Exposure-Weighted Multi-Omics Integration for Pediatric Metabolic Risk Profiling: An Exploratory Study in the HELIX Cohort

A biologically informed framework for early risk detection

Vighnesh Sairaman

August 12, 2025

University of Southern California
sairaman@usc.edu

Research Problem – Context and Gaps

Challenge:

- Pediatric metabolic dysfunction is escalating globally
- Traditional metrics (BMI, waist circumference) reflect late-stage phenotypes
- Environmental exposures shape early-life metabolic programming
- Need for exposure-aware predictive models

Gap:

- Conventional approaches miss underlying molecular mechanisms
- Limited integration of multi-omics data with environmental exposures

Our Solution – Biologically Informed Framework

We propose an integrative approach combining:

- **Environmental exposures:** Air pollutants, lifestyle, built environment
- **Effect-weighted multi-omics risk scores:** Methylome, miRNA, transcriptome, proteome, metabolome
- **Machine learning models:** XGBoost, LASSO, SHAP-based interpretation

Goal: Predict pediatric metabolic risk early using biologically grounded and interpretable models

HELIX Cohort Overview

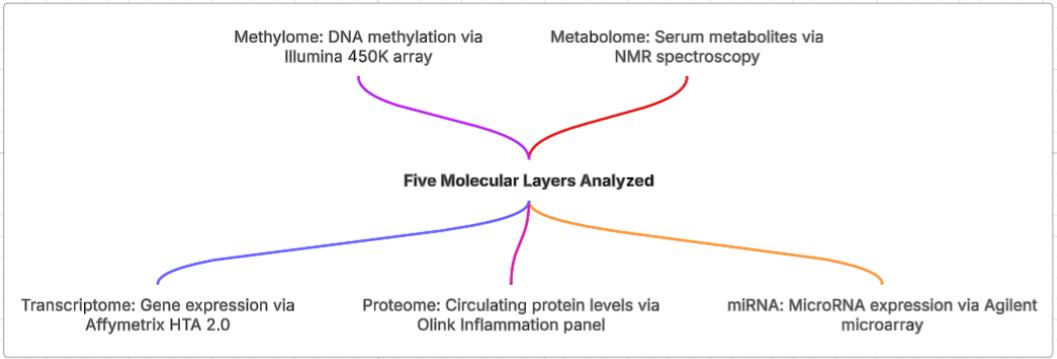
Study Population:

- **1,300+** European children from 6 harmonized birth cohorts
- BiB (UK), EDEN (France), INMA (Spain), KANC (Lithuania), MoBa (Norway), RHEA (Greece)
- Measurements collected at both **prenatal and childhood** stages

Environmental Exposures (n = 217):

- **Air pollution:** NO_2, $PM_{2.5}$, PM_{10}, benzene
- **Toxicants:** Phthalates, phenols, PFASs, heavy metals
- **Lifestyle:** Diet score, physical activity, organic food
- **Built environment:** Green space, walkability, noise
- **Socioeconomic:** Education, housing, financial hardship

Multi-Omics Profiling in HELIX



Time Points: All omics measured in both **pregnancy** and **childhood** windows

Methodological Framework

Risk Score Calculation:

For each omics layer l and child j:

$$\text{Risk}_j^{(l)} = \sum_{i=1}^{n_l} x_{i,j} \cdot \beta_i^{(l)} \cdot (-\log_{10}(p_i^{(l)}))$$

Where:

- $x_{i,j}$: exposure level (z-score)
- $\beta_i^{(l)}$: effect size from HELIX studies
- $p_i^{(l)}$: statistical significance

Composite Score:

$$\text{Risk}_j^{\text{total}} = \sum_{l=1}^{5} w_l \cdot \text{Risk}_j^{(l)}$$

Omics Layer Weights:

- Methylome: 30%
- Proteome: 20%
- Metabolome: 20%
- Transcriptome: 15%
- miRNA: 15%

Binary Outcome:

- Risk scores scaled 0-100%
- Threshold at 75th percentile
- "At-risk" vs "Not at-risk" classification

Machine Learning Models

Two Model Configurations:

a. **Exposure-only:** 217 standardized exposure variables
b. **Combined:** Exposure variables + 5 omics-derived risk scores

LASSO	XGBoost	Random Forest
-> L1 regularization	-> Gradient boosting	-> Bootstrap aggregation
-> Sparse feature selection	-> Nonlinear relationships	-> Ensemble of trees
-> Linear decision boundary	-> Feature interactions	-> Handles noisy features
->High interpretability	->Robust to multicollinearity	-> Variance reduction
Validation: 20-fold stratified cross-validation | SHAP analysis for interpretability | Metrics: AUC, Accuracy, Precision, Recall, F1-Score		

