Multi-Modal Modelling of Preterm Birth Risk in a Black/African-ancestry cohort

Budhachandra Singh Yumkhaibam

Supervised by: Dr. Jacob Luber, Dr. Jennifer Woo (BIOMO Lab)

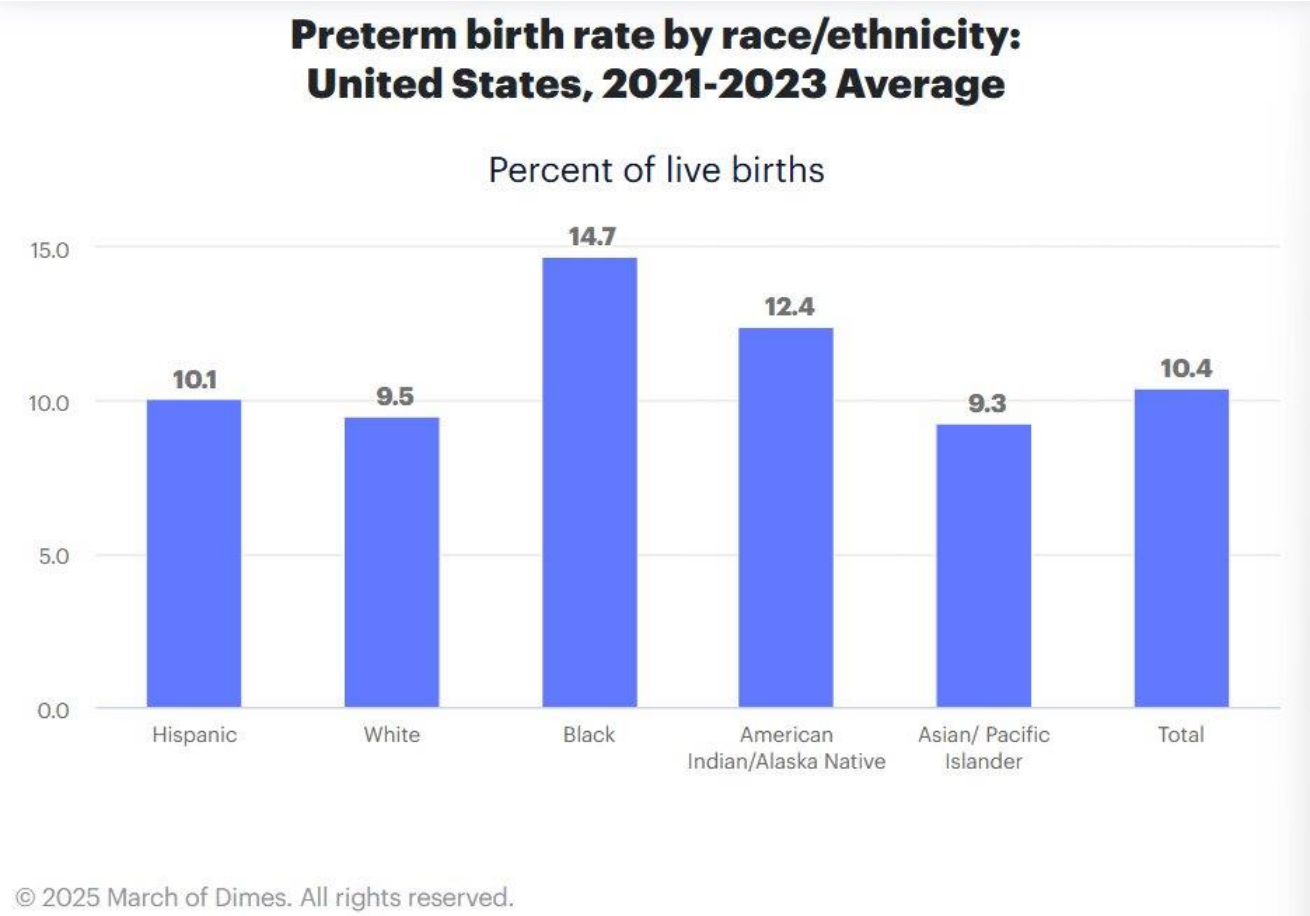
Committee: Dr. Farhad Kamangar, Dr. Sihong He, Dr. Jeffrey Demuth

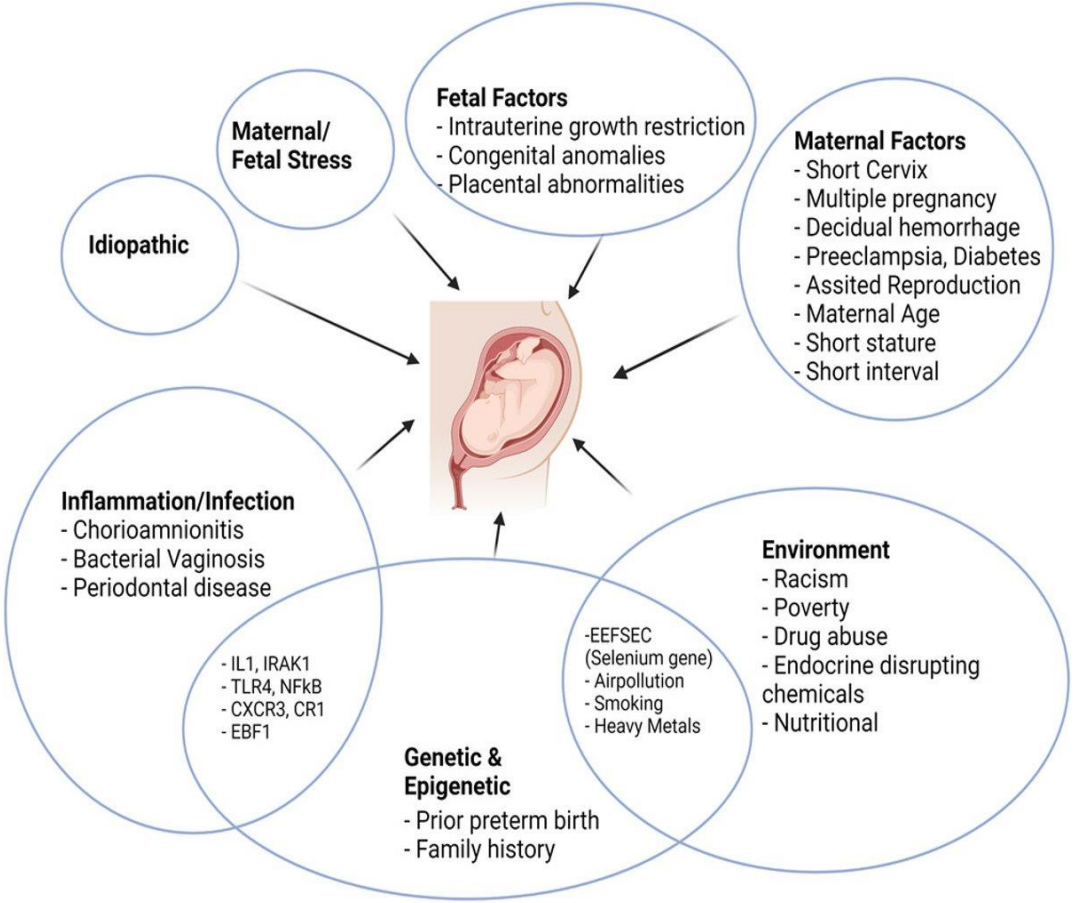
Preterm Birth

Introduction

- Preterm birth is when a baby is born too early, before 37 weeks of pregnancy have been completed. In 2022, preterm birth affected about 1 of every 10 infants born in the United States.
  - Centers for Disease Control and Prevention

- Pathways to preterm birth are multifactorial because PTB is influenced by maternal and fetal genomes.





