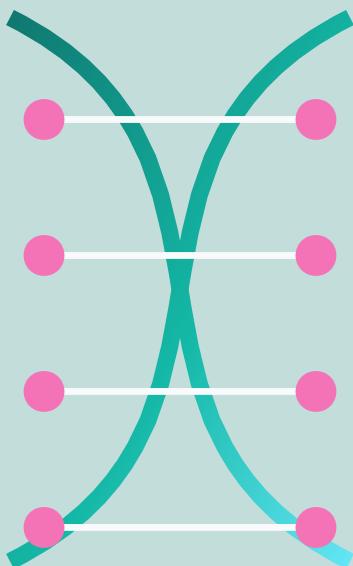


# Classifying Major Depressive Disorder with Exon Sequence Embeddings from DNA Foundation Models



<h3>Classifying Major Depressive Disorder with Exon Sequence Embeddings from DNA Foundation Models · Concept Sketch</h3>  
<p>Genome-scale signal aggregation framing PRS vs. foundation model granularity.</p>

---

# Classifying Major Depressive Disorder with Exon Sequence Embeddings from DNA Foundation Models

**Authors:** Heesun Yoon, Eunji Lee, Heehwan Wang, Jiook Cha, Xin Dai, Shinjae Yoo, Yoonjung Yoonie Joo

**Year:** 2025

**Venue:** BIOKDD '25 (24th International Workshop on Data Mining in Bioinformatics)

---

## 1. Classification

- **Domain Category:**

- **Genomics FM + Application.** This paper applies pretrained DNA foundation models (Caduceus, DNABERT-2) to psychiatric genomics, specifically using exon sequence embeddings to classify Major Depressive Disorder (MDD).

- **FM Usage Type:**

- **Application of existing FMs.** The study leverages two pretrained genomic foundation models—Caduceus (Mamba-based) and DNABERT-2 (Transformer-based)—to generate DNA sequence embeddings, then trains classical ML models on top of these embeddings for MDD classification.

- **Key Modalities:**

- **DNA sequences:** Whole exome sequencing (WES) data from 38 MDD-associated genes
  - **Demographic/clinical features:** Age, sex, household income, genetic principal components
  - **Derived representations:** High-dimensional exon sequence embeddings from foundation models
- 

## 2. Executive Summary

This study introduces a novel framework for classifying Major Depressive Disorder (MDD) using raw DNA sequence embeddings extracted from the exonic regions of 38 MDD-associated genes. Rather than relying on traditional polygenic risk scores (PRS) that aggregate single nucleotide polymorphism (SNP) effects and typically explain only 2-3% of MDD variance, the authors leverage pretrained genomic foundation models—Caduceus (based on the Mamba architecture) and DNABERT-2 (Transformer-based)—to generate context-rich, high-dimensional representations of exon sequences from whole exome sequencing (WES) data. Using UK Biobank data with 10,307 MDD cases and 10,307 healthy controls, they systematically evaluate 15 machine learning pipelines combining three embedding aggregation strategies (PCA, max pooling, mean pooling) with five classifiers (logistic regression, random forest, CatBoost, MLP, 1D-CNN). The best-performing pipeline—CatBoost with max pooling—achieves a mean AUC of 0.851 and AUPRC of 0.812, representing a ~49-60% improvement over traditional PRS approaches. A leave-one-gene-out

(LOGO) analysis reveals that the SOD2 gene contributes most significantly to classification performance despite lacking established GWAS hits in exonic regions, highlighting the model's ability to capture biological signals beyond SNP-level associations. Replication with DNABERT-2 embeddings yields identical top performance (AUC 0.851), confirming the framework's generalizability across different foundation model architectures. This work demonstrates that pretrained sequence models can effectively encode complex genomic context for psychiatric risk prediction, opening new avenues for applying foundation models to mental health genomics.

