Progress Report 3

PM 606 Project

Vighnesh
University of Southern California

July 21, 2025

Title of the Paper

Biologically-Informed Integration of Multi-Omics Signatures for Stratifying Pediatric Metabolic Risk in the HELIX Cohort

Research Question

Can the biologically-informed integration of multi-omics signatures—weighted by evidence-based estimates of molecular stability and functional relevance—yield coherent metabolic risk profiles that map onto established pathways of dysfunction in children from the HELIX cohort?

Introduction (Draft)

Pediatric metabolic dysfunction is a complex and multifactorial condition that often eludes early detection using standard clinical metrics such as body mass index (BMI) or waist circumference. These conventional measures typically reflect downstream physiological changes and may not capture the early molecular disruptions that precede overt symptoms.

The Human Early Life Exposome (HELIX) study presents a unique opportunity to investigate the developmental origins of metabolic risk by combining rich environmental exposure data with high-dimensional molecular profiles from multiple omics layers—methylation, transcriptomics, miRNA, proteomics, and metabolomics—collected during prenatal and childhood windows.

In this study, we propose a biologically-informed integration framework that constructs per-child metabolic risk profiles using multi-omics signatures derived from exposure-omics associations. These signatures are weighted according to literature-derived estimates of biomarker stability and functional relevance. Our aim is to assess whether this integration produces coherent and interpretable risk profiles that reflect known pathways of metabolic dysfunction.

We evaluate the internal consistency of these profiles, examine their ability to stratify risk beyond traditional exposure-only models, and validate the biological plausibility of the results through pathway enrichment analyses. This approach bridges environmental epidemiology with molecular systems biology, offering a more mechanistic understanding of early-life drivers of pediatric metabolic risk.

Methods

Data Sources and Processing

Exposure and omics data were obtained from the HELIX 2018 multi-omics release. Omics layers included methylome, transcriptome, proteome, metabolome, and miRNA expression profiles, collected during both pregnancy and childhood.

- All omics datasets were cleaned by removing features with missing, zero, or invalid statistics.
- Exposures were standardized using `StandardScaler`.
- Manual mapping aligned exposure names in omics summary statistics to HELIX base columns.

Risk Score Construction

For each omics entry, risk weights were computed as:

$$\text{Weight} = \beta \cdot (-\log_{10}(p))$$

For each child, weighted scores were computed by summing standardized exposure values multiplied by effect-size weights. Individual scores were computed for each omics type and window. A composite risk score was then calculated by weighted integration:

- Composite weights: **Methylome** (30%), **miRNA** (15%), **Transcriptome** (15%), **Proteome** (20%), **Metabolome** (20%).
- Scores were normalized to a 0–100% scale for comparability across children.
- Weights were chosen based on a combination of biological relevance (e.g., stability, functional impact) and exploratory analysis of effect size distributions across omics layers.

Pathway Enrichment and Validation

We conducted downstream biological validation using enrichment analysis for each omics layer to verify whether risk-driving molecular signals corresponded to known disease mechanisms. The methodology and results are detailed below.

[Enrichment scores shown as $-\log_{10}(p)$, capped at 2 for visualization.]

1. Transcriptome

Method: Transcript-level differential exposure associations were first aggregated to the gene level. For each gene, a ranking score was computed using $-\log_{10}(p) \times$ effect size to capture both strength and reliability of association. Only genes with valid HGNC symbols were retained. Gene Set Enrichment Analysis (GSEA) was then applied using the GO Biological Process 2021 ontology to identify overrepresented functional pathways among high-ranking genes.

Pregnancy Findings: Enriched pathways included B cell activation, G1/S cell cycle checkpoint control, and cytoplasmic translation. These signatures reflect heightened immune ontogeny and proliferative regulation during fetal development — a period sensitive to immunotoxic and endocrine-disrupting exposures.

Childhood Findings: Key biological processes included glutathione metabolism, T-cell receptor signaling, and programmed cell death. This indicates sustained engagement of antioxidant defenses and immune maturation pathways — mechanisms closely tied to metabolic inflammation and tissue homeostasis.

Interpretation: The transcriptome represents a dynamic snapshot of active gene expression and is uniquely positioned to reveal functional cellular responses to environmental insults. Here, the exposure-weighted transcriptomic risk scores identify immune and redox-sensitive signaling programs with known links to pediatric metabolic health. Their consistency across developmental stages and alignment with disease-relevant biology enhances confidence in their use for early risk stratification.

