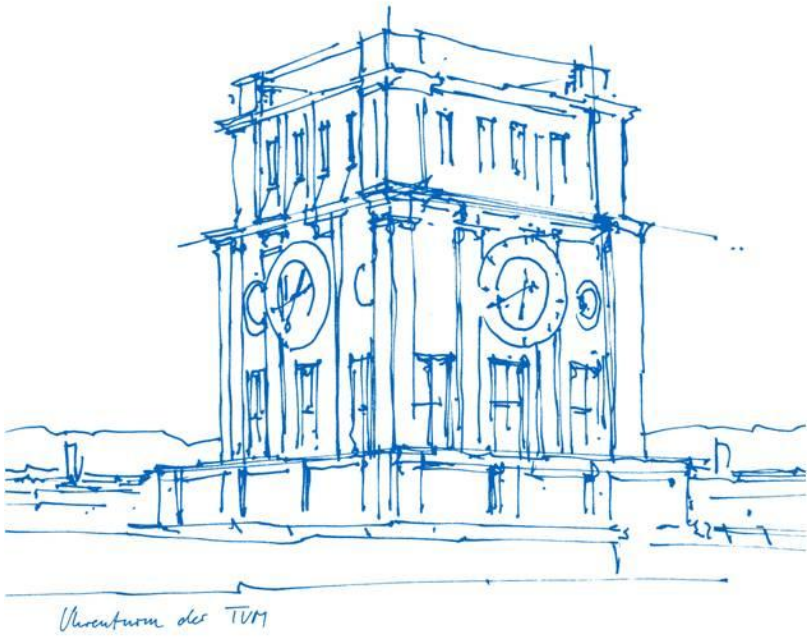
Functional Gene Embeddings (Group A)

Niklas Bühler, Keming Zhang, Marco Basmaji

Munich, February 2024



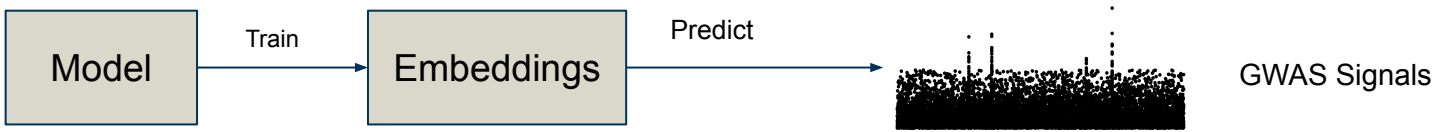
Outline

1. Motivation & Goals
2. Data Description
3. Methods
4. Results
5. Further Research

Motivation & Goals

Motivation & Goals

- **GWAS**: Investigate the effect of common variants on traits and diseases
- **Functional Gene Embeddings**: Numerical vectors capturing gene functions
- **Goal**: Generate useful representations of genes for downstream tasks
- Evaluate quality of embeddings on GWAS signal prediction



Data Description

1. GTEx
2. Tabula Sapiens
3. DepMap Gene Effect
4. RNA Isoform

Data Description

GTEx, The Genotype-Tissue Expression:

- Collected samples from non-diseased tissue across many individuals (50k x 11k dense matrix)

Tabula sapiens:

- Contains single cell transcriptomics data of 483,152 cells across 58,870 genes.

1. G. Consortium, “The gtex consortium atlas of genetic regulatory effects across human tissues,” Science, vol. 369, no. 6509, pp. 1318–1330, 2020.
2. T. S. Consortium*, R. C. Jones, J. Karkanias, M. A. Krasnow, A. O. Pisco,S. R. Quake, J. Salzman, N. Yosef, B. Bulthaup, P. Brown,et al., “The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans,” Science , vol. 376, no. 6594, p. eabl4896, 2022.

Data Description

DepMap Gene Effect dataset:

- Scores that measure the effect size of knocking out a gene.
- Measure the dependency between genes and cell lines of around 17k genes and 500 cell lines.

RNA Isoform Dataset:

- It contains RNA levels in 32 tissues, based on RNA-seq.

1. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, et al. “Defining a cancer dependency map,” Cell vol. 170, no. 3, pp. 564–576, 2017.
2. M.Uhl´en,L.Fagerberg,B.M.Hallstr¨om, C.Lindskog,.Oksvold,A. Mardinoglu, ˚A. Sivertsson, C. Kampf, E. Sj¨ostedt, A. Asplund,et al.,
“Tissue-based map of the human proteome,”Science vol. 347,no. 6220,p. 1260419, 2015.

Data Description

Data Approach:

- All of the datasets were aggregated based on unique Ensembl gene IDs and subsetted according to the protein-coding genes (Around 19k genes).

- GTEx was cut over samples and the Autoencoder was trained over only 5k samples.

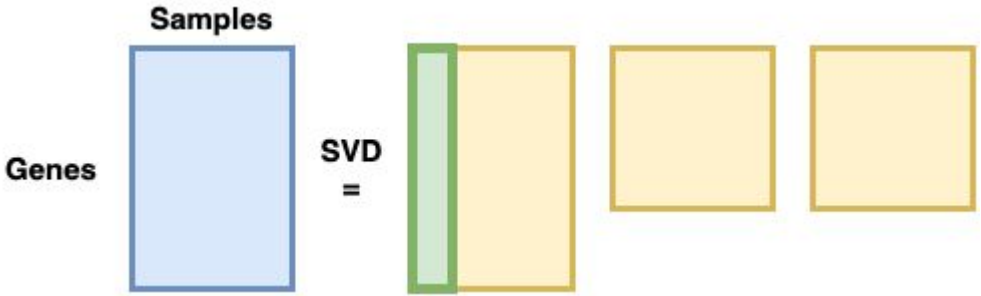
- Tabula Sapiens was aggregated by cell ontology classes (Pseudo bulks).