---

### 3. Problem Setup and Motivation

#### Scientific / practical problem:

- **Predicting MDD risk from genetic data:**
  - Major Depressive Disorder affects ~300 million people worldwide with substantial economic burden (~\$6 trillion projected by 2030)
  - Individual risk prediction is challenging due to extreme polygenicity and high phenotypic heterogeneity
- **Limitations of existing genetic approaches:**
  - **Polygenic Risk Scores (PRS):** Aggregate SNP-level GWAS effects but explain only ~2-3% of MDD phenotypic variance (AUC ~0.53-0.57)
  - **Variant-level focus:** Cannot capture broader sequence-level context, regulatory motifs, structural variants, or long-range dependencies
  - **Exonic regions understudied:** GWAS emphasizes non-coding regions, but functionally important protein-coding sequences may harbor disease-relevant information missed by SNP-based methods
- **Opportunity with DNA foundation models:**
  - Models like DNABERT, DNABERT-2, Caduceus, and Nucleotide Transformer pretrained on large genomic corpora can generate embeddings that capture sophisticated context beyond single variants
  - These models have shown success in regulatory element classification and gene expression prediction but remain unexplored in psychiatric genomics

#### Why this is hard:

- **Polygenicity and small effects:**
  - MDD involves thousands of loci with tiny individual contributions; signal-to-noise ratio is low
- **Data modality mismatch:**
  - Whole exome sequencing (WES) captures exons cost-effectively but misses intronic/intergenic regulatory elements where many GWAS hits lie
- **Embedding aggregation:**
  - 38 genes produce 38 separate high-dimensional embeddings; need effective strategies to combine them without losing gene-specific signals or introducing noise
- **Demographic confounding:**
  - MDD cases differ from controls in age, sex, and socioeconomic status; must control these covariates
- **Model selection and overfitting:**
  - Many hyperparameters and modeling choices; need rigorous nested cross-validation to avoid overoptimistic results

- **Interpretability:**
    - Which genes or exons drive predictions? Black-box embeddings require LOGO-style analyses to identify key contributors
- 

## 4. Data and Modalities

### Dataset:

- **UK Biobank:**
  - 10,307 MDD cases and 10,307 healthy controls (downsampled for balance)
  - Whole Exome Sequencing (WES) data (Exome OQFE variant call files, interim 200k release)
  - Demographic and clinical features: age at recruitment, sex, household income, top 10 genetic principal components
- **MDD case definition:**
  - Participants meeting at least one of three depression criteria (Howard et al. 2018):
    - Broad depression
    - Probable MDD
    - ICD-coded MDD
  - Cases: Single/recurrent probable major depression episodes
- **Healthy control definition:**
  - No bipolar disorder or major depression
  - No history of psychiatric care on admission
  - Never sought professional help for mental distress

### Modalities:

- **Genomic sequences:**
  - FASTA sequences of exonic regions from 38 autosomal MDD-associated genes (genes reported  $\geq 2$  times in prior literature)
  - Genome Reference Consortium Human Build 38 (GRCh38)
  - Individual-level sequences with reference alleles replaced by alternate alleles where applicable
- **Demographics:**
  - Age (mean: cases  $55.4 \pm 8.0$  years, controls  $57.1 \pm 7.8$  years; significant difference  $p < 0.001$ )
  - Sex (cases 64.1% female, controls 46.9% female;  $p < 0.001$ )
  - Household income (controls more likely in higher income brackets;  $p < 0.001$ )
  - Genetic PCs (top 10 PCs to control population structure)

### Preprocessing / representation:

- **Sequence embedding generation:**
  - Primary: **Caduceus-PS** (256 hidden dimensions, 16 MambaDNA layers)
    - Bi-directional Mamba with reverse-complement (RC) equivariance
    - Hidden states averaged across forward and RC representations to obtain RC-invariant embeddings ( $512 \rightarrow 256$  dimensions)
  - Replication: **DNABERT-2** (Transformer-based genomic foundation model)
  - Each of 38 genes produces one 256-dimensional embedding per sample

