aOTUs via feature mapping across multiple layers. Second, the dropout technique employed in fully-connected layers and regularization technique for the weights can further reduce the potential risk of overfitting when the signal is sparse. In all, MDeep has an overall better prediction than other methods across various levels of signal density. To justify the significant improvement of MDeep, we perform a paired wilcoxon test between R^2 of MDeep and R^2 of any other method in each scenario. We find that MDeep achieves an overall statistically significant improvement over any other method in each scenario (pvalue<0.05) except in the "Cluster-S, Signal-L" scenario where MDeep, Lasso and Enet have comparable performance.

Cluster Size: It is noteworthy that reducing cluster size decreases the phylogenetic information. We thus observe that the improved prediction of MDeep over other methods is diminishing as the cluster size becomes small. When signal density is low and the cluster size is small ("Cluster-S, Signal-L"), the prediction performance of MDeep is on par with other sparse regression models, and the improved prediction of MDeep over other methods increases as the cluster size increases. The trend is similar when the signal density is medium or high. Clearly, MDeep benefits more from a large aCluster because the convolutional operation will mostly capture aOTUs from a large aCluster, however, may include irrelevant OTUs when the aCluster is small.

OTU-level β: In reality, the effect size of aOTUs may vary within each aCluster. Thus, we perform an additional simulation study to investigate whether MDeep is robust when β of aOTUs varies within each aCluster. Without loss of generality, we choose the scenario when cluster is large and signal density is high (Cluster-L and Signal-H) for continuous outcome. In order to add the variability to the cluster-level β, we sample OTU-level β in each aCluster from a normal distribution with cluster-level β as mean and 0.1 as standard deviation. As expected, prediction performance remains similarly between cluster-level β and OTU-level β (Fig 2.3A).

Table 2.1: Prediction performance in binary classification (Cluster-L, Signal-L)

	Scenario	NN	RF	Lasso	Enet	MDeep
Sensitivity	S1	0.6620	0.6534	0.7208	0.7290	**0.8188**
	S2	0.6794	0.6730	0.7480	0.7460	**0.7904**
Specificity	S1	0.6692	0.6586	0.7346	0.7350	**0.8254**
	S2	0.6582	0.6508	0.7432	0.7442	**0.772**
Accuracy	S1	0.6656	0.6560	0.7277	0.7320	**0.8221**
	S2	0.6688	0.6619	0.7456	0.7451	**0.7812**
Precision	S1	0.6709	0.6594	0.7387	0.7406	**0.8271**
	S2	0.6702	0.6593	0.7437	0.7455	**0.7777**
MCC	S1	0.3354	0.3139	0.4588	0.4665	**0.6468**
	S2	0.3420	0.3250	0.4935	0.4921	**0.5644**
F1 scores	S1	0.6622	0.6544	0.7251	0.7310	**0.8210**
	S2	0.6712	0.6649	0.7429	0.7443	**0.7827**

2.3.7 Results of prediction for binary outcome

We repeat the same simulations for the binary classification. Compared to regression, we find the overall trends of MDeep are similar in each setting. However, there are some changes of other methods. First, performance of NN and CNN is improved and both are superior to RF (Fig 2.3B). Second, MDeep benefits more by exploiting the phylogenetic information demonstrated by more improved prediction in S1 over S2 between binary classification and regression. In addition to R^2, we also include other metrics (Fig 2.6B, Table 2.1). Based on all evaluation metrics, we conclude that MDeep is consistently superior to other methods when the cluster size is relatively high or signal is relatively dense, and comparable to other methods in other scenarios. Without loss of generality, we use R^2 as the metric to evaluate the significant improvement of MDeep. We find that MDeep achieves an overall statistically significant improvement over any other method in each scenario by using paired wilcoxon signed rank test (pvalue<0.05) except in the “Cluster-S, Signal-L" scenario where MDeep, Lasso and Enet are comparable.

2.3.8 Results of additional data transformation and generation

The simulation results for using CLR as data transformation are presented in Fig 2.8A, B. Though CLR does not necessarily improve the prediction performance compared to relative OTU abundance (Fig 2.3), the overall trend still holds. MDeep still has a clear advantage over other methods when the tree is informative and is comparable to other methods when the tree is non-informative. In addition, simulation results for data generated from negative binomial distribution with considering the microbial correlations have the similar trends as the results from Dirichlet multinomial distribution (Fig 2.8 C, D).

2.4 Application of MDeep in real datasets

We apply MDeep along with other methods aforementioned to three real datasets. The first dataset is acquired from a study that explores how gut microbiome varies across age and topography [49]. The second dataset is obtained from a longitudinal comparative study that investigates the fecal microbiome of monozygotic (MZ) and dizygotic (DZ) twins pairs born in Malawi who became discordant for kwashiorkor [50]. The third dataset is taken from a study which investigates the role for intestinal bacteria in rheumatoid arthritis[51]. To incorporate the phylogenetic information via the convolutional operation, we apply hierarchical agglomerative clustering (HAC) algorithm to the phylogeny-induced correlation structure C, which is inferred from the phylogenetic tree. In this way, phylogenetically related OTUs will be close to each other and irrelevant OTUs will be far apart from each other. HAC is an unsupervised clustering algorithm, which will group OTUs in a ‘bottom-up’ approach, where each OTU is considered as a cluster in the bottom and merged with close OTUs to form new clusters and move up to form a hierarchy. Therefore, no number of clusters needs to be specified and no membership needs to be assigned for OTUs since they will be ordered in a hierarchical structure automatically. In addition, hierarchical clustered OTUs can potentially reflect the taxonomic grouping of OTUs.

2.4.1 Predicting chronological age based on gut microbiome of individuals in USA

In this study, there are 531 individuals (115 from Malawi, 100 from Venezuela, and 316 from the USA). Gut microbiome of all the individuals are profiled using 16S rRNA gene-targeted sequencing and the datasets are deposited to Qiita [52] with assigned study ID 850. We first download and process the datasets using QIMIE, resulting in a total of 14,170 OTUs. The phylogenetic tree is constructed using FastTree [53]. We then use individuals in USA for the prediction task as the sample size is relatively large for the deep learning approach. Subsequently, we carry out a series of data pre-processing steps as described in [54], consisting of i) removing outlier samples; ii) removing less informative and noisy OTUs(OTU prevalence $<$ 10%; median non-zero counts $<$ 10); iii)normalizing zero-inflated OTU counts using GMPR [55]; iv) replacing outlier counts using winsorization with 97% quantile; v) reducing the influence of highly abundant taxa counts by square-root transformation. As a result, we have 308 individuals profiled with 1087 OTUs for model training and testing. The final phylogenetic tree is tuned accordingly with 1087 leaves left and correlation matrix $C_{1087\times 1087}$ is calculated based on the tree.

We treat age as either a continuous outcome or a binary outcome to evaluate the prediction performance for both regression and binary classification. Specifically, for regression, ages of all individuals are directly treated as continuous outcomes; for binary classification, we classify all the individuals to three age groups, namely, baby (age$\leq$3 years, n=54), child (3<age<18 years, n=125), and adult (age$\geq$18 years, n=129) [54]. Particularly, we evaluate the prediction performance in classifying two age groups: "baby vs child" and "child vs adult". We exclude "baby vs adult", an easy case where all methods have superior and indistinguishable prediction performance due to the drastically distinct microbiome composition.

In contrast, adjacent age periods such as “baby vs child" or “child vs adult" are ideal for identifying the difference of the prediction performance among different methods.

We compare MDeep to other methods aforementioned. In addition, we include Ph-CNN as a comparison for binary classification as it is only designed for binary outcome. To incorporate the phylogenetic information via the convolutional operation, we adopt the hierarchical agglomerative clustering algorithm (HAC) to cluster the OTUs based on their phylogenetic correlation, making phylogenetically related OTUs close to each other and irrelevant OTUs far apart from each other. We choose HAC as default clustering algorithm because no number of clusters needs to be specified and hierarchical clustered OTUs could potentially reflect the taxonomic grouping of OTUs. Moreover, we compare the MDeep to CNN without using the phylogenetic information by randomly shuffling OTUs. To evaluate the prediction performance, we randomly split the dataset into two sets with 80% as the training set and the rest 20% as the testing set. Tuning parameter selection and model fitting are performed on the training set and prediction performance is evaluated on the testing set. We repeat the randomization 50 times and report the metrics: R^2 and PMSE for regression; R^2, sensitivity, specificity, accuracy, precision, F1 score, and MCC for binary classification.

For continuous prediction of age, the overall trends of prediction performance of all methods are similar to the simulation study, where MDeep ranks top by achieving the highest R^2 and lowest PMSE, followed by Enet. RF and NN do not perform well. Importantly, MDeep outperforms CNN, which indicates an improved prediction is achieved by utilizing the phylogenetic information (Fig 2.4A, 2.9A). In addition, we perform a paired wilcoxon test between R^2 of MDeep and R^2 of any other method and find that MDeep achieves an overall statistically significant improvement over any other method (pvalue$<$0.05).

For binary classification of age groups, MDeep obtains the highest R^2 in the comparisons of both “baby vs child" and “child vs adult" (Fig 2.4B, C). In addition, other metrics further demonstrate the MDeep's superb performance overall (Fig 2.7B,C and Table 2.2).

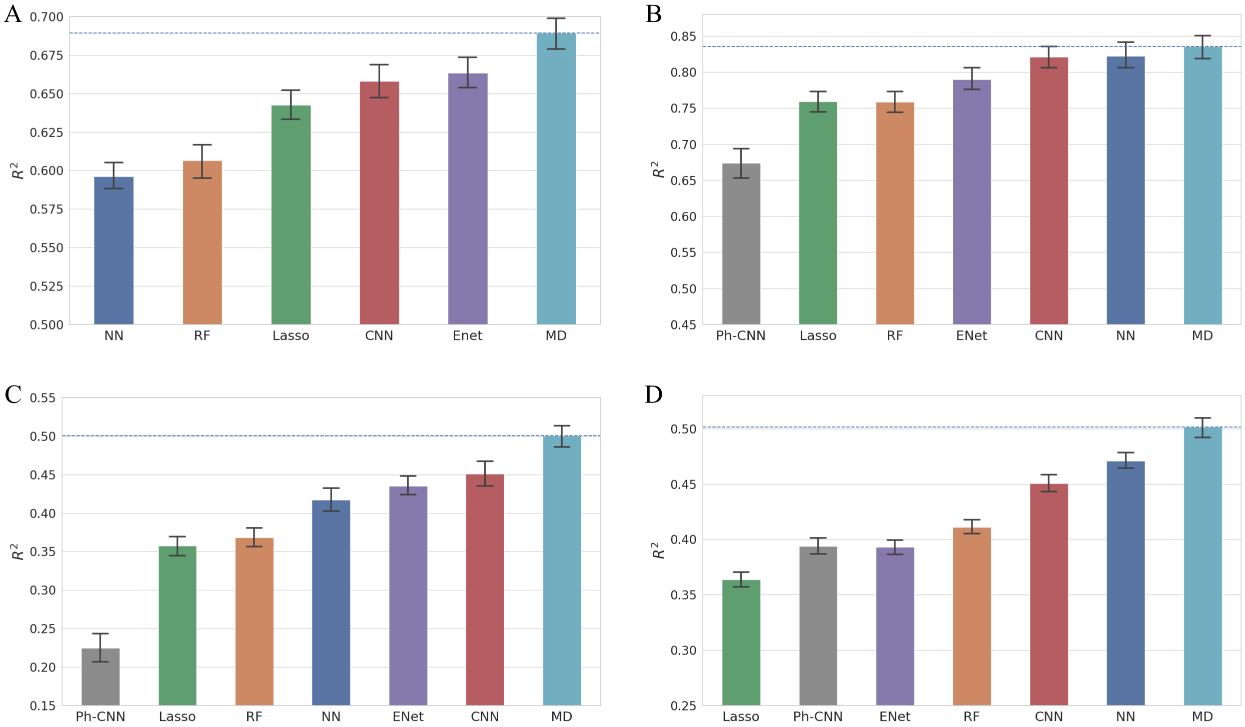

Figure 2.4: Prediction performance of chronological age based on gut microbiome of individuals in USA: (**A**) R^2 for all ages (**B**) R^2 for “Baby vs Child" (**C**) R^2 for “Child vs Adult". Prediction performance of gender based gut microbiome of Malawian twins: (**D**) R^2 for male and female. The blue dashed line the mean value of R^2 for MDeep. The whisker shows the standard deviation of R^2 for each method.

Except Ph-CNN, neural network architecture is consistently robust in the "Baby vs Child" comparison, where MDeep, NN and CNN perform better than the others (Fig 2.4B). In contrast, the relative performance of NN compared to CNN and MDeep is deteriorated significantly in the "Child vs Adult" comparison (Fig 2.4C) because of potential overfitting. Ph-CNN has overall worst performance in both binary classification tasks maybe because of the neural network architecture is suboptimal or the phylogenetic information is not fully exploited. Altogether, MDeep has the best prediction performance overall in predicting age values and classifying age groups. In addition, the accurate prediction of age based on human gut microbiome further validates the hypothesis that the composition of human gut microbiome changes with age [56]. Finally, the paired wilcoxon signed rank test between R^2 of MDeep and R^2 of any other method demonstrates that MDeep achieves a statistically

Table 2.2 Prediction performance of gender based on gut microbiome of Malawian twin pairs. BC represents "Baby vs Child"; CA represents "Child vs Adult".

	Type	Lasso	RF	NN	Enet	CNN	MDeep
Sensitivity	BC	0.9856	0.9856	**0.9920**	0.9784	0.9904	0.9896
	CA	0.7531	0.7669	0.7977	0.7808	0.8123	**0.8254**
Specificity	BC	0.8018	0.8327	0.8691	0.8636	0.8764	**0.8891**
	CA	0.7736	0.7712	0.776	0.8112	**0.8176**	0.8160
Accuracy	BC	0.9294	0.9389	0.9544	0.9433	0.9556	**0.9589**
	CA	0.7631	0.7690	0.7871	0.7957	0.8149	**0.8208**
Precision	BC	0.9216	0.9329	0.9469	0.9438	0.9493	**0.9544**
	CA	0.7826	0.7803	0.7917	0.8165	0.8276	**0.8286**
MCC	BC	0.8334	0.8568	0.8933	0.8674	0.896	**0.9041**
	CA	0.5317	0.5405	0.5774	0.5957	0.634	**0.6459**
F1 scores	BC	0.9516	0.9577	0.9684	0.9601	0.969	**0.9712**
	CA	0.7634	0.7717	0.7924	0.7956	0.8172	**0.8242**

significant improvement of prediction over any other method in the comparison of “child vs adult" (pvalue<0.05), and over Ph-CNN, Lasso, RF and Enet in the comparison of “baby vs child" (pvalue<0.05). To sum up, based on the results of both regression and binary classification, MDeep has the overall best prediction performance for the chronological age.

2.4.2 Predicting gender based on gut microbiome of human twins

One previous research shows that dysbiosis of gut microbiota is involved in metabolic syndrome development, which has a different incidence between men and women [57]. In light of this, we seek to study whether the composition of gut microbiome is different between male and female discordant for kwashiorkor. To achieve this, we treat gender as a binary outcome and classifying gender based on gut microbiome using the dataset collected from twin pairs in Malawi. This dataset is also conducted with 16S rRNA gene-targeted sequencing and deposited in Qiita with study ID 737. We download and process the dataset, resulting in 4321 OTUs profiling a total of 1041 twins (including MZ and DZ) consisting of 483 females, 512 males and 46 samples with missing gender information. After aforementioned data preprocessing steps are performed, we have 995 individuals profiled with 2291 OTUs for the binary classification task. Accordingly, the phylogenetic tree is constructed using FastTree [53] and truncated with the final number of leaves the same as the number of left OTUs. Thus, a correlation matrix $C_{2291\times2291}$ is calculated based on the tree.

Table 2.3 Prediction performance for gender based gut microbiome of human twins with kwashiorkor

	Lasso	RF	NN	ENet	CNN	MDeep
Sensitivity	0.802	0.812	**0.8504**	0.8153	0.8274	0.8227
Specificity	0.7352	0.7468	0.7315	0.802	0.8001	**0.8263**
Accuracy	0.7695	0.7804	0.7915	0.8086	0.8136	**0.8235**
Precision	0.7667	0.7775	0.7743	0.8165	0.8173	**0.8364**
MCC	0.5384	0.5604	0.5864	0.6166	0.6271	**0.6475**
F1 score	0.7827	0.7931	0.8088	0.815	0.8212	**0.8284**

For binary classification of gender, MDeep significantly outperforms other any method by achieving a much higher R^2 (pvalue$<$0.05) via a paired wilcoxon signed rank test. Neural network architectures are superior to sparse regression methods and RF, where MDeep, NN and CNN obtain a higher R^2 (Fig 2.4D). However, Ph-CNN still does not perform well. Moreover, MDeep maintains the overall top performance across various evaluation metrics (Fig 2.7D and Table 2.3). The accurate classification of gender indicates that composition of gut microbiome is different between men and women affected by kwashiorkor. In other words, there is a sex-dependent effect on shaping the gut microbiome.

2.4.3 Predicting rheumatoid arthritis based on human microbiome

Besides traits (e.g. age and gender), we also apply MDeep to a disease-related study, which investigates the role for intestinal bacteria in Rheumatoid arthritis (RA) [50]. In this study, there are four groups of samples, which are new-onset rheumatoid arthritis (NORA), chronic, treated rheumatoid arthritis (CRA), psoriatic arthritis (PsA), and healthy controls

(HLT) with sample size 44, 26, 16 and 28 respectively. The prediction task is the binary classification of NORA from HLT. Since NORA has been found associated with fewer marker taxa (e.g. Prevotella copri), the dataset is helpful to evaluate whether MDeep could maintain robust performance in the unfavorable scenario, where only a fewer marker taxa are associated with the outcome. The dataset is analyzed by MOTHUR [51], resulting in 997 OTUs and a phylogenetic tree built by FastTree [45]. After aforementioned data pre-processing steps are performed, an OTU abundance matrix consisting of 887 OTUs and the corresponding phylogeny-induced correlation structure $C_{887\times887}$ are obtained for model training and testing.

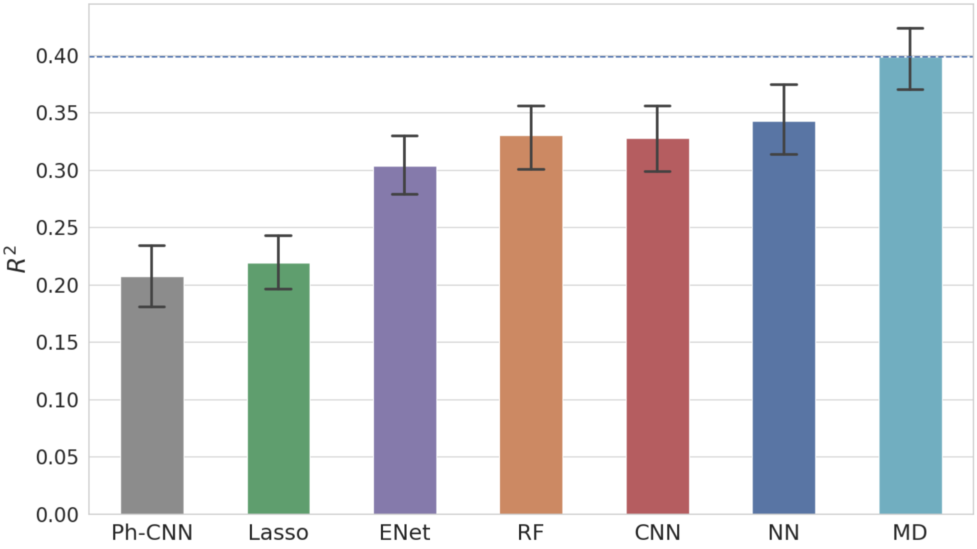

Figure 2.5: Prediction performance of new-onset rheumatoid arthritis on human microbiome: R^2 for new-onset rheumatoid arthritis and healthy. The blue dashed line the mean value of R^2 for MDeep. The whisker shows the standard deviation of R^2 for each method.

For the binary classification for NORA from HLT, MDeep still remains the highest R^2 (2.5). The improvement of MDeep over other methods is significant (pvalue$<$0.05 via a paired Wilcoxon signed-rank test). We also find that overall performance of neural network-based approaches (MDeep, NN, CNN) outperforms sparse regression models (Elastic Net, Lasso and Random Forest). Additionally, MDeep maintains the overall top performance across various evaluation metrics (Fig 2.10 and Table 2.4). The accurate classification of NORA

Table 2.4 Prediction performance of NORA based on human intestinal microbiome

	Lasso	ENet	RF	Ph-CNN	CNN	NN	MDeep
Sensitivity	0.8682	0.8145	**0.8831**	0.7556	0.8473	0.8393	0.8407
Specificity	0.4177	0.6466	0.6208	0.5919	0.6468	0.6651	**0.7166**
Accuracy	0.6686	0.73	0.7629	0.6886	0.7657	0.7686	**0.7857**
Precision	0.699	0.7726	0.7741	0.7468	0.7907	0.7953	**0.8245**
MCC	0.2986	0.463	0.5245	0.7377	0.515	0.5193	**0.5642**
F1 score	0.7542	0.7756	0.8089	0.398	0.8066	0.8063	**0.8188**

from HLT strengthens the claim that MDeep will maintain robust performance even if the signal is sparse, where only a fewer marker taxa are associated with the outcome.

2.4.4 Computational performance

Comparing MDeep and Ph-CNN, the two deep learning models that utilize the phylogenetic information, we find MDeep provides superior computational performance since clustering phylogenetically related OTUs only based on the evolutionary model using HAC. Ph-CNN, on the other hand, relying on MultiDimensional Scaling projection, is much more computationally intensive. To demonstrate this, we benchmark the computational performances of MDeep and Ph-CNN on a server consisting of dual 8-core Xeon Sandy Bridge E5–2670 processors with 64 GB system memory. We use the processed dataset collected from twin pairs in Malawi consisting of 995 individuals profiled with 2291 OTUs. Without loss of generality, we record the time of first 100 epochs in one batch in the training process for classifying gender. We find that Ph-CNN runs slowly, which is as nearly as 90 times of the number of epochs. In contrast, the running time of MDeep is stable and increases little (Table 2.13, Fig 2.8). For example, in first 100 epochs, Ph-CNN takes 8879 seconds, compared to 57 seconds from MDeep, that is, MDeep is 158 times faster than Ph-CNN. Similarly, MDeep is 8 times, 42 times and 121 times faster than Ph-CNN in first 1, 10 and 50 epochs. This observation indicates that the gain of computational efficiency is more evident when the number of epochs increases for MDeep compared to Ph-CNN. Overall, MDeep not only outperforms Ph-CNN in binary classification but also in computational performance.

2.5 Discussion

In this work, we develop a microbiome-based deep learning model, MDeep for predicting both continuous and binary outcomes. The effectiveness of the proposed model relies on its ability to maximize the utilization of the information in the microbiome data, which consists of both phylogenetic tree and the OTU abundance. First, MDeep exploits the phylogenetic information by clustering OTUs based on phylogenetic correlation before the input layer. This strategy induces the local smoothing, which is phylogenetically correlated OTUs will be more likely captured together by the convolutional operation in the local receptive field. Second, MDeep encourages the local connectivity in the feature mapping, where phylogenetically related OTUs at a lower taxonomy are more likely maintained close to each other at a higher taxonomy.

We carry out comprehensive simulations to evaluate the prediction performance with the considerations of cluster size, signal density, and informativeness of the phylogenetic tree. As a result, MDeep favors the scenarios of dense and large-clustered signals where it outperforms other methods and it is still comparable to others as the signal density and cluster size decreases. Especially, MDeep not only has a superior prediction performance to NN and CNN with a similar network architecture, but also superior to sparse regression models and random forest. This observation strengthens the benefit of exploiting phylogenetic information in predictive modeling. Moreover, the same observation persists when the tree is uninformative, which demonstrates the robustness of MDeep. In reality, a noisy or mis-specified tree is not uncommon because of an inappropriate tree construction method or inconsistency between DNA sequence similarity and biological similarity. In the analysis of two real datasets, MDeep keeps the trends to outperform other methods overall. Particularly, MDeep outperforms Ph-CNN, another deep learning method that also utilizes the phylogenetic information, in binary classification. In addition, MDeep computes much faster.

Although CNN has been widely used to improve the prediction performance with a relatively large sample size ranging from thousands to millions, we demonstrate here that CNN can potentially achieve a high prediction accuracy when the sample size is moderate (e.g. hundreds). This is achieved by developing a "shallow" neural network architecture with a few convolutional layers and fully-connected layers.

We use mapped reads of OTUs as the feature representation. Notably, reference-free or alignment-free approach is an alternative feature representation approach of microbiome data. Kmers of raw reads can be directly used as features and phylogenetic tree can be constructed based on Kmers [58, 59]. This approach has the benefit of skipping computationally costly sequence alignments required in OTU-picking, thus making the process of training a prediction model more efficient. The comparison between OTU-based and Kmer-based MDeep needs further investigations.