Methods

1. Generating functional gene embeddings
    a. Principal Components
    b. Autoencoder
    c. Variational Deep Tensor Factorization Model
2. Evaluating the utility of functional gene embeddings

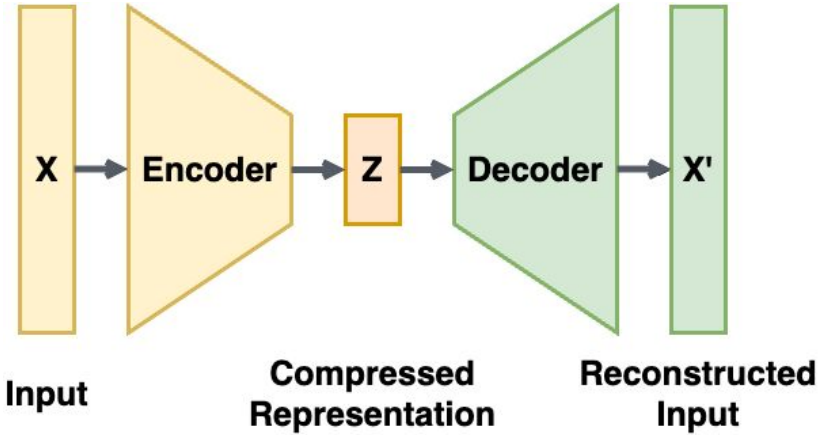
Method – Generating Embeddings – Principal Components

- Principal components that could explain **80-95 percent of variance**



Method – Generating Embeddings – Autoencoder

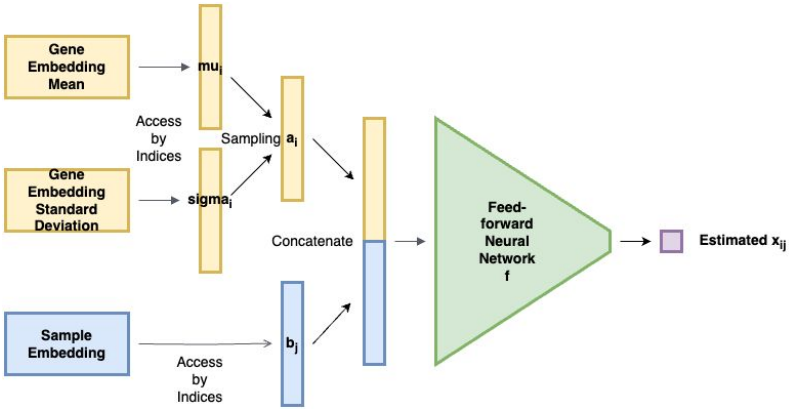
- Basic structure of an autoencoder



- Latent Space: the compressed representation is extracted and serve as **information-dense embeddings** of the genes

Method – Generating Embeddings – Variational Deep Tensor Factorization Model

- Basic structure of Variational Deep Tensor Factorization Model[1]



- Training the gene and sample embeddings to reconstruct the data matrix

1.F. Brechtmann, T. Bechtler, S. Londhe, C. Mertes, and J. Gagneur, “Evaluation of input data modality choices on functional gene embeddings,” NAR Genomics and Bioinformatics, vol. 5, no. 4, p. lqad095, 2023.

Method – Evaluating the Embeddings – GWAS Signal Prediction

- Evaluate the utility of gene embeddings by using them as gene features to **predict genome-wide association study (GWAS) summary statistics**, and investigating how they could **improve the accuracy scores**.
- General process of the evaluation on **30 traits**.

Method – Evaluating the Embedding – GWAS Signal Prediction – Details

- Prediction model: linear regression and XGBoost regression
- Traits considered: 30 traits, mostly maximally independent (see the Supplement)
- Covariates: gene density, effective gene size, inverse of the mean minor allele count of SNPs in the genes, as well as logarithmic values of them
- Cross validation strategy: Leave-One-Chromosome-Out
- GWAS data sources:
  - The Pan UKB study[1]
  - The UKBiobank[2]
  - The publication results of Brechtmann et al[3]

1.Pan-UKB team, “Pan-uk biobank,” 2020.
2.C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J.O’Connell, et al., “The uk biobank resource with deep phenotyping and genomic data,” Nature, vol. 562, no. 7726, pp. 203209, 2018.
3.F. Brechtmann, T. Bechtler, S. Londhe, C. Mertes, and J. Gagneur, “Evaluation of input data modality choices on functional gene embeddings,” NAR Genomics and Bioinformatics, vol. 5, no. 4, p. lqad095, 2023.

Results

1. Embedding Results
2. Significant Differences between Full and Null Models
3. Directly Examining ΔR^2
4. Well-Predictable Traits Improve More with Additional Data
5. Which Embeddings work Well on which Traits?
6. Evaluating with Linear Regression

Embedding Results

- **Eleven embeddings** from **three architectures** and **four data sources**.
- **XGBoost** Regression Model: Better general results (R^2) and more interesting analyses

Embedding Source	Embedding-Generating Model or Method	Dimension
DepMap Dataset	Principal Components	81
DepMap Dataset	Autoencoder	128
DepMap Dataset	VDTFM	128
GTEx Dataset	Principal Components	64
GTEx Dataset	Autoencoder	128
GTEx Dataset	Autoencoder	256
RNA Isoform Dataset	Principal Components	4
RNA Isoform Dataset	Autoencoder	32
Tabula Sapiens Dataset	Principal Components	5
Tabula Sapiens Dataset	Autoencoder	64
Tabula Sapiens Dataset	VDTFM	256