- **Embedding aggregation (three strategies tested):**
    - **PCA:** Feature scaling + PCA to 256 principal components
    - **Max pooling:** Element-wise maximum across all 38 gene embeddings
    - **Mean pooling:** Element-wise mean across all 38 gene embeddings
  - **Final input:**
    - Aggregated gene embeddings (256 dimensions) concatenated with 14 demographic features (age, sex, income, 10 genetic PCs)
- 

## 5. Model / Foundation Model

### Foundation Models Used:

Model	Architecture	Key Features	Purpose
<b>Caduceus-PS</b>	Mamba (structured state space model)	Bi-directional, RC-equivariant, linear complexity, long context (~131k)	Primary embedding generator
<b>DNABERT-2</b>	Transformer (BERT-like)	Bidirectional attention, multi-species genome pretraining	Replication/validation of framework generalizability

### Caduceus Details:

- **Architecture:** BiMamba + MambaDNA layers with parameter sharing for RC equivariance
- **Advantages:**
  - Linear computational complexity (vs quadratic for Transformers) → handles long sequences efficiently
  - Bi-directionality captures upstream and downstream context
  - RC equivariance ensures equivalent representations for forward and reverse-complement strands
- **Configuration:** 256 hidden dimensions, 16 layers, ~131k sequence length capacity
- **Output:** Per-sequence hidden states split in half and averaged to yield RC-invariant 256-d embeddings

### Downstream Classification Models (15 pipelines):

Five classifiers × three aggregation strategies:

1. **Logistic Regression (LR):** Linear baseline
2. **Random Forest (RF):** Ensemble of decision trees
3. **CatBoost (CB):** Gradient boosting with categorical feature support
4. **Multilayer Perceptron (MLP):** Feedforward neural network
5. **1D Convolutional Neural Network (1D-CNN):** Convolutional layers for pattern detection

## Training Setup:

- **Nested cross-validation:**
    - **Outer loop:** 10-fold stratified CV for unbiased model assessment
    - **Inner loop:** 3-fold stratified CV for hyperparameter tuning
  - **Hyperparameter optimization:**
    - **Framework:** Optuna with Bayesian optimization
    - **Pruning strategies:** HyperbandPruner for neural nets, MedianPruner for others
    - **Trials:** 150 for MLP, 100 for 1D-CNN, 300 for LR/RF/CB
  - **Early stopping:**
    - CatBoost: 10 consecutive rounds without AUC improvement
    - MLP/1D-CNN: 200 epochs with binary cross-entropy loss monitoring
    - 10% holdout set for validation
  - **Evaluation metrics:**
    - Primary: **AUC** (threshold-independent, consistent)
    - Secondary: **AUPRC** (for tied AUC values), precision, recall, F1-score
- 

## 6. Multimodal / Integration Aspects (If Applicable)

This work is **primarily unimodal (genomics)** at the embedding level but incorporates **late fusion** of genetic embeddings with demographic/clinical features.

### Modalities integrated:

- **Genomic:** DNA sequence embeddings from foundation models (captures exonic sequence context)
- **Demographic/clinical:** Age, sex, household income, genetic PCs

### How they are integrated:

- **Late fusion via feature concatenation:**
  - Gene embeddings aggregated into a 256-d vector
  - Concatenated with 14 demographic features
  - Combined feature vector fed to downstream classifiers
  - This is analogous to **late fusion** in the integration baseline plan: modality-specific representations (gene sequences → embeddings, demographics as raw features) are combined at the feature level before final prediction

### Why this integration is useful:

- **Demographics as confounders:** Age, sex, and income differ significantly between MDD cases and controls; including them improves model accuracy and reduces spurious associations
- **Genetic PCs control population structure:** Prevent inflation due to ancestry-related confounding
- **Complementary information:** Genetic embeddings capture biological predisposition; demographics provide environmental/social context