Jains et al., 2022, *American Journal of Reproductive Immunology*

Preterm Birth Prediction (PTB)

The Literature~

[1] Adi et al (2021) Plasma proteomic data predict spontaneous delivery (sPTB) and using transcriptomic data (measure of RNA expression levels) of whole blood gene expression predicts PTB

[2] Ngo et al. (2018) Noninvasive blood tests for fetal development predict gestational age and preterm delivery

Motivation

Single Modal usually blood protein levels or maybe ultrasound image etc.

If multi-modal then demographic is mixed or just not reported at all.

There is a lack of study targeted at the most effected population group, non-Hispanic Black women. Our model is multi modal and focused on black cohort.

Data

Collected

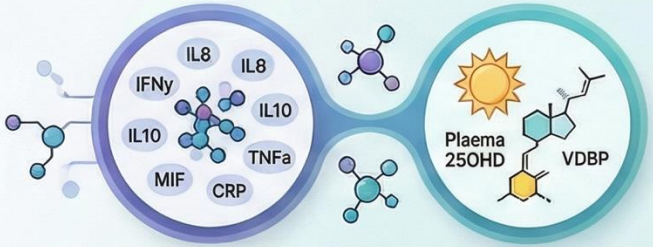
- A retrospective case-cohort study from a large longitudinal study called Biosocial Impact on Black Births (BIBB) that recruited Black pregnant women at Wayne State University, and Ohio State University.
- Center for Epidemiologic Scale for Depression (CES-D) screening for depression.
- Experiences of Discrimination Scale (EOD).
- Plasma cytokine levels of interferon (IFN)-γ, interleukin (IL)-6, IL-8, IL-10, and tumor necrosis factor (TNF)-α.
- Total 25(OH)D and Vitamin D Binding Protein analysis.
- Whole blood for RNA sequencing analysis.
- Illumina SNP array analysis

- Clinical/psyschosocial Data, N=184 patients; 36 (features)
- Cytokine/biomarker Data, N=183 patients; 75 (features)
- Genotype (gwas) Data, N=158 patient records
- RNA-seq Data -> Fastq files from genewiz and ucf, N=70 patient records

Anatomy of a Clinical Study: Key Variables Collected

This infographic breaks down the variables from a clinical study into distinct categories. It shows how the study combines biological markers with physical measurements, health history, and socio-demographic data to create a comprehensive patient profile.

Biological Markers

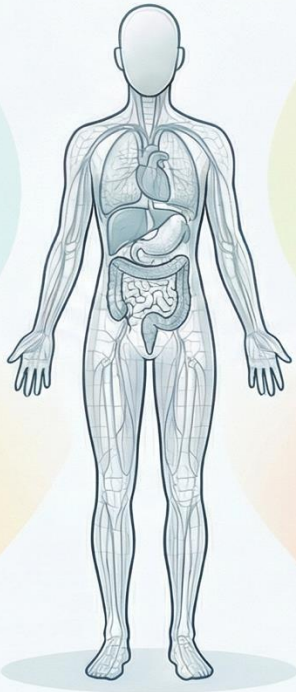


Inflammatory Cytokines

Key proteins that signal inflammation in the body.

Vitamin D Pathway

Vitamin D levels and its primary binding protein were messured.



Patient Health Profile

Physical Metrics

Core body measurements were recorded at the first visit. (e.g., Weight, Height, SMI)

Pre-Existing & Current Conditions

A history of major health issues was documented.

Lifestyle Factors

Relevant habits impacting pregnancy and health were assessed. (e.g., Nicotine Use During Pregnancy)

Socio-Demographic Data

Key life circumstances and background were captured.

Socio-Environmental & Psychological Factors

Psychosocial Assessments

Standardized scales were used to measure mental health and social experiences.

Traditional PTB prediction

Existing Methods

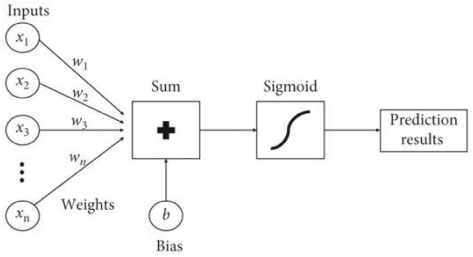
- Ultrasound based methods
- Fetal Fibronectin test (tFN)
- History based prediction like prior PTB

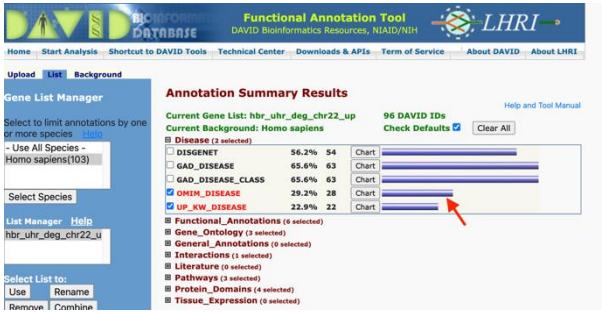
Cleaning Data

- ID harmonization PTID vs ID_N
- Numeric Coercion; handling of missing (dropping vs impute) and standardization
- Data Normalization
- Outlier detection and filtering

Experiments

- Model 1 is a simple model with clinical/psychosocial data + cytokine/biomarker data (no genomic data)
- Model 2 is a model which extends model 1 with genomic data (multi – modal model)
- Model 3 extends Model 2 with RNA sequencing data
- Permutation Enrichment for Differentially expressed genes (PTB vs non PTB)
- DAVID clustering in Differential expression genes (PTB vs non PTB)