2.6 Appendix

2.6.1 Supplementary figures

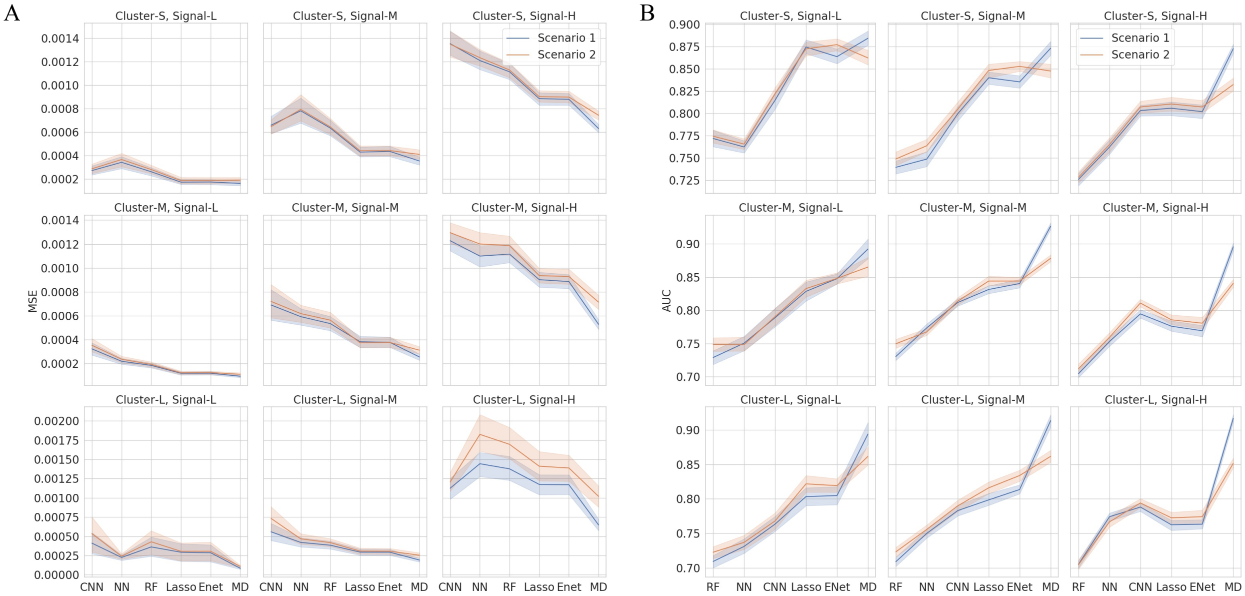

Figure 2.6: Prediction performance measured by PMSE for continuous outcome (**A**) and AUC for binary outcome (**B**) in the simulation study. Scenario1 (S1) represents informative phylogeny and Scenario2 (S2) represents non-informative phylogeny. Cluster-S, -M, and -L represent small, medium and large clusters. Signal-L, -M, and -H represent low, medium and high signal density, respectively.

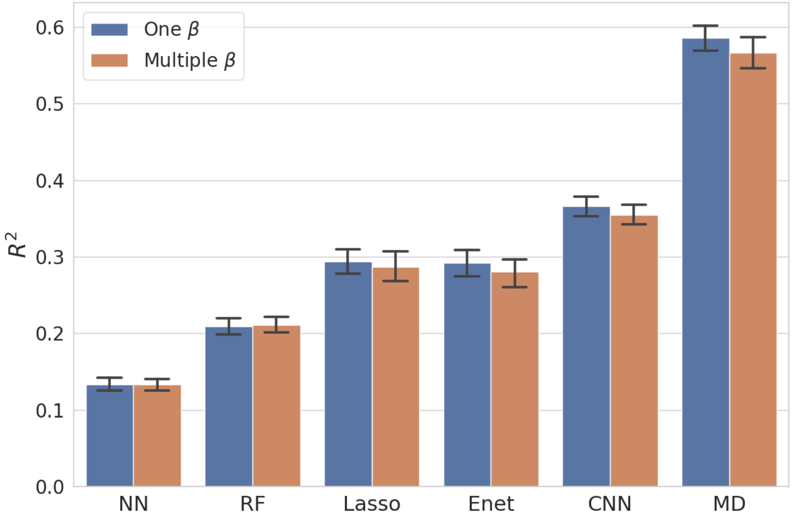

Figure 2.7: Prediction performance measured by R^2 for continuous outcome when the tree is informative when cluster is large and signal density is high (Cluster-L and Signal-H). One cluster-level β indicate aOTUs within a aCluster have the same β value. Multiple OTU-level β indicate aOTUs within a aCluster have different β values, which are generated from a normal distribution with cluster-level β as mean and 0.1 as standard deviation.

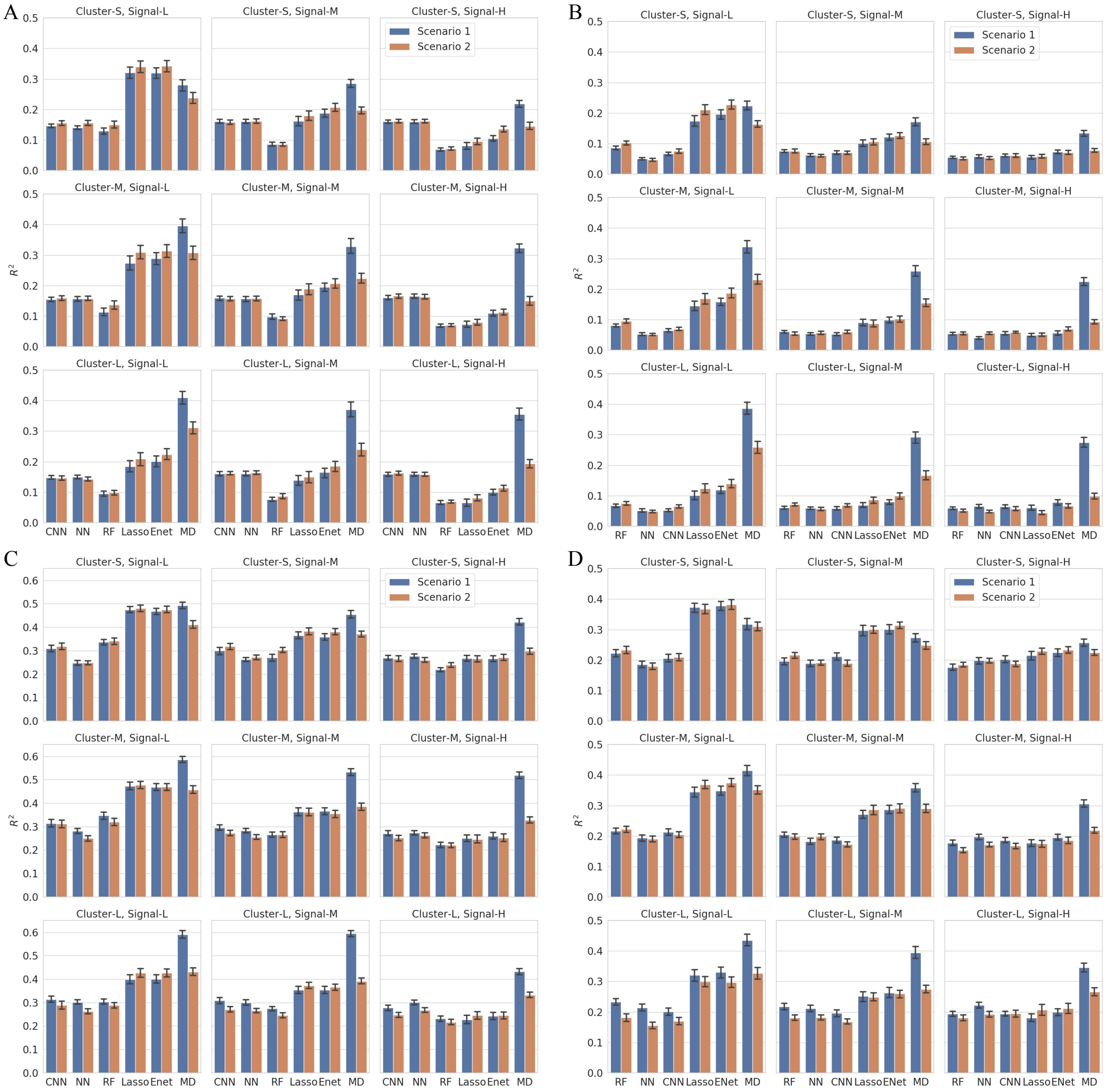

Figure 2.8: R^2 for continuous-outcome (**A**) and binary-outcome(**B**) simulations across different signal levels and scenarios when the abundance of associated OTU clusters is low. S1: phylogeny-informative scenarios, and S2: phylogeny-non-informative scenarios; Cluster-S, -M, and -L represent small, medium and large clusters, and Signal-L, -M, and -H represent low, medium and high signal density, respectively.

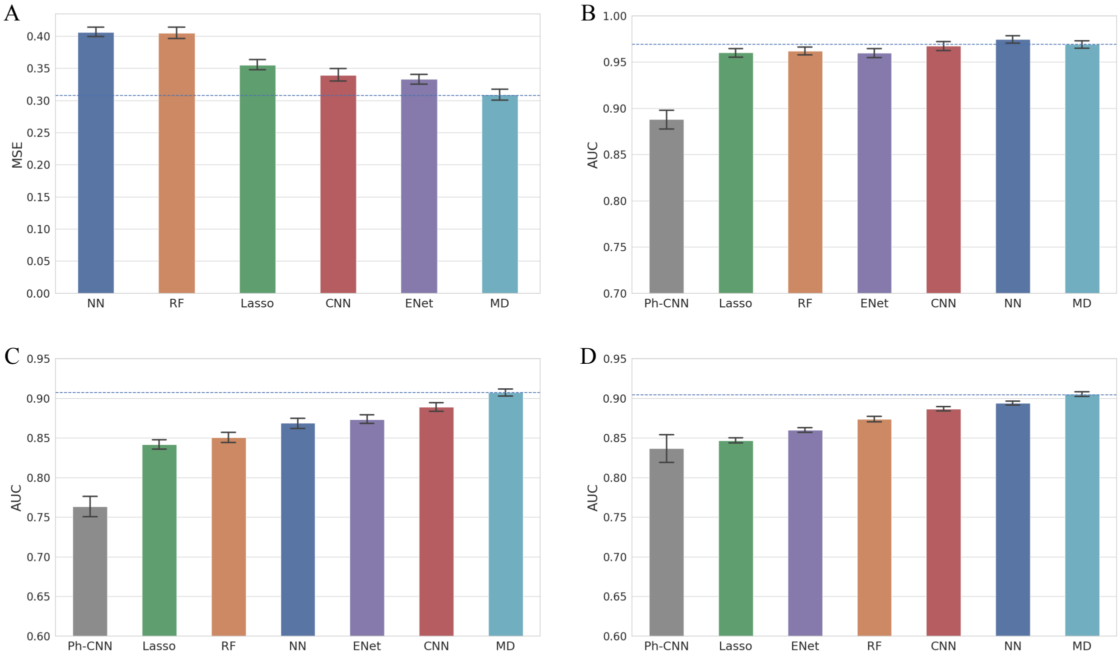

Figure 2.9: Prediction performance of chronological age based on gut microbiome of individuals in USA (**A**, **B**, **C**): (**A**): PMSE for all ages (**B**) AUC for "Baby vs Child" (**C**) AUC for "Child vs Adult". Prediction performance of gender based gut microbiome of Malawian twins (**D**): AUC for male and female. The blue dashed line the mean value of the metric of MDeep.

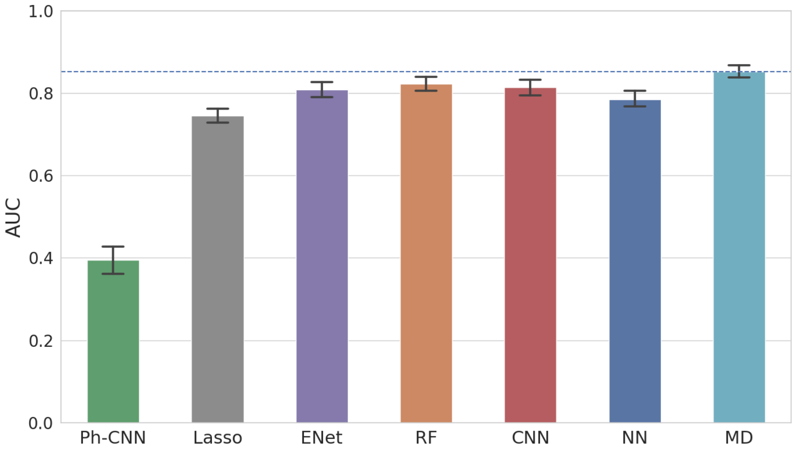

Figure 2.10: Prediction performance of new-onset rheumatoid arthritis on human microbiome: AUC for new-onset rheumatoid arthritis and healthy.

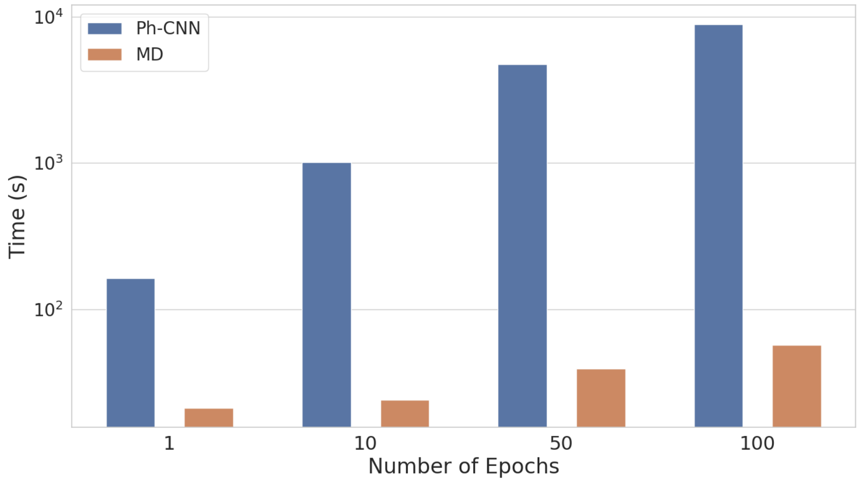

Figure 2.11: Running time comparison between Ph-CNN and MDeep

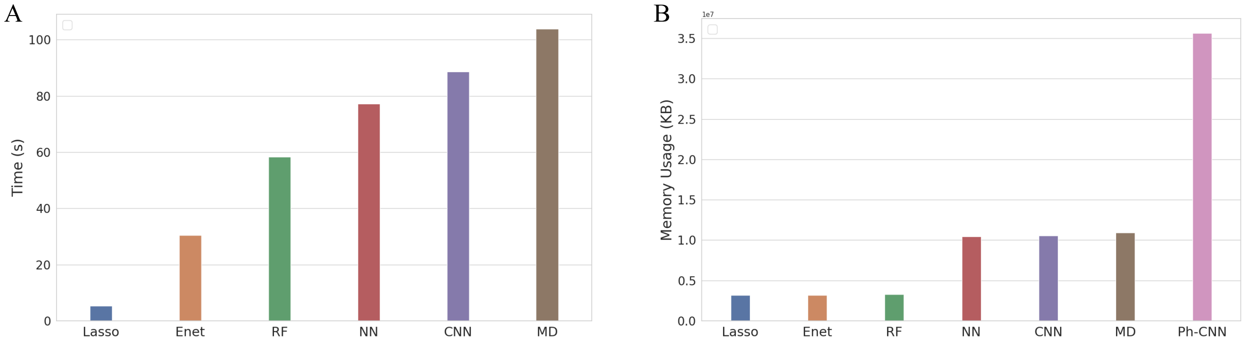

Figure 2.12: (**A**):Running time comparison among all the methods but Ph-CNN. (**B**): Memory usage comparison among all the methods.

2.6.2 Supplementary tables

Table 2.5: Prediction performance in binary classification (Cluster-S, Signal-L). Top performed method in each metric is bold.

	Scenario	NN	RF	CNN	Lasso	ENet	MDeep
Sensitivity	S1	0.692	0.7100	0.7306	0.7906	0.7932	**0.8018**
	S2	0.6844	0.7024	0.7402	0.7814	**0.7906**	0.7744
Specificity	S1	0.6924	0.6978	0.7336	0.7960	0.7848	**0.8032**
	S2	0.6984	0.7128	0.7470	**0.7988**	0.7936	0.7924
Accuracy	S1	0.6922	0.7039	0.7321	0.7933	0.7890	**0.8025**
	S2	0.6914	0.7076	0.7436	0.7901	**0.7921**	0.7834
Precision	S1	0.6956	0.7019	0.7327	0.7975	0.7892	**0.8044**
	S2	0.6966	0.7113	0.7461	**0.7974**	0.7944	0.7915
MCC	S1	0.3877	0.409	0.4658	0.5894	0.5803	**0.6065**
	S2	0.3861	0.4164	0.4889	0.5824	**0.5858**	0.5688
F1 scores	S1	0.6906	0.7049	0.7302	0.7921	0.7897	**0.8021**
	S2	0.6873	0.7058	0.7418	0.7878	**0.7914**	0.7814

Table 2.6: Prediction performance in binary classification (Cluster-S, Signal-M). Top performed method in each metric is bold.

	Scenario	NN	RF	CNN	Lasso	ENet	MDeep
Sensitivity	S1	0.7072	0.6842	0.7162	0.7648	0.7678	**0.7768**
	S2	0.6796	0.6836	0.7178	0.7530	0.7620	**0.7698**
Specificity	S1	0.6462	0.6696	0.7332	0.748	0.7526	**0.7964**
	S2	0.7046	0.6798	0.7302	0.7736	**0.7796**	0.7526
Accuracy	S1	0.6767	0.6769	0.7247	0.7564	0.7602	**0.7866**
	S2	0.6921	0.6817	0.7240	0.7633	**0.7708**	0.7612
Precision	S1	0.6694	0.6744	0.7294	0.7531	0.7571	**0.7956**
	S2	0.7023	0.6827	0.728	0.7703	**0.7775**	0.7590
MCC	S1	0.3575	0.3552	0.4508	0.5146	0.5217	**0.5759**
	S2	0.3887	0.3649	0.4498	0.5279	**0.5429**	0.5247
F1 scores	S1	0.6847	0.678	0.7215	0.7576	0.7616	**0.7842**
	S2	0.6866	0.6816	0.7214	0.7606	**0.7687**	0.7628

Table 2.7: Prediction performance in binary classification (Cluster-S, Signal-H). Top performed method in each metric is bold.

	Scenario	NN	RF	CNN	Lasso	ENet	MDeep
Sensitivity	S1	0.6824	0.6558	0.7222	0.7238	0.7232	**0.7884**
	S2	0.6904	0.6800	0.7246	0.7384	0.7336	**0.7544**
Specificity	S1	0.7048	0.6686	0.7326	0.7374	0.7354	**0.7886**
	S2	0.6830	0.6532	0.7176	0.7278	0.7216	**0.7450**
Accuracy	S1	0.6936	0.6622	0.7274	0.7306	0.7293	**0.7885**
	S2	0.6867	0.6666	0.7211	0.7331	0.7276	**0.7497**
Precision	S1	0.7028	0.6655	0.7322	0.7353	0.7334	**0.7908**
	S2	0.6901	0.6633	0.7208	0.7311	0.7253	**0.7490**
MCC	S1	0.3910	0.3253	0.4572	0.4629	0.4598	**0.5789**
	S2	0.3783	0.3348	0.4442	0.4676	0.4564	**0.5016**
F1 scores	S1	0.6890	0.6597	0.7252	0.7281	0.7272	**0.7882**
	S2	0.6858	0.6701	0.7211	0.7336	0.7285	**0.7501**

Table 2.8: Prediction performance in binary classification (Cluster-M, Signal-L). Top performed method in each metric is bold.

	Scenario	NN	RF	CNN	Lasso	ENet	MDeep
Sensitivity	S1	0.6778	0.6844	0.7192	0.7514	0.7548	**0.8204**
	S2	0.6740	0.6966	0.715	0.7522	0.7636	**0.7914**
Specificity	S1	0.6844	0.6508	0.7116	0.7562	0.7574	**0.8114**
	S2	0.6834	0.6784	0.7152	0.7520	0.7436	**0.7788**
Accuracy	S1	0.6811	0.6676	0.7154	0.7538	0.7561	**0.8159**
	S2	0.6787	0.6875	0.7151	0.7521	0.7536	**0.7851**
Precision	S1	0.6890	0.6635	0.7150	0.7572	0.7577	**0.8154**
	S2	0.6872	0.6862	0.7158	0.7530	0.7540	**0.7833**
MCC	S1	0.3674	0.3368	0.4323	0.5097	0.5138	**0.6336**
	S2	0.3629	0.3768	0.4317	0.5061	0.5086	**0.5730**
F1 scores	S1	0.6786	0.6725	0.7158	0.7524	0.7547	**0.8167**
	S2	0.6754	0.6899	0.7141	0.7513	0.7561	**0.7855**

Table 2.9: Prediction performance in binary classification (Cluster-M, Signal-M). Top performed method in each metric is bold.

	Scenario	NN	RF	CNN	Lasso	ENet	MDeep
Sensitivity	S1	0.6770	0.6858	0.7334	0.7420	0.7470	**0.8358**
	S2	0.6788	0.6892	0.7270	0.7600	0.7648	**0.7834**
Specificity	S1	0.7134	0.6498	0.7266	0.7608	0.7604	**0.8474**
	S2	0.7072	0.6818	0.7398	0.7590	0.7614	**0.7962**
Accuracy	S1	0.6952	0.6678	0.7300	0.7514	0.7537	**0.8416**
	S2	0.693	0.6855	0.7334	0.7595	0.7631	**0.7898**
Precision	S1	0.7071	0.6634	0.7300	0.7569	0.7586	**0.8472**
	S2	0.7036	0.6846	0.7384	0.7603	0.7637	**0.7966**
MCC	S1	0.3959	0.3373	0.4617	0.5042	0.5093	**0.6848**
	S2	0.3902	0.3722	0.4688	0.5207	0.5277	**0.5821**
F1 scores	S1	0.6864	0.6731	0.7303	0.7483	0.7513	**0.8405**
	S2	0.6871	0.6858	0.7310	0.7588	0.7631	**0.7882**

Table 2.10: Prediction performance in binary classification (Cluster-M, Signal-H). Top performed method in each metric is bold.

	Scenario	NN	RF	CNN	Lasso	ENet	MDeep
Sensitivity	S1	0.6836	0.6498	0.7252	0.7140	0.7042	**0.8082**
	S2	0.6938	0.6664	0.7332	0.7088	0.7092	**0.7594**
Specificity	S1	0.6792	0.648	0.7048	0.7024	0.7036	**0.8064**
	S2	0.6796	0.6476	0.7232	0.7178	0.726	**0.7546**
Accuracy	S1	0.6814	0.6489	0.715	0.7082	0.7039	**0.8073**
	S2	0.6867	0.657	0.7282	0.7133	0.7176	**0.7570**
Precision	S1	0.6850	0.6506	0.7130	0.7082	0.7059	**0.8106**
	S2	0.6877	0.6557	0.7276	0.7174	0.7235	**0.7574**
MCC	S1	0.3688	0.2993	0.4327	0.4181	0.4093	**0.6178**
	S2	0.3773	0.3156	0.4577	0.4282	0.4365	**0.5156**
F1 scores	S1	0.6793	0.6486	0.7171	0.7097	0.7037	**0.8075**
	S2	0.6869	0.6594	0.7294	0.7116	0.7152	**0.7572**

Table 2.11: Prediction performance in binary classification (Cluster-L, Signal-M). Top performed method in each metric is bold.

	Scenario	NN	RF	CNN	Lasso	ENet	MDeep
Sensitivity	S1	0.6808	0.6636	0.7098	0.7134	0.7136	**0.8264**
	S2	0.6940	0.6746	0.7106	0.7346	0.7430	**0.7764**
Specificity	S1	0.6836	0.643	0.7094	0.7322	0.7336	**0.8320**
	S2	0.6770	0.6544	0.7176	0.746	0.7456	**0.7768**
Accuracy	S1	0.6822	0.6533	0.7096	0.7228	0.7236	**0.8292**
	S2	0.6855	0.6645	0.7141	0.7403	0.7443	**0.7766**
Precision	S1	0.6876	0.6515	0.7132	0.7293	0.7301	**0.8328**
	S2	0.6843	0.6627	0.7180	0.7447	0.7462	**0.7797**
MCC	S1	0.3693	0.3081	0.4221	0.4479	0.4490	**0.6605**
	S2	0.3743	0.3304	0.4298	0.4821	0.4898	**0.5554**
F1 scores	S1	0.6793	0.6561	0.7091	0.7192	0.7202	**0.8282**
	S2	0.6864	0.6674	0.7130	0.7383	0.7436	**0.7764**

Table 2.12: Prediction performance in binary classification (Cluster-L, Signal-H). Top performed method in each metric is bold.