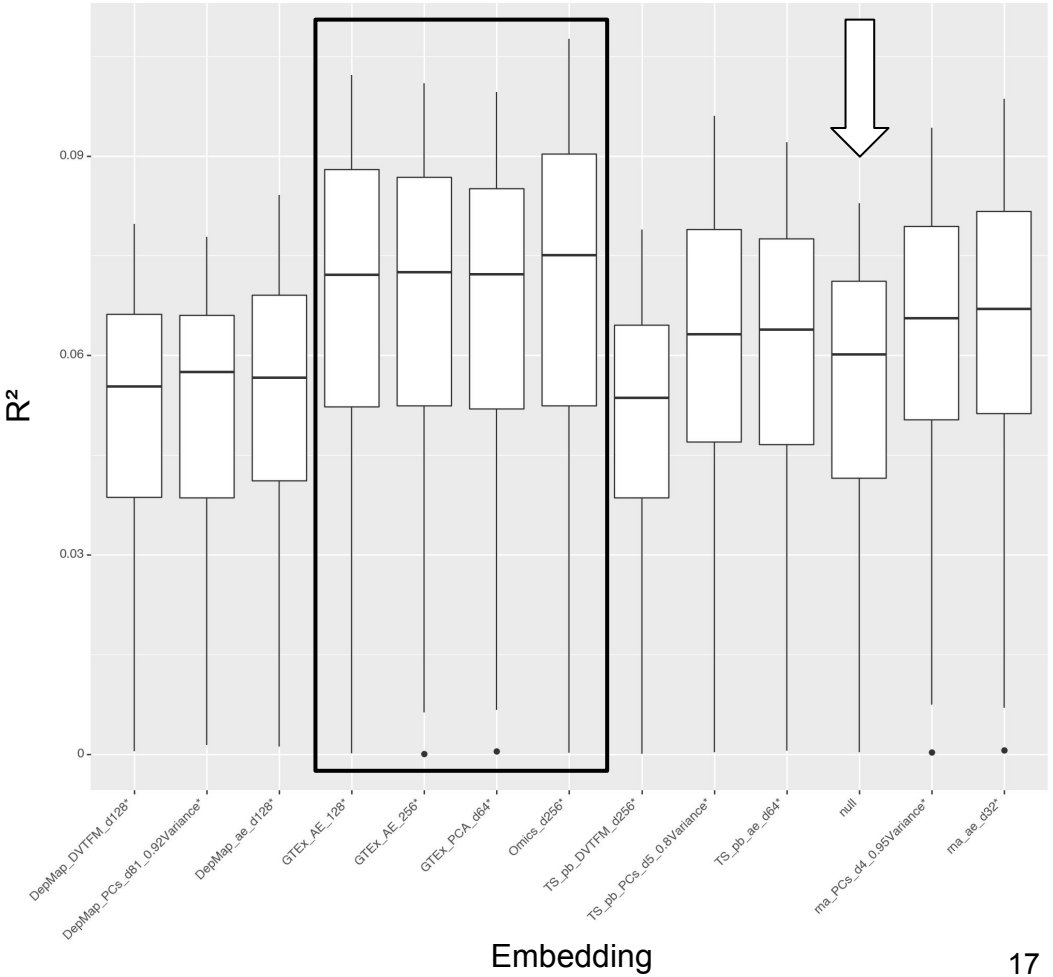
Table: Eleven embeddings from three different architectures and four different data sources.

Significant Differences between Full and Null Models

Comparatively **strong improvements** with **Omics** and **GTEx** embeddings.

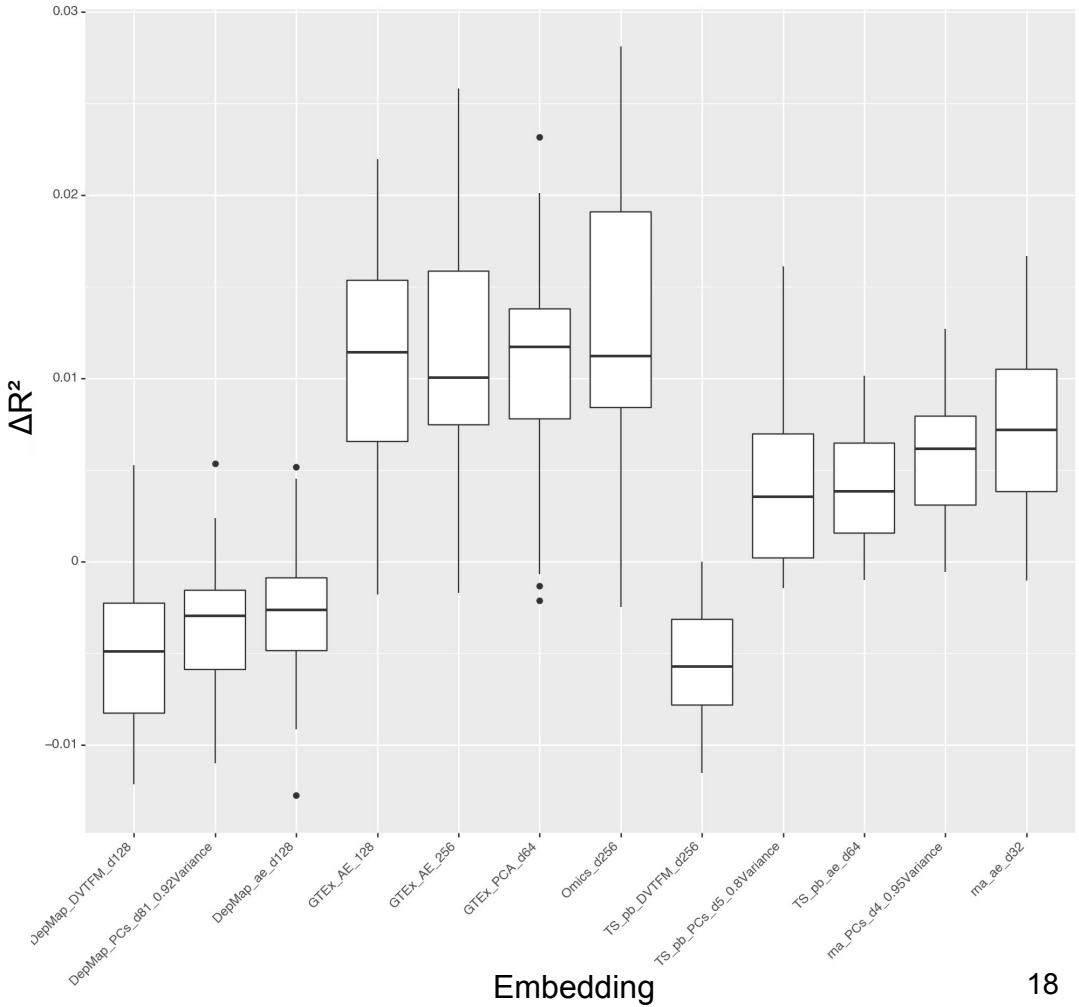
Wilcoxon Signed-Rank Test:

- **Significant differences** (α = 0.05)
- **Bonferroni Correction** (α ≈ 0.0042): still significant

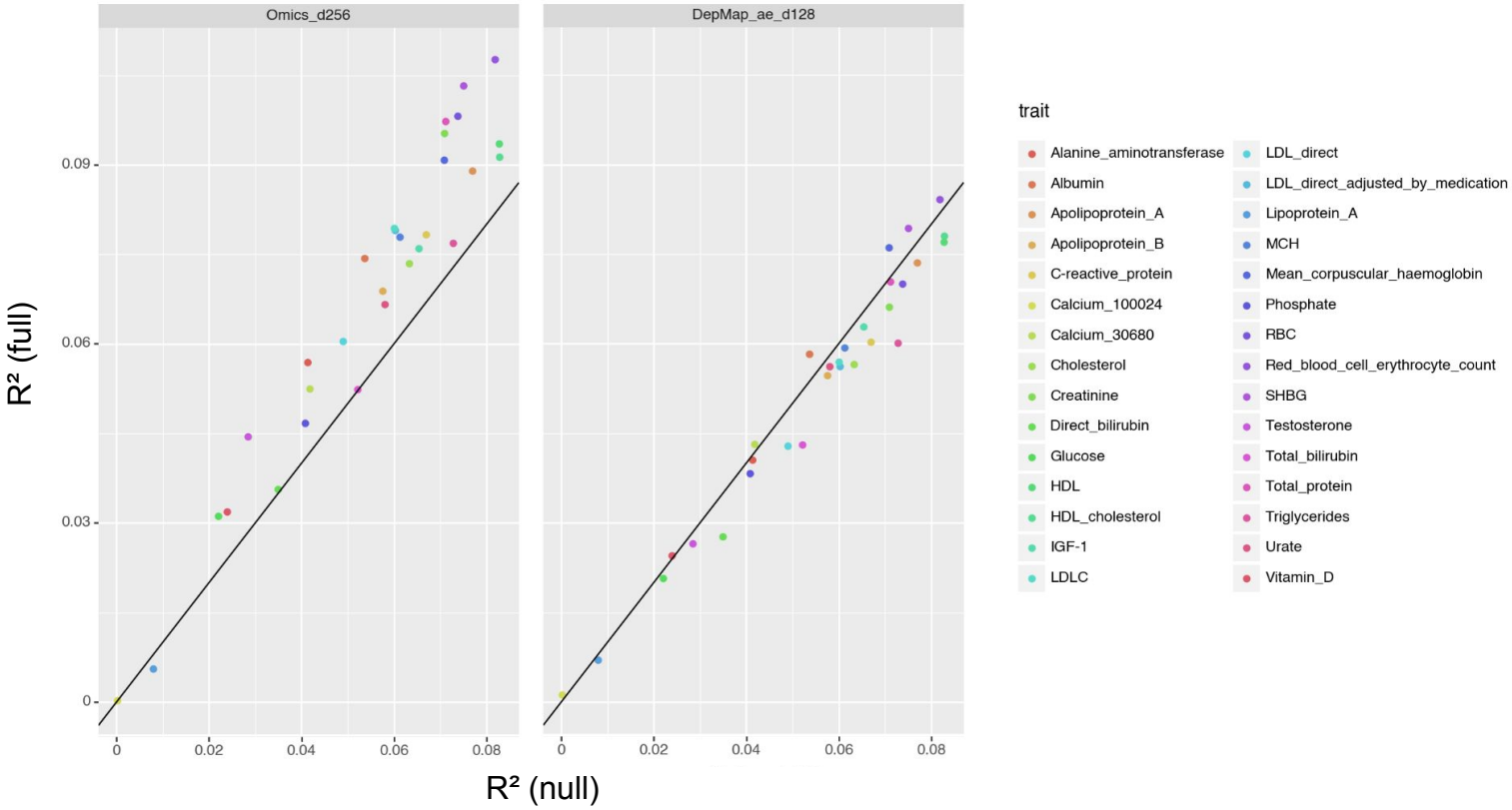


Directly Examining ΔR²

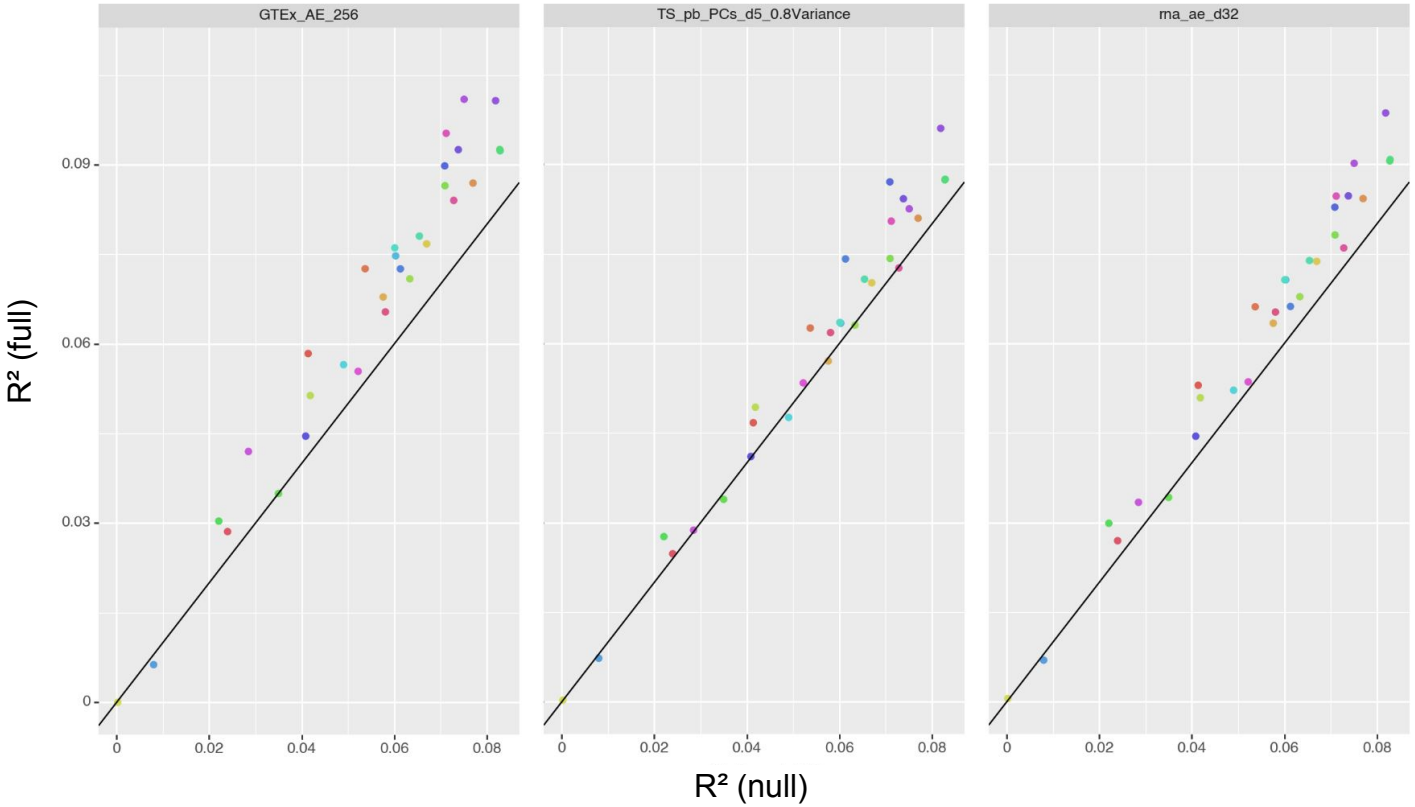
- **Omics** and **GTEx strongest**
- **RNA** and **TS** embeddings **also increased R²** for most traits
- **DepMap** and **VDTFM** embeddings **of TS didn't improve R²** on most traits