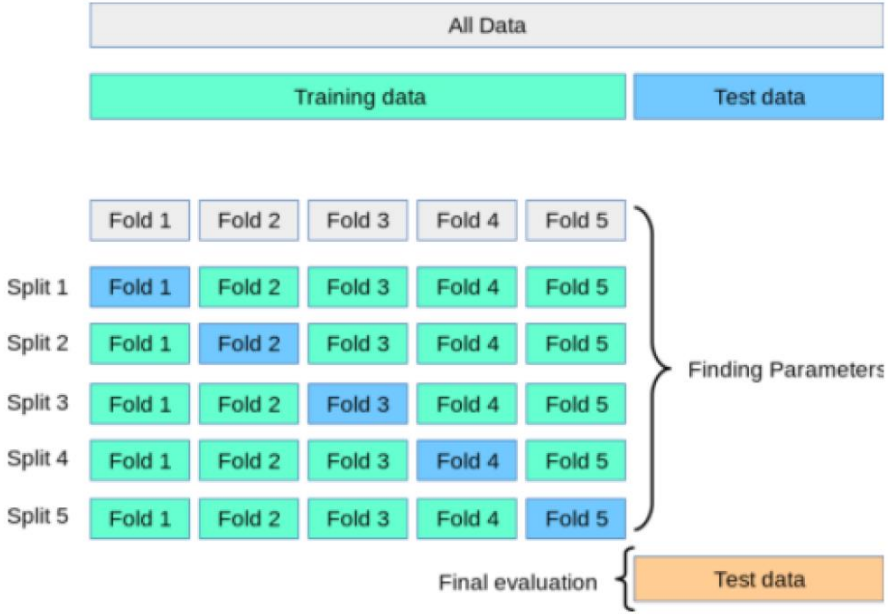


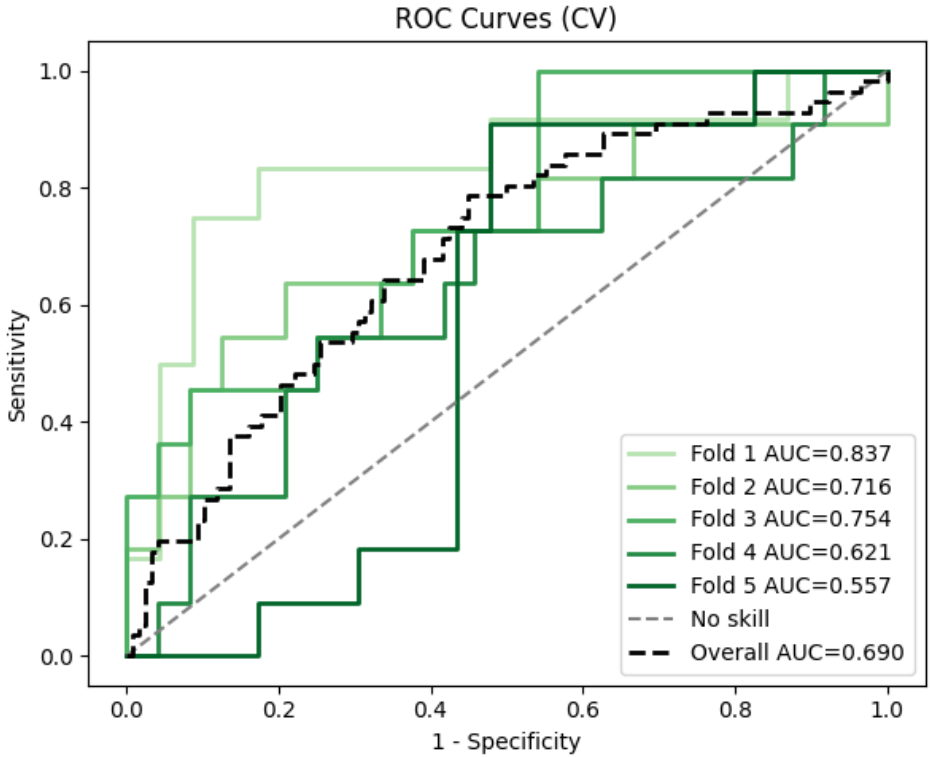
MODEL 1

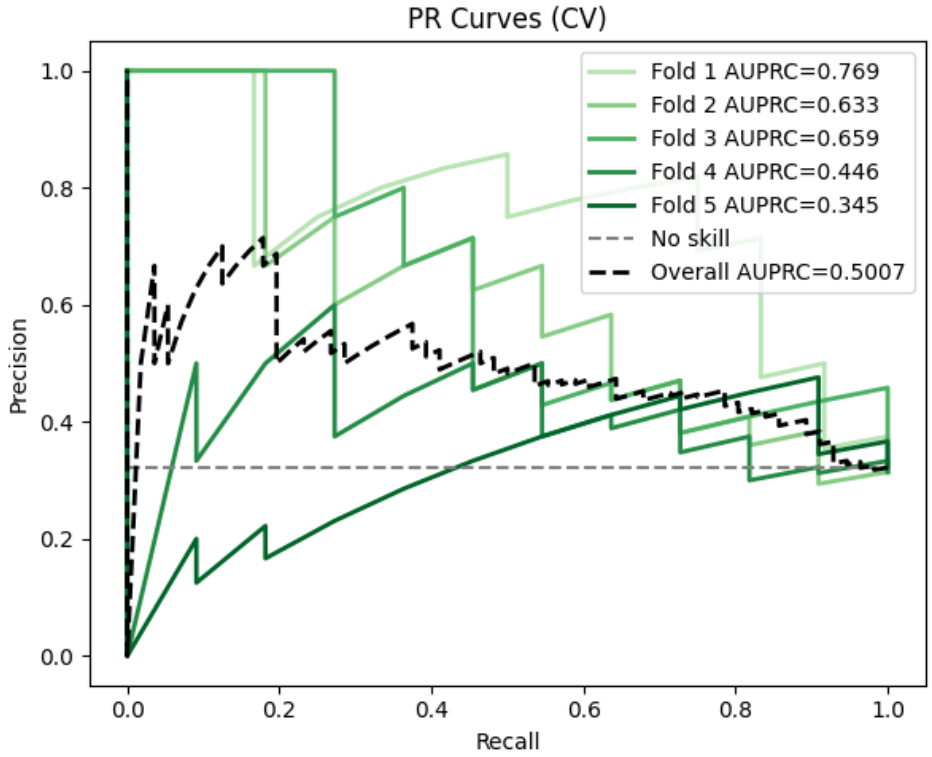
Clinical/Psychosocial + Cytokine/biomarker Data Model

- We used L2-regularized logistic regression and evaluated on Stratified 5-fold Cross-Validation (no separate hold-out).
- Metrics are: accuracy, ROC AUC, PR AUC
- Plots produced are: per fold + overall ROC/PR; SHAP best/worst folds