	Scenario	NN	RF	CNN	Lasso	ENet	MDeep
Sensitivity	S1	0.6986	0.6684	0.7094	0.6952	0.7036	**0.8196**
	S2	0.7058	0.6582	0.7186	0.7008	0.7032	**0.7702**
Specificity	S1	0.6994	0.6310	0.7124	0.6958	0.6984	**0.8400**
	S2	0.6804	0.6214	0.7208	0.7026	0.7098	**0.764**
Accuracy	S1	0.699	0.6497	0.7109	0.6955	0.7010	**0.8298**
	S2	0.6931	0.6398	0.7197	0.7017	0.7065	**0.7671**
Precision	S1	0.7052	0.6462	0.7134	0.698	0.7017	**0.8386**
	S2	0.6935	0.6356	0.7222	0.7026	0.7089	**0.7680**
MCC	S1	0.4036	0.3010	0.4240	0.3930	0.4036	**0.6616**
	S2	0.3921	0.2806	0.4416	0.4049	0.4142	**0.5362**
F1 scores	S1	0.6966	0.6557	0.7095	0.6948	0.7013	**0.8278**
	S2	0.6945	0.6458	0.7185	0.7004	0.7050	**0.7676**

Table 2.13: Running time comparison between Ph-CNN and MDeep (Unit is second)

Num of Epochs	1	10	50	100
Ph-CNN	163	1012	4712	8879
MDeep	**21**	**24**	**39**	**57**

Chapter 3

WEVar: a weighted ensemble learning framework for annotating and prioritizing non-coding genetic variants

3.1 Introduction

Determining the functional consequences of noncoding variants is still a challenge in genetics research. Different from coding variants that have effects on protein coding, the functional consequences of noncoding variants are still elusive. In the past decade, genome-wide association studies (GWASs) have uncovered thousands of genetic variants that influence risk for complex human traits and diseases, among which more than 90% are noncoding. Moreover, quantitative trait locus (QTL) analyses have identified "xQTLs" [60] affecting molecular phenotypes such as eQTLs [61] for gene expression; mQTLs [62] for DNA methylation; aseQTLs [63] for allele-specific expression; dsQTLs [64] for DNase I sensitivity and sQTLs [65] for alternative RNA splicing. However, identification of these causal variants is still difficult because of the limitation of sample size and linkage disequilibrium, which may mask true causal ones. In addition, functional interpretations aimed at delineating these causal genetic variants and biological mechanisms underlying the observed statistical associations have lagged. Therefore, one key task for post-GWAS study is on not only refining the identification of causal variants but also improve the derivation of biological meaning.

Although majority of noncoding variants remains underexplored, the functional consequences of noncoding variants are believed to disrupt the normal regulatory mechanisms in promoter and enhancer regions and therefore impact the downstream gene expression in tissue/cell type specific manner, which may result in the onset of various diseases. For example, the prevalence of TERT promoter mutations has been established in melanoma, gliomas and bladder cancer [66]. Moreover, novel MYB-binding motifs created by somatic mutations in the intergenic region resulted in a super-enhancer upstream of the TAL1 oncogene in a subset of T cell acute lymphoblastic leukaemia [67]. Regarding Alzheimer's disease (AD), though coding variant APOE-$\epsilon4$ is unequivocally the most significant genetic risk factor for AD [68, 69], it does not fully explain the AD risk conferred by APOE and the surrounding regions. Additionally, a combination of risk alleles from multiple variants with small effect sizes results in aggregate effects, thus contributing to a higher AD risk [70, 71]. Since the change of regulatory activities can be measured by the epigenomic profiles, epigenomic profiles are widely used as a hallmark to evaluate the functional consequence of noncoding variants. As an example, a recent data analysis shows that active chromatin marks (H3K27ac and H3K4me1), repressive chromatin marks (H3K9me3 and H3K27me3) have different signals in the neighborhoods of a risk variant, rs3024505 associated with type 1 diabetes, and a benign variant rs114490664 [72]. These examples demonstrates epigenomic information could play an important role in predicting and interpreting noncoding variants, which function in regulatory activities.

The rapid development of massively parallel sequencing technologies enable thousands of “multi-omics" data available at large-scale public consortia such as the Encyclopedia of DNA Elements (ENCODE) [73], Roadmap Epigenomics [74] and the International Human Epigenome Consortium [75]. These consortia collect whole-genome wide sequencing features measuring different biological activities such as histone modifications (e.g. ChIP-seq), methylation (e.g. MeDIP-seq, methylation array, WGBS), chromatin accessibility (DNase-seq) and chromatin interactions (Hi-C) across hundreds of different tissues and cell types. Moreover, it has been shown that many GWAS SNPs associated with many diseases and QTL SNPs of multiple tissues are located within these noncoding regions with coverages from ENCODE and Roadmap Epigenomics datasets [72, 76, 77]. Therefore, the enriched collection of these “multi-omics" data make possible the functional annotation of genetic variants and lead to the development of computational methods for predicting functional variants. Taking advantages of these “multi-omics" data, dozes of computational methods have been developed to annotate and prioritize functional noncoding variants [5, 6, 9, 7, 11, 10, 8]. These methods are developed using different machine learning or statistical models, training variants and variant annotations derived from these “multi-omics" data. Supervised learning approach, such as GWAVA [5], CADD [6], DANN [9], FATHMM-MKL [7]), LINSIGHT [11]), FunSeq2 [10], DIVAN [72] and TIVAN [77], is essentially a binary classification task: putative or experimentally validated variants are labelled as positive and benign variants are labelled as negative. Different from the supervised approaches, a common practice of unsupervised methods such as Eigen[8] is to integrate variant annotations into one functional score, which measures the functional importance. For most of the existing methods, genome-wide precomputed functional scores for all known noncoding variants such as 1000 Genomes Project [78] are provided. Users can obtain these scores by providing a list of variants identifiers or arbitrary genomic regions. Usually, a large score indicates the variant could be more functional.

Nevertheless, it has been shown that prediction performance of existing methods show poor concordance on the state-of-the-art benchmark datasets [79], The potential reasons lie on two aspects. First, they are trained using different training variants and variant annotations to predict functional noncoding variants in different context (e.g. disease, phenotype), making one method trained using variants in one context have suboptimal prediction for variants in another context. Second, these methods utilize specific algorithms tailored to specific scenarios, limiting the generalizability. For example, GWAVA is developed using pathogenetic variants collected from The Human Gene Mutation Database (HGMD) [80] and thus is used to predict pathogenetic variants; FunSeq2 is specifically designed for predicting noncoding regulatory variants in cancer. Therefore, an ensemble approach that combines the predictions of all these methods in a weighted scheme could offer a more powerful approach than each method. The weights of each individual methods, which reflects the contributions of methods in the prediction task, can be adaptively learnt in different context, which improves the generalizability and flexibility.

Here, we developed a supervised learning method WEVar (Weighted Ensemble framework for predicting functional noncoding Variants), which integrates representative scoring methods in a constrained optimization framework. Specifically, the precomputed scores of these methods are treated as features with two constraints: the summation of weights of existing methods are required to be one; a L_2-norm is further imposed on the weights for smoothing the estimates. Apparently, there are several advantages of WEVar. First, compared to individual scoring method using hundreds or thousands of features, WEVar directly utilizes the precomputed scores and thus reduces number of features dramatically. Second, WEVar leverages existing methods by adaptively learning the contributions of each method, thus optimizing the prediction performance. Last but most importantly, WEVar has two modes: "context-free" and "context-dependent". Context-free WEVar is used for predicting variants with unknown context. Context-dependent WEVar can further improve the prediction when variants in training and testing set are in the same context. The "context-dependent" mode allows to accurately predict cell type/tissue-specific functional consequences of noncoding variants, which is of great interest. Using simulation and real data studies, we demonstrate both context-free WEVar and context-dependent WEVar outperform the individual scoring methods on the state-of-the-art benchmark datasets. Importantly, context-dependent WEVar further improves the prediction when the number of training variants is large enough. Furthermore, we find WEVar can prioritize experimentally validated noncoding variants in a LD block.

3.2 Materials

WEVar is developed directly on top of precomputed functional score, which is an optimally integrative metric representing for thousands of functional annotations, from multiple individual scoring methods. Using these integrative functional scores directly will decrease the number of features in the model development and thus avoid the challenge

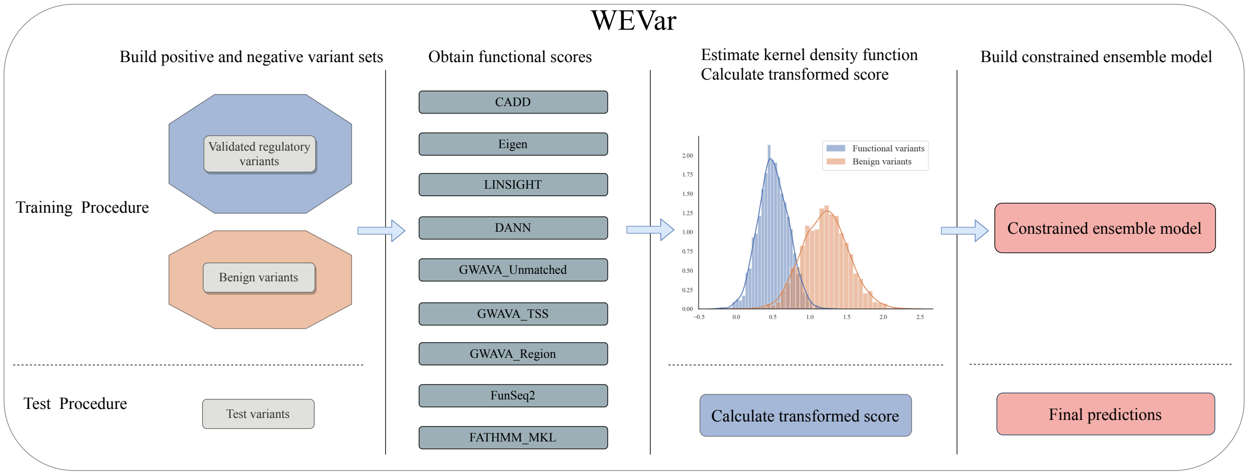

Figure 3.1: Overview of the WEVar. WEVar aims to predict functional noncoding variants, which has two modes: “context-free” and “context-dependent”. For “context-free" mode, the training variant set is chosen from a curated set of functional regulatory variants from diverse context to train a model for functional prediction of variants from unknown or heterogeneous context. For “Context-dependent" mode, the training variant set is selected from one specific context of interest (i.e. disease, tissue, cell type), to train a model for functional prediction of variants from the same context. In the training phase, WEVar compiles the training set with labelled functional and non-functional variants and annotate all variants with precomputed functional scores from representative scoring methods. For each method, the raw scores are transformed using kernel density function (KDE) for both functional and non-functional variant sets respectively. Using these transformed scores as predictive features, a constrained ensemble model is trained. In the testing phase, precomputed functional scores of testing variants are transformed based on the estimated KDE in the training phase and then serve as input features for trained ensemble model to predict the ensemble WEVar score.

high-dimensional data and multicollinearity. We will outline the details of WEVar in the following sessions.

3.2.1 Obtaining precomputed functional scores

We download base-level genome-wide precomputed functional scores from all possible substitutions of single nucleotide variants (SNVs) in the human reference genome (GRCh37/hg19) from scoring methods including CADD [6], DANN [9], FunSeq2 [10], FATHMM-MKL [7], Eigen [8] and LINSIGHT [11]. In addition, we use three sets of precomputed scores from GWAVA (i.e. GWAVA_region, GWAVA_TSS, GWAVA_unmatched) for all SNVs in 1000

Genomes Project [78]. We choose these scoring methods to integrate into WEVar because they are widely used and mostly representative. Since the precomputed score of LINSIGHT is on region level, we assign all variants in the region with the same region-level score. More details for the source of these precomputed scores can be found in Table 3.1.

3.2.2 Assembling variants in training and testing set

For context-free WEVar, the training variant set compiles a curated set of 5,247 causal regulatory variants including i) deleterious or pathogenic noncoding variants from the Human Gene Mutation Database (HGMD) [80] and ClinVar [81] ii) validated regulatory noncoding variants from the OregAnno [82] and iii) candidate causal SNPs for 39 immune and non-immune diseases in the fine-mapping study [83] obtained from Li et al. [84]. The compiled variants are associated with different traits, have functional consequence in different tissues and cell types, and reside in different noncoding regions such as promoters, enhancers, 5'UTRs and 3'UTRs, making them ideal as a training set to predict functional consequence of noncoding variants from unknown or heterogeneous context. Accordingly, we collect six state-of-the-art benchmark independent variant sets from a wide range of context. Among them, three variant sets are collected from Li et al. [84], which include experimentally validated regulatory variants, expression quantitative trait loci (eQTL) [85] (FDR<0.1%) and allelic imbalanced SNPs [86] (FDR<0.1%) Moreover, GWAS significant noncoding SNPs are collected from NHGRI-EBI GWAS Catalog [87] (pvalue$<10^{-5}$). Furthermore, two collected regulatory variant sets are validated by massively parallel reporter assays (MPRAs) in GM12878 lymphoblastoid cells [88] and K562 leukemia cells [89]. For context-free WEVar, these variant sets are used for independent testing. For context-dependent WEVar, we divide each variant set into ten folds with nine-folds as training set and one-fold as testing set.

3.3 Methods

3.3.1 Statistical learning framework of WEVar

The workflow of WEVar is illustrated in Figure 3.1, which consists of four steps: (i) Creating the compiled training variant set (ii) Obtaining the precomputed functional scores for training variants (iii) Transforming the functional scores (iv) Training a constraint ensemble model.

Creating compiled training variant set

Depending on the purpose, we compile the training set for either ‘context-free‘ or “context-dependent" WEVar, as described in the section “Assembling variants in training and testing set".

Obtaining precomputed functional scores for training variants

Precomputed functional scores are retrieved from representative scoring methods including CADD, DANN, Eigen, FunSeq2, FATHMM-MKL, LINSIGHT and GWAVA for variants in the training set, as described in the section “Obtaining precomputed functional scores".

Transforming precomputed functional scores

Precomputed functional scores of integrated scoring methods are on different scales, which may result in different effect sizes of weight estimates by WEVar. However, the resulted different weight estimates are not due to different contributions of integrated scoring methods but because of the systematic bias induced by score scale. Therefore, it is important to perform a normalization step to make functional scores from different scales comparable. To integrate different scores are on the same scale, for each jth scoring method, we estimate two probability density functions (PDF) using kernel density estimation (KDE) based on the empirical distribution of the normalized scores for positive variant set and negative set respectively. As a result, PDF of the positive set denoted as $\mathbf{p}_j(s|+)$ approximates the probability that a variant will have a prediction score s given the variant is functional (+), while PDF of the negative set denoted as $\mathbf{p}_j(s|-)$ approximates the probability that a variant will have the same prediction score s given the variant is nonfunctional (-). Therefore, we use the ratio of two PDFs for the given ith variant, which is essentially the Bayes factor, to represent the likelihood the variant is functional versus nonfunctional. To stabilize the scale of the likelihood, we further take a logarithm of the ratio as the transformed score x_{ij}^N as:

$$x_{ij}^N = \mathbf{log}\frac{\mathbf{p}_j(x_{ij}|+)}{\mathbf{p}_j(x_{ij}|-)} \tag{3.1}$$

where x_{ij} is the raw functional score of the ith variant in the jth scoring method; $\mathbf{p}_j(x_{ij}|+)$ and $\mathbf{p}_j(x_{ij}|-)$ are probability density of x_{ij} in positive and negative set respectively.

Training a constraint ensemble model

Using the transformed scores, we will fit a constraint ensemble model, which is essentially a <u>C</u>onstrained <u>P</u>enalized <u>L</u>ogistic <u>R</u>egression model. Let $\mathbf{x}^N \in \mathbb{R}^p$ be the transformed scores of a variant for all scoring methods and $y \in \{-1, +1\}$ be the variant label. The conditional probability of the variant being functional given $\mathbf{x}^N$ can be formulated as:

$$\mathbf{p}(y = 1|\mathbf{x}^N) = \frac{1}{1 + \exp(-y(\mathbf{w}^\top \mathbf{x}^N) + b)} \tag{3.2}$$

where $\mathbf{w} \in \mathbb{R}^p$ is a weight vector, which contains the regression coefficients, and $b \in R$ is the intercept. The likelihood function for n variants from both positive and negative set is defined as $\prod_{i=1}^n \mathbf{p}(y_i|\mathbf{x}_i^N)$. The objective function, which is the average of negative log-likelihood, is defined as:

$$f(\mathbf{w}, \mathbf{b}) = -\frac{1}{n}\log\prod_{i=1}^{n}\mathbf{p}(y_i|\mathbf{x}_i^N) \tag{3.3}$$

By minimizing the objective function, we can estimate $\mathbf{w}$ and b as:

$$\underset{\mathbf{w},b}{\text{minimize}} \quad f(\mathbf{w}, b) \tag{3.4}$$

We further apply two constraints to the log-likelihood function. First, the weight of each scoring method is larger or equal to 0, indicating all scoring methods will contribute neutrally or positively to the prediction. Second, the sum of all weights equals to 1, which is a reasonable assumption for the summation of contributions from all scoring methods. In addition, to leverage all scoring methods by avoiding a sparse solution, we add an L_2-norm to the objective function. Finally, we have the L_2-norm regularized objective function with the two constraints as:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & f(\mathbf{w}, b) + \lambda||\mathbf{w}||_2 \\ \text{subject to} \quad & \sum_{j=1}^{p} \mathbf{w}_j = 1 \\ & \mathbf{w}_j \geq 0, \ j = 1, \dots, p. \end{aligned} \tag{3.5}$$

where $\lambda \geq 0$ is the tuning parameter for L_2-norm, which can be optimized from cross-validation in the training phase.

To minimize the loss function with equality and inequality constraints, we first rewrite the loss function as the standard form:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & f(\mathbf{w}, b) + \lambda||\mathbf{w}||_2 \\ \text{subject to} \quad & h_k(\mathbf{w}) \leq 0, k = 1, \dots, p. \\ & l(\mathbf{w}) = 0 \end{aligned} \tag{3.6}$$

We then introduce Generalized Lagrange function to relax two constraints, which is formulated as:

$$\mathcal{L}(\mathbf{w}, b, \alpha, \beta) = f(\mathbf{w}, b) + \lambda||\mathbf{w}||_2 + \sum_{k=1}^{p} \alpha_k h_k(\mathbf{w}) + \beta l(\mathbf{w}) \tag{3.7}$$

In this way, the dual problem is easier to solve compared with the primal problem. The primal solution can be constructed from the dual solution as:

$$g(\alpha, \beta) = \min \mathcal{L}(\mathbf{w}, b, \alpha, \beta) \tag{3.8}$$

The Lagrange dual function can be considered as a pointwise maximization of some affine functions so it is always concave. The dual problem is always convex even if the primal problem is not convex, which can be easily solved by gradient-based methods.

Testing phase

In the testing phase, given variants be annotated precomputed functional scores from all scoring methods, which will be further transformed through the estimated KDE in the training phase. The transformed scores will serve as input features for trained ensemble model to predict the ensemble WEVar score.

Implementation

We adopt the SciPy [90], a Python scientific computing library, to perform the kernel density estimations, and CVXPY [91], a Python-embedded modeling library for convex optimization, to estimate constrained weights from the objective function.

Software availability

WEVar is implemented in a standalone software toolkit available at (`https://github.com/lichen-lab/WEVar`), which mainly consists of i) a compiled data package including precomputed scores for all SNVs (GRCh37/hg19) in 1000 Genomes Project across all integrated scoring methods; ii) a model package of pre-trained context-free and context-dependent WEVar models; and iii) a Python software package to perform the functional prediction using pre-trained models or re-train a new model. To use a pre-trained model, WEVar will take compiled data package and genomic coordinates of testing variants as input. Alternatively, WEVar will take compiled data package and genomic coordinates of training variants to re-train a new WEVar model.

3.4 Results

First, we will perform a simulation study to evaluate the accuracy of weight estimation by WEVar for all integrated scoring methods and investigate whether the prediction performance of WEVar is improved compared to individual scoring method. Second, we will evaluate the context-free functional prediction and context-dependent functional prediction on the state-of-the-art benchmark datasets respectively. Third, we will apply WEVar to prioritize experimentally validated causal regulatory variants in multiple risk loci associated with multiple traits.

3.4.1 Evaluation of WEVar in a simulation study

Evaluation metrics

The performance of all scoring methods is evaluated using area under the receiver operating characteristics curve (AUROC), the area under the precision-recall curve (AUPR) and Pearson correlation between predicted and true labels (COR). AUROC and AUPR are metrics based on the ranks of the predicted scores. COR has the additional ability to measure how the predicted values are correlated with the true labels. Using different probability cutoffs, AUROC measures the trade-off between the true positive rate and false positive rate. AUPR compares the trade-off between the true positive rate and precision. AUROC is preferred for balanced class, whereas AUPR is more appropriate for imbalanced class. Since we have both balanced and unbalanced testing datasets, we present both metrics.

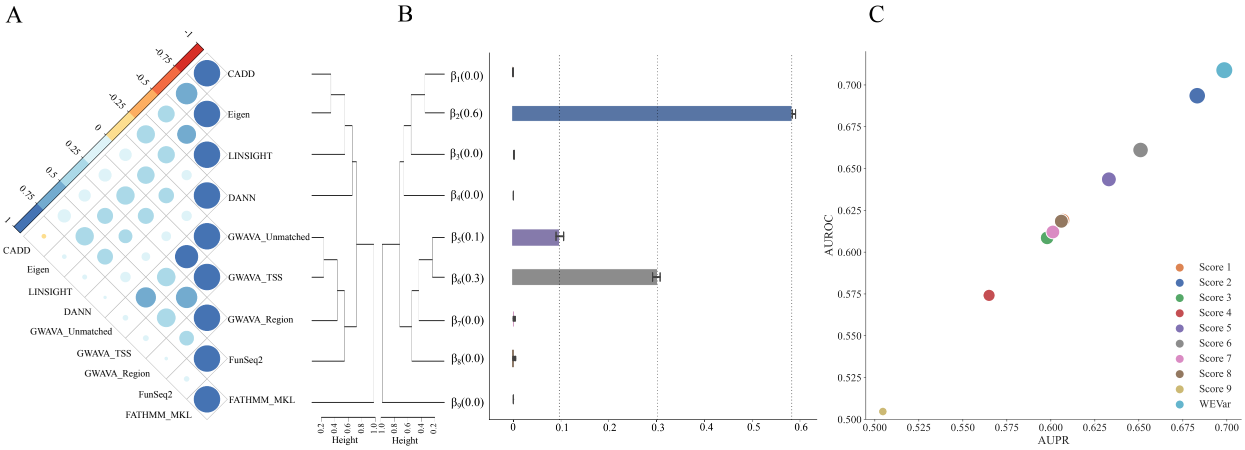

Figure 3.2: (A) Pairwise Pearson correlations between precomputed functional scores among scoring methods for the integrated causal regulatory variants collected from Li et al. [84]. (B) Average regression coefficient estimated by WEVar in the training phase in 50 simulations. (C) Average prediction performance by WEVar on the independent testing datasets. X axis presents AUPR; Y axis presents AUROC; the bubble size represents COR. AUPR, AUROC and COR are averaged in the testing phase in 50 simulations.

Simulating correlated functional scores and variant labels

We conduct a simulation study to evaluate whether WEVar can estimate contribution of each individual scoring method accurately and whether WEVar can improve prediction performance compared to each individual scoring method. Since the functional scores of different methods have an overall positive correlation (Figure **??**A), we simulate functional scores of all scoring methods with consideration of the score correlation. Using the simulated scores, we generate a total $10,000$ variants with an equal size of functional and nonfunctional variants in the training set. Similarly, we independently generate an equal number of $10,000$ variants in the testing set for prediction evaluation. We then apply WEVar to retrain a model in the training set and predict WEVar scores in the testing set. Using WEVar scores and true labels in the testing set, we will calculate AUROC, AUPR and COR. We repeat the whole procedure 50 times and obtain the average of all evaluation metrics.

