Progress Report 1

PM 606 Project

Vighnesh
University of Southern California

June 19, 2025

Provisional Title

Integrating Environmental Exposures and Molecular Omics to Predict Pediatric Metabolic Health Risk Using the HELIX Cohort

Background

The HELIX (Human Early Life Exposome) study is a large European cohort integrating environmental exposure data and multi-omics profiles in children. This enables the study of how early-life exposures and biological pathways jointly affect health outcomes. We focus on predicting pediatric metabolic health risk — a complex, multifactorial phenotype — and aim to evaluate whether molecular data (transcriptomics, methylomics, metabolomics) add predictive value beyond environmental exposures alone.

Research Question

Can molecular omics data improve the prediction of pediatric metabolic health risk beyond environmental exposures alone?

Data Overview

- **Exposure Data:** Air pollution (NO_2, $PM_{2.5}$, PM_{10}), physical activity, socioeconomic variables, green space metrics.
- **Omics Data:** Transcriptomics, methylomics, and metabolomics from the HELIX 2018 multi-omics release.
- **Outcome:** A binary label derived from standardized z-scores (BMI, waist circumference, HOMA-IR, systolic BP).

Descriptive Summaries

Exploratory analysis included:

- Distribution and summary statistics for key exposure variables.
- Correlation matrix among exposure variables.
- Missingness map and imputation strategy under development.
- Z-score based thresholding for metabolic risk.

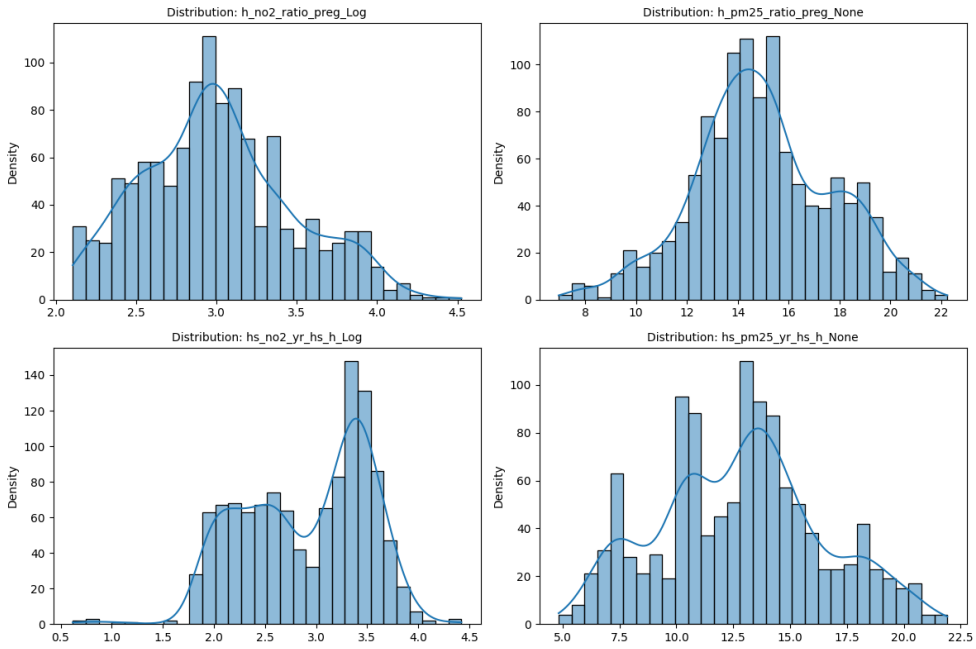

Figure 1: Distribution of Selected Exposure Variables

Scientific Justification

Previous studies on HELIX have often examined single exposure-outcome relationships. Our work advances this by building integrated predictive models using both environmental and omic features. This has clinical and commercial implications for pediatric health stratification tools.

Analysis Plan

We compare three supervised learning models:

1. **Exposure-only model**
2. **Omics-only model**
3. **Combined model (Exposures + Omics)**

Algorithms Used:

- Logistic Regression (baseline)
- LASSO (for feature selection)
- XGBoost (performance and explainability via SHAP)

Evaluation Metrics:

- Accuracy, Precision, Recall, ROC AUC
- SHAP values for feature importance

Preliminary Results

Exposure-Only Model

- **XGBoost:**
  - Accuracy: 0.87
  - Precision: 0.87
  - Recall: 0.96
  - ROC AUC: 0.96
- **Confusion Matrix:**

Table 1: Confusion Matrix - XGBoost (Exposure-Only)

