Gene-Brain CCA Analysis: Concise Report

Author: Allie Date: January 14, 2026 Dataset: UK Biobank (N=4,218)

Executive Summary

This study investigated whether combining genetic embeddings (from DNABERT-2 foundation model) with brain imaging (fMRI) data improves Major Depressive Disorder (MDD) prediction.

Key Findings:

* Gene-only prediction achieves AUC 0.759 (holdout)
* fMRI adds no predictive value (early fusion AUC 0.762, +0.003)
* Unsupervised CCA/SCCA underperforms direct supervised learning by 17-23 AUC points
* Full 768-D embeddings improve performance by +29% vs scalar pooling

Korean Summary (한국어 요약)

본 연구는 DNABERT-2 foundation model의 genetic embeddings과 brain imaging (fMRI) 데이터를 결합하여 Major Depressive Disorder (MI

주요 결과:

* Gene-only 예측은 AUC 0.759 달성 (holdout)
* fMRI는 예측 가치를 추가하지 않음 (early fusion AUC 0.762, +0.003)
* Unsupervised CCA/SCCA는 direct supervised learning보다 17-23 AUC points 낮은 성능
* Full 768-D embeddings은 scalar pooling 대비 +29% 성능 향상

Dataset Overview

Metric	Value
Total subjects	4,218
MDD cases	1,735 (41.1%)
Controls	2,483 (58.9%)
Gene features	111 genes x 768-D
fMRI features	180 brain ROIs

Methods Summary

Experiment 1: Two-Stage CCA/SCCA

Stage 1 (Unsupervised): CCA/SCCA finds gene-brain correlations Stage 2 (Supervised): Predict MDD from canonical variates

Gene reduction strategies:

* Mean pooling: Average of 768 dimensions
* Max pooling: Maximum of 768 dimensions

Experiment 2: Leakage-Safe Pipelines

Pipeline A: Interpretable SCCA on scalar genes Pipeline B: Supervised prediction with full 768-D embeddings

Key Results

Experiment 1: Mean vs Max Pooling

Metric	Mean Pooling	Max Pooling
Stage 1 p-value	0.040 (sig)	0.995 (n.s.)
Gene-only AUC	0.588	0.505

Mean pooling preserves more predictive information. Max pooling destroys the genetic signal.

Experiment 2: Pipeline B Results

Model	Holdout AUC	Note
gene_only_logreg	0.759	Best
early_fusion_logreg	0.762	Marginal
fmri_only_logreg	0.559	Chance
cca_joint_logreg	0.546	Weak
scca_joint_logreg	0.566	Poor

Master Comparison

Experiment	Best AUC	Key Insight
Exp1 Mean Pool	0.588	Mean preserves signal
Exp1 Max Pool	0.522	Max loses signal
Exp2 Pipeline B	0.762	Full embeddings best
Yoon et al.	0.851	Reference (N=29k)

Scientific Conclusions

Finding	Evidence
Gene-brain coupling weak	r=0.368, p=0.04

CCA/SCCA hurts prediction	0.566 vs 0.759 AUC
fMRI adds no value	AUC 0.50-0.56
Full embeddings essential	+29% improvement

Clinical Implications

English

* Brain imaging does not improve genetic prediction of MDD
* Foundation model embeddings must be preserved (not pooled)
* Gene-brain alignment is statistically real but clinically irrelevant

Korean (한국어)

* Brain imaging은 MDD의 genetic 예측을 향상시키지 않음
* Foundation model embeddings은 보존되어야 함
* Gene-brain alignment은 통계적으로 실재하나 임상적으로 무관

Recommendations

Immediate Next Steps

* Gene curation - Filter to Yoon's 38 genes (Expected AUC: 0.80-0.84)
* Remove PCA bottleneck - Use LASSO on full 85K features (Expected AUC: 0.78-0.82)
* Match methodology - Implement 10-fold nested CV

Future Directions

* Expand sample size (target N>10,000)
* Test alternative brain features (network-specific)
* Explore fMRI foundation models (BrainLM)
* Supervised feature selection for interpretability

향후 방향 (Korean)

* Sample size 확장 (N>10,000 목표)
* 대체 brain features 테스트
* fMRI foundation models 탐색 (BrainLM)
* 해석 가능성을 위한 supervised feature selection

Technical Glossary

Term	Definition	Korean
AUC	Area Under ROC Curve	ROC 곡선 아래 면적

CCA	Canonical Correlation Analysis	정준상관분석
SCCA	Sparse CCA	희소 정준상관분석
Foundation Model	Pre-trained neural network	대규모 사전 훈련 신경망
fMRI	Functional MRI	기능적 자기공명영상
PCA	Principal Component Analysis	주성분 분석
Holdout Set	Fixed test set	고정 테스트 셋
MDD	Major Depressive Disorder	주요우울장애

End of Report / 보고서 끝