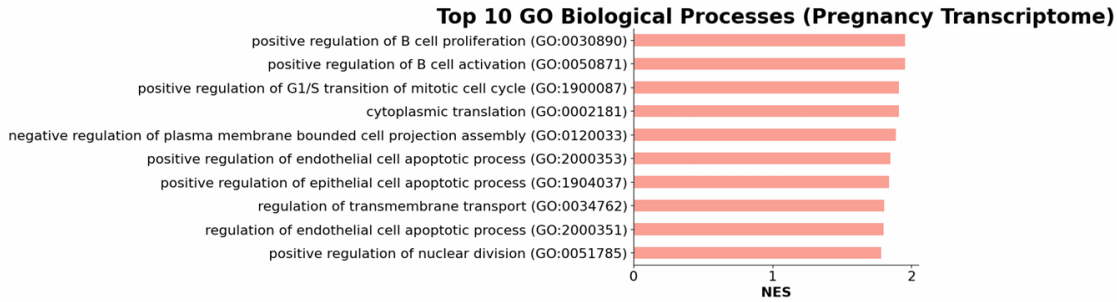

Figure 1: Transcriptome Pregnancy Enrichment — Top GO Biological Processes

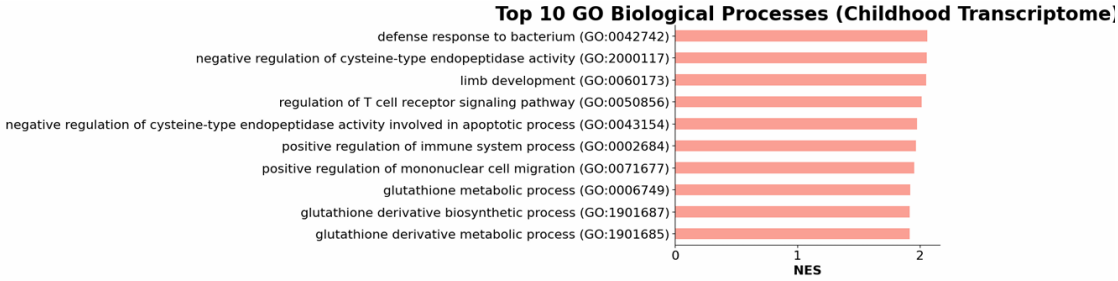

Figure 2: Transcriptome Childhood Enrichment — Top GO Biological Processes

2. Methylome

Method: Each CpG probe associated with environmental exposures was mapped to genes using the UCSC RefGene annotation table. Since multiple CpGs can map to the same gene, we retained only the most significant CpG (highest $-\log_{10}(p) \times$ effect score) per gene to avoid redundancy. Gene Set Enrichment Analysis (GSEA) was then performed on the resulting ranked gene list using the GO Biological Process 2021 library.

Pregnancy Findings: Top pathways included regulation of T-cell differentiation, neurodevelopmental patterning, and epithelial cell polarity—suggesting epigenetic priming of immune function and tissue architecture during fetal development.

Childhood Findings: Key enriched terms centered around synaptic signaling, renal and reproductive tract development, and embryonic morphogenesis, pointing to persistent epigenetic regulation of organ systems beyond birth.

Interpretation: DNA methylation is a long-term molecular memory of early environmental exposures. The enriched pathways in both prenatal and childhood periods support the concept of developmental epigenetic programming, where environmental factors shape immune and neuronal susceptibility. These signatures are particularly compelling as early predictors of metabolic dysfunction, given the central role of immune and neuroendocrine regulation in energy balance and inflammation.