Well-Predictable Traits Improve More with Additional Data

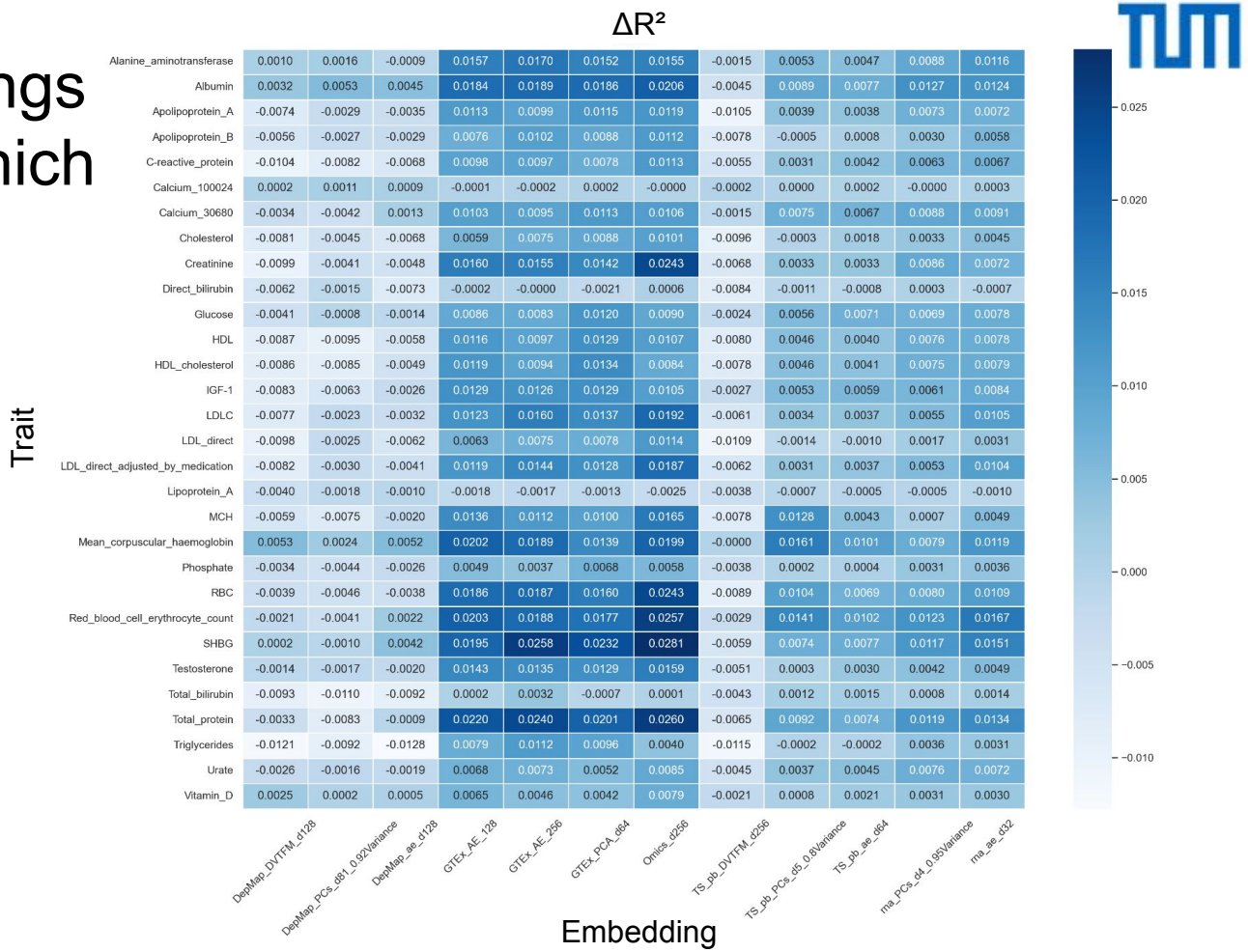


Well-Predictable Traits Improve More with Additional Data



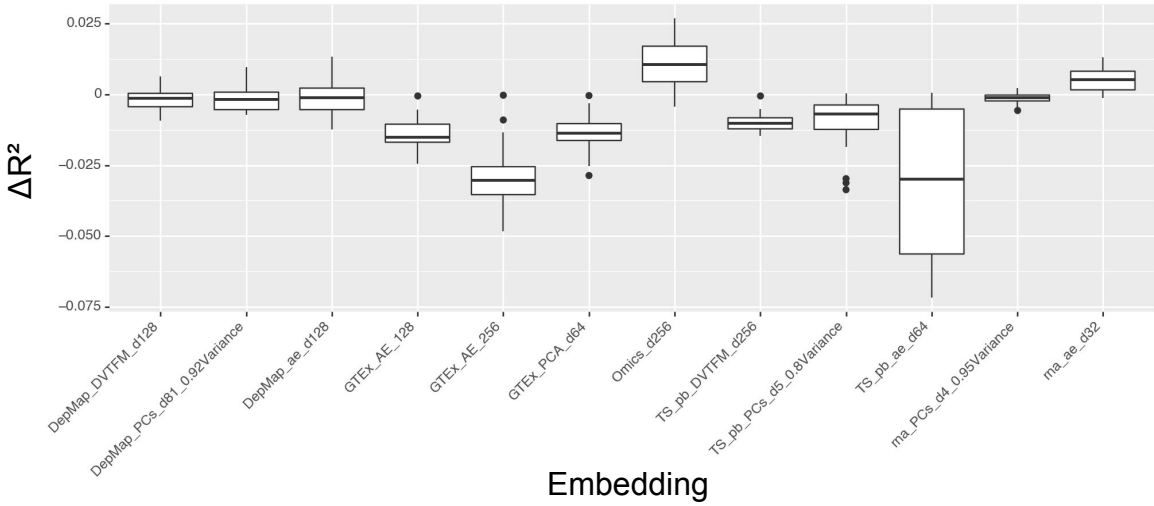
Which Embeddings work Well on Which Traits?

- **VDTFM didn't help** across datasets
- **Some traits didn't improve** (e.g. Lipoprotein A)
- **Others greatly improved through GTEx** and **Omics** (e.g. SHBG)



Evaluating with Linear Regression

- Almost all embeddings **perform worse**
- **Exception: DepMap**, but still only slight increase
- **Omics robust** against regression model selection
- **RNA Autoencoder** also **robust** in both setups, but not as good
- Hypothesis: **Assumptions of linear regression model** might not hold



Further Research

1. Comparable Dimensions of Embeddings

2. Embeddings Generated from VDTFM

3. Combining Embeddings from Different Data Sources

References

[1] F. Brechtmann, T. Bechtler, S. Londhe, C. Mertes, and J. Gagneur, “Evaluation of input data modality choices on functional gene embeddings,” NAR Genomics and Bioinformatics, vol. 5, no. 4, p. lqad095, 2023.
[2] A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, et al., “Defining a cancer dependency map,” Cell, vol. 170, no. 3, pp. 564–576, 2017.
[3] G. Consortium, “The gtex consortium atlas of genetic regulatory effects across human tissues,” Science, vol. 369, no. 6509, pp. 1318–1330, 2020.