Specifically, using the integrated causal regulatory variant set collected from Li et al. [84], we calculate a $p \times p$ variance-covariance matrix R of precomputed functional scores among all integrated scoring methods, where p is the number of scoring methods. We cluster these methods based on Pearson correlation and find that these methods have different levels of disagreement, indicating that performance of these methods show poor concordance on the benchmark dataset (Figure **??**A). Not surprisingly, GWAVA_Unmatched, GWAVA_Region and GWAVA_TSS are clustered together since they use the same positive training variant set. Surprisingly, FATHMM-MKL has the lowest correlation with all the other methods. Indeed, this observation highlights the rationale why a weighted ensemble strategy proposed by WEVar is essential to improve the prediction because it is able to upweight the scoring methods fit in current context while down-weight the unfit others. We further perform Cholesky decomposition on R as:

$$R = C \cdot C^{\top} \tag{3.9}$$

where C is a $p \times p$ lower triangular matrix with real positive diagonal entries. To maintain the correlations of simulated scores, we generate the correlated functional scores X as the product between $C^{\top}$ and random variable d, which is sampled from an independent normal distribution as:

$$X = d \cdot C^{\top}, \quad d \sim N(0,1). \tag{3.10}$$

where x_{ij} as the functional score of ith noncoding variant in jth scoring method. η_i, which is the weighted average score of ith variant, can be generated as:

$$\eta_i = \sum_{j=1}^{p} x_{ij} \cdot \beta_j \tag{3.11}$$

where β_j is the weight associated with jth method. Without loss of generality, we manually assign 0.6 to β_2, 0.3 to β_6, 0.1 to β_5, and 0 to the rest. We then perform inverse logit transformation to η_i to obtain probability π_i, based on which the binary label y_i for ith

variant is generated from a Bernoulli distribution as:

$$y_i \sim \text{Bern}(\pi_i), \quad \text{where } \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \tag{3.12}$$

Results of the simulation study

In the simulation study, we will evaluate whether WEVar can truly discover the contributions of individual scoring method by comparing the estimated regression coefficients ($\hat{\beta}$) with the assigned true values (β). To fit a WEVar model, the optimal tuning parameter for L_2-norm is selected using fivefold cross-validation (5-CV), where the whole training set is divided into five-folds, where four-folds is used to train the model and one-fold is used to obtain the evaluation metric i.e. AUROC. The optimal tuning parameter is chosen based on the average AUROC from 5-CV, and a final model is fitted using the whole training set with the optimal tuning parameter. To evaluate the performance of the final model an independent testing set, we use all evaluation metrics AUROC, AUPR and COR.

As a result, we find that the estimated weights are nearly unbiased to the underlying truths (Figure **??**B), which suggests that WEVar can discover the contribution of each individual scoring method correctly when the functional scores of these methods are correlated. With accurate contribution estimation, WEVar can significantly improve the prediction performance in the independent testing (Figure **??**C) by achieving the highest AUROC, AUPR and COR. Overall, the simulation results validate the benefit of exploiting different scoring methods in an integrative weighted scheme.

3.4.2 Context-free functional prediction

Overview of context-free WEVar

We first introduce context-free WEVar, which is trained using integrated causal regulatory variants collected from Li et al. [84]. We call this WEVar mode “context-free" because

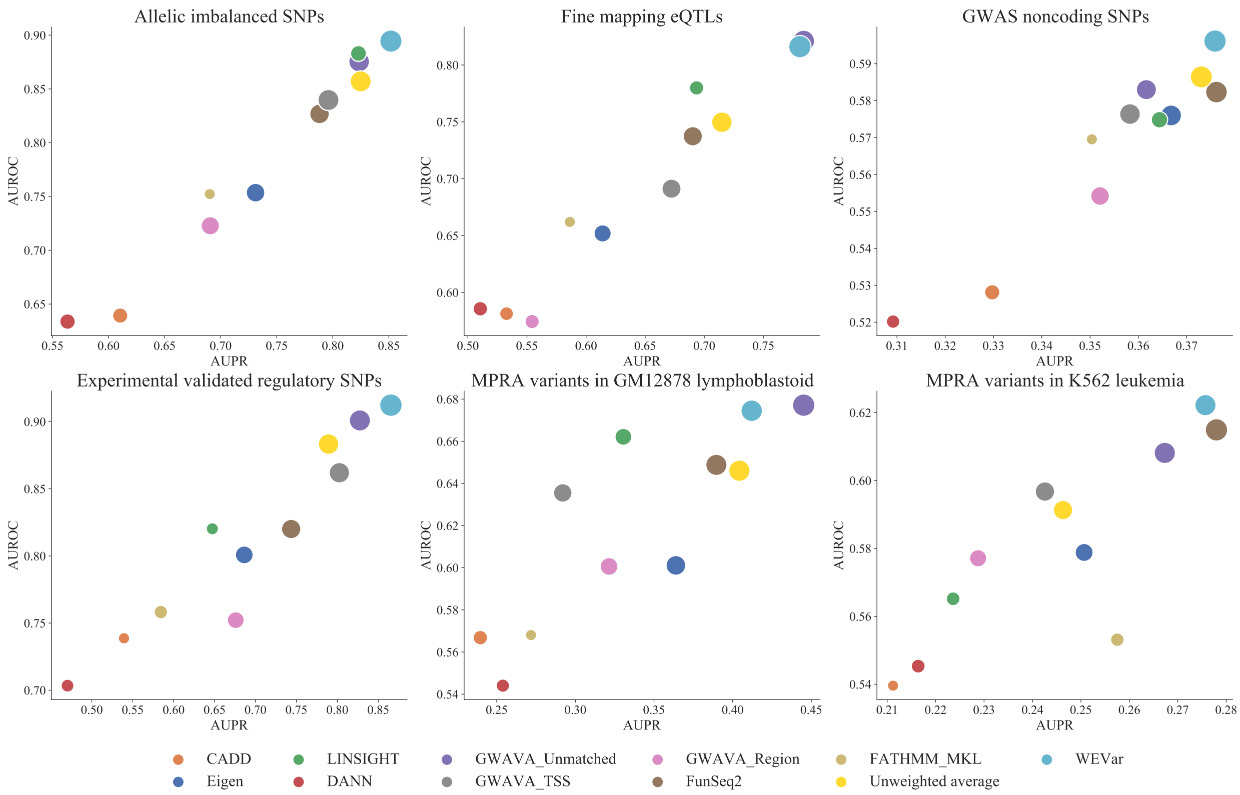

Figure 3.3: Evaluation of context-free WEVar and integrated scoring methods. Context-free WEVar is trained using the integrated functional regulatory variants collected by Li et al. [84], which include variants in HGMD, ClinVar, OregAnno and fine-mapping candidate causal SNPs for 39 immune and non-immune diseases with a total of 5,247 positive variants and 55,923 negative variants. Context-free WEVar is tested on the state-of-the-art benchmark datasets, which include i) Allelic imbalanced SNPs in chromatin accessibility with a total of 8,592 positive variants and 9,678 negative variants (Allelic imbalanced SNPs); ii) Uniformly processed fine-mapping eQTLs from 11 studies with a total of 31,118 positive variants and 36,540 negative variants (Fine mapping eQTLs); iii) GWAS noncoding SNPs with a total of 19,797 positive variants and twice number of negative variants (GWAS SNPs) [**IWscore**]; iv) Manually curated experimentally validated regulatory SNPs with a total of 76 positive variants and 156 negative variants (Experimentally validated regulatory SNPs); v) MPRA validated variants in lymphoblastoid cells with a total of 693 positive variants and 2,772 negative variants (MPRA variants in GM12878 lymphoblastoid); vi) MPRA validated variants in erythrocytic leukemia cells with a total of 342 positive variants and 1,368 negative variants (MPRA variants in K562 leukemia). We further remove variants on sex chromosome or with missing precomputed scores. X axis presents AUPR; Y axis presents AUROC; the bubble size represents COR.

these variants are not limited to a specific context but have a broad definition of functionality across a wide range of context. These variants are either experimentally validated or

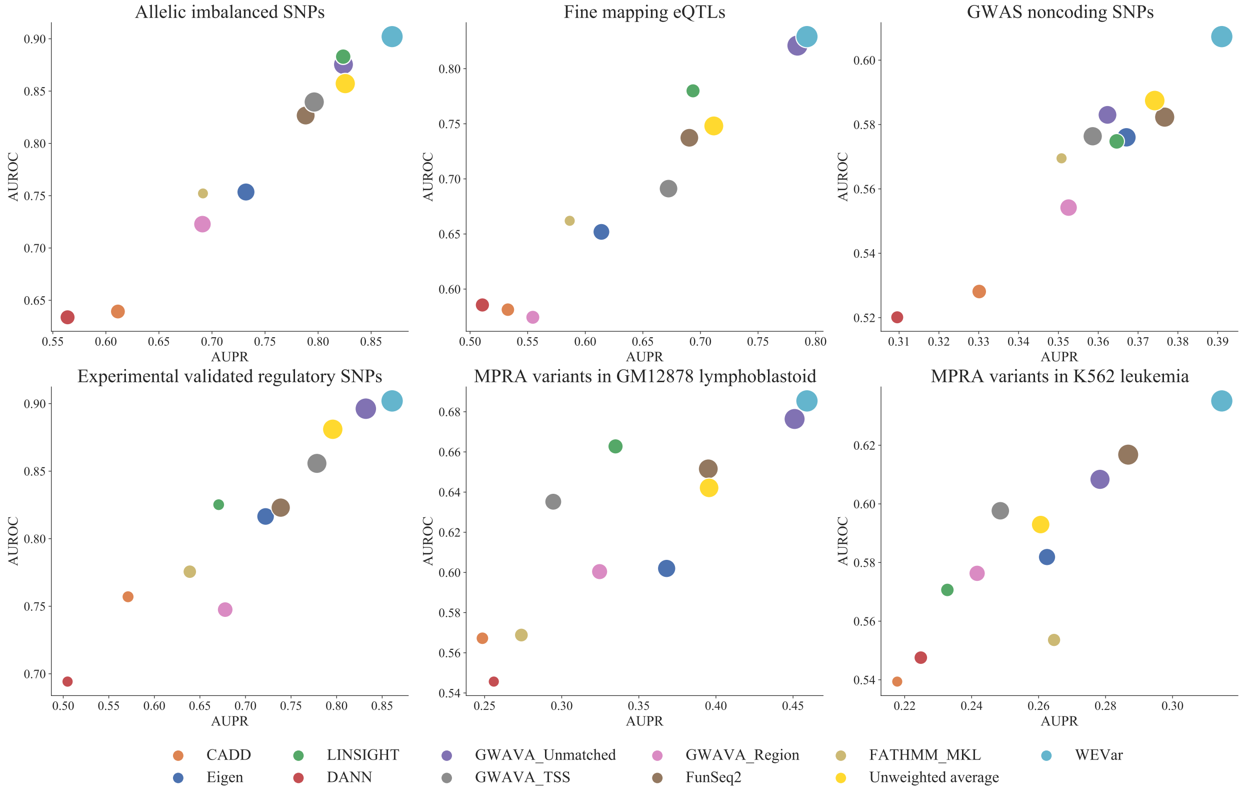

Figure 3.4: Evaluation of context-dependent WEVar and integrated scoring methods on state-of-the-art benchmark datasets, which include Allelic imbalanced SNPs, Fine mapping eQTLs, GWAS noncoding SNPs, Experimentally validated SNPs, MPRA validated variants in GM12878 lymphoblastoid cells and MPRA validated variants in K562 leukemia cells. We further remove variants on sex chromosome or with missing precomputed scores. To restrict the training and testing variants are from the same context, for each dataset, we randomly split the dataset into ten-folds with nine-folds as the training set and one-fold as the testing set. Context-dependent WEVar is trained on the nine-folds and independently evaluated on the left one-fold. AUC, AUCPR and COR are calculated and averaged in the ten replicates for each method. X axis presents AUPR; Y axis presents AUROC; the bubble size represents COR.

highly putative causal variants associated with different diseases, molecular phenotypes or clinical outcomes, which are located in different noncoding regions such as promoters, enhancers, 5'UTRs and 3'UTRs. The diverse context and widespread genomic locations of these variants make it potentially powerful to predict functional noncoding variants when the context is unknown or heterogeneous. To demonstrate the generality of context-free

WEVar, we evaluate it on the independent benchmark datasets containing noncoding variants of different functionalities and from diverse context. We also remove any duplicated variants overlapped with training dataset from each independent testing dataset, which can prevent potential overfitting. To verify the effectiveness of the weight strategy, besides all scoring methods WEVar integrates, we also include “Unweighted average" as a comparison, which is the unweighted average of min-max normalized precomputed functional scores from the integrated methods. In the training phase of WEVar, tuning parameter for L_2-norm is selected using 5-CV. For all methods, AURPC, AUCPR and COR are reported on each independent testing dataset.

Results of functional prediction between context-free WEVar and integrated scoring methods

We start to compare the prediction performance between WEVar and its integrated scoring methods on three datasets, which consist of putatively functional variants based on statistical association (Figure 3.3). The first dataset, which is produced by Maurano et al. [86] and processed by Li et al. [84], contains 8,592 significant allelic imbalanced SNPs of chromatin accessibility (FDR<0.1) as the positive set and 9,678 frequency-matched background SNPs around nearest transcription start sites of randomly selected genes as the negative set. We observe that WEVar obtains the largest AUROC, AUPR and COR (0.894, 0.852, 0.644) with substantial improvements over each individual scoring method (Table 3.2). Following WEVar, LINSIGHT, GWAVA_Unmatched and Unweighted average have an overall comparable performance. However, the COR of LINSIGHT is much lower (0.255) compared to GWAVA_Unmatched (0.535) and Unweighted average (0.559). Surprisingly, FATHMM_MKL also has the lowest COR (0.053). Moreover, CADD and DANN, which utilize the same training set, have comparable but poorest performance among all methods (CADD: 0.639, 0.610, 0.228; DANN: 0.634, 0.563, 0.236). Interestingly, the prediction performance of GWAVA_Unmatched, GWAVA_TSS and GWAVA_Region are discordant even if they use the same positive training set (GWAVA_Unmatched: 0.875, 0.823, 0.535; GWAVA_TSS: 0.840, 0.796 0.559; GWAVA_Region: 0.723, 0.691, 0.382).

The second dataset consists of eQTLs in 11 studies across 7 tissues identified from Brown et al. [85] and processed by Li et al. [84]. The positive set consists of 31,118 significant eQTL SNPs (FDR<0.1) and the negative set contains 36,540 frequency-matched background SNPs around nearest TSS of randomly selected genes. We observe that WEVar has the largest COR and comparable AUROC and AUPR to GWAVA_Unmatched (WEVar: 0.816, 0.781, 0.509; GWAVA_Unmatched: 0.821, 0.781, 0.476) (Table 3.3). Moreover, both WEVar and GWAVA_Unmatched have clearly advantages over other scoring methods. For example, they improve nearly 0.04 AUROC and 0.09 AUPR over LINSIGHT, and 0.07 AUROC and 0.07 AUPR over Unweighted average. Particularly, there is substantial improvement of nearly 0.1 COR to Unweighted average and over 0.3 to LINSIGHT. Notably, the relative performance of GWAVA_Region drops dramatically and it has the lowest AUROC (0.574). FATHMM_MKL still has the lowest COR (0.047) followed by CADD and DANN (0.126, 0.1500).

The third dataset collects 19,797 GWAS significant noncoding SNPs from NHGRI-EBI GWAS Catalog [87] as positive set and twice number of variants in the negative set, which are randomly sampled from all noncoding variants in 1000 Genomes project with minor allele frequency (MAF) $\geq$ 5% [**IWscore**]. The relative prediction performance of all methods are similar to the first dataset of allelic imbalanced SNPs. WEVar outperforms all scoring methods by obtaining the highest AUROC, AUPR, and COR. FATHMM_MKL have the lowest COR, while CADD and DANN have the lowest AUROC and AUPR (Table 3.4).

In addition to the three datasets comprised of putatively functional noncoding variants derived from association analyses, we compare the prediction performance between WEVar and all scoring methods on three datasets consisting of experimentally validated regulatory variants. The first dataset include 81 experimentally validated regulatory SNPs curated by Li et al. [84]. We find the trends of prediction performance for all methods still holds similarly to allelic imbalanced SNPs and GWAS significant noncoding SNPs, where WEVar obtains the largest AUROC, AUPR and COR (0.912, 0.865, 0.718) followed by GWAVA_Unmatched (0.901, 0.828, 0.649) and Unweighted average(0.883, 0.789, 0.617) (Table 3.5).

The other two datasets contain processed causal regulatory variants validated by MPRAs in two cell lines [92]. The first MPRA dataset includes 665 variants with genomic loci annotation in Ensembl database as positive set, which are selected out of 842 expression-modulating variants that show significantly differential allelic expression in GM12878 lymphoblastoid cells [88]. The negative set contains 2,772 control variants tested by MPRA but neither allele showed significant effects on expression (Bonferroni corrected pvalue>0.1). The second MPRA dataset consists of 339 positive variants that cause significant change of expression via targeted motif disruption in enhancers in K562 erythrocytic leukemia cells (pvalues<0.05) [89]. The negative set contains 1,359 control variants without causing significant change (pvalues>0.1). As a result, WEVar has comparable performance with top-performed GWAVA_Unmatched in predicting MPRA validated regulatory variants in GM12878 lymphoblastoid cells (WEVar: 0.674, 0.412, 0.286 vs GWAVA_Unmatched: 0.677, 0.445, 0.317) (Table 3.6). WEVar achieves largest AUROC and AUPR in predicting MPRAs validated regulatory variants in K562 leukemia cells (Table 3.7).

Clearly, context-free WEVar has the overall best performance on the state-of-the-art independent testing datasets, which demonstrate its robustness and generality to predict functional noncoding variants across a wide range of context. Following WEVar, GWAVA_Unmatched, Unweighted average and FunSeq2 have superior performance to others. In contrast, CADD, DANN and FATHMM_MKL perform poorly. Particularly, FATHMM_MKL suffers from a low COR. Notably, integrating scores in a weighted scheme indeed boosts the prediction performance as demonstrated by the improvement of WEVar over Unweighted average.

3.4.3 Context-dependent functional prediction

Overview of context-dependent WEVar

Different from context-free functional prediction, context-dependent functional prediction happens when a context-dependent WEVar is trained and the training and testing variants are from the same context. We develop “context-dependent" mode for WEVar because functional variants are usually studied in a cell type/tissue-specific way. The context-matching between training and testing variants may improve the prediction power. We demonstrate the prediction performance of context-free WEVar first, followed by a comparison between context-free and context-dependent WEVar to demonstrate the advantage for WEVar by being context-dependent.

Results of functional prediction between context-dependent WEVar and integrated scoring methods

We use the same benchmark datasets to evaluate context-free functional prediction. To restrict the training and testing variants from the same context, we randomly split each dataset into ten-folds with nine-folds as the training set and one-fold as the testing set. Tuning parameter for L_2-norm is selected in the training set using 5-CV with AUROC as the evaluation metric. A final context-dependent WEVar is fitted using the whole training set with the selected tuning parameter and makes the functional prediction on the testing set. AUROC, AUPR and COR are calculated by comparing prediction scores and true labels of variants in the testing set. We use leave-one-fold-out by selecting nine-folds as training set and one-fold as testing set ten times. Accordingly, the whole procedure is repeated ten times and all evaluation metrics are reported as average.

We observe that context-dependent WEVar outperforms all scoring methods by obtaining the highest AUROC, AUPR and COR across all the benchmark datasets (Figure 3.4 and

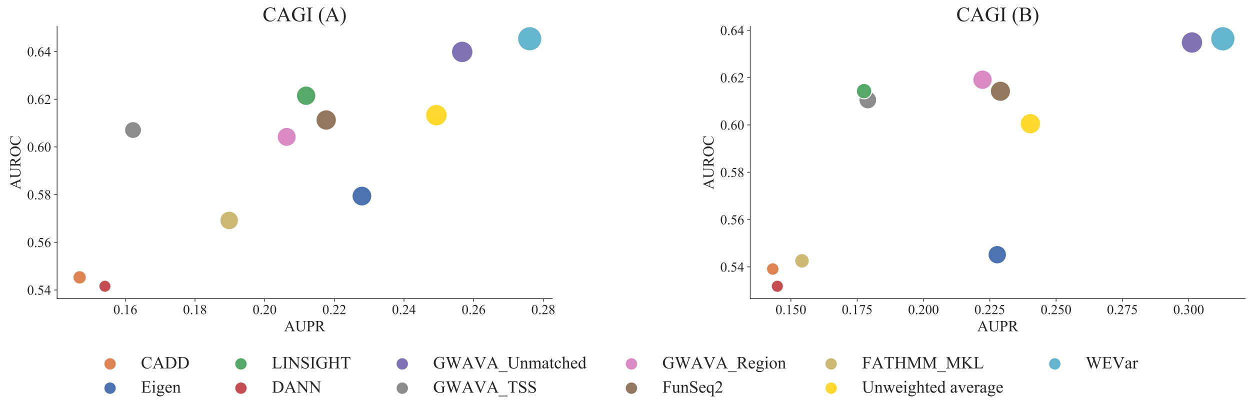

Figure 3.5: Prediction performance comparison between context-dependent WEVar and integrated scoring methods on the CAGI benchmark datasets. In CAGI, 2,873 SNVs with 345 as positive set and 2,528 as negative set. The testing set contains a total of 2,808 SNVs with 348 positive variants and 2,460 negative variants. We further remove SNVs on sex chromosome or with missing precomputed scores in both sets. (A) Context-dependent WEVar is first trained on the training set and evaluated on the testing set. (B) Similarly, we switch the training and testing set and perform an additional independent evaluation. The figure presents the AUPR, AUROC, and COR. X axis presents AUPR; Y axis presents AUROC; bubble size represents COR.

Table 1.2-1.7). Moreover, we observe similar trends between context-dependent and context-free functional prediction, where WEVar, GWAVA_Unmatched and Unweighted average are the top-performed methods, while CADD, DANN and FATHMM_MKL have overall poor performance.

To further objectively gauge the performance of context-dependent WEVar, we utilize the training and testing variant set in the first part of challenge of Critical Assessment of Genome Interpretation eQTL challenge (CAGI) [93] derived from MPRA validated regulatory variants from GM12878 lymphoblastoid cells [88]. The variants selected by CAGI show significant level of transcriptional activity for either of two alleles. Specifically, the level of transcriptional activity is measured by differential abundance of transcripts versus plasmid input. Based on the FDR cutoff 0.01, a binary label is generated to indicate whether or not at least one of the two alleles of the variant exhibits a significantly high transcriptional activity (i.e. labeling 1 if FDR<0.01, otherwise, 0). As a result, the training set consists a total of 2,873 SNVs with 345 as positive set and 2,528 as negative set. The testing set contains a total of 2,808 SNVs with 348 positive variants and 2,460 negative variants. We further remove SNVs on sex chromosome or with missing precomputed scores in both sets. Besides following the original training and testing procedure, we further carry out an additional comparison by switching the training and testing set.

Consistent with our previous findings, context-dependent WEVar has superior performance to other scoring methods in both comparisons by achieving the highest AUROC, AUPR and COR, followed by GWAVA_Unmatched and Unweighted average (Figure 3.5, Table 3.9 3.8). Moreover, CADD and DANN have the overall poorest performance. The additional independent evaluation further strengthens the advantage of context-dependent WEVar in predicting functional noncoding variants by benefiting from matched context in training and testing set.