Model Performance – Exposure-Only Models

Features: 217 standardized environmental exposure variables

Model	AUC	Accuracy	Recall
Logistic Regression	0.867	0.771	0.878
XGBoost	0.958	0.870	0.956
Random Forest	0.916	0.847	0.978

Key Insights:

- Strong performance from XGBoost (AUC: 0.958, Recall: 0.956) indicates non-linear exposure-risk relationships.
- Random Forest achieves the highest recall (0.978) but slightly lower AUC, suggesting broader but less precise classification.
- Logistic Regression underperforms non-linear methods due to limited ability to model interactions.
- Serves as the baseline to assess added value from omics-informed features.

Model Performance – Combined Omics + Exposure Models

Features: 217 exposure variables + 5 omics-derived risk scores (Methylome, miRNA, Transcriptome, Proteome, Metabolome)

Model	AUC	Accuracy	F1-Score
Logistic Regression	0.997	0.968	0.936
XGBoost	0.986	0.940	0.872
Random Forest	0.961	0.875	0.671

Key Insights:

- Logistic Regression achieves near-perfect discrimination (AUC: 0.997), showing that omics-weighted scores provide highly separable risk classes.
- Gains in performance reflect the biological grounding of omics scores, which capture immune, metabolic, and signaling pathway disruptions.
- Enhanced interpretability: each omics layer's contribution can be visualized at the child level.

Feature Importance – LASSO + SHAP Convergence

LASSO Feature Selection:

- Optimal penalty parameter (λ): 2.7826
- Non-zero coefficients retained: 65 out of 217+ features
- Balanced trade-off between predictive performance and sparsity

Top Predictive Features:

Chemical Exposures:

- `hs_pfna_c_Log2`: PFNA (perfluorononanoic acid)
- `h_PM_Log`: Particulate matter ($PM_{2.5}$/PM_{10})
- `hs_pfoa_m_Log2`: PFOA (perfluorooctanoic acid)
- `hs_cd_m_Log2`: Cadmium (heavy metal)

Omics Risk Scores:

- `total_metabo_risk`: Metabolomic disruption
- `protein_child_risk`: Childhood proteomic alterations
- `total_transcript_risk`: Gene expression shifts
- `mirna_child_risk`: MicroRNA dysregulation

Biological Interpretation of Top Features

Chemical Exposures:

- **PFAS (PFNA, PFOA):** Persistent organic pollutants linked to endocrine disruption, lipid metabolism dysregulation, and immune modulation.
- **Cadmium (Cd):** Heavy metal with strong evidence for mitochondrial toxicity and impaired insulin signaling.
- **Air Pollution ($PM_{2.5}$/PM_{10}):** Induces oxidative stress, systemic inflammation, and vascular dysfunction.

Omics Risk Scores:

- **Metabolomics:** Amino acid and neurotransmitter pathway imbalance — markers of metabolic inflexibility.
- **Proteomics:** Cytokine activity and biosynthetic pathway activation indicating chronic low-grade inflammation.
- **Transcriptomics & miRNA:** Immune system activation, cell cycle regulation, and apoptotic pathway engagement.

Pathway Enrichment Analysis: Methods

Biological Validation Approach:

- Features ranked by composite score: $\beta \cdot (-\log_{10}(p))$
- Layer-specific enrichment pipelines (GSEA, miEAA, MetaboAnalyst)
- Separate analysis for prenatal vs childhood periods

Multi-Omics Integration Strategy:

- Four complementary molecular layers analyzed
- Period-specific enrichment patterns identified
- Cross-platform validation of biological themes
- Statistical significance across all data types

Pathway Enrichment Analysis: Results

Transcriptome:

- B cell activation
- T-cell receptor signaling
- Glutathione metabolism
- Cell cycle regulation

Proteome:

- Cytokine production
- Phosphorylation cascades
- Transcriptional regulation
- Biosynthetic pathways

Methylome:

- T-cell differentiation
- Neurodevelopmental patterning
- Synaptic signaling
- Embryonic morphogenesis

Metabolome:

- Amino acid metabolism
- Neurotransmitter transport
- Na/Cl-dependent transporters
- Membrane translocation

Clinical Relevance: Consistent enrichment across four major domains: immune activation, oxidative stress, biosynthetic regulation, and neurodevelopment—all hallmarks of metabolic dysfunction.