## **Relation to the integration baseline plan:**

- **Late fusion approach:**
    - Aligns with the plan’s preference for preserving modality-specific signals before integration
    - Genetic embeddings are computed independently (via pretrained FMs), then concatenated with demographics
  - **Robustness and evaluation:**
    - Nested CV with proper train/test splits mirrors the plan’s emphasis on evaluation discipline
    - Leave-one-gene-out analysis with Wilcoxon signed-rank test + FDR correction aligns with plan’s recommendation for attribution via LOGO and statistical rigor
  - **Genetics embedding hygiene:**
    - Use of RC-equivariant Caduceus and RC-averaging reflects the plan’s citation of Caduceus for proper genetics handling
    - Deterministic embedding generation ensures reproducibility
  - **Future escalation:**
    - Current work uses simple concatenation; could extend to two-tower contrastive learning (genetic encoder vs demographic encoder) or attention-based fusion if performance plateaus
- 

## **7. Experiments and Results**

### **Tasks / benchmarks:**

- **Binary classification:** MDD cases (n=10,307) vs healthy controls (n=10,307) in UK Biobank
- **Primary evaluation:** 15 modeling pipelines (5 classifiers  $\times$  3 aggregation strategies)
- **Replication:** Same 15 pipelines with DNABERT-2 embeddings
- **Interpretability:** Leave-one-gene-out (LOGO) analysis to rank gene contributions

### **Baselines:**

- **Traditional PRS:** Historical MDD PRS models explain ~2-3% variance, AUC ~0.53-0.57
- **Within-study baselines:** All 15 pipeline combinations serve as internal comparisons

### **Key findings:**

#### **1. Primary Caduceus Results (Table 2):**

- **Best performance:** CatBoost + max pooling
  - Mean AUC: **0.851  $\pm$  0.009**
  - Mean AUPRC: **0.812  $\pm$  0.013**
- **Tied second-best:** Logistic Regression + max pooling
  - Mean AUC: **0.851  $\pm$  0.010**
  - Mean AUPRC: **0.809  $\pm$  0.015** (slightly lower)
- **Third:** 1D-CNN + max pooling
  - Mean AUC: **0.849  $\pm$  0.010**

- **Aggregation strategy impact:**
  - **Max pooling:** Consistently best across all classifiers
  - **PCA:** Intermediate performance
  - **Mean pooling:** Worst performance across all classifiers (e.g., 1D-CNN dropped to AUC  $0.638 \pm 0.010$ )
- **Performance gain over PRS:**
  - ~49-60% improvement in AUC relative to traditional PRS (0.851 vs 0.53-0.57)

## 2. DNABERT-2 Replication:

- **Best performance:** CatBoost + max pooling
  - Mean AUC: **0.851** (identical to Caduceus)
  - Confirms framework generalizability across different FM architectures (Mamba vs Transformer)
- **Aggregation differences:**
  - Mean pooling performed comparably to max pooling with DNABERT-2 (unlike Caduceus)
  - Suggests optimal aggregation may be FM-specific

## 3. LOGO Analysis (Figure 3):

- **SOD2 gene:**
  - **Highest contribution:** Median  $\Delta\text{AUC} = 0.190$  (IQR: 0.185-0.196)
  - **Only statistically significant gene** after FDR correction (Wilcoxon  $W=0$ ,  $p_{\text{FDR}}=0.038$ )
  - Removing SOD2 drastically reduces classification performance despite no established GWAS hits in its exonic regions
- **SOD2 biological relevance:**
  - Encodes superoxide dismutase 2 (mitochondrial enzyme)
  - Involved in oxidative stress regulation
  - Oxidative stress implicated in MDD pathophysiology
- **Demographics ablation:**
  - Removing demographic features:  $\Delta\text{AUC} = 0.040$  (0.037-0.045), significant ( $p_{\text{FDR}}=0.038$ )
  - Confirms demographics contribute meaningfully but less than top genes

## 4. Cohort Demographics:

- **Age:** Cases younger than controls (55.4 vs 57.1 years,  $p<0.001$ )
  - **Sex:** Cases more female (64.1% vs 46.9%,  $p<0.001$ )
  - **Income:** Controls more likely in higher income brackets ( $p<0.001$ )
  - All differences significant → justifies including as covariates
- 