Besides improving the functional prediction, another important characteristic of WEVar is that it can identify the informative predictors that play the major contribution to the functional prediction among all integrated scoring methods. Consequently, we find that sets of informative predictors are different across benchmark datasets (Figure 3.6, Table 3.10). In most cases, WEVar identifies a parsimonious set of scoring methods that dominate the functional prediction especially FunSeq2 and GWAVA_Unmatched are two ubiquitous major contributors. Moreover, GWAVA_TSS is an additional major contributor for Allele imbalanced SNPs, Experimentally validated regulatory SNPs and integrated causal regulatory variants used by context-free WEVar. Regarding MPRA validated regulatory variants in GM12878 lymphoblastoid cells, Eigen is the additional method that has a major contribution. Similarly, GWAVA_Region and Eigen are two additional major contributors for two comparisons for CAGI training and testing variants. However, for GWAS noncoding SNPs and MPRA validated regulatory variants in K562 leukemia cells, there is a ubiquitous solution, where the contributions of all methods are relative uniform. These findings demonstrate that considering context-specificity in WEVar leads to different weight estimates and

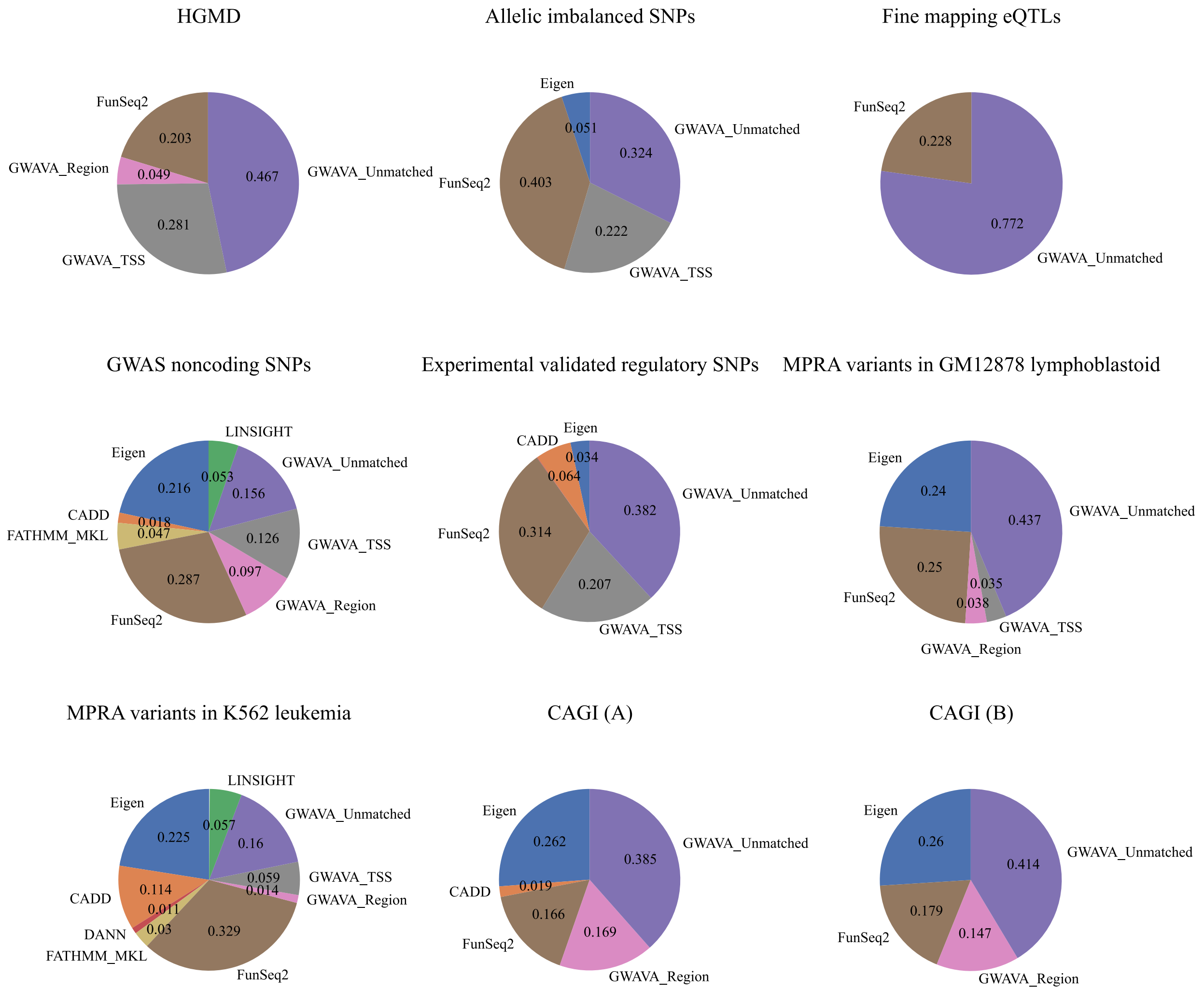

Figure 3.6: Weight estimation for all benchmark datasets. WEVar identifies a parsimonious set of scoring methods that play major contribution to functional prediction for most datasets. The only exceptions are GWAS noncoding SNPs and MPRA variants in K562 leukemia cells, where there is a universal solution.

result in different sets of informative predictors. These observations also suggest that it is important to obtain an optimal weights when integrating different scoring methods, as the non-uniform weights estimated by WEVar lead improved functional prediction across benchmark datasets. Additionally, this point has been validated by both simulation and real data applications that WEVar outperforms the Unweighted average.

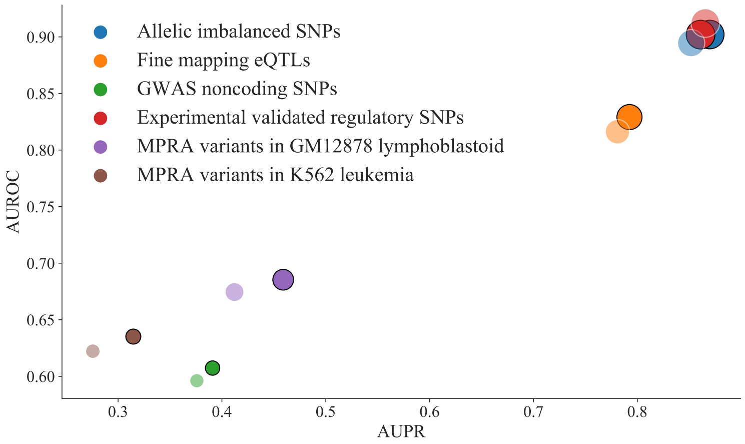

Figure 3.7: Prediction performance comparison between context-free WEVar and context-dependent WEVar across six independent testing datasets. X axis presents AUPR; Y axis presents AUROC; bubble size represents COR; solid bubble represents context-dependent WEVar; transparent bubble represents context-free WEVar.

Results of comparison between context-free and context-dependent functional prediction

We hypothesize that considering context-specificity and context-matching context between training and testing variants in “context-dependent" WEVar will likely improve the predictive power for functional prediction. To validate this hypothesis, we directly compare the results of functional predictions between context-free and context-dependent WEVar on the aforementioned state-of-the-art benchmarking datasets (Figure 3.7, Table 3.2-3.7).

For MPRA validated variants in GM12878 lymphoblastoid cells, context-dependent WEVar significantly outperforms context-free WEVar with large performance gain in around 5% AUPR and 8% COR but modest gain in AUROC. Similarly, context-dependent WEVar also achieves a large improvement by increasing about 4% AUPR and 4% COR but slightly improvement of AUROC for MPRA validated variants in K562 leukemia cells. Moreover, the improvement of context-dependent WEVar is evident demonstrated by nearly 5% and 3% increase in COR but slightly increase in AUROC and AUPR for both Fine mapping eQTLs and Allele imbalanced SNPs. In addition, context-dependent WEVar has a modest improvement of all metrics for GWAS noncoding SNPs. However, there is a lack of improvement on Experimentally validated regulatory SNPs, which could be explained by the small sample size of training set. This observation indicates that a large training set is necessary to improve the predictive power for context-dependent functional prediction. Overall, the comparisons between context-dependent and context-free WEVar validate the hypothesize that considering context-specificity and context-matching will improve the functional prediction. However, this improvement depends on the availability of enough sample size for training a robust context-dependent WEVar.

3.4.4 Prioritization of causal regulatory variants by WEVar on benchmarking datasets

To demonstrate the application of WEVar in studying complex traits, we apply genome-wide functional scores of all noncoding variants in 1000 Genomes Project precomputed by context-free WEVar for fine-mapping analysis in risk loci. The diverse benchmarking datasets are generated from different experiments and study different traits, which are able to test the robustness of WEVar in prioritizing causal regulatory variants in risk loci.

Noncoding variants modulating gene expression

We evaluate WEVar on reported "expression-modulating variants" (emVars), which have been validated to show differential gene expression between alleles, from the MPRA study in GM12878 lymphoblastoid cells [88]. To assess whether these emVars with a strong linkage to GWAS SNPs can be prioritized by WEVar score, we create an extended LD block (r^2 >0.2) utilizing ldproxy [**ldlink**] to extract variants from all reference populations within the LD block, which are further assigned WEVar score.

Consequently, WEVar is able to prioritize emVars in exampled LD blocks (Figure 3.8 and Table 3.11). For example, emVar rs4790718 (chr17:4870893) scores higher than three

LD-linked GWAS SNPs rs1060431 (chr17:4840868, pvalue=2x10^{-26}), rs6065 (chr17:4836381, pvalue=2x10^{-12}) and rs571461910 (chr17:4869143, pvalue=3.98x10^{-9}), which are mapped to SPAG7 and associated with Platelet counts. Similarly, emVar rs922483 (chr8:11351912) is successfully prioritized by the highest score among all LD-linked variants including GWAS SNP rs2736340 (chr8:11343973, pvalue=6.03x10^{-20}) associated with Systemic lupus erythematosus. Moreover, emVar rs56316188 (chr8:59323811) scores higher than GWAS SNP rs2859998 (chr8:59324162, pvalue=1x10^{-7}), which is mapped to UBXN2B and associated with narcolepsy with cataplexy. Additionally, emVar rs306587 (chr10:30722908) is prioritized among LD-linked variants including one GWAS SNP rs1042058 (chr10:30728101, pvalue=6x10^{-11}). Overall, these examples demonstrate that WEVar can successfully prioritize experimentally validated regulatory variants that modulate gene expression among LD-linked putatively causal GWAS SNPs, indicating that WEVar can potentially aid the fine mapping analysis.

Causal regulatory variants associated with Schizophrenia

Schizophrenia, typically diagnosed in the late teens years to early thirties, is a mental disorder characterized by disruptions in thought processes, perceptions, emotional responsiveness, and social interactions. Schizophrenia is one of the top 15 leading causes of disability worldwide [94, 95] and estimated international prevalence of schizophrenia among non-institutionalized persons is 0.33% to 0.75% [96]. Although GWAS has identified numerous noncoding schizophrenia-associated variants hypothesized to affect gene transcription, the causal regulatory variants are still elusive. To experimentally evaluate the regulatory potential of these GWAS SNPs and LD-linked variants, a recent study [97] screens several schizophrenia loci from a large GWAS cohort-Schizophrenia Working Group of the Psychiatric Genomics Consortium, using MPRA experiments in both K562 leukemia cells and SK-SY5Y neuroblastoma cells.

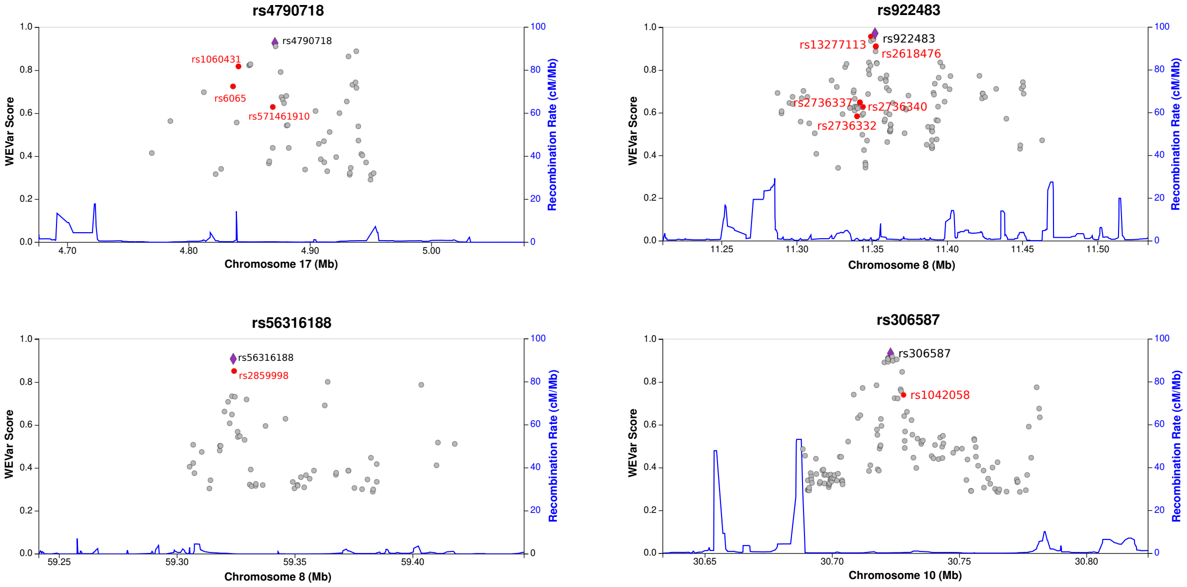

Figure 3.8: WEVar can prioritize noncoding variants modulating gene expression. The noncoding variants are "expression-modulating variants" (emVars) identified from MPRA study in GM12878 lymphoblastoid cells. We create an extended LD block (r^2 >0.2) to include emVars and LD-linked GWAS SNPs. As a result, rs4790718 (chr17:4870893) scores higher than three LD-linked GWAS SNPs rs1060431 (chr17:4840868, pvalue=2x10^{-26}), rs6065 (chr17:4836381, pvalue=2x10^{-12}) and rs571461910 (chr17:4869143, pvalue=3.98x10^{-9}), which are mapped to SPAG7 and associated with Platelet counts. Similarly, rs922483 (chr8:11351912) is successfully prioritized by the highest score among all variants including GWAS SNP rs2736340 (chr8:11343973, pvalue=6.03x10^{-20}) associated with Systemic lupus erythematosus. Moreover, rs56316188 (chr8:59323811) scores higher than GWAS SNP rs2859998 (chr8:59324162, pvalue=1x10^{-7}), which is mapped to UBXN2B and associated with narcolepsy with cataplexy. Additionally, rs306587 (chr10:30722908) is prioritized among LD-linked variants including one GWAS SNP rs1042058 (chr10:30728101, pvalue=6x10^{-11}). emVars are marked purple. LD-linked GWAS SNPs are marked red.

We apply context-free WEVar functional scores to discover causal regulatory variants associated with Schizophrenia. Briefly, we define "causal regulatory variants" as variants with significant differential expression between two alleles with a FDR cutoff 0.2. For each causal regulatory variant, we extend the risk locus by considering all variants in LD (r^2 >0.2). We further obtain precomputed context-free WEVar score for all variants in the risk locus. As a result, WEVar successfully prioritizes causal regulatory variants in the risk loci by assigning them the highest WEVar score (Figure 3.9 and Table 3.12). For example, rs34877519 (chr3:2554612) is successfully prioritized by obtaining the score higher than any variant in the risk locus including GWAS SNPs rs11708578 (chr3:2515894, pvalue=7.08x10^{-11}) and rs17194490 (chr3:2547786, pvalue=1.00x10^{-11}); rs7927437 (chr11:123395987) receives the highest score among all variants in the risk locus including GWAS SNP rs77502336 (chr11:123394636, pvalue=3.98x10^{-10}); rs7779548 (chr7:137074540) scores higher than any variant in the risk locus including GWAS SNP rs3735025 (chr7:7:137074844, pvalue=3.98x10^{-12}); rs6498914 (chr16:63699425) obtains the highest score among all variants in the risk locus including GWAS SNP rs2018916 (chr16:63700508, pvalue=7.08x10^{-9}). Overall, these findings demonstrates that causal regulatory variants are not necessary the GWAS lead SNPs but the LD-linked variants. In addition, WEVar is a powerful tool in post-GWAS analysis to pinpoint the causal regulatory variants in the risk loci, which cannot be identified by a standard GWAS approach.

Causal regulatory variants associated with multiple traits and validated by multiple platforms

We benchmark WEVar on state-of-the-art datasets, which are generated from multiple studies for different traits such as Cleft lip/palate, heart, hair color and breast cancer and consists of regulatory variants experimentally validated by different functional assays. Similar to previous analyses, we define the risk locus by considering all variants in LD (r^2 >0.2) for each regulatory variant. Consequently, WEVar is able to prioritize regulatory variants in each risk locus (Figure 3.10 and Table 3.13).

Specifically, rs6801957 (chr3:38767315, pvalue=9x10^{-9}), located in intronic region of SCN10A, has been validated by BAC reporter system and 4C-seq to modulate cardiac SCN5A expression [98]. Consistent with the experimental validation, WEVar assigns the highest score to rs6801957 in the risk locus, which also includes multiple GWAS SNPs rs6795970 (chr3:38766675, pvalue=1x10^{-58}), rs7433306 (chr3:38770639, pvalue=1x10^{-14}), rs6790396 (chr3:38771925, pvalue=2x10^{-39}), rs6599255 (chr3:38796415, pvalue=2x10^{-}10)

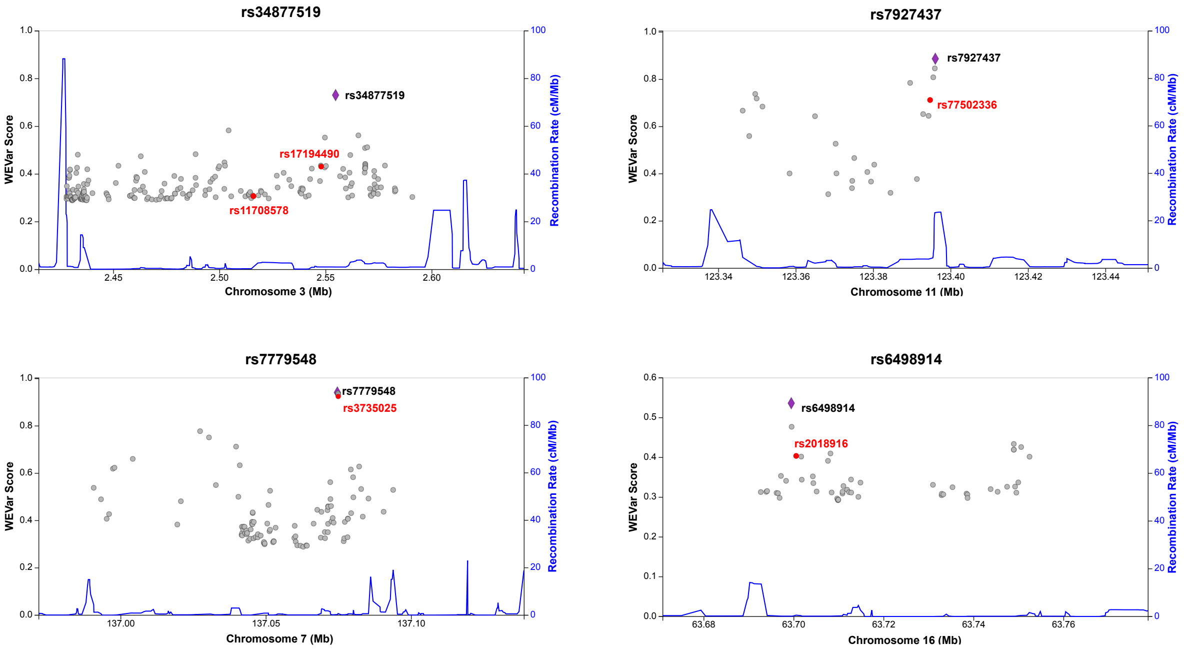

Figure 3.9: WEVar can prioritize causal regulatory variants associated with Schizophrenia. causal regulatory variants are defined as variants with significant differential expression between two alleles (FDR<0.2) in MPRA experiments in both K562 leukemia cells and SK-SY5Y neuroblastoma cells. For each causal regulatory variant, we extend the risk locus by considering all variants in LD (r^2 >0.2). As a result, rs34877519 (chr3:2554612) is successfully prioritized by obtaining the score higher than any variant in the risk locus including GWAS SNPs rs11708578 (chr3:2515894, pvalue=7.08x10^{-11}) and rs17194490 (chr3:2547786, pvalue=1.00x10^{-11}); rs7927437 (chr11:123395987) receives the highest score among all variants in the risk locus including GWAS SNP rs77502336 (chr11:123394636, pvalue=3.98x10^{-10}); rs7779548 (chr7:137074540) scores higher than any variant in the risk locus including GWAS SNP rs3735025 (chr7:7:137074844, pvalue=3.98x10^{-12}); rs6498914 (chr16:63699425) achieves the highest score among all variants in the risk locus including GWAS SNP rs2018916 (chr16:63700508, pvalue=7.08x10^{-9}). The causal regulatory variants validated by MPRA are marked purple. LD-linked GWAS SNPs are marked red.

rs6798015 (chr3:38798836, pvalue=2x10^{-12}) and rs10428132 (chr3:38777554, pvalue=1x10^{-68}). We further evaluate another variant rs227727 (chr17:54776955, pvalue=7.3x10^{-8}), which is mapped to 17q22 NOG locus and found associated with Cleft lip/palate. The NSCL/P-associated allele of rs227727 significantly decreases the nearby enhancer activity compared to the unassociated allele, which is experimentally validated by quantitative reporter assays transfected with a luciferase reporter vector [100]. Similarly, rs227727 is prioritized with

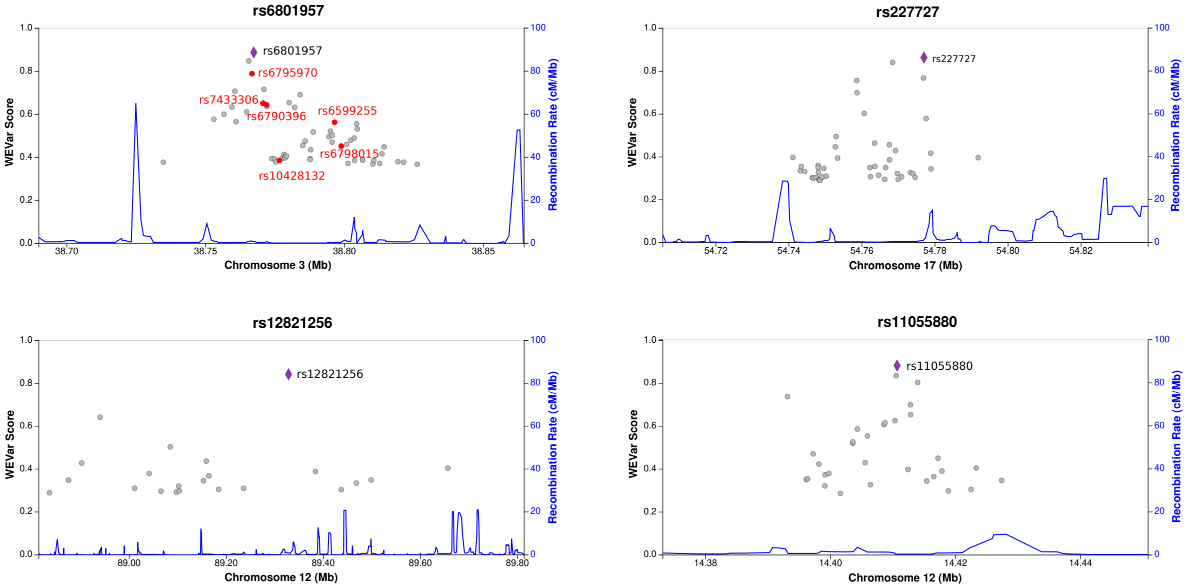

Figure 3.10: WEVar prioritizes causal regulatory variants associated with multiple traits and validated by multiple platforms. We benchmark WEVar on state-of-the-art datasets, which are generated from different studies and consists of regulatory variants experimentally validated by different functional assays. We define the risk locus by considering all variants in LD (r^2 >0.2) for each validated causal regulatory variant. As a result, rs6801957 (chr3:38767315, pvalue=9x10^{-9}), which is validated to modulate cardiac SCN5A expression [98], has been assigned the highest score in the risk locus, which also includes multiple GWAS SNPs rs6795970 (chr3:38766675, pvalue=1x10^{-58}), rs7433306 (chr3:38770639, pvalue=1x10^{-14}), rs6790396 (chr3:38771925, pvalue=2x10^{-39}), rs6599255 (chr3:38796415, pvalue=2x10^{-}10) rs6798015 (chr3:38798836, pvalue=2x10^{-12}) and rs10428132 (chr3:38777554, pvalue=1x10^{-68}). The NSCL/P-associated allele of rs227727 (chr17:54776955, pvalue=7.3x10^{-8}, which is mapped to 17q22 NOG locus and found associated with Cleft lip/palate, significantly decreases the nearby enhancer activity compared to the unassociated allele. Similarly, rs227727 is prioritized by obtaining the highest score in the risk locus. rs12821256 (chr12:89328335, pvalue=4x10^{-30}) is located in a regulatory enhancer in the upstream of lncRNA LINC02458, which has been experimentally validated to alter the binding site of lymphoid enhancer-binding factor1 (LEF1) transcription factor. Again, rs12821256 scores highest in the risk locus, which supports the experimental finding. A breast cancer risk SNP rs11055880 (chr12:14410734), which resides in an intergenetic enhance, has been validated by CRISPR-Cas9 approach [99] to have endogenous regulatory activities on expression of ATF7IP. Consistently, rs11055880 obtains the highest score among all variants in the risk locus. The studied variant is marked purple, and the LD-linked variants are marked red.

the highest score in the risk locus. The next evaluated variant rs12821256 (chr12:89328335, pvalue=4x10^{-30}) is located in a regulatory enhancer in the upstream of lncRNA LINC02458.

It has been experimentally validated that rs12821256 is associated with hair color by altering the binding site of lymphoid enhancer-binding factor1 (LEF1) transcription factor. The altered binding site of LEF will reduce LEF responsiveness and enhancer activity in cultured human keratinocytes [101]. Again, rs12821256 scores highest in the risk locus, which is supported by the experimental finding. The last investigated variant is a breast cancer risk SNP rs11055880 (chr12:14410734), which resides in an intergenetic enhancer and validated by CRISPR-Cas9 approach [99] to have endogenous regulatory activities on expression of ATF7IP. Consistently, rs11055880 obtains the highest score among all variants in the risk locus.