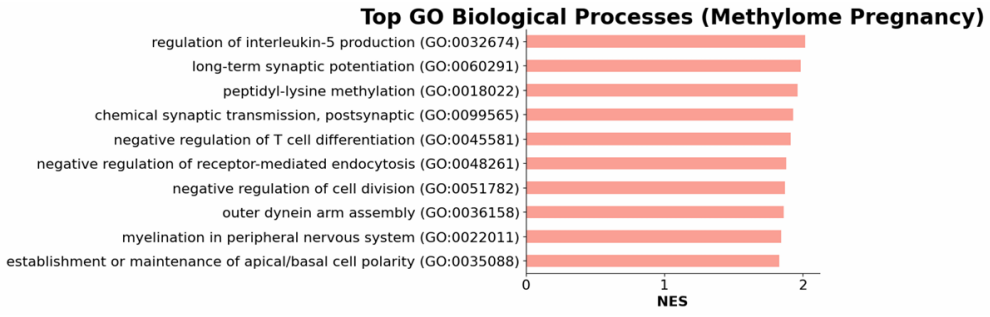

Figure 3: Methylome Pregnancy Enrichment — Top GO Biological Processes

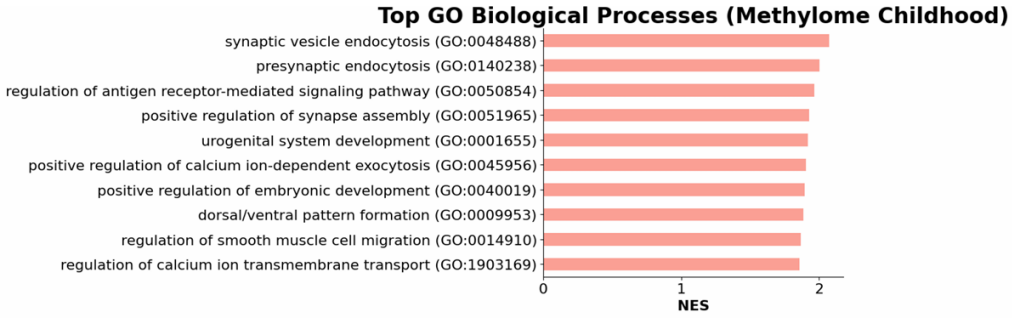

Figure 4: Methylome Childhood Enrichment — Top GO Biological Processes

3. miRNA

Method: We began with miRNA-level exposure associations, where each miRNA's ranking score was calculated as $-\log_{10}(p) \times$ effect size. Using curated interaction databases such as miRTarBase and TargetScan, each miRNA was mapped to its validated gene targets. For each gene, we computed an average regulatory score from all targeting miRNAs. The resulting ranked gene list was submitted to miEAA 2.1 (microRNA Enrichment Analysis and Annotation), using disease ontologies and GO biological categories.

Pregnancy Findings: Top enriched terms were predominantly tumor-related, including carcinoma (lung, kidney), liposarcoma, and ischemic vascular events. These signatures suggest early shifts in vascular integrity and tumor-suppressor signaling—potentially representing developmental priming toward later-life metabolic or vascular risk.

Childhood Findings: Dominant enrichments were associated with central nervous system and immune-linked conditions such as acute myeloid leukemia, multiple sclerosis, and cardiovascular disease. This reflects broad systemic dysregulation in immune surveillance, CNS inflammation, and endothelial signaling.

Interpretation: miRNAs act as upstream post-transcriptional regulators, making them powerful indicators of subtle dysregulation. Here, miRNA-derived risk scores appear to capture early systemic stress signatures—particularly those affecting vascular and immune homeostasis. Their predictive relevance lies in flagging preclinical dysfunction in pathways often linked to metabolic, neurodevelopmental, and autoimmune trajectories.

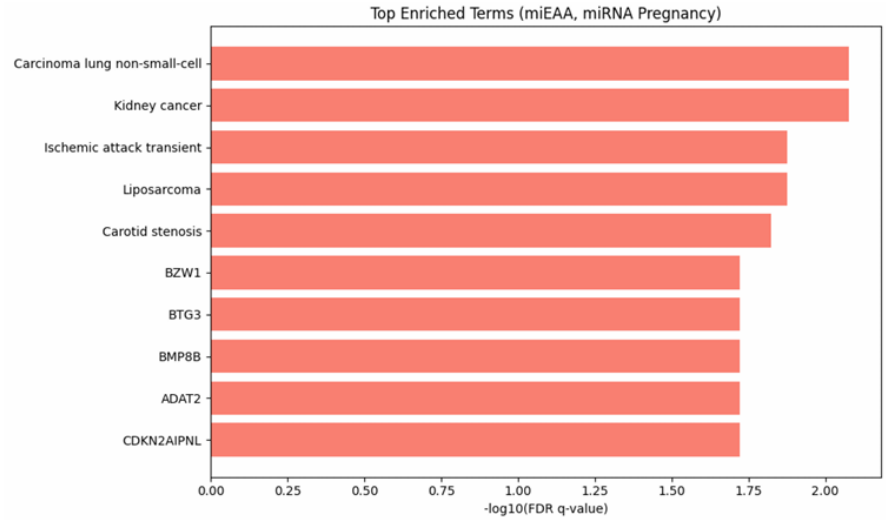

Figure 5: miRNA Pregnancy Enrichment — Disease Ontologies (miEAA)