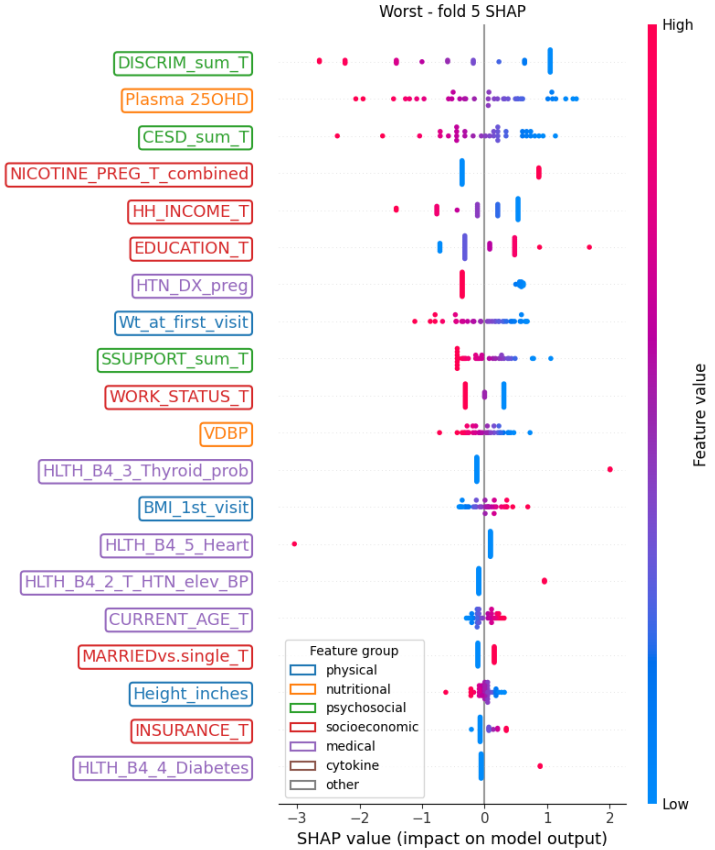
K fold Cross validation visualized

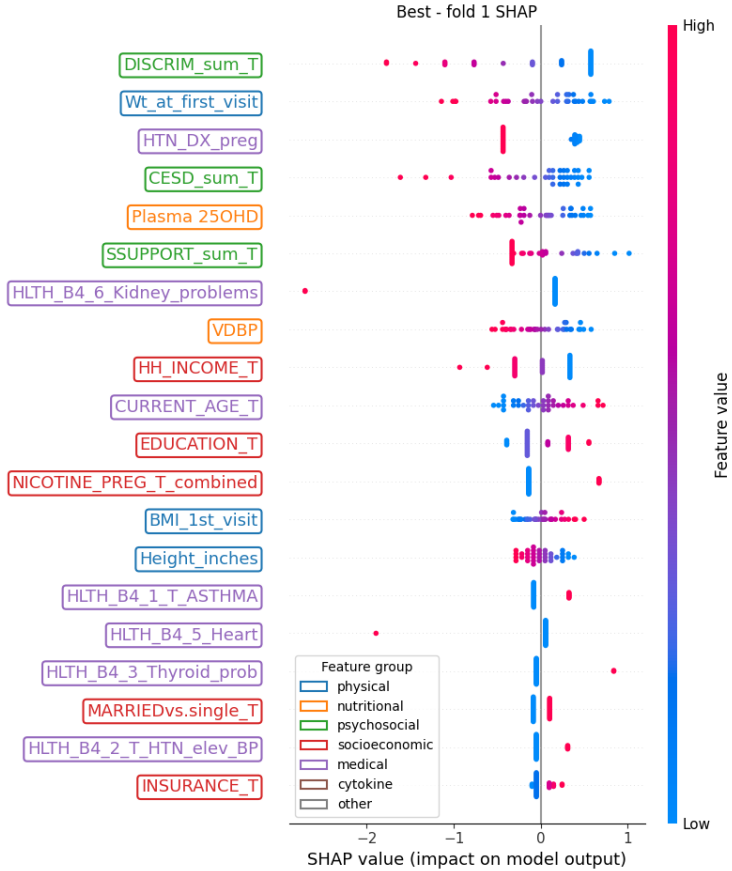






Mean Accuracy: ~66


Worst - fold 5 SHAP
DISCRIM_sum_T
Plasma 25OHD
CESD_sum_T
NICOTINE_PREG_T_combined
HH_INCOME_T
EDUCATION_T
HTN_DX_preg
Wt_at_first_visit
SSUPPORT_sum_T
WORK_STATUS_T
VDBP
HLTH_B4_3_Thyroid_prob
BMI_1st_visit
HLTH_B4_5_Heart
HLTH_B4_2_T_HTN_elev_BP
CURRENT_AGE_T
MARRIEDvs.single_T
Height_inches
INSURANCE_T
HLTH_B4_4_Diabetes
Feature group
physical
nutritional
psychosocial
socioeconomic
medical
cytokine
other
−3
−2
−1
0
1
2
SHAP value (impact on model output)
High
Feature value
Low


Best - fold 1 SHAP
DISCRIM_sum_T
Wt_at_first_visit
HTN_DX_preg
CESD_sum_T
Plasma 25OHD
SSUPPORT_sum_T
HLTH_B4_6_Kidney_problems
VDBP
HH_INCOME_T
CURRENT_AGE_T
EDUCATION_T
NICOTINE_PREG_T_combined
BMI_1st_visit
Height_inches
HLTH_B4_1_T_ASTHMA
HLTH_B4_5_Heart
HLTH_B4_3_Thyroid_prob
MARRIEDvs.single_T
HLTH_B4_2_T_HTN_elev_BP
INSURANCE_T
Feature group
physical
nutritional
psychosocial
socioeconomic
medical
cytokine
other
−2
−1
0
1
SHAP value (impact on model output)
High
Feature value
Low

- Conclusion of first model
  - Discrimination have strong signal.
  - Features for which we can provide "active" intervention like VitD, Plasma25OHD, Weight factors etc. are shown to have high signals.
  - Non-Linear relations e.g. weight at first visit, age of patient etc
  - Linear relations e.g. Nicotine use, Asthma history
  - High variance in weakest vs strongest model. It could be more stable.