Model Interpretability: SHAP Analysis

SHAP (SHapley Additive exPlanations):

- Quantifies marginal contribution of each feature to individual predictions
- TreeSHAP optimized for ensemble models
- Local accuracy and consistency guaranteed

Key Insights from SHAP Analysis:

Exposure-Only Model:

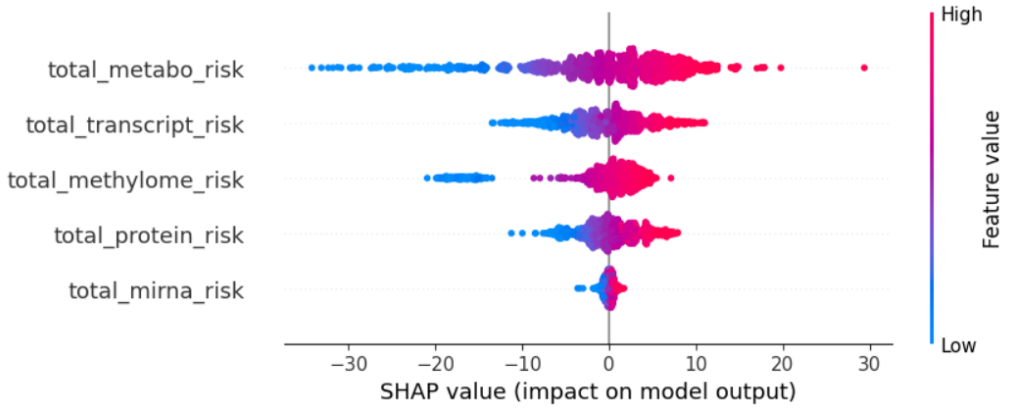
- Modest feature impacts (SHAP range: ±5)
- Surface-level environmental associations
- Limited mechanistic context
- Primary drivers: PFAS, arsenic, walkability

Combined Model:

- Stronger feature impacts (SHAP range: ±30)
- Biologically informed risk scores dominate
- `total_metabo_risk`: highest impact
- `total_transcript_risk`: second highest

Mechanistic Interpretation from SHAP:

- Omics-derived features capture molecular-level perturbations
- Environmental exposures provide upstream context
- Combined approach reveals exposure-to-biology cascade

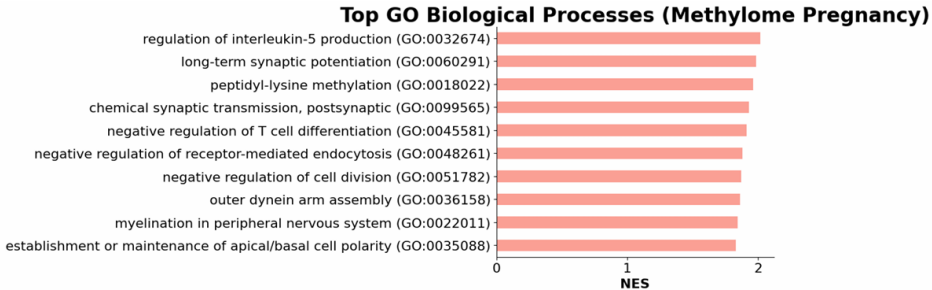


Biological Plausibility

Cross-Layer Validation:

- Pathway enrichment confirms biological relevance of selected features
- Consistent themes across five omics layers and two life stages
- Features map to established mechanisms of metabolic disease

Clinical Significance: These pathways are well-established drivers of pediatric obesity, insulin resistance, and cardiovascular disease.



Mechanistic Pathways Identified:

Immune System:

- B and T-cell activation
- Cytokine production cascades
- Inflammatory protein signaling
- Immune surveillance pathways

Metabolic Regulation:

- Amino acid metabolism
- Energy substrate processing
- Mitochondrial function
- Glucose homeostasis

Oxidative Stress:

- Glutathione metabolism
- Antioxidant defenses
- Free radical scavenging
- Cellular protection mechanisms

Neurodevelopment:

- Synaptic signaling
- Neurotransmitter transport
- Brain morphogenesis
- Neuroinflammatory responses

Scientific Contributions: Methodological Innovation

Methodological Innovation:

- Exposure-weighted multi-omics risk score construction across five layers (methylome, transcriptome, proteome, metabolome, miRNA)
- Manual mapping of omics-exposure associations to harmonized HELIX base variables
- Weighted integration of omics layers based on biological stability and functional relevance
- Internal validation via pathway enrichment as a biological plausibility check

Technical Advances:

- Dimensionality reduction through LASSO with convergence to SHAP-identified features
- 20-fold stratified cross-validation for robust performance estimation
- Risk stratification into minimal, moderate, and high tiers with zoned visualizations
- Integration of biological annotation pipelines (GSEA, miEAA, MetaboAnalyst) into ML workflow

Scientific Contributions: Biological Insights & Translation

Biological Insights:

- PFAS, cadmium, and particulate air pollution emerge as convergent high-impact exposures
- Omics signatures implicate immune activation, oxidative stress, biosynthesis, and neurodevelopment pathways
- Prenatal and childhood molecular profiles show consistent enrichment patterns across life stages
- Methylome and proteome layers contribute most to high-risk profiles

Translational Potential:

- Framework adaptable to other exposome-omics datasets and age groups
- Potential for early-life screening and targeted prevention
- Supports mechanistic evidence base for environmental health policy
- Enables integration into precision pediatric medicine approaches

Understanding

- **Novel Framework:** First study to combine harmonized exposome data with biologically weighted, multi-omics-derived risk scores for pediatric metabolic health stratification in the HELIX cohort.
- **Impact:** This framework bridges environmental exposure science with molecular epidemiology, providing a template for risk stratification that could transform pediatric metabolic disease prevention.

Future Directions – Research Extensions

Research Extensions:

- Validate framework using **longitudinal outcomes** (e.g., zBMI, HOMA-IR, lipid profiles).
- Replicate across **multi-ethnic cohorts** with comparable exposome–omics data (e.g., ALSPAC, Project Viva).
- Integrate clinical endpoints for **predictive modeling beyond risk stratification**.
- Apply to intervention studies to assess modifiability of identified high-risk pathways.

Immediate Next Steps:

a. External validation in independent datasets.
b. Prospective tracking of clinical outcomes over time.
c. Cost-effectiveness assessment for public health implementation.

Future Directions – Technical Extensions

Technical Extensions:

- Develop **dynamic risk score updating** with new exposure or omics data.
- Incorporate **real-time exposure monitoring** (e.g., wearable sensors).
- Build mobile health tools for **personalized feedback and prevention guidance**.
- Explore **federated learning** for cross-cohort integration without data sharing.

Key Findings – Primary Results

Primary Results:

a. **Superior Performance:** Combined omics-exposure models achieved AUC = 0.997 vs 0.958 for exposure-only.
b. **Feature Efficiency:** 5 omics scores + select exposures outperformed 217 raw exposure variables.
c. **Biological Validity:** Pathway enrichment confirmed mechanistic relevance across all omics layers.
d. **Interpretable Predictions:** SHAP analysis revealed clear exposure-to-biology pathways.

Key Findings – Environmental Molecular Drivers

Critical Environmental Factors:

- PFAS chemicals (PFNA, PFOA): endocrine disruption, bioaccumulation.
- Air pollution ($PM_{2.5}$, PM_{10}): oxidative stress, inflammation.
- Heavy metals (cadmium): toxic accumulation, metabolic interference.

Molecular Signatures:

- Metabolomic disruption: energy metabolism, amino acid processing.
- Proteomic alterations: immune signaling, inflammatory cascades.
- Transcriptomic changes: gene expression, cellular stress responses.

Clinical Translation – Early Warning Short-Term Pathway

Early Warning System:

- Identifies at-risk children before clinical symptoms.
- Provides mechanistic insight for targeted interventions.
- Enables personalized preventive strategies.

Implementation Pathway – Short-term (1–2 years):

- Clinical validation studies.
- Biomarker panel optimization.
- Healthcare provider training.
- Pilot screening programs.

Clinical Translation – Long-Term Pathway Public Health Impact

Implementation Pathway – Long-term (3–5 years):

- Population-wide screening.
- Electronic health record integration.
- Public health policy applications.
- Environmental intervention studies.

Public Health Impact:

- Reduced healthcare costs through prevention.
- Decreased pediatric obesity rates.
- Environmental exposure reduction policies.
- Health equity improvements through early identification.

Conclusions

We demonstrated that:

a. Biologically informed omics-exposure integration significantly enhances pediatric metabolic risk prediction
b. Effect-weighted risk scores provide both statistical power and mechanistic interpretability
c. Multi-layer pathway enrichment validates the biological relevance of predictive features

Clinical Significance:

- Enables early identification of at-risk children before clinical manifestation
- Provides actionable molecular targets for intervention
- Supports precision medicine approaches in pediatric metabolic health

Next Steps: External validation, prospective clinical studies, and translation to clinical practice.

Thank you for your attention

Questions and Discussion

Vighnesh Sairaman
University of Southern California
sairaman@usc.edu