Environmental Science & Technology

pubs.acs.org/est

Viewpoint

Small-Sample Learning for Next-Generation Human Health Risk Assessment: Harnessing AI, Exposome Data, and Systems Biology

Tianxiang Wu, Lu Zhao, Mengyuan Ren, Song He, Le Zhang, Mingliang Fang, and Bin Wang*

Cite This: *Environ. Sci. Technol.* 2025, 59, 5−10

Read Online

KEYWORDS: *risk assessment, small-sample learning, exposome, big data, transfer learning, multimodal learning, systems biology*

Conducting human health risk assessments (HRA) related to environmental exposure presents significant challenges, particularly when applying traditional epidemiological methods to exposome big data. These data encompass a broad spectrum of exogenous environmental factors (e.g., persistent organic pollutants, endocrine disruptors, metals, ambient air pollutants, bacteria, and noise) and endogenous biological information (e.g., the epigenome, transcriptome, proteome, metabolome, gut microbiota, inflammation, and oxidative stress).[1] For example, the human exposome database (HExpMetDB) compiled >20 000 chemicals and prioritized 13 441 chemicals on the basis of the probabilistic hazard quotient and 7770 chemicals on the basis of the risk index.[2] Our recently developed exposome database included ∼119 million exposures and 17 186 disease subtypes from well-established sources (see the ExposomeX database at http://www.exposomex.cn/#/database101). Additionally, we proposed TOXRIC (https://toxric.bioinforai.tech/), a database with comprehensive toxicological data, standardized attribute data, practical benchmarks, informative visualization of molecular representations, and an intuitive function interface.[3] The TOXRIC database contains 113 372 chemicals across 13 toxicity categories, 1474 toxicity end points covering *in vivo* and *in vitro* end points, and 39 feature types. To obtain the causal relationship, these methods often demand substantial time and resources due to the requirement of a large sample size, long-term follow-ups, extensive data collection, and high-cost experimental analysis. Given these constraints, there is an increasing necessity to explore alternative methodologies that can effectively evaluate health risks with a limited sample size. We propose that small-sample learning can potentially address practical concerns, such as recruitment feasibility and the high