MODEL 2

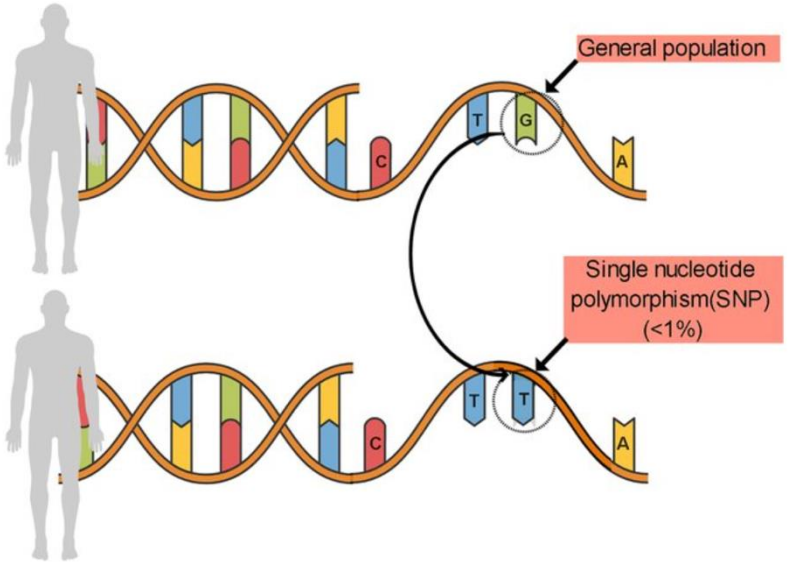
Model 1 + Genomic Data Model

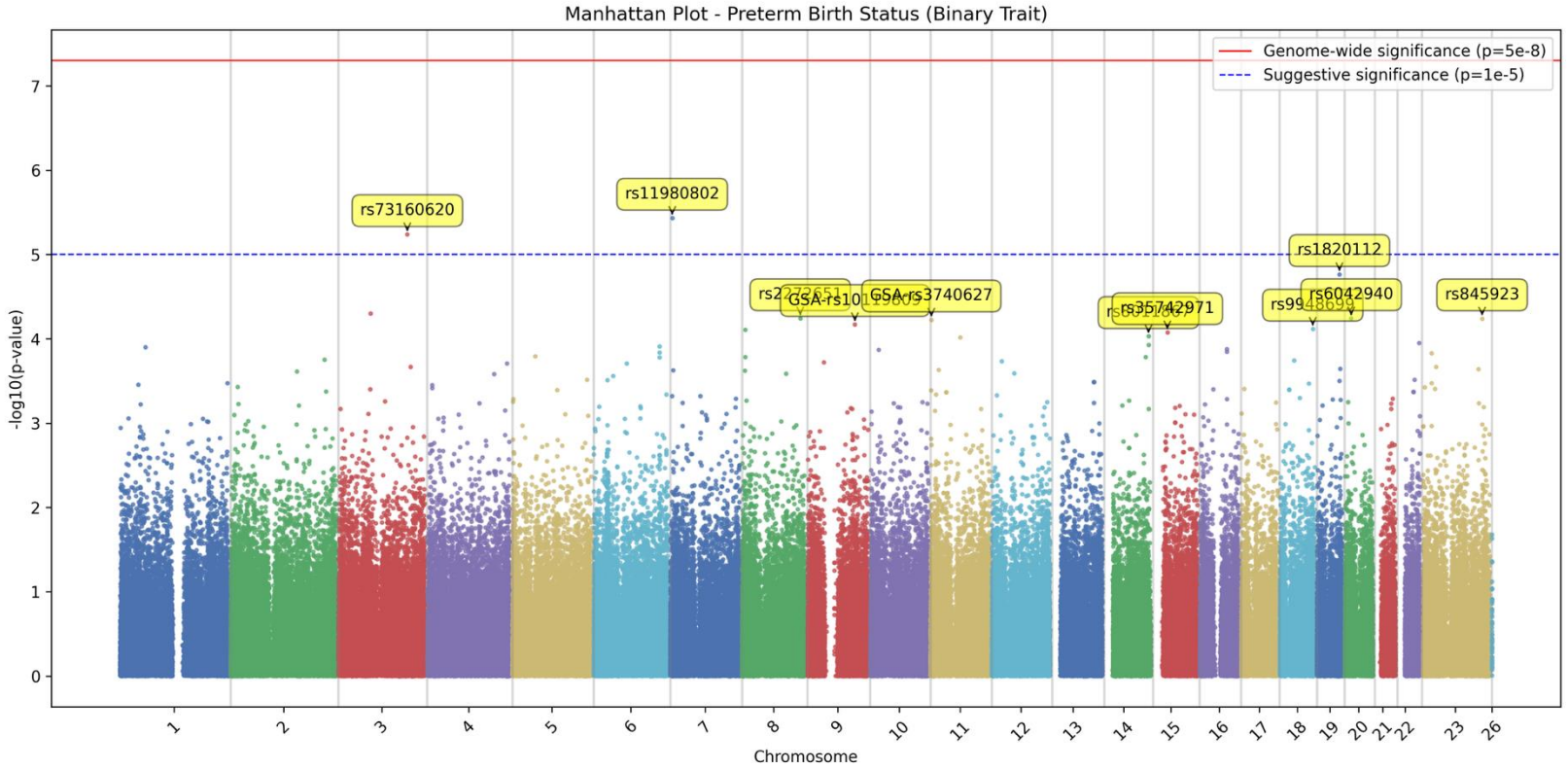
- We used L2-regularized logistic regression and evaluated on Stratified 5-fold Cross-Validation (no separate hold-out).
- Metrics are: accuracy, ROC AUC, PR AUC
- Plots produced are: per fold + overall ROC/PR; SHAP best/worst folds

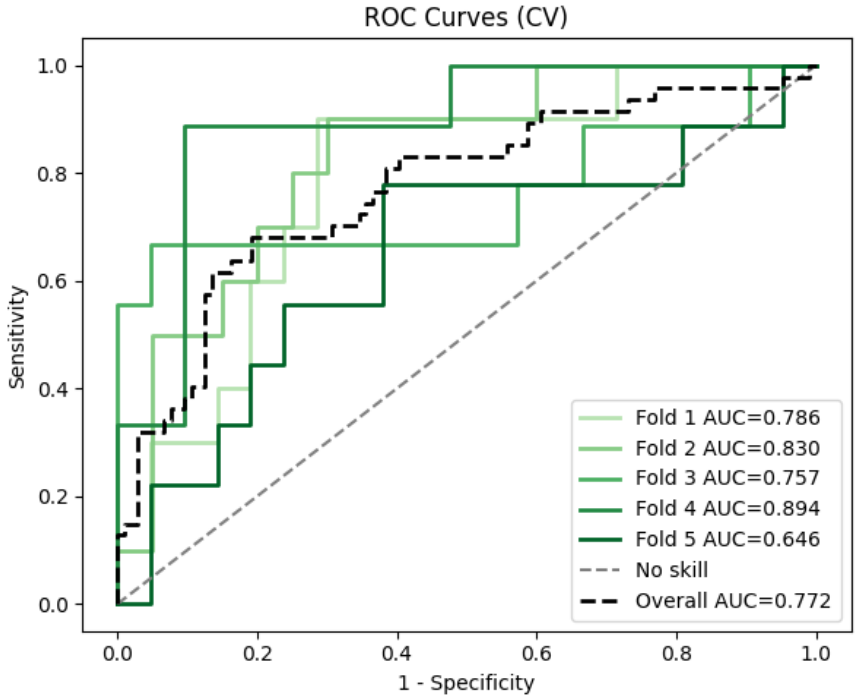
The Genomic Data

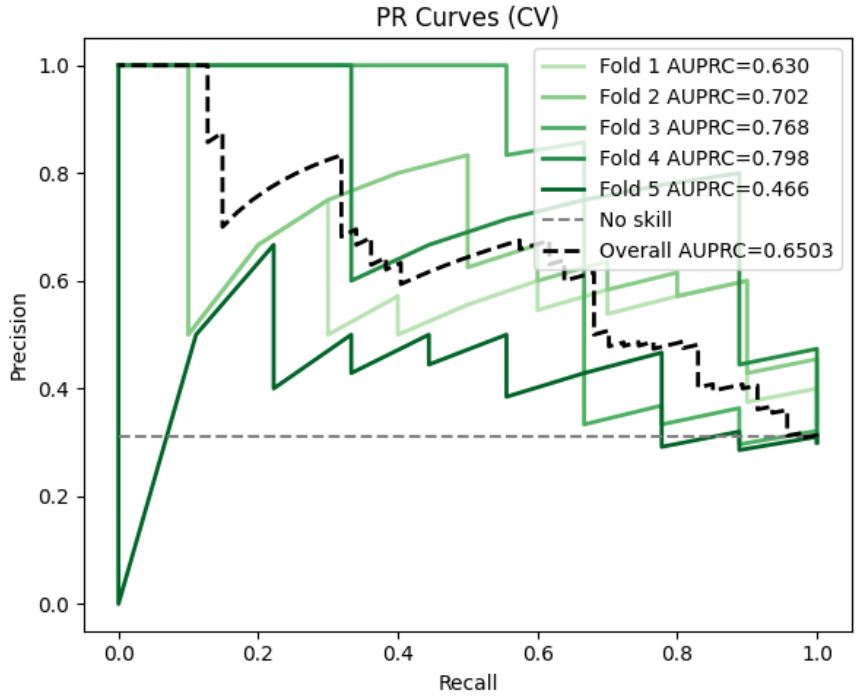
GWAS analysis

- Principal Components of
- GWAS (suggestive) significant SNPs
- Only 2 SNPs were significant after correction for multiple testing.
- SNPs: rs73160620 and rs11980802