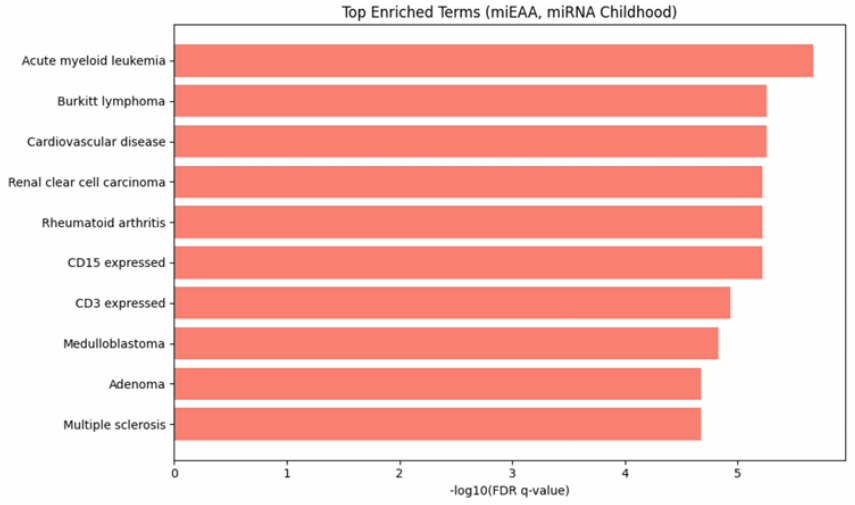

Figure 6: miRNA Childhood Enrichment — Disease Ontologies (miEAA)

4. Proteome

Method: Protein-level exposure effect sizes were combined with statistical significance to compute ranking scores ($-\log_{10}(p) \times$ effect). Only proteins with valid gene symbols or UniProtKB identifiers were retained. Gene Set Enrichment Analysis (GSEA) was performed using the GO Biological Process 2021 database to identify enriched functional pathways.

Pregnancy Findings: Top pathways involved positive regulation of cytokine production, protein phosphorylation, and cellular proliferation. These suggest immune priming, intracellular signaling, and cell growth during fetal development.

Childhood Findings: Enriched terms included regulation of transcription, biosynthetic processes, and cell migration—pointing to active metabolic remodeling and tissue development during postnatal growth.

Interpretation: Proteomic risk scores capture shifts in immune signaling and biosynthesis that are foundational to metabolic regulation. The stage-specific enrichment (cytokine activity in utero vs. biosynthesis postnatally) highlights dynamic proteomic adaptation in response to early environmental exposures. These molecular footprints likely precede and influence the onset of metabolic dysfunction, further supporting their inclusion in the composite risk stratification model.

