Discovery Time Machine Platform (DTMP)

An AI discovery engine for human developmental exposure-disease relationships

Google.org Impact Challenge: AI for Science | Child Health and Development Studies (fiscal sponsor: Public Health Institute) | $3M over 36 months | Collaborator: Prof. Janine LaSalle and team (UC Davis / MIND Institute)

The opportunity

Life in utero shapes health across generations and is implicated in current epidemics of infertility, early-onset cancer, and brain disorders. The Child Health and Development Studies (CHDS) is the only prospective human cohort with both maternal and paternal prenatal biospecimens in its founding generation (1960s births), spanning roughly 70,000 people across four generations with 60-plus years of multi-omic data. The archive sits across decades of legacy programmatic files and cannot be queried through any existing platform, so a national-treasure dataset remains locked away from discovery.

The solution

The DTMP is a governed, queryable interface over the full CHDS archive paired with a layer of AI agents that mine the data for molecular pathways of generational disease risk. A neural Mission Control agent reasons across CHDS records and public scientific databases, delegating to specialist symbolic agents that run statistical tests against raw data: Rodin for metabolomics, association tests, and pathway enrichment. Every hypothesis must be grounded in both mechanistic literature and prospective CHDS evidence before it advances. A Hypothesis Ledger tracks each candidate from inference to confirmed finding, and a public Atlas accumulates confirmed exposure, target, and biomarker results with full provenance.

Architecture runs on four layers: a data tier (a governed BigQuery environment under a HIPAA BAA, with a Family_ID spine linking exposures, molecular data, and outcomes across four generations), an agent tier, a logic tier (the Hypothesis Ledger), and a result tier (the Atlas). The codebase is portable to any longitudinal cohort without requiring CHDS data access.

Why it will work

CHDS data already map directly onto existing drugs. Prenatal DDT predicts HER2-positive breast cancer, a target hit by six FDA-approved therapies. Prenatal pesticides predict midlife Aβ42/40 ratio, where lecanemab and donanemab are approved for early Alzheimer's. A parathion metabolite in archived 1960s serum disrupts estrogen synthesis and predicts breast cancer within 15 years, plus lethal prostate cancer with race-specific signatures. Two published proof-of-concept studies (Go et al. 2023; Richardson et al. 2025) confirm CHDS biospecimens yield publication-ready outputs.

Underlying assets: 15,000+ families enrolled at Kaiser Oakland (1959-1967), followed across four generations, with GC-MS organochlorine and PCB panels, LC-HRMS untargeted metabolomics, LC-MS/MS PFAS, steroid hormone and cytokine panels, 450k methylation arrays, WGBS methylomes from newborn dried blood spots and cell-free DNA, mammography and structural MRI, and Kaiser clinical records, all linked to the California Cancer and Death Registries.

Team and partners

Barbara Cohn (CHDS / PHI Research Program Director) leads scientific direction and data governance. **Sarah Daniels** (Engineered Resilience) is PI for scientific design and AI development. **Nickilou Krigbaum** leads CHDS-side coordination; **Boris Minasenko** (Rodin developer) runs metabolomics and statistical pipelines; **Janine LaSalle** (UC Davis / MIND Institute) generates new WGBS and cell-free DNA methylomes feeding 2C Bioscience translation. An ML Engineer and Data Engineer will be hired to build the agentic system and pipeline infrastructure.

Budget

Budget category	Allocation (USD)
PHI personnel and staffing (Cohn, Krigbaum, Minasenko, ML Engineer, Data Engineer)	$1,439,748
Contractor (Engineered Resilience) and UC Davis subcontracts	$671,802
Indirect costs and project contingency	$399,456
Cloud infrastructure and AI compute (BigQuery, Gemini, storage)	$299,000
Dissemination, partnerships, and Scientific Advisory Board	$189,994
Total request, 36 months	**$3,000,000**

Timeline and milestones

Months 1-12. Formal CHDS data inventory using LLM-assisted parsing of legacy SAS, R, and Python code; encrypted participant ID crosswalk; PHI legal, DUA, and registry non-redisclosure audit; data manifest for the first tranche (N=4,000); three-tier web interface tested on synthetic data; full system and differential-privacy architecture co-designed with Google engineers during the Accelerator. Outcome: a working public-level natural-language query interface on simulated data, with the architecture released openly.

Months 13-24. First real-data module goes live (metabolomics, neurodegenerative biomarkers); the AI agent layer is added; the first complete discovery loop runs on real data; new WGBS sequencing is generated on first-generation pregnancy cell-free DNA and later-generation blood; DUA workflows are automated. Outcome: the first governed AI system running specialist statistical agents against a three-generation prospective human cohort.

Months 25-36. Full pipeline runs across all CHDS modules for the 4,000-person pilot; four access tiers and four user-type interfaces deploy; cost-recovery hosting activates; LaSalle's epigenomic findings advance 2C Bioscience. Outcome: at least 20 exposure-disease associations confirmed in prospective human data, a populated public Atlas, and third-party validation that the architecture generalizes beyond CHDS.

Ethics, open source, and sustainability

Only aggregate features at minimum cell size k≥10 leave the PHI-controlled environment; BigQuery differential privacy and Cloud Sensitive Data Protection enforce complementary guarantees, with the threat model reviewed by PHI's IRB. Open-source outputs ship under Apache 2.0 (DTMP architecture, agent codebase, harmonization pipeline, prompt templates), CC-BY 4.0 (differentially private aggregate feature matrices), and open-access publications; individual-level participant data stays restricted under HIPAA and consent. Sustainability comes from an NIH R01 for expansion (submitted in month 34), tiered paid access and pharmaceutical or research agreements for cost recovery, and a community-maintained free tier.

Case study: translation in practice

Arc: archived 1960s sample → prenatal methylome → confirmed exposure-disease signature → biomarker → product at 2C Bioscience

The epigenomic track demonstrates the end-to-end translation the platform is built to produce. Janine LaSalle, Professor at UC Davis and Director of the MIND Institute's Epigenomics Core has shown a strong methylome overlap between placenta and fetal brain, justifying placental cell-free DNA as a proxy for prenatal programming. Her company, 2C Bioscience, is translating these epigenomic biomarker discoveries into clinical diagnostic products for early identification of neurodevelopmental risk. Running WGBS on cell-free DNA from banked 1960s pregnancy blood recovers the prenatal epigenome decades after collection, and her DMR finder and Comethyl pipelines surface methylation signatures that predict midlife outcomes such as Alzheimer's. Confirmed signatures feed her start-up, 2C Bioscience, which is converting CHDS-derived epigenomic biomarkers into clinical products: a peptide to improve IVF embryo implantation and a first-in-class program for prevention of infertility and autism.