	Predicted 0	Predicted 1
Actual 0	28	13
Actual 1	4	86

LASSO Logistic Regression

- Accuracy: 0.78
- Precision: 0.81
- Recall: 0.89

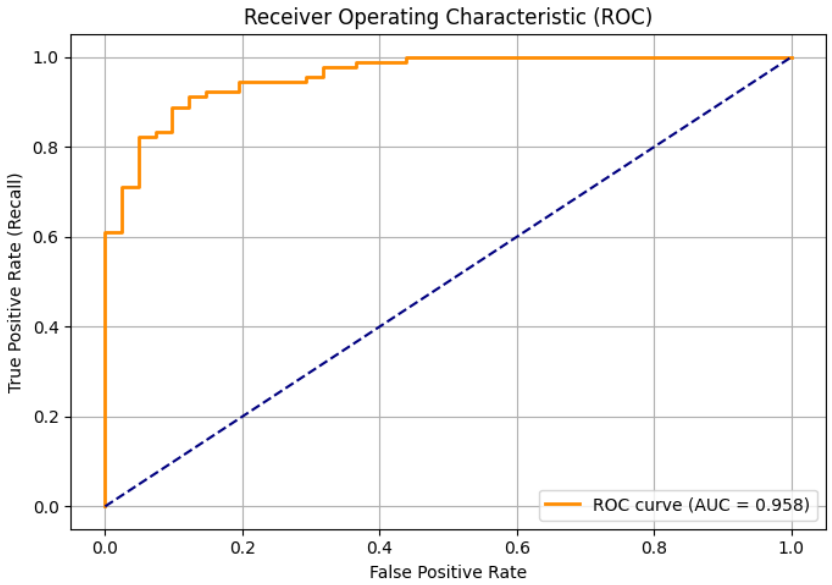

Figure 2: ROC Curve - Exposure-Only Models

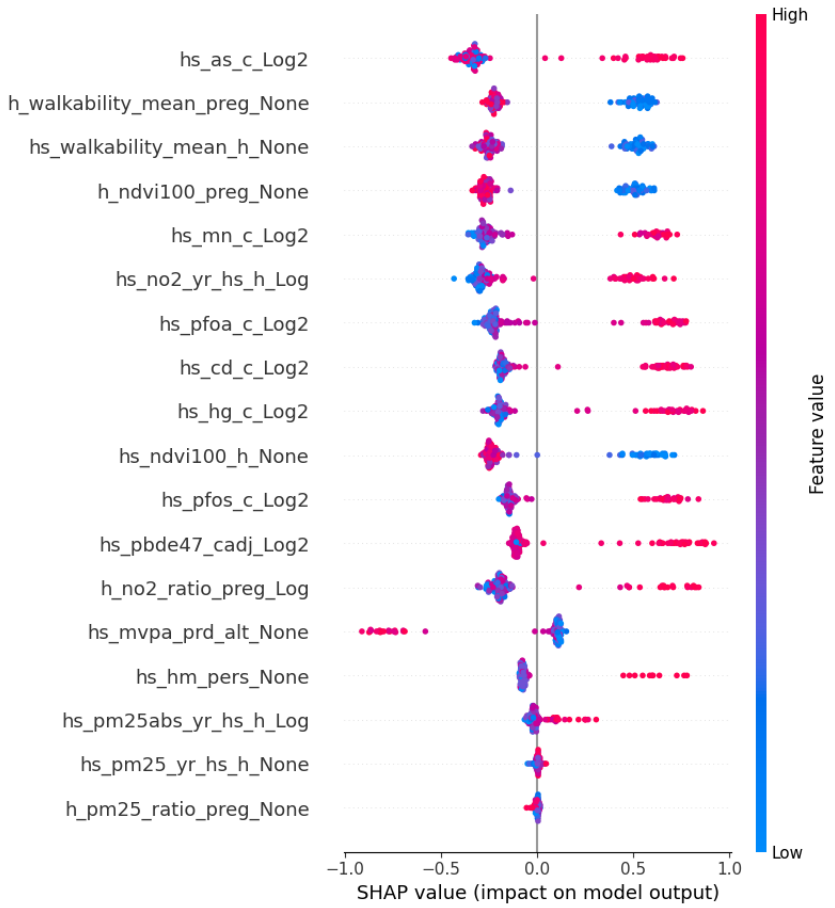

Figure 3: SHAP Summary Plot - Top Exposure Predictors

Interpretation of Exposure-Only Model

The exposure-only model, trained on air pollution, built environment, and activity-related variables, demonstrates strong discriminatory power in predicting high pediatric metabolic risk. The XGBoost classifier achieved an accuracy of 87% and a ROC AUC of 0.96, indicating that the model is highly effective at separating children with elevated metabolic risk from those at normal risk based on exposures alone.

The confusion matrix reveals a high true positive rate (recall = 96%), suggesting the model is particularly good at correctly identifying at-risk children. However, there is a moderate false positive rate (13 cases where low-risk children were predicted to be high-risk), which may reflect overlapping exposure profiles or threshold misalignment. This trade-off is typical in clinical screening tools where sensitivity is prioritized.

The SHAP summary plot further revealed which environmental factors most contributed to risk prediction. Key predictors included long-term NO_2 exposure and $PM_{2.5}$ levels during pregnancy, aligning with prior literature linking air pollution to childhood adiposity and insulin resistance. This supports the biological relevance of the model and builds confidence in the interpretability of the features.

While these results are promising, they likely capture only part of the risk profile. Environmental exposures alone cannot explain the full variability in pediatric metabolic health. The addition of omics features in subsequent models will test whether underlying molecular signatures offer additive predictive value beyond the exposome.

Next Steps

- Complete omics data preprocessing (normalization, QC)
- Train and evaluate omics-only model

- Train and evaluate combined model
- Compare models using statistical tests (e.g., DeLong test for AUC)
- Begin writing Methods and Results sections for Progress Report 2