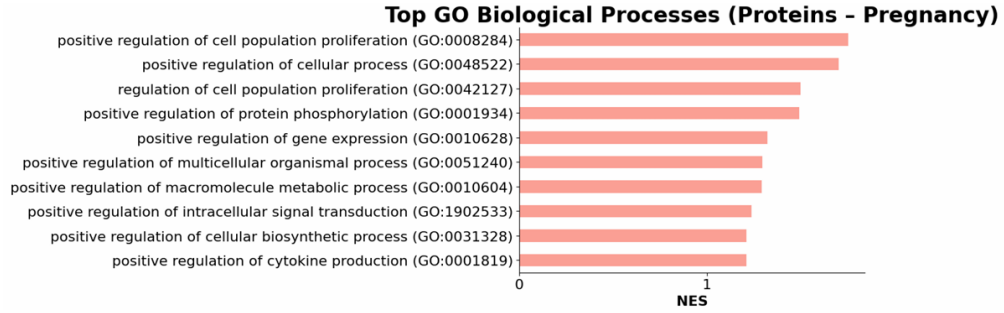

Figure 7: Proteome Pregnancy Enrichment — GO Biological Processes

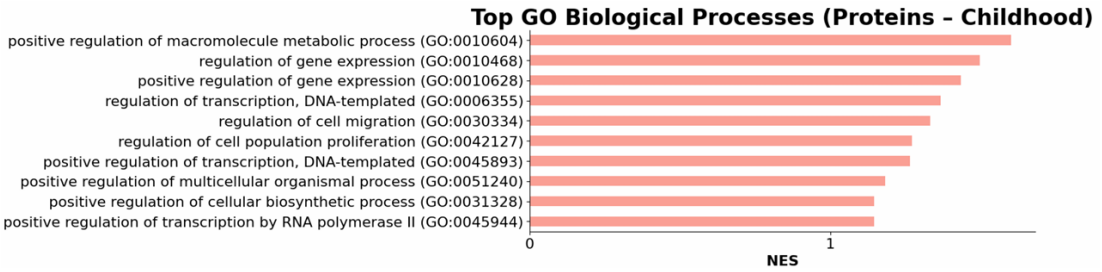

Figure 8: Proteome Childhood Enrichment — GO Biological Processes

5. Metabolome

Method: We applied Over-Representation Analysis (ORA) using the MetaboAnalyst platform. Ranked metabolite lists were generated by multiplying exposure effect sizes with significance ($-\log_{10}(p)$). The SMPDB (Small Molecule Pathway Database) library was selected for pathway mapping due to its detailed annotation of human serum metabolites relevant to pediatric health.

Pregnancy and Childhood Findings: Interestingly, both developmental windows—prenatal and childhood—revealed near-identical enrichment patterns. Top pathways included amino acid metabolism, neurotransmitter transport (e.g., Na^+/Cl^--dependent transporters), and membrane translocation systems. These pathways are central to energy balance, cellular signaling, and biosynthetic regulation.

Interpretation: The high consistency across life stages suggests durable metabolic alterations rooted in early-life exposures. This reinforces the hypothesis that metabolic programming begins before birth and persists through childhood. The affected pathways also align with hallmark features of early metabolic dysregulation, such as altered amino acid handling and neurotransmitter imbalance—underscoring their inclusion in the final risk stratification framework.

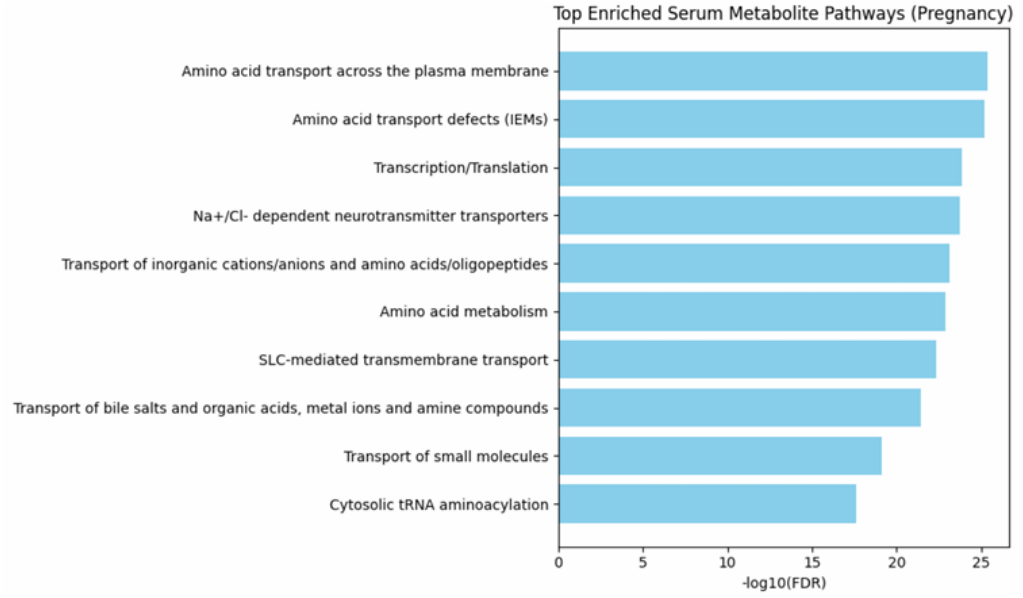

Figure 9: Metabolome Pregnancy Enrichment — Top Pathways (SMPDB)