Overall, the consistency between experimental validations and prioritization results based on WEVar score demonstrates the capability and robustness of WEVar to prioritize functional noncoding variants in a LD-linked risk locus. The robustness is reflected by the successful prioritization of heterogeneous variants, which are located in various genomic regions, associated with different traits, and validated by different functional assays.

3.5 Discussion

In this work, we develop a statistical learning framework “WEVar" to predict functional noncoding variants by integrating representative scoring methods in an optimized weighted scheme. The development of WEVar is motivated by the existing gap of strong discordant performance of existing methods on state-of-the-art benchmark datasets, as shown by the inconsistent prediction performance of these methods on the integrated causal regulatory SNPs (Figure 3.2A).

Overall, the advantages of WEVar lies on several aspects. First, existing approaches, either supervised or unsupervised, are developed using thousands of functional annotations derived from multi-omics data deposited in large national consortia such as ENCODE and Roadmap Epigenomics. Different from existing methods, WEVar is developed on top of these methods by directly utilizing genome-wide precomputed functional scores, which collapse multi-dimensional functional annotations into a single score. Therefore, without losing information of functional annotations, direct application of the functional scores of existing approaches significantly reduces the dimensionality of feature space in model development of WEVar. Second, WEVar will identify informative predictors in an optimized weighted scheme and thus can leverage the advantages of different approaches, which likely lead to improved prediction performance compared to each integrated individual scoring method. Third, WEVar offers two modes: “context-free” and “context-dependent". Each mode has its favorite scenario. We adopt a comprehensive training set [84], which integrates curated causal SNPs, located in different genomic regions, collected from different sources and associated with different traits to develop context-free WEVar. The large sample size, diverse context and genomic locations as well as heterogeneous trait association of these training variants make context-free WEVar powerful to predict functional noncoding variants with unknown or heterogeneous context. In contrast, training variant set of context-dependent WEVar is derived from the same context i.e. tissue-, cell type-, disease-specific. The context-specificity of training set makes context-dependent WEVar prefer the scenario when noncoding variants in training and testing set are from the same context, which may lead to improvement of functional prediction.

We perform a real data-based simulation study by considering the inherent correlations of precomputed functional scores among integrated scoring methods. The results demonstrate that WEVar outperforms individual scoring method and can estimate the contributions of integrated scoring methods accurately, which may explain the improved performance of WEVar. Next, we evaluate context-free functional prediction and context-dependent functional prediction respectively on state-of-the-art benchmark datasets, which include three variant sets containing putatively causal regulatory variants derived from statistical associations (i.e. Allelic imbalanced SNPs, Fine mapping eQTLs, GWAS significant noncoding

SNPs), and three datasets consisting of experimentally validated regulatory variants (i.e. Experimentally validated regulatory SNPs, MPRA validated variants in GM12878 lymphoblastoid cells, MPRA validated variants in K562 leukemia cells). Besides evaluating context-dependent WEVar in each benchmark dataset by dividing it into training and testing set, we adapt an independent training and testing set from CAGI. Consequently, both context-free and context-dependent WEVar achieve an overall improvement of functional prediction compared to integrated scoring methods across all datasets. Specifically, WEVar outperforms Unweighted average, indicating the benefit of exploiting the optimized contributions of individual scoring method. GWAVA_Unmatched and Unweighted average are top-performed. In contrast, DANN, CADD and FATHMM_MKL always perform poorly. By comparing context-free and context-dependent WEVar on the same benchmark datasets, we find that context-dependent WEVar improve the functional prediction compared to context-free WEVar except for Experimentally validated regulatory SNPs possibly to the small sample size of training set. This observation indicates that being context-dependent improves the functional prediction and a large sample size is needed for make this improvement.

Another important characteristic of WEVar is that it can identify predictors that play major contribution to the functional prediction. As a result, major contributors are different across benchmark datasets. In most cases, WEVar identifies a parsimonious set of scoring methods that dominate the functional prediction especially FunSeq2 and GWAVA_Unmatched are two ubiquitous major contributors. However, for GWAS noncoding SNPs and MPRA validated regulatory variants in K562 leukemia cells, there is a ubiquitous solution, where the contributions of all methods are relative uniform. These findings demonstrate that both estimated weights and major contributors vary from context to context. Thus, it is important to obtain an optimal weights when integrating different scoring methods, as the non-uniform weights estimated by WEVar lead improved functional prediction across benchmark datasets. Additionally, this point is validated by both simulation and real data applications that WEVar outperforms the unweighted average of functional scores.

To demonstrate the application of WEVar in complex traits, we apply WEVar in the fine mapping analysis to evaluate whether it can successfully prioritize causal regulatory variants among LD-linked noncoding variants. By using precomputed WEVar score directly, variants assigned the highest score in a risk locus is considered to be prioritized. By using three benchmarking datasets of experimentally validated regulatory variants, we find that WEVar can prioritize regulatory variants modulating gene expression in GM12878 lymphoblastoid cells, associated with Schizophrenia and multiple traits such as Cleft lip/palate, heart, hair color and breast cancer. These findings demonstrate that WEVar can prioritize functional noncoding variants in risk loci and therefore alleviate the limitation of current GWAS, where the true causal SNPs may be masked by LD.

WEVar is a flexible approach, which can be further extended and improved by both integrated scoring methods and training variant set. In the current implementation, we include several representative scoring methods that are most popular in this field. With the rapid development post-GWAS analysis, there are other powerful methods developed or developing can be integrated into WEVar to further improve the prediction performance. The flexibility of WEVar is also reflected on the training variant set. With the affordability and popularity of functional assays such as massively parallel reporter assays (MPRAs) and clustered regularly interspaced short palindromic repeats (CRISPR)-based gene editing, more experimentally validated functional variants can be discovered and integrated into WEVar to improve the predictive power.

3.6 Appendix

3.6.1 Supplementary tables

Table 3.1: Download link for precomputed scores from individual scoring methods

GWAVA	ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/
CADD	http://cadd.gs.washington.edu/download
DANN	https://cbcl.ics.uci.edu/public_data/DANN/
FATHMM-MKL	https://github.com/HAShihab/fathmm-MKL
FunSeq2	http://funseq2.gersteinlab.org/downloads
Eigen	https://xioniti01.u.hpc.mssm.edu/v1.0/
LINSIGHT	https://github.com/CshlSiepelLab/LINSIGHT

Table 3.2: Allelic imbalanced SNPs

	Context-free			Context-dependent		
Functional Score	AUROC	AUPR	COR	AUROC	AUPR	COR
Eigen	0.754	0.731	0.410	0.754	0.732	0.410
CADD	0.639	0.610	0.228	0.639	0.611	0.228
DANN	0.634	0.563	0.236	0.634	0.564	0.236
FATHMM_MKL	0.752	0.690	0.053	0.752	0.692	0.053
FunSeq2	0.827	0.788	0.467	0.827	0.788	0.467
GWAVA_Region	0.723	0.691	0.382	0.723	0.691	0.382
GWAVA_TSS	0.840	0.796	0.559	0.840	0.796	0.559
GWAVA_Unmatched	0.875	0.823	0.535	0.875	0.824	0.535
LINSIGHT	0.883	0.823	0.255	0.883	0.824	0.257
Unweighted average	0.857	0.825	0.559	0.857	0.826	0.559
WEVar	**0.894**	**0.852**	**0.644**	**0.902**	**0.870**	**0.698**

Table 3.3: Fine mapping eQTLs

	Context-free			Context-dependent		
Functional Score	AUROC	AUPR	COR	AUROC	AUPR	COR
Eigen	0.652	0.614	0.256	0.652	0.614	0.256
CADD	0.581	0.533	0.126	0.581	0.533	0.126
DANN	0.586	0.511	0.150	0.586	0.511	0.150
FATHMM_MKL	0.662	0.586	0.047	0.662	0.586	0.047
FunSeq2	0.737	0.690	0.348	0.737	0.690	0.348
GWAVA_Region	0.574	0.554	0.142	0.574	0.554	0.142
GWAVA_TSS	0.691	0.672	0.341	0.691	0.672	0.341
GWAVA_Unmatched	**0.821**	**0.784**	0.476	0.821	0.784	0.476
LINSIGHT	0.780	0.693	0.146	0.780	0.694	0.146
Unweighted average	0.750	0.715	0.412	0.748	0.712	0.408
WEVar	0.816	0.781	**0.509**	**0.829**	**0.792**	**0.539**

Table 3.4: GWAS noncoding SNPs

	Context-free			Context-dependent		
Functional Score	AUROC	AUPR	COR	AUROC	AUPR	COR
Eigen	0.576	0.367	0.119	0.576	0.367	0.119
CADD	0.528	0.330	0.052	0.528	0.330	0.052
DANN	0.520	0.309	0.034	0.520	0.310	0.034
FATHMM_MKL	0.570	0.350	0.015	0.570	0.351	0.015
FunSeq2	0.582	0.376	0.132	0.582	0.377	0.132
GWAVA_Region	0.554	0.352	0.091	0.554	0.353	0.091
GWAVA_TSS	0.576	0.358	0.117	0.576	0.359	0.117
GWAVA_Unmatched	0.583	0.362	0.109	0.583	0.362	0.109
LINSIGHT	0.575	0.364	0.064	0.575	0.365	0.064
Unweighted average	0.586	0.373	0.137	0.587	0.374	0.138
WEVar	**0.596**	**0.376**	**0.142**	**0.607**	**0.391**	**0.168**

Table 3.5: Experimental validated regulatory SNPs

	Context-free			Context-dependent		
Functional Score	AUROC	AUPR	COR	AUROC	AUPR	COR
Eigen	0.801	0.686	0.501	0.816	0.722	0.502
CADD	0.739	0.539	0.288	0.757	0.571	0.323
DANN	0.703	0.470	0.329	0.694	0.505	0.305
FATHMM_MKL	0.758	0.584	0.347	0.776	0.639	0.364
FunSeq2	0.820	0.744	0.567	0.823	0.739	0.567
GWAVA_Region	0.752	0.676	0.464	0.748	0.678	0.430
GWAVA_TSS	0.862	0.803	0.613	0.856	0.778	0.599
GWAVA_Unmatched	0.901	0.828	0.649	0.896	0.832	0.663
LINSIGHT	0.820	0.647	0.308	0.825	0.671	0.318
Unweighted average	0.883	0.789	0.617	0.881	0.796	0.614
WEVar	**0.912**	**0.865**	**0.718**	**0.902**	**0.861**	**0.705**

Table 3.6: MPRA variants in GM12878 lymphoblastoid

	Context-free			Context-dependent		
Functional Score	AUROC	AUPR	COR	AUROC	AUPR	COR
Eigen	0.601	0.364	0.228	0.602	0.368	0.228
CADD	0.567	0.240	0.087	0.567	0.249	0.087
DANN	0.544	0.254	0.060	0.546	0.256	0.060
FATHMM_MKL	0.568	0.272	0.022	0.569	0.274	0.112
FunSeq2	0.649	0.390	0.275	0.652	0.395	0.278
GWAVA_Region	0.601	0.321	0.169	0.600	0.325	0.168
GWAVA_TSS	0.635	0.292	0.190	0.635	0.295	0.190
GWAVA_Unmatched	**0.677**	**0.445**	**0.317**	0.676	0.451	0.316
LINSIGHT	0.662	0.331	0.144	0.663	0.335	0.144
Unweighted average	0.646	0.404	0.280	0.642	0.396	0.270
WEVar	0.674	0.412	0.286	**0.685**	**0.459**	**0.363**

Table 3.7: MPRA variants in K562 leukemia

	Context-free			Context dependent		
Functional Score	AUROC	AUPR	COR	AUROC	AUPR	COR
Eigen	0.579	0.251	0.109	0.582	0.263	0.111
CADD	0.540	0.211	0.047	0.539	0.218	0.047
DANN	0.545	0.216	0.068	0.548	0.225	0.070
FATHMM_MKL	0.553	0.258	0.065	0.554	0.265	0.067
FunSeq2	0.615	0.278	**0.165**	0.617	0.287	0.165
GWAVA_Region	0.577	0.229	0.100	0.576	0.242	0.100
GWAVA_TSS	0.597	0.243	0.126	0.598	0.249	0.128
GWAVA_Unmatched	0.608	0.267	0.150	0.608	0.278	0.152
LINSIGHT	0.565	0.224	0.066	0.571	0.233	0.069
Unweighted average	0.591	0.246	0.125	0.593	0.261	0.128
WEVar	**0.622**	**0.276**	0.149	**0.635**	**0.315**	**0.185**

Table 3.8: CAGI (A)

Functional Score	AUROC	AUPR	COR
Eigen	0.5794	0.2279	0.1582
CADD	0.5453	0.1469	0.0579
DANN	0.5416	0.1541	0.0428
FATHMM_MKL	0.5692	0.1898	0.1367
FunSeq2	0.6113	0.2177	0.1696
GWAVA_Region	0.6042	0.2063	0.1436
GWAVA_TSS	0.6071	0.1622	0.1085
GWAVA_Unmatched	0.6398	0.2567	0.1908
LINSIGHT	0.6215	0.2119	0.1507
Unweighted average	0.6133	0.2493	0.1927
WEVar	**0.6454**	**0.2761**	**0.2507**

Table 3.9: CAGI (B)

Functional Score	AUROC	AUPR	COR
Eigen	0.5452	0.2278	0.1427
CADD	0.5391	0.1432	0.0441
DANN	0.5318	0.1449	0.0396
FATHMM_MKL	0.5425	0.1542	0.0738
FunSeq2	0.6142	0.2290	0.1790
GWAVA_Region	0.6191	0.2223	0.1627
GWAVA_TSS	0.6105	0.1790	0.1287
GWAVA_Unmatched	0.6348	0.3012	0.2076
LINSIGHT	0.6143	0.1776	0.0877
Unweighted average	0.6005	0.2403	0.1835
WEVar	**0.6364**	**0.3129**	**0.2786**

Table 3.10: Estimated weights for all scoring methods integrated by WEVar on benchmark datasets

Dataset	Eigen	CADD	DANN	FATHMM_MKL	FunSeq2	GWAVA_Region	GWAVA_TSS	GWAVA_Unmatched	LINSIGHT
HGMD	0.000	0.000	0.000	0.000	0.203	0.049	0.281	0.467	0.000
Allelic imbalance SNPs	0.051	0.000	0.000	0.000	0.403	0.000	0.222	0.324	0.000
Fine mapping eQTLs	0.000	0.000	0.000	0.000	0.228	0.000	0.000	0.772	0.000
GWAS noncoding SNPs	0.216	0.018	0.000	0.047	0.287	0.097	0.126	0.156	0.053
Experimental validated regulatory SNPs	0.034	0.064	0.001	0.000	0.314	0.000	0.207	0.382	0.000
MPRA variants in GM12878 lymphoblastoid	0.240	0.000	0.000	0.000	0.250	0.038	0.035	0.437	0.000
MPRA variants in K562 leukemia	0.225	0.114	0.011	0.030	0.329	0.014	0.059	0.160	0.057
CAGI (A)	0.262	0.019	0.000	0.000	0.166	0.169	0.000	0.385	0.000
CAGI (B)	0.260	0.000	0.000	0.000	0.179	0.147	0.000	0.414	0.000

Table 3.11: WEVar scores for emVars and LD-linked GWAS SNPs

emVars		LD-linked GWAS SNPs					
Variant ID	WEVar score	Variant ID	WEVar score	Variant ID	WEVar score	Variant ID	WEVar score
rs4790718	**0.9277**	rs1060431	0.8191	rs6065	0.7259	rs571461910	0.6301
rs922483	**0.9737**	rs2736340	0.6289				
rs56316188	**0.9096**	rs2859998	0.8530				
rs306587	**0.9349**	rs1042058	0.7411				

Table 3.12: WEVar scores for causal regulatory variants and LD-linked GWAS SNPs associated with Schizophrenia

Regulatory variants		LD-linked GWAS SNPs			
Variant ID	WEVar score	Variant ID	WEVar score	Variant ID	WEVar score
rs34877519	**0.7312**	rs17194490	0.4319	rs11708578	0.3068
rs7927437	**0.8873**	rs77502336	0.7125		
rs7779548	**0.9416**	rs3735025	0.9255		
rs6498914	**0.5364**	rs2018916	0.4039		

Table 3.13: WEVar scores for regulatory variants associated with multiple traits and LD-linked GWAS SNPs

Regulatory variants		LD-linked GWAS SNPs					
Variant ID	WEVar score	Variant ID	WEVar score	Variant ID	WEVar score	Variant ID	WEVar score
rs6801957	**0.8879**	rs6795970	0.7892	rs7433306	0.6514	rs6790396	0.6444
		rs6599255	0.5627	rs6798015	0.4522	rs10428132	0.3957
rs227727	**0.8643**						
rs12821256	**0.8426**						
rs11055880	**0.8825**						

Chapter 4

DeepMFIVar: a deep multimodal learning framework for functional interpretation of genetic variants

4.1 Introduction

Identifying functional variants underlying disease risk is currently limited by the challenge of interpreting the functional consequences of genetic variants. In the past ten years, thousands of loci associated with the risk of human disease have been identified by Genome-wide association studies (GWAS) [102]. Moreover, quantitative trait locus (QTL) analysis have uncovered “xQTLs” [60] affecting molecular phenotypes such as eQTLs [61] for gene expression; mQTLs [62] for DNA methylation; aseQTLs [63] for allele-specific expression; dsQTLs [64] for DNase I sensitivity and sQTLs [65] for alternative RNA splicing. However, these studies are limited by the sample size and linkage disequilibrium [103, 104, 105], which will mask true causal variants from the neural ones. Thus, integrating biological meaning to this GWAS studies plays an important role in increasing the resolution of disease risk associated region and interpreting GWAS results [106, 107, 108, 109]. Recent studies in developing computational models that predict TF binding, chromatin accessibility, and histone modification from the genome sequence in the near region provide novel frameworks for understanding the functional consequences of non-coding genetic variants. Most of these models leverage advances in deep learning to predict the functional consequences by taking the DNA sequence context. These computation approaches only rely on the reference genome and fail to model the contribution of genetic variance. However, the genetic variation and confounding factors (e.g., gender age) driving the epigenetic signal are essential for molecular phenotypes (See Fig. 4.1), and introducing genetics and confounding factors into

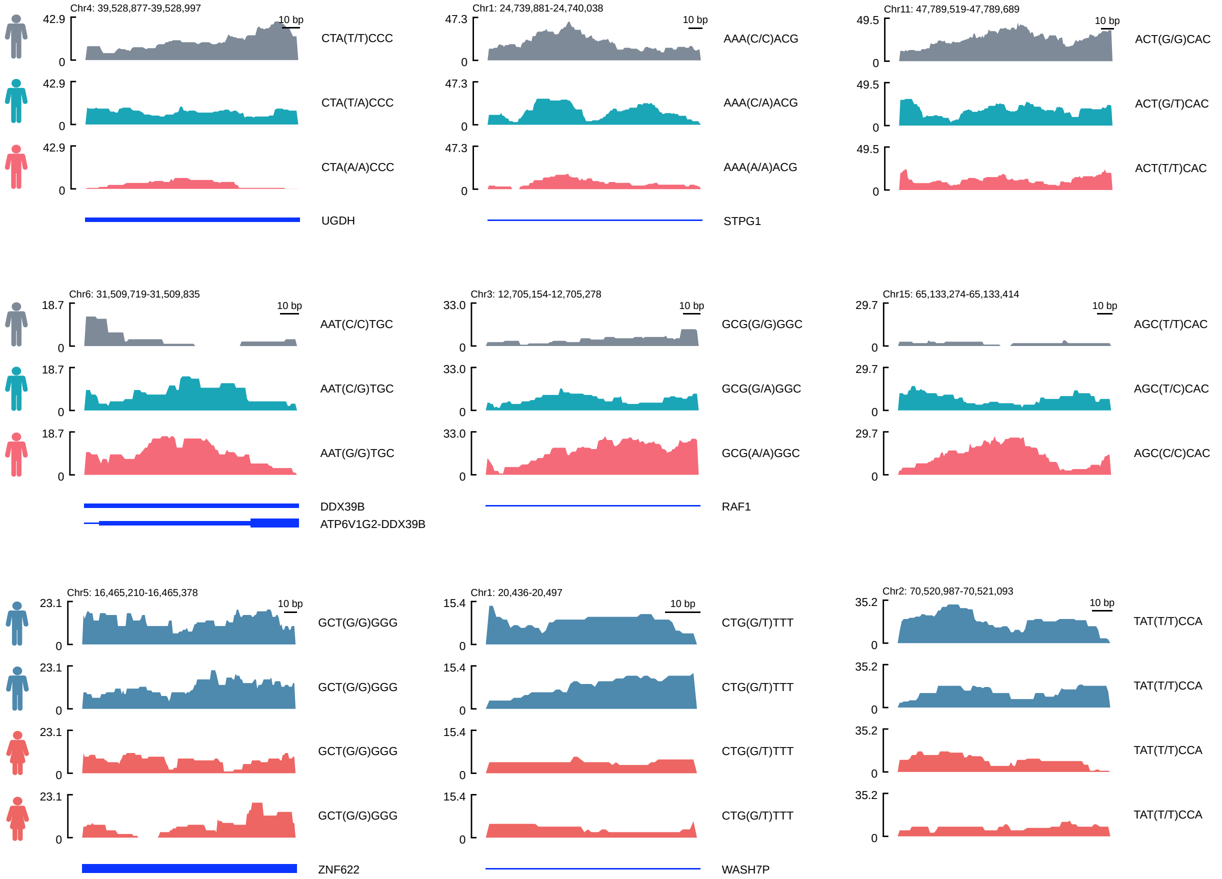

Figure 4.1: Histone modification (H3K9ac) from epigenetic assay (ChIP-seq) across multiple individuals and genomic regions.

the computational model have the potential to improve prediction accuracy and increase the power of variant impact predictions.

Here we introduce a deep multimodal learning framework for functional interpretation of genetic variants (DeepMFIVar). Our framework is designed by (i) building a novel deep multimodal learning framework by taking both DNA sequence context and patient-level clinical and pathologic phenotypic information, (ii) performing model training on multiple epigenetic experiments, (iii) integrating whole genome sequencing to create a personal genome sequence for each individual, (iv) modeling quantitative variation across different individuals in the epigenetic signals.

4.2 Materials

4.2.1 Epigenomic datasets from human frontal cortex for aging and Alzheimer's disease

Epigenomic datasets comprise profiling assay experiments for DNA methylation for 202 individuals and ChIP-seq experiments for histone modifications (H3K9ac) for 196 individuals. All individuals are part of the Religious Order Study (ROS) or the Memory and Aging Project (MAP), two cohort studies of aging that includes brain donation at the time of death. The clinical and pathologic phenotypic data of these individuals are also available from "ROS-MAP" studies. The original data is downloaded from the Rush Alzheimer's Disease Center website (https://www.radc.rush.edu/). DNA methylation dataset consists of methylation ratio at $415,848$ discrete CpG dinucleotides. These methylation profiles are generated by the Illumina HumanMethylation450 beadset and a sample of the dorsolateral prefrontal cortex obtained from each individual[110]. Histone modification (H3K9ac) dataset comprises $2,269,524$ H3K9ac regions. Peaks are detected for each sample individually by MACS2 using the broad peak option. We employ a series of ChIP-seq quality measures to remove low-quality samples, which results in 669 individual samples[110]. A combination of different filters are applied to remove low-quality H3K9ac domains, i.e., i) the ChIP counts are less than control counts, ii) P-value of Poisson test are less than 0.05. After quality control, $141,807$ out of $2,269,524$ remains.

4.2.2 Whole Genome Sequencing (WGS)

Individuals' brain tissue DNA are acquired from the ROSMAP study. We download the individual genotype calls in genomic VCF files. The reference genome sequence is modified to generate the personal genome sequence based on variant calling files. We only consider the single nucleotide polymorphisms and ignore the insertion and deletion.

4.3 Methods

4.3.1 Constructing personal genome and clinical outcome as input

We construct the personal genome sequence using the GRCH37 reference genome with sites modified according to the biallelic SNPs in the variant calling file (VCF) of the ROSMAP study. For the homozygous alternate sites, we directly replace the reference allele with the alternative allele. The heterozygous sites are represented by the IUPAC nucleotide codes. For example, an A/G heterozygote is coded by the characters 'R'. Only biallelic SNPs are considered, so we have six additional characters. Each homozygous site is one-hot encoded into 4×1 matrix, with rows corresponding to A, G, C, and T. Heterozygous sites are encoded with a value of 0.5 in the two corresponding rows.

We extract the genome sequence with a fixed distance from the CpGs site in the methylation island and from the center of the H3K9ac domain for DNA methylation and histone modification (H3K9ac), respectively. The extracted genome sequence are matched with the methylation ratio, and histone modification reads. We use 1000 bp and 2000 bp sequence length to extract regions from the personal genomes for DNA methylation and histone modification, respectively. We collect seven clinical and pathologic phenotypic features for each individual. The detailed information is described in Table 4.2. We further apply minimax-normalization on these phenotypic features.

