PUBHOMICS: A Multispecies Biological Dataset to Catalyze AI-Driven Toxicity Assessment for Environmental and Public Health[1]

Daniel Chinwendu Ukaegbu

Abstract

Environmental and public health remain under served by the recent data revolution that enabled major AI advances in drug discovery. Existing toxicity datasets are biased toward drug-like molecules and are fragmented across repositories, limiting their use for machine learning and cross-species translation. We propose **PUBHOMICS**, a scalable, openly shareable dataset capturing transcriptional responses to environmentally relevant chemical perturbations across cell types, organs, and species. PUBHOMICS will expand chemical coverage to classes absent from existing resources, enable AI models to predict transcriptomic responses to novel exposures, and support mechanism-based toxicity prediction with cross-species translation for regulatory decision-making. By advancing exposomics toward causation and providing a foundation for New Approach Methodologies (NAMs), PUBHOMICS aims to accelerate regulatory adoption and enable "benign-by-design" strategies that bridge exposure science with systems biology.

1 Introduction & Motivation

In the life sciences, well-annotated, open datasets have driven advances in drug discovery enabling AI to predict ligand–receptor binding, design new compounds, and optimize therapeutics. These gains were possible because biological data was standardized, accessible, and interoperable (Liu et al., 2024). However, environmental and public health remain underserved by this data revolution. Despite the vast burden of environmental exposures, most of the tens of thousands of chemicals in use today especially Persistent Organic Pollutants (POPs), heavy metals, plasticizers, flame retardants, and agricultural chemicals remain untested for safety (CDC, 2024; Judson et al., 2009). Existing large toxicity datasets, largely from pharma, are biased toward small, drug-like molecules (Kretschmer et al., 2025; Seal et al., 2025). However, the data that does exist, although very limited (von Borries et al., 2023), is fragmented across different databases and is not structured for machine learning (Ajisafe et al., 2025; HESI, 2025). Beyond the paucity of the data, the lack of a unified framework for collecting and harmonizing toxicity data, particularly for cross-species, mechanistic datasets that capture gene–environment interactions, has become a critical bottleneck (Motsinger-Reif et al., 2024; Sekatcheff et al., 2024). With NIH and global regulators pushing to replace animal testing with New Approach Methodologies (NAMs) (FDA, 2025; NIH, 2025), the absence of foundational, ML-ready datasets threatens progress in AI-enabled chemical risk assessment for public and environmental health protection (FDA, 2025; Fortin et al., 2023).

2 Goal of the Proposal

We propose PUBHOMICS, a scalable, openly shareable dataset capturing transcriptional responses to environmental and public health–relevant chemical perturbations across cell types, organs, and species for developing AI tools for chemical hazard assessment. This dataset will include transcriptomics profiles and rich metadata on exposure duration, timing, and conditions for chemicals spanning real-world chemical and biological space relevant to protect public health.

2.1 Why This Dataset Matters

Recent AI foundation models like scGPT and Geneformer are advancing genomic analysis. In environmental health, generative models such as AnimalGAN and TransOrGAN demonstrate strong performance in AI-driven chemical hazard assessment (Chen et al., 2023; Li et al., 2023). However, these models rely predominantly on baseline expression, pharmaceutical, or oncology datasets or less diverse cell lines (Chen & Zou, 2023; Sheinin et al., 2025), limiting their depth and coverage for public health applications due to data constraints. PUBHOMICS will fill this gap by delivering the first large-scale, harmonized transcriptomics resource capturing biological responses to environmentally relevant chemicals across species. It will expand chemical coverage to include classes of toxicants absent from existing datasets, enable AI models to predict transcriptomic responses to novel exposures, and support mechanism-based toxicity prediction with cross-species translation for regulatory decision-making. By mapping environmental exposures to disease pathways such as cancer, autoimmune disorders, and developmental diseases, PUBHOMICS will advance exposomics research beyond correlation toward causation, offering enhanced opportunities to explore environmental exposure–disease relationships (Sarigiannis et al., 2025; Sillé, 2020; Wan et al., 2025). It will also provide a critical foundation for scaling and validating New Approach Methodologies (NAMs), accelerating development of other data layers and regulatory adoption (Sillé et al., 2020, 2025), and empower the chemical industry to implement "benign-by-design" strategies that bridge exposure science with systems biology (Maertens, 2022; Nielsen & Moon, 2013).

3 Data-Creation Pathway

Phase 1 – Integration & Harmonization (3–7 Months). We will curate millions of transcriptomic profiles for ~2000 environmentally relevant chemicals, currently fragmented across databases such as GEO, ToxCast, and CTD, as well as published literature. Automated pipelines will be employed to extract and harmonize data into common gene identifiers and units, perform quality control to mitigate batch effects and exclude low-quality samples, and annotate chemicals using standardized ontologies.
Phase 2 – Targeted Data Generation (6–18 Months). We will generate transcriptomics profiles for 1,250 compounds (50 per class across 25 major classes) using an in-house clustering approach ensuring chemical diversity coverage. Primary zebrafish screening with dose–response testing will be integrated with human cell validation for priority hits using PLATE-seq transcriptomics.