[4] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, et al., “Tissue-based map of the human proteome,” Science, vol. 347, no. 6220, p. 1260419, 2015.
[5] T. T. S. Consortium*, R. C. Jones, J. Karkanias, M. A. Krasnow, A. O. Pisco, S. R. Quake, J. Salzman, N. Yosef, B. Bulthaup, P. Brown, et al., “The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans,” Science, vol. 376, no. 6594, p. eabl4896, 2022.
[6] M. Wainberg, R. A. Kamber, A. Balsubramani, R. M. Meyers, N. Sinnott-Armstrong, D. Hornburg, L. Jiang, J. Chan, R. Jian, M. Gu, et al., “A genome-wide atlas of co-essential modules assigns function to uncharacterized genes,” Nature genetics, vol. 53, no. 5, pp. 638–649, 2021.
[7] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.
[8] L. Biewald, “Experiment tracking with weights and biases,” 2020. Software available from wandb.com.
[9] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.
[10] Pan-UKB team, “Pan-uk biobank,” 2020.
[11] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O’Connell, et al., “The uk biobank resource with deep phenotyping and genomic data,” Nature, vol. 562, no. 7726, pp. 203–209, 2018.
[12] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma, “Magma: generalized gene-set analysis of gwas data,” PLoS computational biology, vol. 11, no. 4, p. e1004219, 2015.
[13] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, ACM, Aug. 2016.