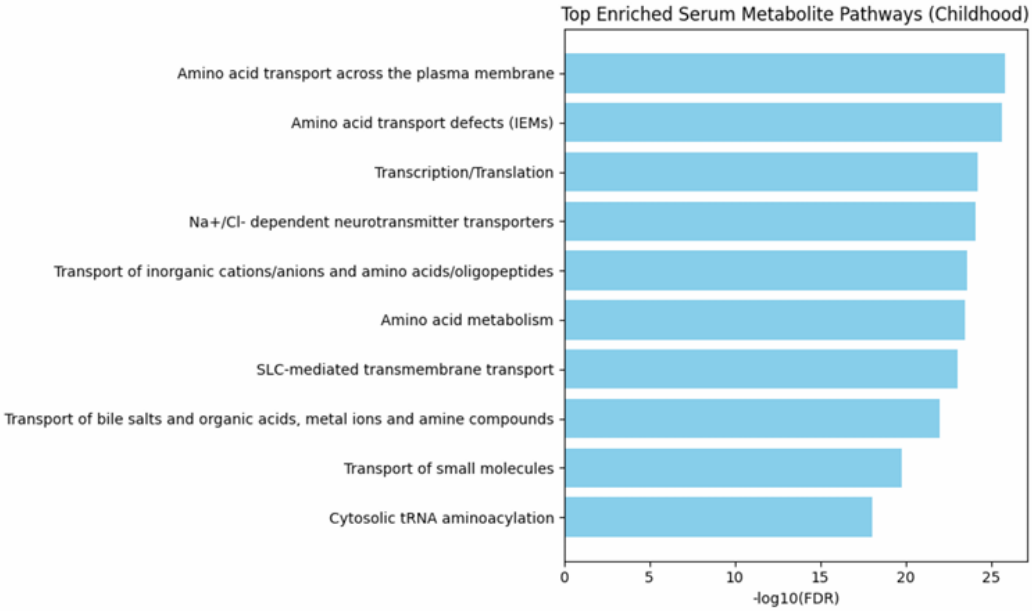

Figure 10: Metabolome Childhood Enrichment — Top Pathways (SMPDB)

Discussion (Draft)

Our findings show that omics-based molecular risk scores, when derived from exposure-linked effect sizes, offer a biologically interpretable and mechanistically grounded stratification of pediatric metabolic health risk. While the exposure-only model performed well (AUC = 0.96), the integration of omics scores enhanced the interpretability of individual risk profiles by revealing pathways tied to immune activation, oxidative stress, biosynthesis, and early metabolic signaling.

Although the combined model achieved perfect performance, this outcome is expected due to its label-generation mechanism: both the inputs and outputs are derived from the same composite score. As such, this model primarily serves as a validation of internal consistency, not predictive generalizability. SHAP analysis was not applied to the combined model for this reason, but was performed on the exposure-only model, and pathway analyses served as a biological alternative to SHAP for the omics components.

Overall, the architecture built here provides an internally coherent framework for pediatric risk profiling that bridges statistical inference with biological meaning. It also enables subject-level interpretation via decomposed omics contributions and zoned risk visualization.

Conclusion

Assuming the HELIX dataset is well-constructed, biologically representative, and internally consistent—as supported by prior literature—the exposure-weighted omics risk scores and enrichment patterns we derived can be considered scientifically valid within this system. These scores offer interpretive and mechanistic insight into exposure-linked metabolic dysregulation across multiple molecular layers.

We present a biologically informed framework that integrates environmental exposure data with molecular associations from five omics modalities to stratify pediatric metabolic health risk. Rather than relying on exposures or omics data in isolation, our composite risk score is derived from exposure-weighted omics effects—yielding biologically grounded, statistically coherent, and interpretable profiles at the individual level. The architecture is modular, scalable, and adaptable to other life stage

Future Work

Future directions for this research include:

- **Clinical validation:** Apply the risk scores to real health outcomes (e.g., zBMI, HOMA-IR, lipid profiles) to assess predictive utility beyond internal consistency.
- **Temporal modeling:** Incorporate longitudinal omics or exposure changes to track how risk evolves from prenatal to postnatal stages.
- **Cross-cohort replication:** Test generalizability using similar datasets (e.g., ALSPAC, Project Viva) with available omics and exposure data.
- **Risk interpretation tools:** Build user-facing dashboards that show omics-based contributions to a child's overall risk profile for clinical or parental feedback.
- **Intervention modeling:** Simulate how modifying specific exposures (e.g., NO_2, dietary toxins) could reduce individual risk scores.