4 Cost & Scalability Strategy

Phase 1 – Data Integration ($25,000; 3–7 months). Covers cost for bioinformatics/data science effort, cloud computing, data collection and access fees, quality control/validation, and data housing.
Phase 2 – New Data Generation ($977,500; 18 months–2 years). Includes PLATE-Seq experiments for 12,500 samples ($625,000–$937,500 at $50–$75/sample), zebrafish procurement and facility costs ($40,000), human cell line validation assays for priority compounds, and personnel/overhead. This budget prioritizes high-throughput, cost-efficient transcriptomics profiling using PLATE-Seq's low per-sample cost compared to traditional RNA-seq (85–90% cost savings) while ensuring a robust experimental design.

5 References

Ajisafe, O. M., Adekunle, Y. A., Egbon, E., Ogbonna, C. E., & Olawade, D. B. (2025). The role of machine learning in predictive toxicology: A review of current trends and future perspectives. *Life Sciences*, 378, 123821. https://doi.org/10.1016/j.lfs.2025.123821

BUILDING A ROADMAP FOR AI-ENABLED HUMAN AND ENVIRONMENTAL HEALTH PROTECTION Executive Summary. (2025). https://hesiglobal.org/wp-content/uploads/2025/07/HESIGlobal-AI-OneHealth-pdf

CDC. (2024, May 17). National Report on Human Exposure to Environmental Chemicals. National Biomonitoring Program. https://www.cdc.gov/biomonitoring/resources/national-exposure-report.html

Chapter 1. Green Chemistry for Green Toxicology. (2022). *RSC Green Chemistry Series*, 1–30. https://doi.org/10.1039/9781839164392-00001

Chen, X., Roberts, R., Liu, Z., & Tong, W. (2023). A generative adversarial network model alternative to animal studies for clinical pathology assessment. *Nature Communications*, 14(1), 7141. https://doi.org/10.1038/s41467-023-42933-9

Chen, Y. T., & Zou, J. (2023). GENEPT: A SIMPLE BUT HARD-TO-BEAT FOUNDATION MODEL FOR GENES AND CELLS BUILT FROM CHATGPT. *bioRxiv*. https://doi.org/10.1101/2023.10.16.562533

Keenan, A. B., Jenkins, S. L., Jagodnik, K. M., et al. (2018). The Library of Integrated Network-Based Cellular Signatures NIH Program. *Cell Systems*, 6(1), 13–24. https://doi.org/10.1016/j.cels.2017.11.001

Kretschmer, F., Seipp, J., Ludwig, M., Klau, G. W., & Böcker, S. (2025). Coverage bias in small molecule machine learning. *Nature Communications*, 16(1). https://doi.org/10.1038/s41467-024-55462-w

Li, T., Roberts, R., Liu, Z., & Tong, W. (2023). TransOrGAN: An Artificial Intelligence Mapping of Rat Transcriptomic Profiles between Organs, Ages, and Sexes. *Chemical Research in Toxicology*, 36(6), 916–925. https://doi.org/10.1021/acs.chemrestox.3c00037

Liu, H., Chen, P., Zhai, X., et al. (2024). PPB-Affinity: Protein-Protein Binding Affinity dataset for AI-based protein drug discovery. *Scientific Data*, 11(1). https://doi.org/10.1038/s41597-024-03997-4

Liu, T., Hwang, L., Burley, S., et al. (2024). BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Research*, 53(D1), D1633–D1644. https://doi.org/10.1093/nar/gkae1075

Maertens, A. (2022). Green Toxicology. Royal Society of Chemistry. https://doi.org/10.1039/9781839164392

Motsinger-Reif, A. A., Reif, D. M., Akhtari, F. S., et al. (2024). Gene-environment interactions within a precision environmental health framework. *Cell Genomics*, 4(7), 100591. https://doi.org/10.1016/j.xgen.2024.100591

Nielsen, D. R., & Moon, T. S. (2013). From promise to practice. *EMBO Reports*, 14(12), 1034–1038. https://doi.org/10.1038/embor.2013.178

NIH to prioritize human-based research technologies. (2025, April 29). National Institutes of Health (NIH). https://www.nih.gov/news-events/news-releases/nih-prioritize-human-based-research-technologies

Seal, S., Mahale, M., García-Ortegón, M., et al. (2025). Machine Learning for Toxicity Prediction Using Chemical Structures: Pillars for Success in the Real World. *Chemical Research in Toxicology*. https://doi.org/10.1021/acs.chemrestox.5c00033

Sekatcheff, E. D., Jeong, J., & Choi, J. (2024). Bridging the Gap Between Human Toxicology and Ecotoxicology Under One Health Perspective. *Environmental Toxicology and Chemistry*. https://doi.org/10.1002/etc.5940

Sheinin, R., Sharan, R., & Madi, A. (2025). scNET: learning context-specific gene and cell embeddings by integrating single-cell gene expression data with protein–protein interactions. *Nature Methods*. https://doi.org/10.1038/s41592-025-02627-0

Sillé, F. (2020). The exposome – a new approach for risk assessment. *ALTEX*, 3–23. https://doi.org/10.14573/altex.2001051

von Borries, K., Holmquist, H., Kosnik, M., et al. (2023). Potential for Machine Learning to Address Data Gaps in Human Toxicity and Ecotoxicity Characterization. *Environmental Science & Technology*, 57(46), 18259–18270. https://doi.org/10.1021/acs.est.3c05300

Wan, M., Simonin, E. M., Johnson, M. M., et al. (2025). Exposomics: a review of methodologies, applications, and future directions in molecular medicine. *EMBO Molecular Medicine*. https://doi.org/10.1038/s44321-025-00191-w