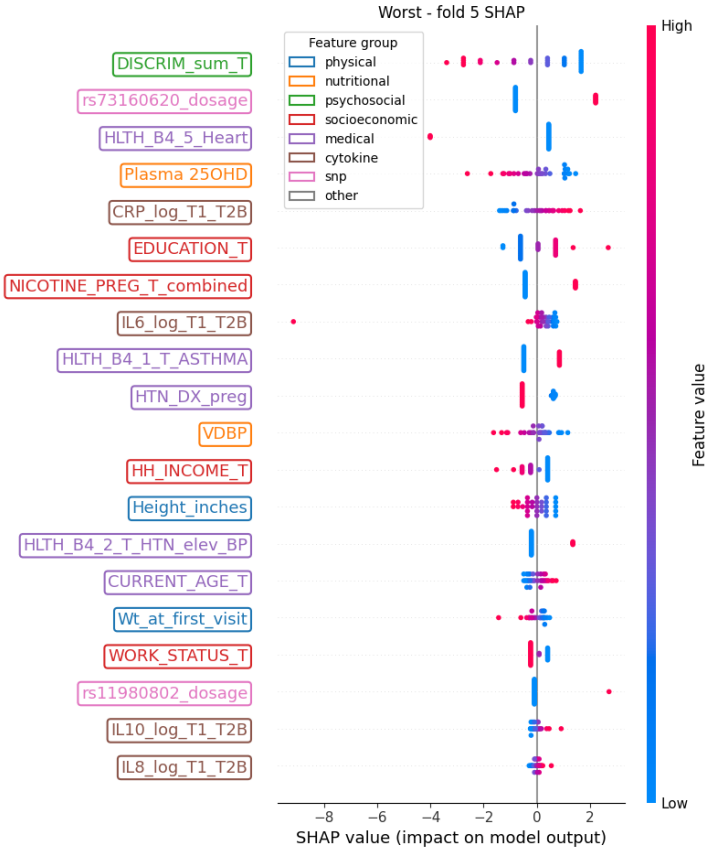


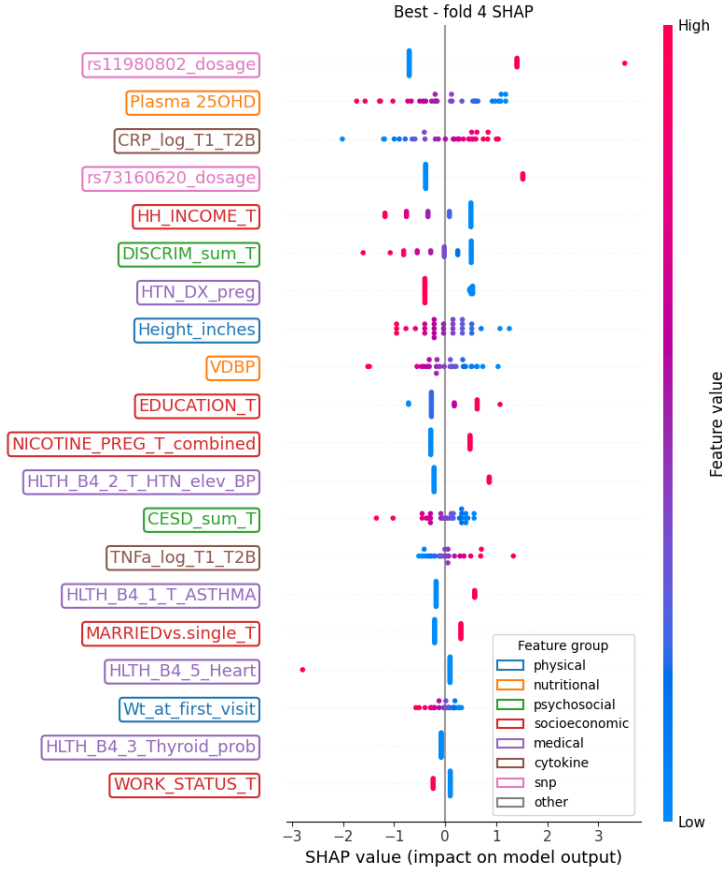

Manhattan Plot - Preterm Birth Status (Binary Trait)
Genome-wide significance (p=5e-8)
Suggestive significance (p=1e-5)
-log10(p-value)
Chromosome
rs73160620
rs11980802
rs1820112
rs6042940
rs845923
GSA-rs3740627
rs35742971





Mean Accuracy: ~72

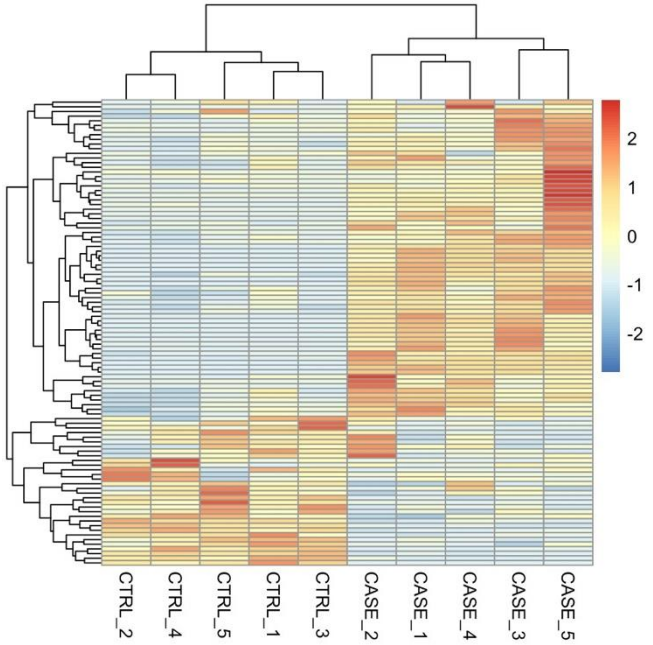

Worst - fold 5 SHAP
DISCRIM_sum_T
rs73160620_dosage
HLTH_B4_5_Heart
Plasma 25OHD
CRP_log_T1_T2B
EDUCATION_T
NICOTINE_PREG_T_combined
IL6_log_T1_T2B
HLTH_B4_1_T_ASTHMA
HTN_DX_preg
VDBP
HH_INCOME_T
Height_inches
HLTH_B4_2_T_HTN_elev_BP
CURRENT_AGE_T
Wt_at_first_visit
WORK_STATUS_T
rs11980802_dosage
IL10_log_T1_T2B
IL8_log_T1_T2B
Feature group
physical
nutritional
psychosocial
socioeconomic
medical
cytokine
snp
other
High
Feature value
Low
−8
−6
−4
−2
0
2
SHAP value (impact on model output)