## 8. Strengths, Limitations, and Open Questions

### Strengths:

- **Substantial performance improvement:** 49-60% AUC gain over traditional PRS demonstrates value of sequence-level context
- **Novel application domain:** First study to apply DNA foundation model embeddings to psychiatric disorder classification

- **Rigorous methodology:**
  - Nested cross-validation prevents overfitting
  - Multiple FM architectures tested (Caduceus + DNABERT-2) → validates generalizability
  - Statistical testing (Wilcoxon + FDR correction) for LOGO analysis
- **Biological interpretability:** LOGO analysis identifies SOD2 as key gene, linking to known oxidative stress mechanisms in MDD
- **Foundation model comparison:** Shows Mamba and Transformer architectures yield similar top performance, suggesting robustness
- **Embedding aggregation insights:** Max pooling consistently superior for Caduceus; strategy choice matters

### **Limitations:**

- **Exon-only analysis:**
  - Focus on protein-coding regions despite most GWAS hits lying in non-coding regulatory elements
  - Intronic and intergenic regions (much larger proportion of genome) not included
- **Gene-level interpretation only:**
  - LOGO analysis identifies contributing genes but not specific exons or base-pair regions
  - Finer-grained attribution (e.g., attention maps, saliency) not explored
- **Single-ancestry cohort:**
  - UK Biobank participants primarily of European ancestry
  - Transferability to other populations uncertain
- **No external validation:**
  - Framework not tested on independent cohorts outside UK Biobank
  - Generalization across datasets unknown
- **Demographic differences:**
  - Cases and controls differ significantly in age, sex, income → potential residual confounding even after covariate adjustment
- **Limited gene set:**
  - Only 38 pre-selected MDD-associated genes; genome-wide exome analysis could yield further insights

### **Open Questions and Future Directions:**

- **Expand genomic coverage:**
  - Include intronic and intergenic regions to capture regulatory elements where many GWAS hits reside
  - Test framework on whole-genome sequencing (WGS) data
- **Multimodal integration:**
  - Combine DNA embeddings with transcriptomics (RNA-seq), epigenomics (methylation), proteomics
  - Integrate with brain imaging (fMRI, sMRI) for brain-genomics multimodal models
- **Fine-grained interpretability:**
  - Exon-level or base-pair-level attribution via attention weights, integrated gradients, or saliency maps
  - Link specific sequence motifs to MDD risk
- **External validation:**
  - Test on independent cohorts (e.g., other biobanks, clinical samples)
  - Evaluate cross-ancestry performance and fairness

- **Generative and design applications:**
    - Can FMs be used to design therapeutic sequences or identify novel targets?
    - Generate synthetic exon variants and predict MDD risk changes
  - **Integration with PRS:**
    - Combine exon embeddings with genome-wide PRS to capture both local and global genetic risk
    - Test whether the two signals are complementary or redundant
- 

## 9. Context and Broader Impact

### Position in the genomics FM landscape:

- This work sits at the intersection of **genomic foundation models** and **psychiatric genomics**
- **Genomics FMs** (DNABERT, Caduceus, Nucleotide Transformer, Evo) have primarily been applied to:
  - Regulatory element classification (promoters, enhancers)
  - Variant effect prediction (pathogenicity)
  - Gene expression forecasting
- **Psychiatric genomics** has relied on:
  - GWAS to identify risk loci
  - PRS for individual-level risk aggregation
  - Traditional ML (SVM, random forests) with handcrafted genomic features

### Relation to well-known ideas:

- **Analogy:** “Like using BERT embeddings for text classification, but for DNA sequences predicting psychiatric risk”
- **Transfer learning paradigm:** Pretrain large models on diverse genomic data, then fine-tune (or use embeddings) for specific downstream tasks
- **Contrastive to PRS:** PRS are linear sums of SNP effects (shallow, interpretable); FM embeddings are high-dimensional, context-rich representations (deep, powerful but less transparent)