4.3.2 DeepMFIVar

DeepMFIVar model consists of two major components: 1) Modality Embedding Subnetworks that take input features and output a rich modality embedding. 2) Tensor Fusion Layer explicitly that models the unimodal and bimodal interactions by using a Cartesian product from modality embeddings.

The personal genomics subnetwork learns a rich representation of the personal DNA sequence (Fig 4.2). To formally define our personal genomic subnetwork $\mathcal{U}_g$, let $\mathbf{x} =$

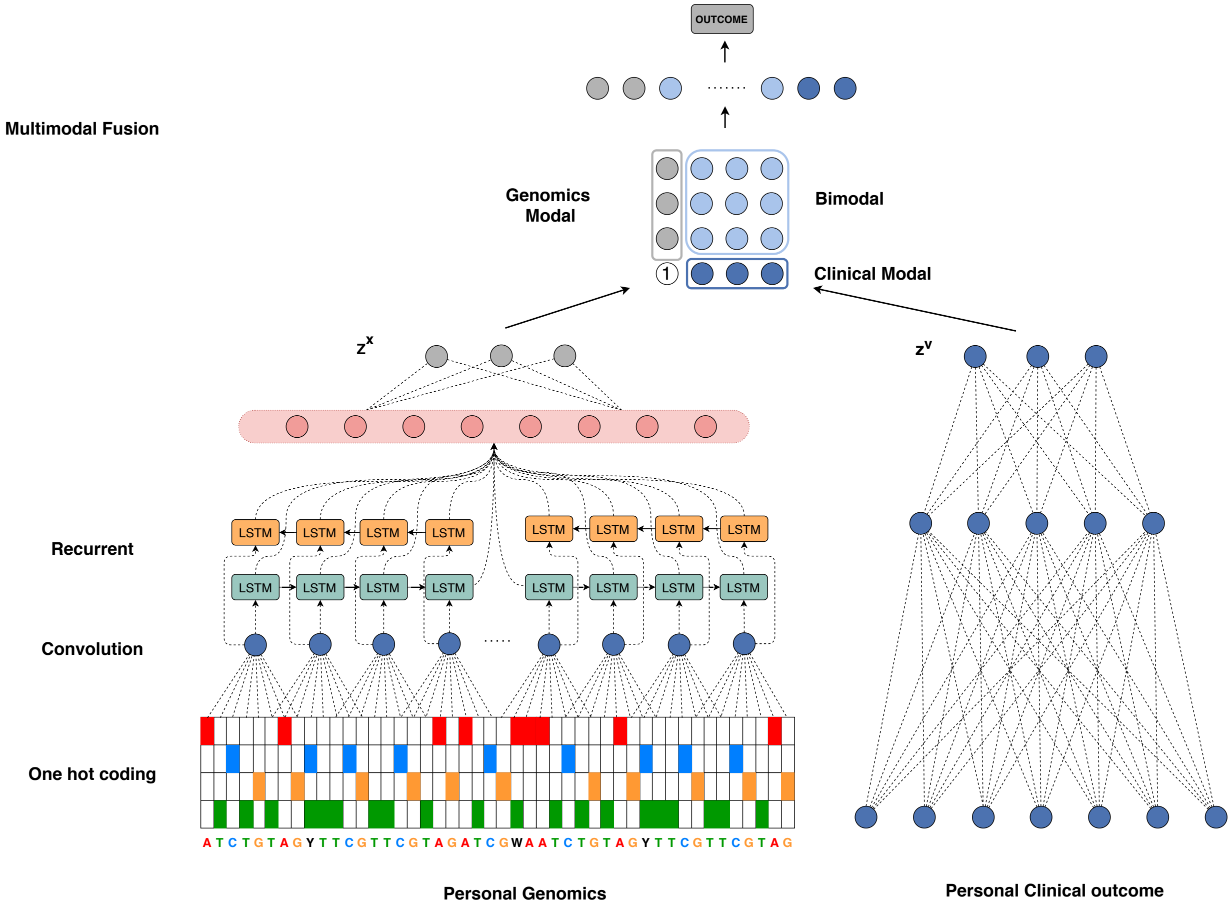

Figure 4.2: Graphical illustration of the DeepMFIVar model

$\{x_1, x_2, x_3, \cdots, x_m\}$, where m is the length of DNA sequence, be the one-hot encoding representation of personal genomics.

A convolution layer acts as a motif scanner across the input matrix to produce a feature map X_t for each convolution kernel. A BLSTM network[111] with a forget gate [112] is employed to learn the orientation and spatial distances dependent motif representations,

$\mathbf{h}_x = \{h_1, h_2, h_3, \cdots, h_{T_x}\}$ according to the following LSTM formulation.

$$\begin{pmatrix} i \\ f \\ o \\ m \end{pmatrix} = \begin{pmatrix} \text{sigmoid} \\ \text{sigmoid} \\ \text{sigmoid} \\ \tanh \end{pmatrix} \mathbf{W}_{g_d} \begin{pmatrix} \mathrm{X_t} \mathbf{W}_{g_c} \\ h_{t-1} \end{pmatrix} \tag{4.1}$$

$$c_t = f \odot c_{t-1} + i \odot m \tag{4.2}$$

$$h_t = o \otimes tanh(c_t) \tag{4.3}$$

$$\mathbf{h_x} = [h_1; h_2; h_3; \cdots; h_{T_x}] \tag{4.4}$$

$\mathbf{h}_x$ is a matrix of sequence feature representations, and is then used as input to a fully-connected network that generates sequence embedding $\mathbf{z}^x$

$$\mathbf{z}^x = \mathcal{U}_g(\mathbf{x}; \mathbf{W}_g) \tag{4.5}$$

where $\mathbf{W}_g$ is the set of all weights in the $\mathcal{U}_g$ network (including $\mathbf{W}_{g_d}, \mathbf{W}_{g_c}$, and $\mathbf{W}_{g_{fc}}$).

Clinical subnetwork is a simple fully-connected neural network $\mathcal{U}_c$. Let $\mathbf{v} = \{v_1, v_2, v_3, \cdots, v_p\}$ be the clinical and pathologic phenotypic features for each individual. $\mathbf{v}$ is passed to a fully-connected network to generate clinical embedding $\mathbf{z}^v$:

$$\mathbf{z}^v = \mathcal{U}_c(\mathbf{v}; \mathbf{W}_c) \tag{4.6}$$

We build a fusion layer that disentangles unimodal and bimodal dynamics by modeling each of them explicitly, which is defined as the following by the Cartesian product:

$$\left\{ (\mathbf{z}^x, \mathbf{z}^v) \quad | \quad \mathbf{z}^x \in \begin{bmatrix} \mathbf{z}^x \\ 1 \end{bmatrix}, \mathbf{z}^v \in \begin{bmatrix} \mathbf{z}^v \\ 1 \end{bmatrix} \right\} \tag{4.7}$$

The extra constant scalar with value 1 is used to generate the unimodal and bimodal dynamics. This definition is equivalent to a differentiable outer product between sequence representation $\mathbf{z}^x$ and the clinical representation $\mathbf{z}^v$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}^x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^v \\ 1 \end{bmatrix} \tag{4.8}$$

$$= \begin{bmatrix} \mathbf{z}^x & \mathbf{z}^x \otimes \mathbf{z}^v \\ 1 & \mathbf{z}^v \end{bmatrix} \tag{4.9}$$

where $\otimes$ indicates the outer product between vectors. $\mathbf{Z}$ is a 2D matrix of all possible combination of unimodal embeddings with three distinct subregions. The subregions $\mathbf{z}^x$ and $\mathbf{z}^v$ are unimodal representations from modality embedding subnetworks forming unimodal interations in tensor fusion layer. The other subregion $\mathbf{z}^x \otimes \mathbf{z}^v$ captures bimodal interactions in tensor fusion layer. After the multimodal fusion, $\mathbf{Z}$ will be flatten and fed to a fully connect neural network which predicts outcome.

4.3.3 Model training and testing

DeepMFIVar model are trained using the Adam algorithm [43] with a minibatch size of 256 to minimize the mean square error loss function on the training set. The weights in convolutional and fully-connected layers are initialized by randomly Xavier uniform distribution, and the orthogonal initialization is applied to initialize the weights in LSTM layers. Validation loss is evaluated at the end of each training epoch to monitor convergence. To improve generalization and to prevent overfitting, we leverage the dropout-layer technique and L2 regularization for weights in the cost function. Each fully connected layer is followed by a dropout layer to avoid overfitting [44]. Specifically, 50% hidden neurons in the fully connected layer are randomly dropped out. The DeepMFIVar model is implemented by Pytorch[113].

Due to the strong positive correlation of epigenomic signals across individuals, we conduct a conservative method to split the dataset into training, validation, and testing sets to guarantee the independence of the three partitions. The training set is composed of samples on chromosome 1-8 from 60% of individuals. The samples on chromosome 17-22 from 20% of individuals are used for hyper-parameter tuning and model selection, and the samples on chromosome 9-16 from the last 20% are held-out for testing.

4.3.4 Variant effect prediction

Given a sequence variant, we predict the epigenomic signals within 1000bp with both of the alternative and reference alleles. We evaluate the impact of the variant by calculating the change of the predicted epigenomic signals:

$$\delta = \mathcal{F}(\text{ref}) - \mathcal{F}(\text{alt}) \tag{4.10}$$

and the change of log odds:

$$\delta_{\text{log}} = \log \frac{\mathcal{F}(\text{ref})}{1 - \mathcal{F}(\text{ref})} - \log \frac{\mathcal{F}(\text{alt})}{1 - \mathcal{F}(\text{alt})} \tag{4.11}$$

where $\mathcal{F}$ is the DeepMFIVar model, and δ and δ_{log} indicate the predicted functional score. We obtain the coordinates and alleles of biallelic SNVs from whole-genome sequencing (WGS) from gnomAD r2.1.1. We remove the multi-allelic sites, which results in a total of 217 million biallelic SNVs. For evaluating variant effects, the DeepMFIVar model is trained on the dataset including all the chromosomes and individuals.

4.4 Results

We first evaluate the performance of DeepMFIVar to predict quantitative signals (i.e., DNA methylation and H3K9ac) from epigenetic experiments with sequence context and clinical and pathologic phenotypic data. We further assess the performance of predicting

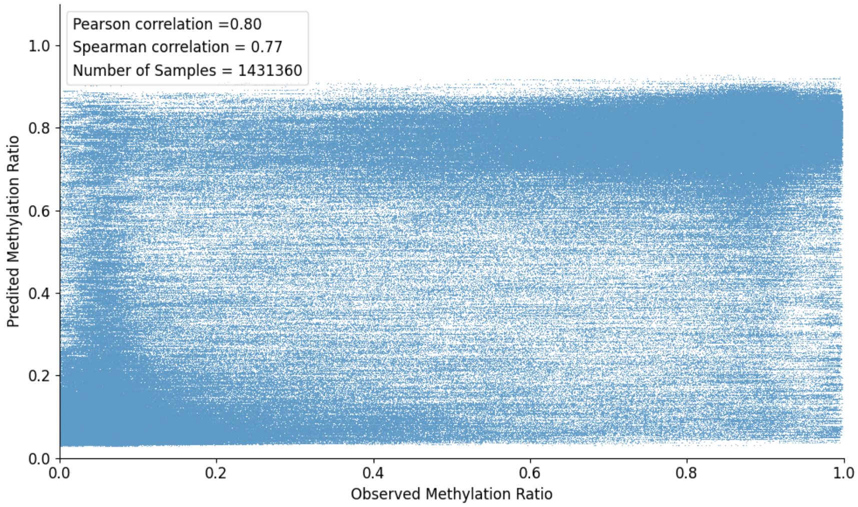

Figure 4.3: Predicted DNA Methylation compared to observed DNA Methylation evaluated on the test set

the impact of functional variants on epigenetic signals. Last, we evaluate the performance of DeepMFIVar on data imputation of the RRBS DNA methylation. The epigenetic signal prediction performance of DeepMFIVar is evaluated using Pearson correlation and Spearman correlation between predicted signals and observed signals. The performance of prioritizing functional and imputing DNA methylation ratio is assessed by using the area under the receiver operating characteristics curve (AUROC).

4.4.1 DeepMFIVar predicts epigenetic signals from sequence context and clinical outcome

DeepMFIVar combines the quantitative epigenetic signal across multiple individuals with whole-genome sequencing into a single supervised machine learning task (Fig 4.2). We first evaluate the ability of DeepMFIVar to predict the DNA methylation ratio of a CpG site from its flanking sequence and clinical outcome. We train the DeepMFIVar model for 10 epochs, evaluating the mean square error on the validation set at the end of each epoch to monitor training progress. We report the prediction results on the test dataset, where the individuals and chromosomes are excluded from the training dataset (Table 4.1). The

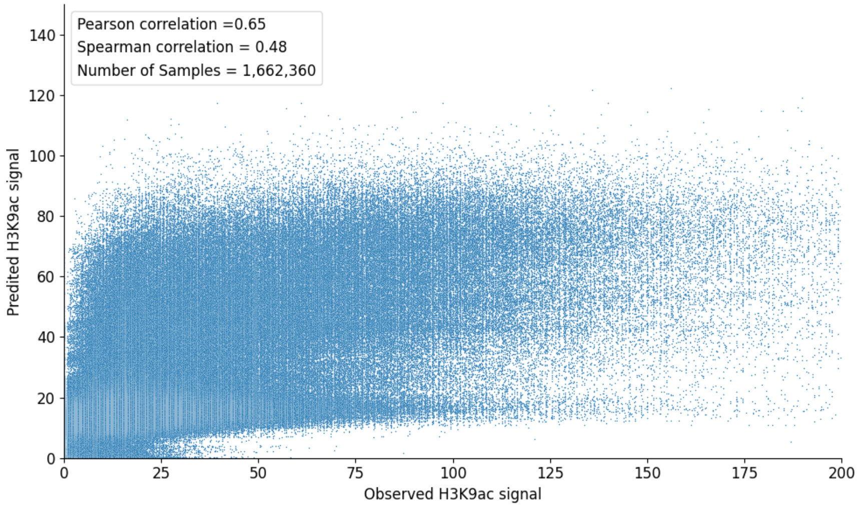

Figure 4.4: Predicted H3K9ac signal compared to observed H3K9ac signal evaluated on the test set

predicted DNA methylation ratio shows strong concordance with the observed methylation ratio in the test set (Spearman correlation = 0.77, Pearson correlation =0.80)(Figure 4.3). Similar to the evaluation of DNA methylation ratio predictions, we train the model on the H3K9ac dataset for 10 epochs and to monitor training progress by the mean square error on the validation dataset. The predicted H3K9ac signal also shows strong correlation with the observed signals in the test set (Spearman correlation = 0.48, Pearson correlation =0.65)(Figure 4.4).

4.4.2 Imputation of RRBS DNA methylation

Reduced representation bisulfite sequencing (RRBS) has been widely used to analyze the genome-wide methylation profiles on a single nucleotide level. However, RRBS is still challenging and expensive, compared with the Illumina Methylation Assay. Here, we use the DeepMFIVar model trained on the dataset from Illumina microarray to impute the RRBS DNA methylation ratio. The 50 RRBS datasets of immortal cell lines, including GM12878 and the WGBS dataset of GM12878, are obtained from the authors of Zeng et al. ([114]). The datasets are downloaded from ENCODE website `https://www.encodeproject.org/`,

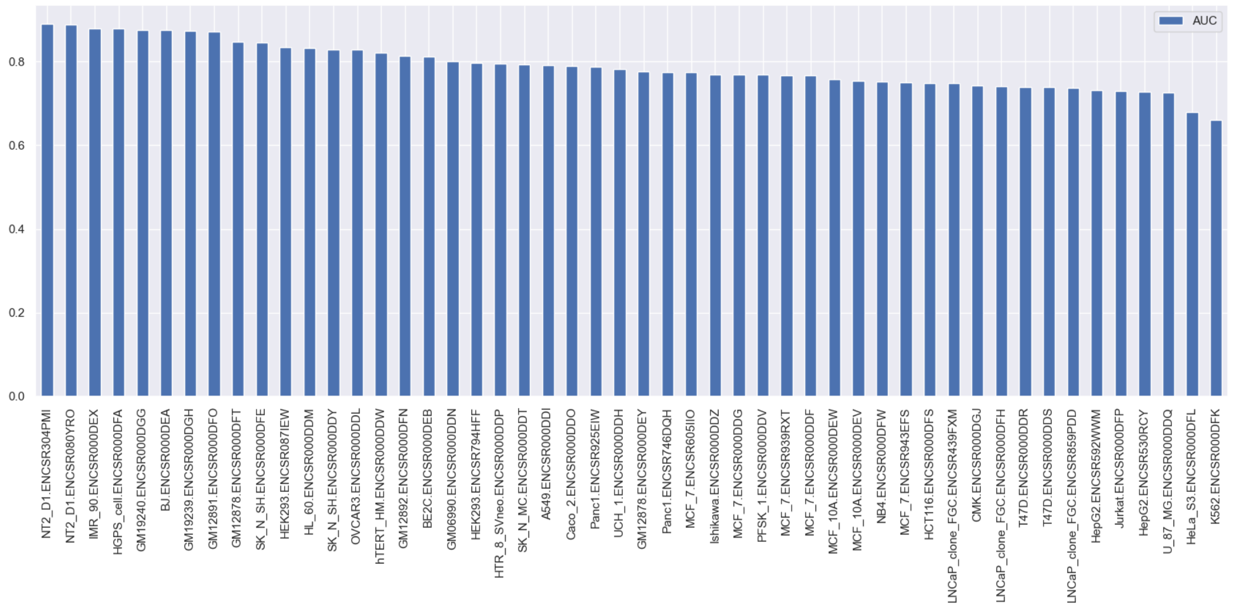

Figure 4.5: AUC for DNA Methylation imputation for 50 ENCODE RRBS datasets

and multiple replicates for the same experiments are merged. We train our DeepMFIVar model on the DNA methylation dataset we previously described in section 4.2.1, and test on 50 ENCODE RRBS datasets. The results demonstrate that DeepMFIVar trained on microarray DNA methylation dataset can accurately impute thos from RRBS by having the average AUC 0.79 (Figure 4.5).

4.5 Appendix

4.5.1 Supplementary figures

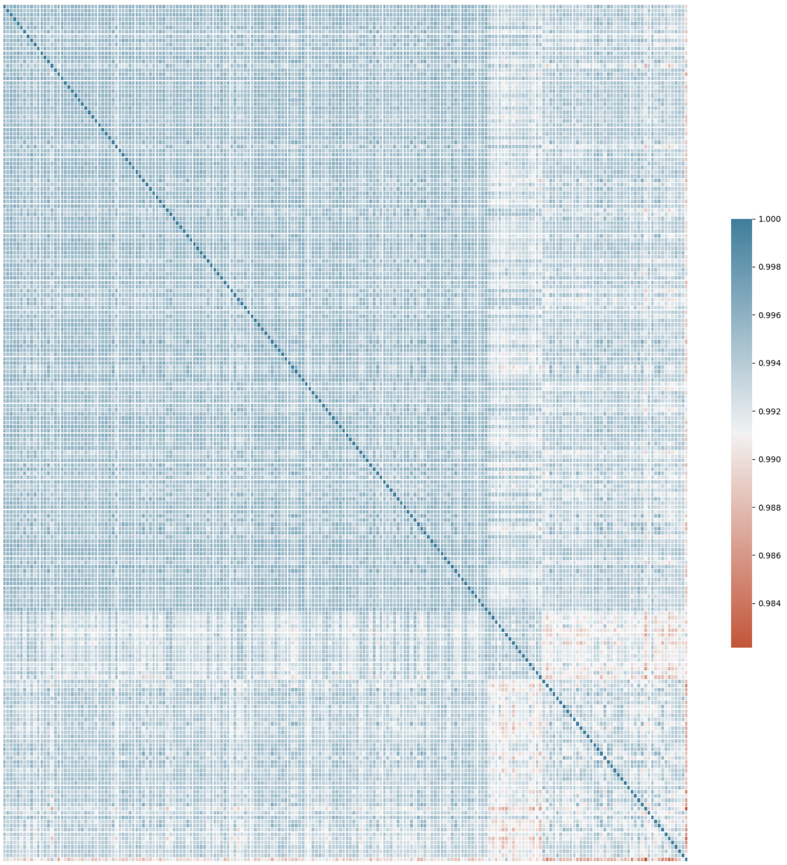

Figure 4.6: Pearson correlation heatmap of DNA methylation ratio across individuals

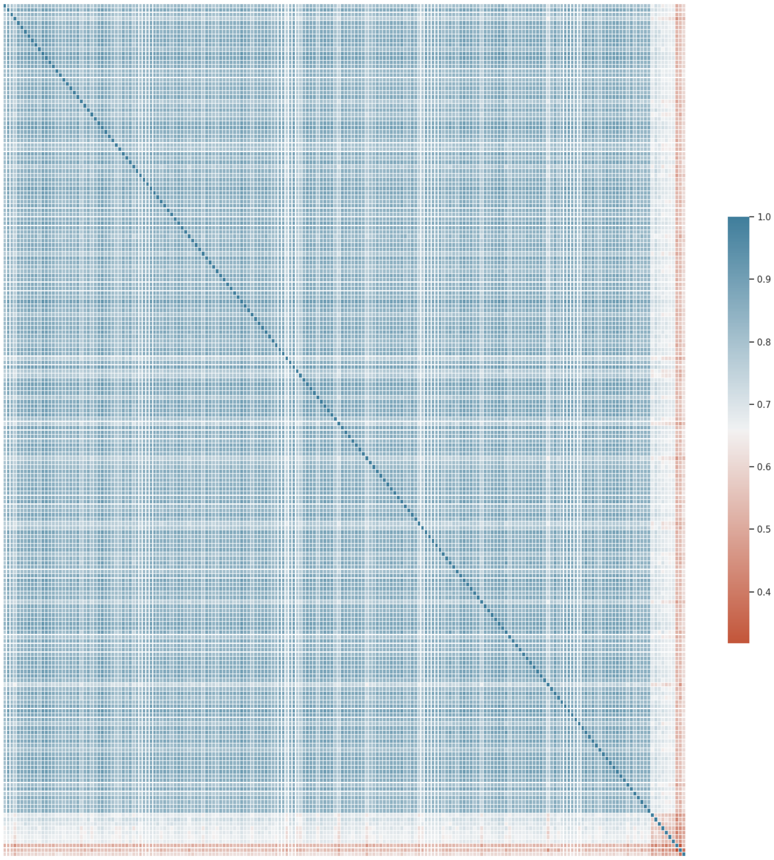

Figure 4.7: Pearson correlation heatmap of histone modification (H3K9ac) across individuals

4.5.2 Supplementary tables

Table 4.1: Summary of DNA methylation and histone modification datasets

Dataset	Set	Samples	Chromsomes	Number of Regions
Methylation	Training	122	chr1-8	7,790,310
	Validation	40	chr17-22	1,431,360
	Testing	40	chr9-16	1,447,120
H3K9ac	Training	118	chr1-8	7,686,284
	Validation	39	chr17-22	1,368,549
	Testing	39	chr9-16	1,620,801

Table 4.2: List of clinical and pathologic phenotypic outcome for each individual.

N	Features	Type
1	Gender	Categorical
2	Education	Continuous
3	Age at death	Continuous
4	APOE genotype	Categorical
5	Braak Stage	Categorical
6	Clinical cognitive diagnosis summary at last visit	Categorical
7	Final clinical consensus diagnosis	Categorical

Chapter 5

Conclusion and Future Work

In the era of ”Big Data”, the large-scale dataset generated by high-throughput sequencing provides an opportunity to use more sophisticated modeling to explore the numerous disease/ phenotypic associations with the human microbiome and genome. In this dissertation, we present a collection of machine learning approaches for analyzing genetic and genomic data.

Chapter 2 describes a novel deep learning approach named MDeep, which is designed based on CNN, for performing either regression or binary classification. The novelty of MDeep lies in its ability to exploit the phylogenetic tree, which is an important prior on microbiome data. A comprehensive simulation study evaluates the performance of MDeep along with other competing methods, considering factors such as cluster size, signal density, and informativeness of the phylogenetic tree. As a result, MDeep favors scenarios with large clusters and high signal density. Overall, MDeep is superior to other methods when the tree is informative and still achieves a robust performance when the tree is uninformative. Experimental results demonstrate that, for both regression and binary classification, MDeep can achieve competitive performance, compared with classic CNN, Ph-CNN, which is another CNN-based method utilizing the phylogenetic information, and other state-of-the-art machine learning methods.

In Chapter 3, we propose a supervised learning framework by integrating the pre-computed scores from representative existing scoring methods, which will benefit from each individual method by automatically learning the relative contribution of each method and produce an ensemble score for the final prediction. The framework consists of two modes. The first “context-free” mode is trained using known causal variants from a wide range of contexts and is applicable to predict variants of unknown context. The second “context-dependent” mode further improves the prediction when the training and testing variants are from the same context. We evaluate the framework using both simulation and real datasets. The results demonstrate WEVar outperforms each individual method. Moreover, we show that the ensemble score successfully prioritizes experimentally validated non-coding variants.