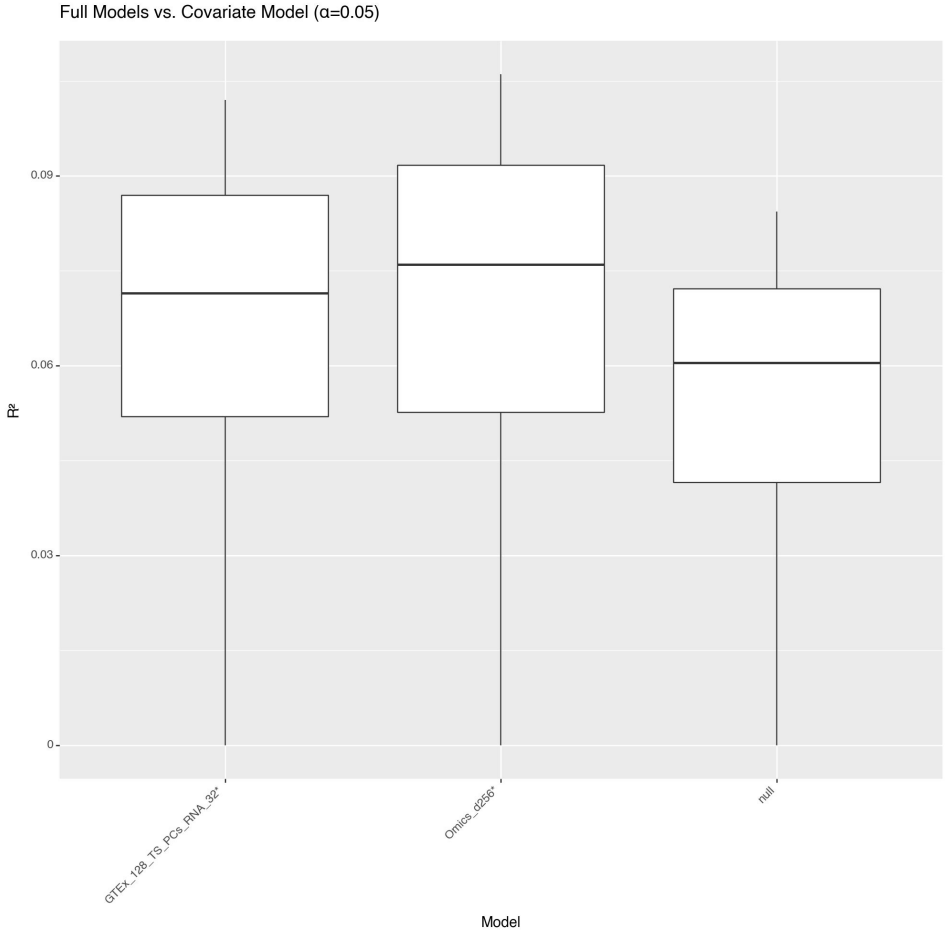
Thank You for the Attention…

Supplement

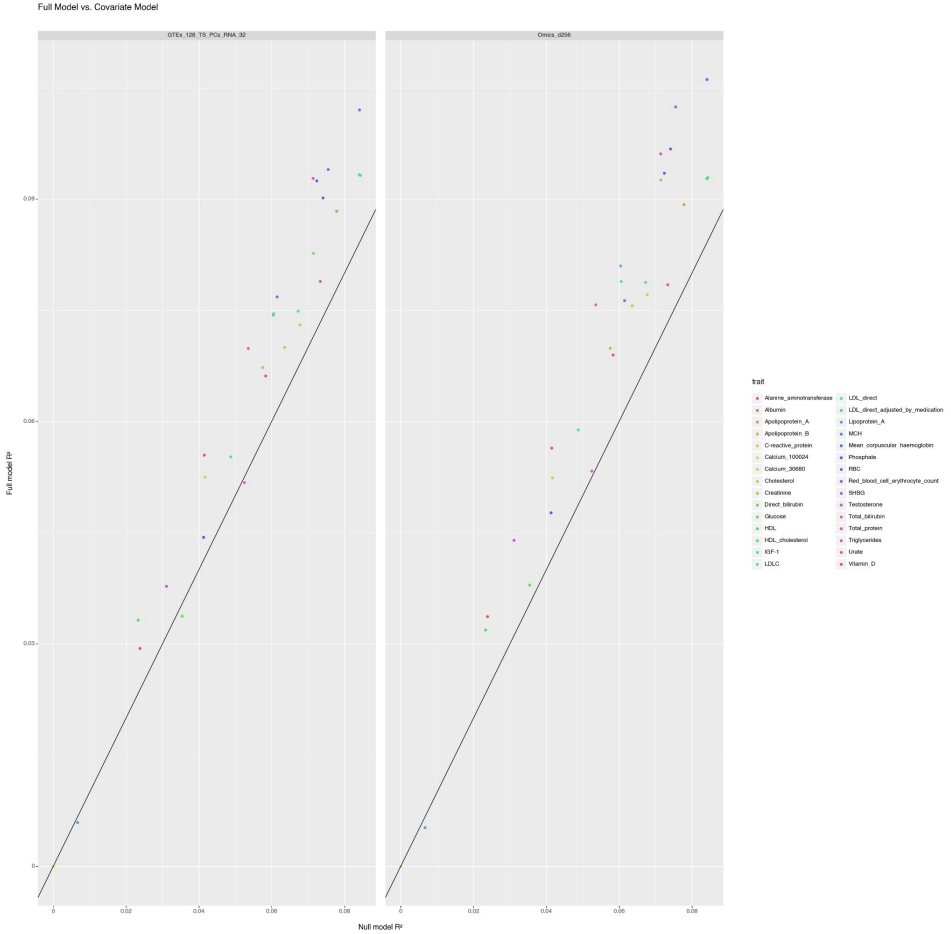
- Traits selected in GWAS Signal Prediction

Trait	Max. ind. set	Trait	Max. ind. set	Trait	Max. ind. set
Alanine Amino-transferate	True	Albumin	True	Apolipoprotein A	True
Apoliprotein B	True	C-reactive Protein	True	Calcium 30680	True
Calcium 100024	True	Cholesterol	True	Creatinine	True
Direct Bilirubin	False	Glucose	Flase	HDL Cholesterol	True
HDL	False	IGF-1	True	LDL Direct Adjusted by Medication	Flase
LDL Direct	True	LDLC	False	Lipoprotein	False
MCH	False	Mean Corpuscular Haemoglobin	False	Phosphate	True
RBC	False	Red Blood Cell Ery-throcyte Count	False	SHBG	True
Testosterone	Flase	Total Bilirubin	True	Total Protein	True
Triglycerides	True	Urate	False	Vitamin D	True