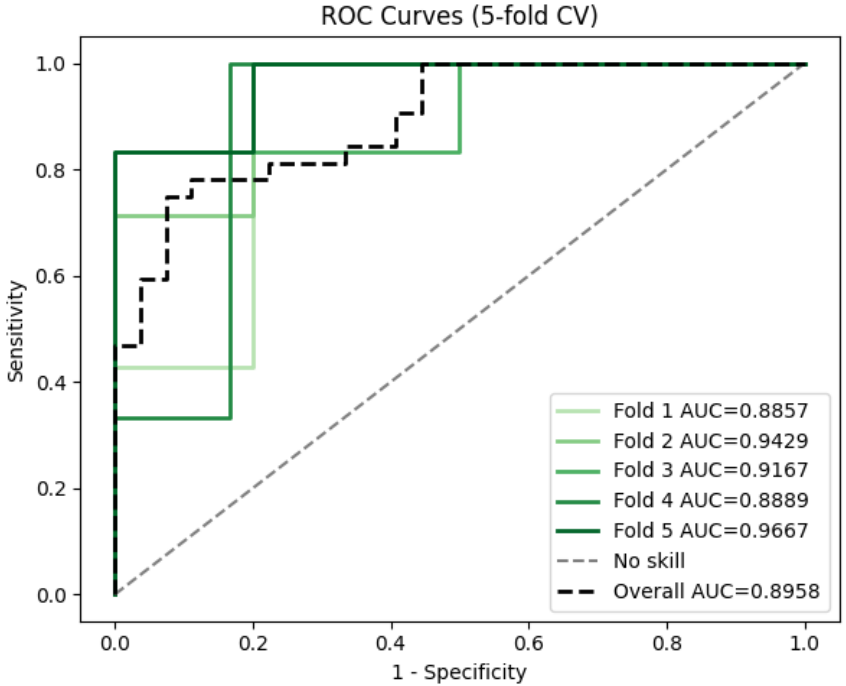

Best - fold 4 SHAP
rs11980802_dosage
Plasma 25OHD
CRP_log_T1_T2B
rs73160620_dosage
HH_INCOME_T
DISCRIM_sum_T
HTN_DX_preg
Height_inches
VDBP
EDUCATION_T
NICOTINE_PREG_T_combined
HLTH_B4_2_T_HTN_elev_BP
CESD_sum_T
TNFa_log_T1_T2B
HLTH_B4_1_T_ASTHMA
MARRIEDvs.single_T
HLTH_B4_5_Heart
Wt_at_first_visit
HLTH_B4_3_Thyroid_prob
WORK_STATUS_T
Feature group
physical
nutritional
psychosocial
socioeconomic
medical
cytokine
snp
other
High
Feature value
Low
−3
−2
−1
0
1
2
3
SHAP value (impact on model output)

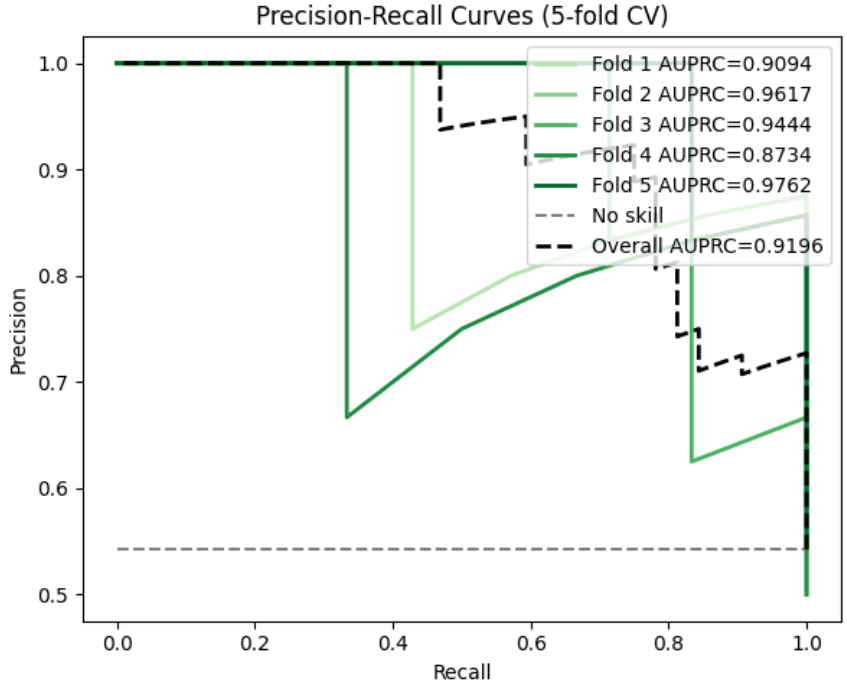
MODEL 3

Model 2 + RNA sequencing data

- TPM of RNA-seq data
  - Filtered top 10 genes :
    - 'RN7SL1', 'RN7SK','RN7SL2', 'MIR3648-1','RMRP', 'RPPH1', 'MT-ATP8', 'MALAT1', 'MT-ND4L','ALAS2'

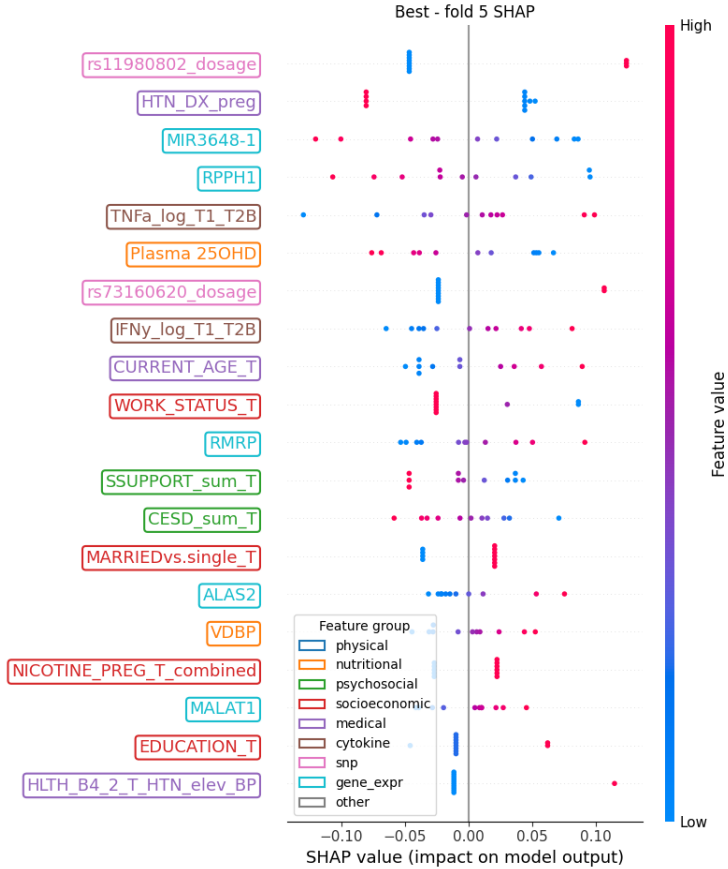






Mean Accuracy: ~80


Worst - fold 4 SHAP
Plasma 25OHD
RPPH1
HTN_DX_preg
CESD_sum_T
CURRENT_AGE_T
CRP_log_T1_T2B
rs73160620_dosage
MT-ND4L
TNFa_log_T1_T2B
MIR3648-1
MARRIEDvs.single_T
IL10_log_T1_T2B
ALAS2
MT-ATP8
rs11980802_dosage
HLTH_B4_3_Thyroid_prob
HLTH_B4_2_T_HTN_elev_BP
RN7SL1
WORK_STATUS_T
INSURANCE_T
Feature group
physical
nutritional
psychosocial
socioeconomic
medical
cytokine
snp
gene_expr
other
−0.15 −0.10 −0.05 0.00 0.05 0.10 0.15
SHAP value (impact on model output)
High
Feature value
Low


Best - fold 5 SHAP
rs11980802_dosage
HTN_DX_preg
MIR3648-1
RPPH1
TNFa_log_T1_T2B
Plasma 25OHD
rs73160620_dosage
IFNy_log_T1_T2B
CURRENT_AGE_T
WORK_STATUS_T
RMRP
SSUPPORT_sum_T
CESD_sum_T
MARRIEDvs.single_T
ALAS2
VDBP
NICOTINE_PREG_T_combined
MALAT1
EDUCATION_T
HLTH_B4_2_T_HTN_elev_BP
Feature group
physical
nutritional
psychosocial
socioeconomic
medical
cytokine
snp
gene_expr
other
−0.10 −0.05 0.00 0.05 0.10
SHAP value (impact on model output)
High
Feature value
Low