### Connection to the integration baseline plan:

- **Genetics embedding hygiene:**
  - Paper explicitly cites Caduceus for RC-equivariance, aligning with the plan’s recommendation for proper DNA handling
  - RC-averaging and deterministic tokenization ensure stable embeddings
- **Late fusion baseline:**
  - Concatenating gene embeddings with demographics mirrors the plan’s preference for late fusion under heterogeneous semantics
- **Evaluation rigor:**
  - Nested CV, LOGO with Wilcoxon + FDR, proper train/test splits align with plan’s robustness discipline
- **Future escalation:**
  - If exon embeddings prove valuable, next steps could include:
    - Two-tower contrastive learning (genetic encoder vs demographic/clinical encoder)
    - Attention-based fusion across genes (hub tokens à la TAPE)

- Integration with brain FMs (BrainLM, Brain-JEPA) for brain-genomics multimodal models

### Why this paper is a useful reference:

- **First application:** Demonstrates feasibility and value of applying DNA FMs to psychiatric disorders
  - **Methodological template:** Provides reusable pipeline for embedding extraction → aggregation → ML classification
  - **Performance benchmark:** Establishes AUC 0.851 as a target for future MDD genetic prediction studies
  - **Gene discovery:** LOGO analysis identifies SOD2 despite absence of exonic GWAS hits, suggesting FM embeddings capture signals missed by SNP-level approaches
  - **Foundation for multimodal work:** Genetic embeddings from this study could serve as one modality in larger brain-genomics integration projects
- 

## 10. Key Takeaways (Bullet Summary)

### Problem:

- Traditional polygenic risk scores (PRS) for MDD explain only 2-3% of variance (AUC ~0.53-0.57), limiting predictive utility
- PRS rely on SNP-level GWAS associations and miss broader sequence context, regulatory motifs, and long-range dependencies
- DNA foundation models (Caduceus, DNABERT-2) trained on large genomic corpora remain unexplored in psychiatric genomics

### Method / Model:

- **Framework:** Extract exon sequence embeddings from 38 MDD-associated genes using pretrained DNA foundation models
- **Primary FM:** Caduceus-PS (Mamba-based, bi-directional, RC-equivariant, 256-d embeddings per gene)
- **Aggregation:** Test PCA, max pooling, and mean pooling to combine 38 gene embeddings
- **Downstream:** Train 15 ML pipelines (5 classifiers × 3 aggregation strategies) with nested cross-validation
- **Data:** UK Biobank WES data, 10,307 MDD cases + 10,307 controls, adjusted for demographics and genetic PCs
- **Replication:** Validate framework generalizability with DNABERT-2 embeddings

### Results:

- **Best performance:** CatBoost + max pooling achieves **AUC  $0.851 \pm 0.009$ , AUPRC  $0.812 \pm 0.013$**
- **Improvement over PRS:** ~49-60% AUC gain (0.851 vs 0.53-0.57)
- **Replication:** DNABERT-2 achieves identical top AUC (0.851), confirming robustness across FM architectures
- **Key gene:** LOGO analysis identifies **SOD2** as most significant contributor ( $\Delta\text{AUC } 0.190$ ), linking to oxidative stress pathways

- **Aggregation matters:** Max pooling consistently best for Caduceus; mean pooling severely degrades performance

### Why it matters:

- **First psychiatric genomics application:** Demonstrates DNA foundation models can effectively predict complex psychiatric disorders
  - **Paradigm shift:** Moves beyond SNP-level associations to sequence-level context, capturing signals missed by GWAS
  - **Substantial performance gain:** Large improvement over traditional PRS suggests promise for clinical risk prediction
  - **Methodological contribution:** Provides reusable framework for applying genomic FMs to any polygenic disorder
  - **Foundation for integration:** Exon embeddings could be combined with brain imaging, transcriptomics, or epigenomics in future multimodal models
- 

Generated via custom pipeline · 2025-11-18