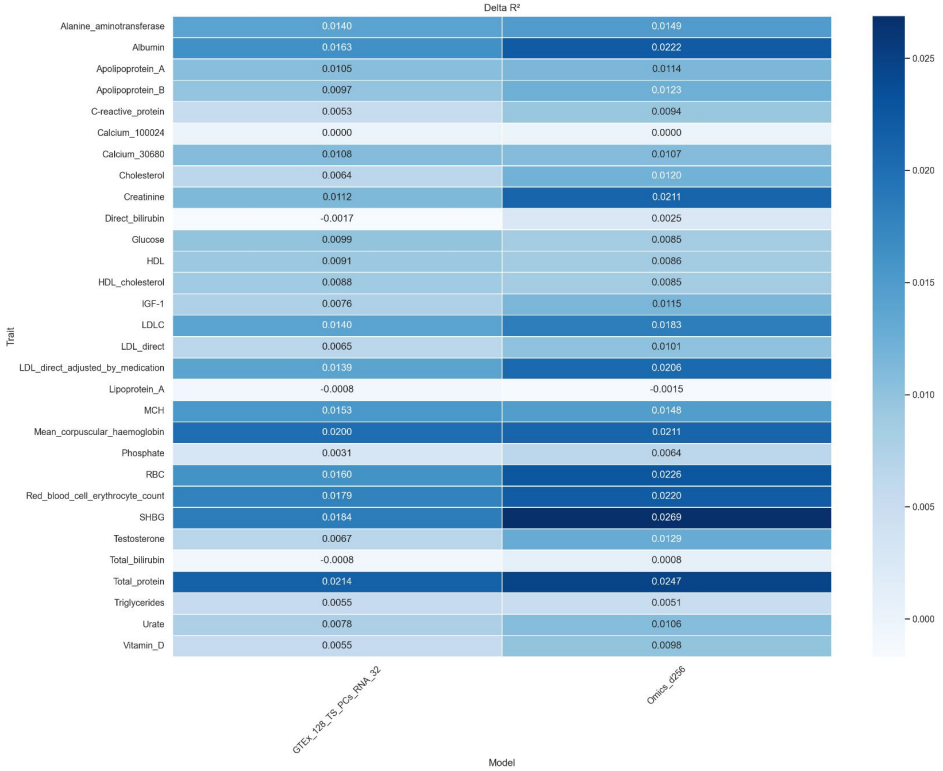
Supplement


Full Models vs. Covariate Model (α=0.05)
0.09
0.06
0.03
0
R²
GTEx_128_TS_PCs_RNA_32*
Omics_d256*
null
Model

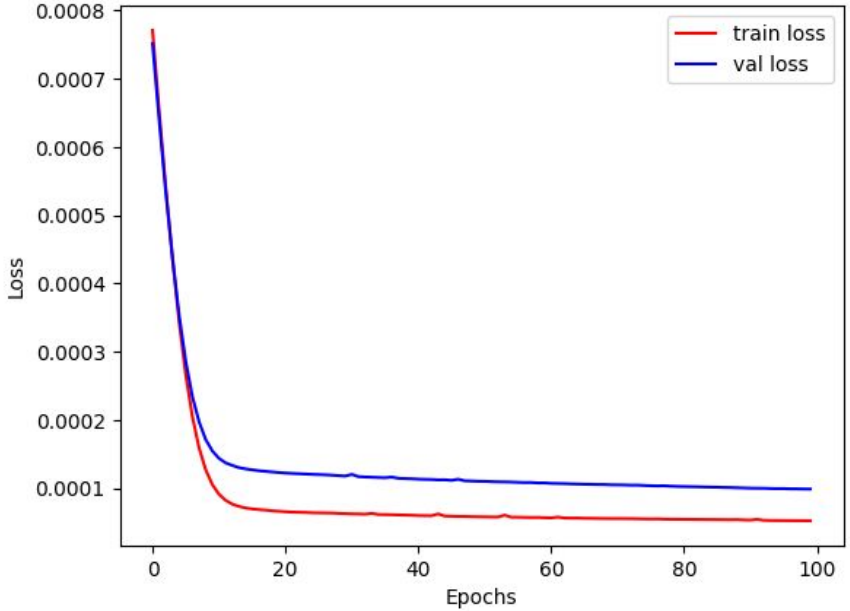
Supplement


Full Model vs. Covariate Model
GTEx_128_TS_PCs_RNA_32
Omics_d256
Full model R²
Null model R²
0
0.02
0.04
0.06
0.08
trait
Alanine_aminotransferase
Albumin
Apolipoprotein_A
Apolipoprotein_B
C-reactive_protein
Calcium_100024
Calcium_30680
Cholesterol
Creatinine
Direct_bilirubin
Glucose
HDL
HDL_cholesterol
IGF-1
LDLC
LDL_direct
LDL_direct_adjusted_by_medication
Lipoprotein_A
MCH
Mean_corpuscular_haemoglobin
Phosphate
RBC
Red_blood_cell_erythrocyte_count
SHBG
Testosterone
Total_bilirubin
Total_protein
Triglycerides
Urate
Vitamin_D

Supplement


Delta R²
Trait
Alanine_aminotransferase 0.0140 0.0149
Albumin 0.0163 0.0222
Apolipoprotein_A 0.0105 0.0114
Apolipoprotein_B 0.0097 0.0123
C-reactive_protein 0.0053 0.0094
Calcium_100024 0.0000 0.0000
Calcium_30680 0.0108 0.0107
Cholesterol 0.0064 0.0120
Creatinine 0.0112 0.0211
Direct_bilirubin -0.0017 0.0025
Glucose 0.0099 0.0085
HDL 0.0091 0.0086
HDL_cholesterol 0.0088 0.0085
IGF-1 0.0076 0.0115
LDLC 0.0140 0.0183
LDL_direct 0.0065 0.0101
LDL_direct_adjusted_by_medication 0.0139 0.0206
Lipoprotein_A -0.0008 -0.0015
MCH 0.0153 0.0148
Mean_corpuscular_haemoglobin 0.0200 0.0211
Phosphate 0.0031 0.0064
RBC 0.0160 0.0226
Red_blood_cell_erythrocyte_count 0.0179 0.0220
SHBG 0.0184 0.0269
Testosterone 0.0067 0.0129
Total_bilirubin -0.0008 0.0008
Total_protein 0.0214 0.0247
Triglycerides 0.0055 0.0051
Urate 0.0078 0.0106
Vitamin_D 0.0055 0.0098
0.025
0.020
0.015
0.010
0.005
0.000
GTEx_v28_TS_PCs_RNA_32
Omics_d226
Model

Supplement


train loss
val loss
Loss
Epochs
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
20
40
60
80
100