Conclusion of Model 1-3

- Stability of the model improved over the first one.
- TPM genes have high signal overall.
- VDBP remains clear indicator and significant.
- Beat strong baselines trained on similar data.
- Accuracy Improvement with each added modality:
  - clinical/psychosocial + cytokine = 65 %
  - clinical/psychosocial + cytokine + SNPs = 72%
  - clinical/psychosocial + cytokine + SNPs + TPMs = 80%

Study (1st author, year)	Journal	N total	N PTB	N term	Data types	Best / main model	Key reported metrics
Nowak 2025	Biological Research for Nursing	20	12	8	Early-pregnancy maternal blood: DNA methylation + mRNA (multi-omic)	PLS-DA (exploratory multivariate model)	R^2 = 0.92, Q^2 = 0.78 (no AUC/accuracy)
Kloska 2025	Scientific Reports	50	28	22	Routine blood labs (CBC, CRP, etc.) + questionnaire (lifestyle, obstetric history)	Linear SVM (boosted)	Acc 82%, Prec 83%, Recall 86%, F1 84% (5-fold CV)
Han 2025	Comput. Struct. Biotechnol. J.	150	48	102	Third-trimester maternal serum untargeted metabolomics	XGBoost with bootstrap	AUROC 0.85 (95% CI 0.57–0.99)
Tobin 2023	Metabolomics	100 mother–infant dyads	~50	~50	Maternal plasma + maternal & infant dried blood spots (untargeted metabolomics), stratified by ART regimen	Random forest (per treatment group)	Accuracy 95.5%, 95.7%, 80.7% for 3 ART groups (AUC not numeric in text)
Jehan 2020	JAMA Network Open	81	39	42	Early-pregnancy plasma cfRNA + plasma proteomics + urine metabolomics (multi-omic)	Integrated multi-omics ML model	AUROC 0.83 (95% CI 0.72–0.91); cfRNA 0.73, metabolomics 0.59, proteomics 0.75

#	Study (year)	N total (PTB / term)	Data types	Maternal age	Race / ethnicity	Best / main model	Key reported metrics	Metric type (short)
1	2025	174	PTB/term: n/a	Clinical + plasma cytokines (T1/T2B)	Not reported	Logistic Regresion	Acc, ROC, AUC	Acc 0.65 ±0.11; ROC AUC 0.70 ±0.11; PR AUC 0.57 ±0.17
2	2025	151	PTB/term: n/a	Clinical + cytokine + rs73160620, rs11980802	Not reported	Logistic Regresion	Acc, ROC, AUC	Acc 0.73 ±0.04; ROC AUC 0.78 ±0.09; PR AUC 0.67 ±0.13
3	2025	59	PTB/term: n/a	Clinical + cytokine + SNPs + top-var RNA-seq TPM	Not reported	Logistic Regresion	Acc, ROC, AUC	Acc 0.80 ±0.09; **ROC AUC 0.92 ±0.03; PR AUC 0.93 ±0.04**

Permutation Enrichment Testing

```
Testing genes:
  ENSG00000136235 (GPNMB)  score=17.138
  ENSG00000169429 (CXCL8)  score=10.733
  ENSG00000130203 (APOE)  score=10.547
  ENSG00000105976 (MET)  score=10.207
  ENSG00000116285 (ERRFI1)  score=10.173
  ENSG00000108691 (CCL2)  score=10.062
  ENSG00000120875 (DUSP4)  score=10.005
  ENSG00000164684 (ZNF704)  score=10.035

Permutation enrichment
  Observed score:        88.899
  Empirical p-value:     0.0001
  Wald-style z-score:    3.719
  First 10 perm draws:   [38.57030068 38.88480728 43.72113354 42.02346892 37.47502869 34.07793069
 38.00828933 43.15553654 37.68880598 36.71734313]
```

DAVID Clustering analysis

Cluster	Cluster Enrichment Score	Category	Term	Count	List Total	Pop Hits	Pop Total	%	P-Value	Benjamini	Fold Enrichment
1	1.36	KEGG_PATHWAY	Malaria	3	5	50	8534	37.50	2E-04	1.16E-02	102.41
1	1.36	UP_KW_DOMAIN	Signal	5	6	4415	14625	62.50	3.15E-02	1.89E-01	2.76
1	1.36	GOTERM_CC_DIRECT	extracellular region	4	8	2313	20795	50.00	3.41E-02	1E+00	4.50
1	1.36	GOTERM_BP_DIRECT	G protein-coupled receptor signaling pathway	3	8	927	19478	37.50	4.05E-02	8.6E-01	7.88
1	1.36	UP_KW_CELLULAR_COMPONENT	Secreted	4	8	2217	18049	50.00	4.42E-02	4.42E-01	4.07
1	1.36	GOTERM_CC_DIRECT	extracellular space	3	8	1867	20795	37.50	1.25E-01	1E+00	4.18
1	1.36	UP_KW_PTM	Glycoprotein	4	8	4844	14316	50.00	4.41E-01	1E+00	1.48
1	1.36	UP_KW_PTM	Disulfide bond	3	8	3956	14316	37.50	6.18E-01	1E+00	1.36
2	0.98	GOTERM_BP_DIRECT	positive chemotaxis	3	8	52	19478	37.50	1.46E-04	3.7E-02	140.47
2	0.98	UP_KW_DOMAIN	Signal	5	6	4415	14625	62.50	3.15E-02	1.89E-01	2.76
2	0.98	GOTERM_CC_DIRECT	plasma membrane	4	8	5597	20795	50.00	2.84E-01	1E+00	1.86
2	0.98	UP_KW_PTM	Glycoprotein	4	8	4844	14316	50.00	4.41E-01	1E+00	1.48
2	0.98	UP_SEQ_FEATURE	CARBOHYD:N-linked (GlcNAc...) asparagine	3	8	4423	20675	37.50	4.61E-01	1E+00	1.75
2	0.98	GOTERM_CC_DIRECT	membrane	3	8	5415	20795	37.50	5.81E-01	1E+00	1.44
2	0.98	UP_KW_DISEASE	Disease variant	3	4	3955	4859	37.50	9.09E-01	1E+00	0.92
3	0.25	GOTERM_CC_DIRECT	plasma membrane	4	8	5597	20795	50.00	2.84E-01	1E+00	1.86
3	0.25	UP_KW_PTM	Phosphoprotein	6	8	8435	14316	75.00	3.97E-01	1E+00	1.27
3	0.25	UP_KW_CELLULAR_COMPONENT	Membrane	3	8	8353	18049	37.50	9.09E-01	1E+00	0.81
3	0.25	UP_SEQ_FEATURE	REGION:Disordered	4	8	13820	20675	50.00	9.56E-01	1E+00	0.75

Conclusions

- Our model successfully incorporate clinical/psychosocial, cytokines and genomic data for a targeted cohort and achieve strong baselines
- TPM has very high signals in genes
- Model sees complicated relations not easily interpretable to clinicians
- Without genomic signals, Discrimination is the highest contributor in SHAP model 1

Limitation of the Study

- Sample size is small
- Some sample had to be discarded to avoid introducing noise
- Model Stability can still be improved
- Other types of architecture for model which could give better overall metrics had to be dropped in favour of interpretability and stability.

Future Directions

- More experimentation with genomic data addition
- Increase sample size and rerun the experiments to verify generalizability
- Test on different models
- Better interpretation of the Permutation Enrichments and DAVID clustering

Thank You!