In Chapter4, we develop a deep multimodal model to accurately predict locus-specific epigenetic signals (DNA methylation and histone modification) by taking DNA sequence and patient-level clinical outcomes. Given the predicted epigenetic signal for the reference and alternative alleles at a locus, we generate a functional score for the 210 million variants observed in previous sequencing projects. We demonstrate that DeepMFIVar can accurately predict locus-specific epigenetic signals using DNA sequence and clinical information, and DeepMFIVar is capable of prioritizing variants for downstream experiments

Bibliography

[1] CM Robinson and JK Pfeiffer. “Viruses and the microbiota”. In: *Annual review of virology* 1 (2014), pp. 55–69.

[2] MJ Li, Z Liu, P Wang, MP Wong, MR Nelson, JPA Kocher, M Yeager, PC Sham, SJ Chanock, Z Xia, et al. “GWASdb v2: an update database for human genetic variants identified by genome-wide association studies”. In: *Nucleic acids research* 44.D1 (2015), pp. D869–D876.

[3] C Melton, JA Reuter, DV Spacek, and M Snyder. “Recurrent somatic mutations in regulatory regions of human cancer genomes”. In: *Nature genetics* 47.7 (2015), p. 710.

[4] M Kellis, B Wold, MP Snyder, BE Bernstein, A Kundaje, GK Marinov, LD Ward, E Birney, GE Crawford, J Dekker, et al. “Defining functional DNA elements in the human genome”. In: *Proceedings of the National Academy of Sciences* 111.17 (2014), pp. 6131–6138.

[5] GR Ritchie, I Dunham, E Zeggini, and P Flicek. “Functional annotation of noncoding sequence variants”. In: *Nature methods* 11.3 (2014), p. 294.

[6] P Rentzsch, D Witten, GM Cooper, J Shendure, and M Kircher. “CADD: predicting the deleteriousness of variants throughout the human genome”. In: *Nucleic acids research* 47.D1 (2018), pp. D886–D894.

[7] HA Shihab, MF Rogers, J Gough, M Mort, DN Cooper, IN Day, TR Gaunt, and C Campbell. “An integrative approach to predicting the functional effects of non-coding and coding sequence variation”. In: *Bioinformatics* 31.10 (2015), pp. 1536–1543.

[8] I Ionita-Laza, K McCallum, B Xu, and JD Buxbaum. “A spectral approach integrating functional genomic annotations for coding and noncoding variants”. In: *Nature genetics* 48.2 (2016), p. 214.

[9] D Quang, Y Chen, and X Xie. “DANN: a deep learning approach for annotating the pathogenicity of genetic variants”. In: *Bioinformatics* 31.5 (2014), pp. 761–763.

[10] Y Fu, Z Liu, S Lou, J Bedford, XJ Mu, KY Yip, E Khurana, and M Gerstein. “FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer”. In: *Genome biology* 15.10 (2014), p. 480.

[11] YF Huang, B Gulko, and A Siepel. “Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data”. In: *Nature genetics* 49.4 (2017), p. 618.

[12] EB Hollister, C Gao, and J Versalovic. “Compositional and functional features of the gastrointestinal microbiome and their effects on human health”. In: *Gastroenterology* 146.6 (2014), pp. 1449–58.

[13] SR Gill, M Pop, RT DeBoy, PB Eckburg, PJ Turnbaugh, BS Samuel, JI Gordon, DA Relman, CM Fraser-Liggett, and KE Nelson. “Metagenomic analysis of the human distal gut microbiome”. In: *Science* 312.5778 (2006), pp. 1355–9.

[14] PJ Turnbaugh, RE Ley, MA Mahowald, V Magrini, ER Mardis, and JI Gordon. “An obesity-associated gut microbiome with increased capacity for energy harvest”. In: *Nature* 444.7122 (2006), pp. 1027–31.

[15] HL Cash, CV Whitham, CL Behrendt, and LV Hooper. “Symbiotic bacteria direct expression of an intestinal bactericidal lectin”. In: *Science* 313.5790 (2006), pp. 1126–30.

[16] LV Hooper, TS Stappenbeck, CV Hong, and JI Gordon. “Angiogenins: a new class of microbicidal proteins involved in innate immunity”. In: *Nat Immunol* 4.3 (2003), pp. 269–73.

[17] D Bouskra, C Brezillon, M Berard, C Werts, R Varona, IG Boneca, and G Eberl. “Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis”. In: *Nature* 456.7221 (2008), pp. 507–10.

[18] AJ Macpherson and NL Harris. “Interactions between commensal intestinal bacteria and the immune system”. In: *Nat Rev Immunol* 4.6 (2004), pp. 478–85.

[19] KJ Pflughoeft and J Versalovic. “Human microbiome in health and disease”. In: *Annu Rev Pathol* 7 (2012), pp. 99–122.

[20] J Kuczynski, CL Lauber, WA Walters, LW Parfrey, JC Clemente, D Gevers, and R Knight. “Experimental and analytical tools for studying the human microbiome”. In: *Nat Rev Genet* 13.1 (2011), pp. 47–58.

[21] JG Caporaso, J Kuczynski, J Stombaugh, K Bittinger, FD Bushman, EK Costello, N Fierer, AG Pena, JK Goodrich, JI Gordon, et al. “QIIME allows analysis of high-throughput community sequencing data”. In: *Nat Methods* 7.5 (2010), pp. 335–6.

[22] TK Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[23] R Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[24] CH Zhang. “Nearly unbiased variable selection under minimax concave penalty”. In: *The Annals of statistics* 38.2 (2010), pp. 894–942.

[25] H Zou and T Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

[26] D Knights, EK Costello, and R Knight. “Supervised classification of human microbiota”. In: *FEMS microbiology reviews* 35.2 (2011), pp. 343–359.

[27] A Statnikov, M Henaff, V Narendra, K Konganti, Z Li, L Yang, Z Pei, MJ Blaser, CF Aliferis, and AV Alekseyenko. “A comprehensive evaluation of multicategory classification methods for microbiomic data”. In: *Microbiome* 1.1 (2013), p. 11.

[28] E Pasolli, DT Truong, F Malik, L Waldron, and N Segata. “Machine learning meta-analysis of large metagenomic datasets: tools and biological insights”. In: *PLoS computational biology* 12.7 (2016), e1004977.

[29] G Zeller, J Tap, AY Voigt, S Sunagawa, JR Kultima, PI Costea, A Amiot, J Böhm, F Brunetti, N Habermann, et al. “Potential of fecal microbiota for early-stage detection of colorectal cancer”. In: *Mol Syst Biol* 10 (2014), p. 766.

[30] JU Scher, A Sczesnak, RS Longman, N Segata, C Ubeda, C Bielski, T Rostron, V Cerundolo, EG Pamer, SB Abramson, et al. “Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis”. In: *Elife* 2 (2013), e01202.

[31] C Manichanh, N Borruel, F Casellas, and F Guarner. “The gut microbiota in IBD”. In: *Nat Rev Gastroenterol Hepatol* 9.10 (2012), pp. 599–608.

[32] E Le Chatelier, T Nielsen, J Qin, E Prifti, F Hildebrand, G Falony, M Almeida, M Arumugam, JM Batto, S Kennedy, et al. “Richness of human gut microbiome correlates with metabolic markers”. In: *Nature* 500.7464 (2013), pp. 541–6.

[33] M Wang, C Tai, W E, and L Wei. “DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants”. In: *Nucleic Acids Res* 46.11 (2018), e69.

[34] B Alipanahi, A Delong, MT Weirauch, and BJ Frey. “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning”. In: *Nat Biotechnol* 33.8 (2015), pp. 831–8.

[35] F Luo, M Wang, Y Liu, XM Zhao, and A Li. “DeepPhos: prediction of protein phosphorylation sites with deep learning”. In: *Bioinformatics* (2019).

[36] C Yang, L Yang, M Zhou, H Xie, C Zhang, MD Wang, and H Zhu. "LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning". In: *Bioinformatics* 34.22 (2018), pp. 3825–3834.

[37] B Yang, F Liu, C Ren, Z Ouyang, Z Xie, X Bo, and W Shu. "BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone". In: *Bioinformatics* 33.13 (2017), pp. 1930–1936.

[38] R Singh, J Lanchantin, G Robins, and Y Qi. "DeepChrome: deep-learning for predicting gene expression from histone modifications". In: *Bioinformatics* 32.17 (2016), pp. i639–i648.

[39] C Angermueller, HJ Lee, W Reik, and O Stegle. "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning". In: *Genome Biol* 18.1 (2017), p. 67.

[40] J Zhou and O Troyanskaya. "Predicting effects of noncoding variants with deep learning-based sequence model". In: *Nat Methods* 12.10 (2015), pp. 931–4.

[41] D Fioravanti, Y Giarratano, V Maggio, C Agostinelli, M Chierici, G Jurman, and C Furlanello. "Phylogenetic convolutional neural networks in metagenomics". In: *BMC Bioinformatics* 19.Suppl 2 (2018), p. 49.

[42] E Martins and T Hansen. "Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data". In: *The American Naturalist* 149.4 (1997), pp. 646–667.

[43] DP Kingma and J Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[44] N Srivastava, G Hinton, A Krizhevsky, I Sutskever, and R Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[45] ES Charlson, J Chen, R Custers-Allen, K Bittinger, H Li, R Sinha, J Hwang, FD Bushman, and RG Collman. “Disordered microbial communities in the upper respiratory tract of cigarette smokers”. In: *PloS one* 5.12 (2010), e15216.

[46] GB Gloor, JM Macklaim, V Pawlowsky-Glahn, and JJ Egozcue. “Microbiome datasets are compositional: and this is not optional”. In: *Frontiers in microbiology* 8 (2017), p. 2224.

[47] S Hawinkel, F Mattiello, L Bijnens, and O Thas. “A broken promise: microbiome differential abundance methods do not control the false discovery rate”. In: *Briefings in bioinformatics* 20.1 (2019), pp. 210–221.

[48] A Cougoul, X Bailly, and EC Wit. “MAGMA: inference of sparse microbial association networks”. In: *BioRxiv* (2019), p. 538579.

[49] T Yatsunenko, FE Rey, MJ Manary, I Trehan, MG Dominguez-Bello, M Contreras, M Magris, G Hidalgo, RN Baldassano, AP Anokhin, et al. “Human gut microbiome viewed across age and geography”. In: *nature* 486.7402 (2012), p. 222.

[50] MI Smith, T Yatsunenko, MJ Manary, I Trehan, R Mkakosya, J Cheng, AL Kau, SS Rich, P Concannon, JC Mychaleckyj, et al. “Gut microbiomes of Malawian twin pairs discordant for kwashiorkor”. In: *Science* 339.6119 (2013), pp. 548–54.

[51] JU Scher, A Sczesnak, RS Longman, N Segata, C Ubeda, C Bielski, T Rostron, V Cerundolo, EG Pamer, SB Abramson, et al. “Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis”. In: *elife* 2 (2013), e01202.

[52] A Gonzalez, JA Navas-Molina, T Kosciolek, D McDonald, Y Vázquez-Baeza, G Ackermann, J DeReus, S Janssen, AD Swafford, SB Orchanian, et al. “Qiita: rapid, web-enabled microbiome meta-analysis”. In: *Nat Methods* 15.10 (2018), pp. 796–798.

[53] MN Price, PS Dehal, and AP Arkin. “FastTree: computing large minimum evolution trees with profiles instead of a distance matrix”. In: *Mol Biol Evol* 26.7 (2009), pp. 1641–50.

[54] J Xiao, L Chen, S Johnson, Y Yu, X Zhang, and J Chen. “Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model”. In: *Frontiers in microbiology* 9 (2018), p. 1391.

[55] J Xiao, L Chen, S Johnson, Y Yu, X Zhang, and J Chen. “Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model”. In: *Frontiers in microbiology* 9 (2018), p. 1391.

[56] T Odamaki, K Kato, H Sugahara, N Hashikura, S Takahashi, Jz Xiao, F Abe, and R Osawa. “Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study”. In: *BMC Microbiol* 16 (2016), p. 90.

[57] JA Santos-Marcos, C Haro, A Vega-Rojas, JF Alcala-Diaz, H Molina-Abril, A Leon-Acuña, J Lopez-Moreno, BB Landa, M Tena-Sempere, P Perez-Martinez, et al. “Sex Differences in the Gut Microbiota as Potential Determinants of Gender Predisposition to Disease”. In: *Mol Nutr Food Res* 63.7 (2019), e1800870.

[58] F Bertels, OK Silander, M Pachkov, PB Rainey, and E van Nimwegen. “Automated reconstruction of whole-genome phylogenies from short-sequence reads”. In: *Mol Biol Evol* 31.5 (2014), pp. 1077–88.

[59] H Fan, AR Ives, Y Surget-Groba, and CH Cannon. “An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data”. In: *BMC Genomics* 16 (2015), p. 522.

[60] B Ng et al. “An xQTL map integrates the genetic architecture of the human brain’s transcriptome and epigenome”. In: *Nat Neurosci* 20.10 (2017), pp. 1418–1426.

[61] JK Pickrell, JC Marioni, AA Pai, JF Degner, BE Engelhardt, E Nkadori, JB Veyrieras, M Stephens, Y Gilad, and JK Pritchard. “Understanding mechanisms underlying human gene expression variation with RNA sequencing”. In: *Nature* 464.7289 (2010), pp. 768–72.

[62] JR Gibbs, MP van der Brug, DG Hernandez, BJ Traynor, MA Nalls, SL Lai, S Arepalli, A Dillman, IP Rafferty, J Troncoso, et al. "Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain". In: *PLoS Genet* 6.5 (2010), e1000952.

[63] EB Josephs, YW Lee, JR Stinchcombe, and SI Wright. "Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression". In: *Proc Natl Acad Sci U S A* 112.50 (2015), pp. 15390–5.

[64] JF Degner, AA Pai, R Pique-Regi, JB Veyrieras, DJ Gaffney, JK Pickrell, S De Leon, K Michelini, N Lewellen, GE Crawford, et al. "DNase I sensitivity QTLs are a major determinant of human expression variation". In: *Nature* 482.7385 (2012), pp. 390–4.

[65] A Takata, N Matsumoto, and T Kato. "Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci". In: *Nat Commun* 8 (2017), p. 14519.

[66] PJ Killela, ZJ Reitman, Y Jiao, C Bettegowda, N Agrawal, J Diaz L. A., AH Friedman, H Friedman, GL Gallia, BC Giovanella, et al. "TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal". In: *Proc Natl Acad Sci U S A* 110.15 (2013), pp. 6021–6.

[67] MR Mansour, BJ Abraham, L Anders, A Berezovskaya, A Gutierrez, AD Durbin, J Etchin, L Lawton, SE Sallan, LB Silverman, et al. "Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element". In: *Science* 346.6215 (2014), pp. 1373–7.

[68] AM Saunders, WJ Strittmatter, D Schmechel, PH George-Hyslop, MA Pericak-Vance, SH Joo, BL Rosi, JF Gusella, DR Crapper-MacLachlan, MJ Alberts, et al. "Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease". In: *Neurology* 43.8 (1993), pp. 1467–72.

[69] WJ Strittmatter, AM Saunders, D Schmechel, M Pericak-Vance, J Enghild, GS Salvesen, and AD Roses. “Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease”. In: *Proc Natl Acad Sci U S A* 90.5 (1993), pp. 1977–81.

[70] JC Bis, X Jian, BW Kunkle, Y Chen, KL Hamilton-Nelson, WS Bush, WJ Salerno, D Lancour, Y Ma, AE Renton, et al. “Whole exome sequencing study identifies novel rare and common Alzheimer’s-Associated variants involved in immune response and transcriptional regulation”. In: *Mol Psychiatry* (2018).

[71] JC Lambert, CA Ibrahim-Verbaas, D Harold, AC Naj, R Sims, C Bellenguez, AL DeStafano, JC Bis, GW Beecham, B Grenier-Boley, et al. “Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease”. In: *Nat Genet* 45.12 (2013), pp. 1452–8.

[72] L Chen, P Jin, and ZS Qin. “DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles”. In: *Genome Biol* 17.1 (2016), p. 252.

[73] EP Consortium. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), pp. 57–74.

[74] C Roadmap Epigenomics, A Kundaje, W Meuleman, J Ernst, M Bilenky, A Yen, A Heravi-Moussavi, P Kheradpour, Z Zhang, J Wang, et al. “Integrative analysis of 111 reference human epigenomes”. In: *Nature* 518.7539 (2015), pp. 317–30.

[75] HG Stunnenberg, C International Human Epigenome, and M Hirst. “The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery”. In: *Cell* 167.7 (2016), p. 1897.

[76] L Chen and ZS Qin. “Using DIVAN to assess disease/trait-associated single nucleotide variants in genome-wide scale”. In: *BMC Res Notes* 10.1 (2017), p. 530.

[77] L Chen, Y Wang, B Yao, A Mitra, X Wang, and X Qin. “TIVAN: tissue-specific cis-eQTL single nucleotide variant annotation and prediction”. In: *Bioinformatics* 35.9 (2019), pp. 1573–1575.

[78] 1GP Consortium et al. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (2012), p. 56.

[79] L Liu, MD Sanderford, R Patel, P Chandrashekar, G Gibson, and S Kumar. “Biological relevance of computationally predicted pathogenicity of noncoding variants”. In: *Nat Commun* 10.1 (2019), p. 330.

[80] PD Stenson, M Mort, EV Ball, K Shaw, AD Phillips, and DN Cooper. “The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine”. In: *Human genetics* 133.1 (2014), pp. 1–9.

[81] MJ Landrum, JM Lee, GR Riley, W Jang, WS Rubinstein, DM Church, and DR Maglott. “ClinVar: public archive of relationships among sequence variation and human phenotype”. In: *Nucleic acids research* 42.D1 (2014), pp. D980–D985.

[82] OL Griffith, SB Montgomery, B Bernier, B Chu, K Kasaian, S Aerts, S Mahony, MC Sleumer, M Bilenky, M Haeussler, et al. “ORegAnno: an open-access community-driven resource for regulatory annotation”. In: *Nucleic acids research* 36.suppl_1 (2007), pp. D107–D113.

[83] KKH Farh, A Marson, J Zhu, M Kleinewietfeld, WJ Housley, S Beik, N Shoresh, H Whitton, RJ Ryan, AA Shishkin, et al. “Genetic and epigenetic fine mapping of causal autoimmune disease variants”. In: *Nature* 518.7539 (2015), p. 337.

[84] MJ Li, Z Pan, Z Liu, J Wu, P Wang, Y Zhu, F Xu, Z Xia, PC Sham, JPA Kocher, et al. “Predicting regulatory variants with composite statistic”. In: *Bioinformatics* 32.18 (2016), pp. 2729–2736.

[85] CD Brown, LM Mangravite, and BE Engelhardt. "Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs". In: *PLoS genetics* 9.8 (2013).

[86] MT Maurano, E Haugen, R Sandstrom, J Vierstra, A Shafer, R Kaul, and JA Stamatoyannopoulos. "Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo". In: *Nature genetics* 47.12 (2015), p. 1393.

[87] A Buniello, JAL MacArthur, M Cerezo, LW Harris, J Hayhurst, C Malangone, A McMahon, J Morales, E Mountjoy, E Sollis, et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Res* 47.D1 (2019), pp. D1005–D1012.

[88] R Tewhey, D Kotliar, DS Park, B Liu, S Winnicki, SK Reilly, KG Andersen, TS Mikkelsen, ES Lander, SF Schaffner, et al. "Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay". In: *Cell* 165.6 (2016), pp. 1519–1529.

[89] P Kheradpour, J Ernst, A Melnikov, P Rogov, L Wang, X Zhang, J Alston, TS Mikkelsen, and M Kellis. "Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay". In: *Genome research* 23.5 (2013), pp. 800–811.

[90] E Jones, T Oliphant, P Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001–.

[91] S Diamond and S Boyd. "CVXPY: A Python-Embedded Modeling Language for Convex Optimization". In: *Journal of Machine Learning Research* 17.83 (2016), pp. 1–5.

[92] Z He, L Liu, K Wang, and I Ionita-Laza. "A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs". In: *Nature communications* 9.1 (2018), pp. 1–12.

[93] A Kreimer, H Zeng, MD Edwards, Y Guo, K Tian, S Shin, R Welch, M Wainberg, R Mohan, NA Sinnott-Armstrong, et al. “Predicting gene expression in massively parallel reporter assays: A comparative study”. In: *Hum Mutat* 38.9 (2017), pp. 1240–1250.

[94] B Moreno-Kustner, C Martin, and L Pastor. “Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses”. In: *PLoS One* 13.4 (2018), e0195687.

[95] S Saha, D Chant, J Welham, and J McGrath. “A systematic review of the prevalence of schizophrenia”. In: *PLoS Med* 2.5 (2005), e141.

[96] GBD Disease, I Injury, and C Prevalence. “Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016”. In: *Lancet* 390.10100 (2017), pp. 1211–1259.

[97] L Myint, R Wang, L Boukas, KD Hansen, LA Goff, and D Avramopoulos. “A screen of 1,049 schizophrenia and 30 Alzheimer’s-associated variants for regulatory potential”. In: *Am J Med Genet B Neuropsychiatr Genet* 183.1 (2020), pp. 61–73.

[98] M van den Boogaard et al. “A common genetic variant within SCN10A modulates cardiac SCN5A expression”. In: *J Clin Invest* 124.4 (2014), pp. 1844–52.

[99] S Liu, Y Liu, Q Zhang, J Wu, J Liang, S Yu, GH Wei, KP White, and X Wang. “Systematic identification of regulatory variants associated with cancer risk”. In: *Genome Biol* 18.1 (2017), p. 194.

[100] EJ Leslie, MA Taub, H Liu, KM Steinberg, DC Koboldt, Q Zhang, JC Carlson, JB Hetmanski, H Wang, DE Larson, et al. “Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci”. In: *Am J Hum Genet* 96.3 (2015), pp. 397–411.

[101] CA Guenther, B Tasic, L Luo, MA Bedell, and DM Kingsley. “A molecular basis for classic blond hair color in Europeans”. In: *Nat Genet* 46.7 (2014), pp. 748–52.

[102] PM Visscher, NR Wray, Q Zhang, P Sklar, MI McCarthy, MA Brown, and J Yang. “10 years of GWAS discovery: biology, function, and translation”. In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[103] SL Spain and JC Barrett. “Strategies for fine-mapping complex traits”. In: *Human molecular genetics* 24.R1 (2015), R111–R119.

[104] LA Hindorff, P Sethupathy, HA Junkins, EM Ramos, JP Mehta, FS Collins, and TA Manolio. “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits”. In: *Proceedings of the National Academy of Sciences* 106.23 (2009), pp. 9362–9367.

[105] JK Pritchard and M Przeworski. “Linkage disequilibrium in humans: models and data”. In: *The American Journal of Human Genetics* 69.1 (2001), pp. 1–14.

[106] M Claussnitzer, SN Dankel, KH Kim, G Quon, W Meuleman, C Haugen, V Glunk, IS Sousa, JL Beaudry, V Puviindran, et al. “FTO obesity variant circuitry and adipocyte browning in humans”. In: *New England Journal of Medicine* 373.10 (2015), pp. 895–907.

[107] G Kichaev, WY Yang, S Lindstrom, F Hormozdiari, E Eskin, AL Price, P Kraft, and B Pasaniuc. “Integrating functional data to prioritize causal variants in statistical fine-mapping studies”. In: *PLoS Genet* 10.10 (2014), e1004722.

[108] JK Pickrell. “Joint analysis of functional genomic data and genome-wide association studies of 18 human traits”. In: *The American Journal of Human Genetics* 94.4 (2014), pp. 559–573.

[109] HK Finucane, B Bulik-Sullivan, A Gusev, G Trynka, Y Reshef, PR Loh, V Anttila, H Xu, C Zang, K Farh, et al. “Partitioning heritability by functional annotation

using genome-wide association summary statistics". In: *Nature genetics* 47.11 (2015), p. 1228.

[110] PL De Jager, Y Ma, C McCabe, J Xu, BN Vardarajan, D Felsky, HU Klein, CC White, MA Peters, B Lodgson, et al. "A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research". In: *Scientific data* 5 (2018), p. 180142.

[111] S Hochreiter and J Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[112] FA Gers, J Schmidhuber, and F Cummins. "Learning to forget: Continual prediction with LSTM". In: (1999).

[113] A Paszke, S Gross, S Chintala, G Chanan, E Yang, Z DeVito, Z Lin, A Desmaison, L Antiga, and A Lerer. "Automatic differentiation in PyTorch". In: (2017).

[114] H Zeng and DK Gifford. "Predicting the impact of non-coding variants on DNA methylation". In: *Nucleic acids research* 45.11 (2017), e99–e99.