Published: January 2, 2025

https://doi.org/10.1021/acs.est.4c11832
Environ. Sci. Technol. 2025, 59, 5−10

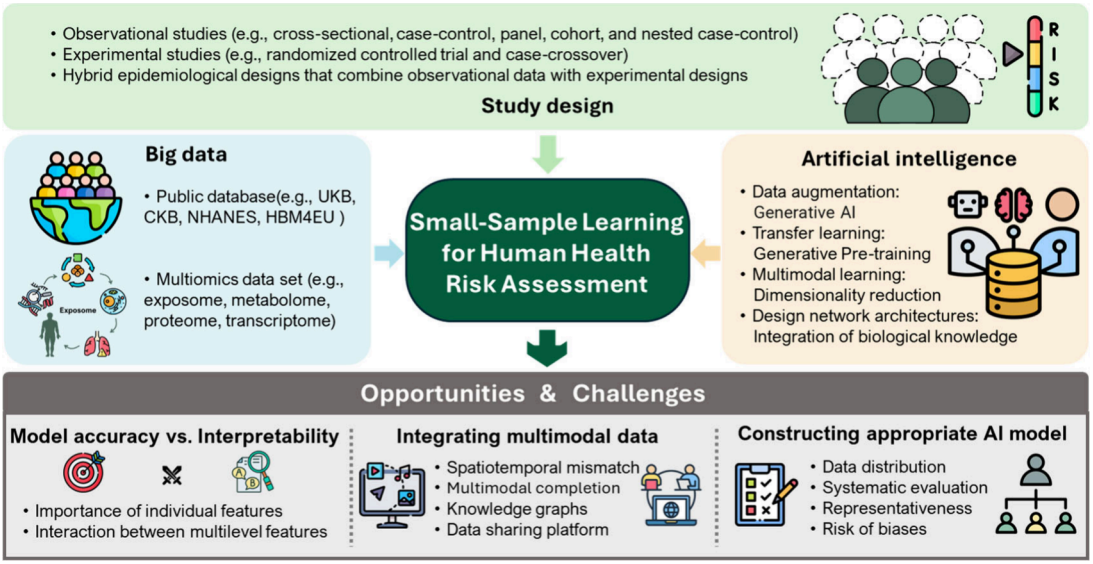

Figure 1. Schematic diagram of small-sample learning for human health risk assessment (HRA) of environmental exposure. To facilitate the related HRA modeling, three aspects should be considered: study design, big data collection, and artificial intelligence. The development of future models should prioritize balancing accuracy with interpretability, integrating multimodal data effectively, and constructing robust, generalizable AI models that incorporate complexity and prior biological knowledge while addressing biases in the data.

costs of sample analysis. The ability to derive meaningful insights from a smaller sample size can enhance the efficiency of HRA, especially in scenarios in which large populations are difficult to access. This is particularly relevant for assessing exposure to rare environmental contaminants or when individual variations in susceptibility are pronounced. Moreover, with the advent of big data and artificial intelligence (AI) technologies, reliable statistical analyses integrating high-dimensional data of external and internal exposure information are urgently needed to develop the next-generation HRA model for screening risk factors, determining effect size, and quantifying disease burden in environmentally complex exposure scenarios.[1] This viewpoint underscores the importance of choosing appropriate epidemiological design and developing innovative strategies supported by big data and AI to advance our understanding of environmental health risks and improve public health outcomes (see Figure 1). The following methods can be explored more thoroughly to facilitate such a study.

First, determining the appropriate epidemiological design for balancing evidence strength, sample size, and available resources is highly critical. Various epidemiological study designs can be utilized, including observational (e.g., cross-sectional, case control, panel, and cohort) and experimental (e.g., randomized controlled trial and case crossover) studies, or some hybrid epidemiological designs that combine observational data with experimental or quasi-experimental designs to improve evidence strength (e.g., propensity score matching and instrumental variable techniques). For example, if we want to predict the incidence of some health outcomes, the prospective design, like a cohort or longitudinal study, should be adopted. Also, repeated measures can be conducted if more participants are not easy to recruit. In addition, previous knowledge about the critical time window and sensitive population can help us recruit the appropriate sample size for developing the HRA model. For instance, Stelzer et al. conducted a longitudinal study with 63 women who went into labor spontaneously and achieved an accurate prediction of labor timing by using a stacked generalization (SG) method integrating all three modalities (metabolome, proteome, and immunome).[4]

Second, data-driven modeling strategies require the support of exposome big data in the following two key respects.

(I) Utilizing large-scale and reliable human biomonitoring databases can provide prior knowledge on the association between environmental exposure and health outcome. Over the past few decades, several large-scale human biomonitoring studies have been established across various countries and regions to conduct epidemiological research on lifestyle, environmental, and genetic factors that contribute to major diseases. Many of these studies are committed to open data and resource sharing, providing data in a fair and transparent manner. These include projects such as the UK Biobank (UKB), Human Biomonitoring for Europe (HBM4 EU), National Health and Nutrition Examination Survey (NHANES), Human Health Exposure Analysis Resource (HHEAR), China Health and Retirement Longitudinal Study (CHARLS), the Environmental Influences on Child Health Outcomes (ECHO), etc. These databases provide comprehensive data on biospecimens, genetics, imaging, medical health records, lifestyle, air pollution, and chemical exposures, offering invaluable resources for early career researchers to explore and test hypotheses regarding the impact of environmental exposures on human diseases. For example, Argentieri and colleagues developed a biological age prediction model using proteomics data from the UKB (N = 45 441) by identifying 204 key plasma proteins from 2897 proteins for constructing a proteomic age clock.[5] Validation in independent cohorts from China (N = 3977) and Finland (N = 1990) based on the same 204 proteins demonstrated high predictive accuracy. By leveraging biomarkers with strong generalizability identified in large populations, this approach facilitates accurate risk evaluations in smaller cohorts,

https://doi.org/10.1021/acs.est.4c11832
Environ. Sci. Technol. 2025, 59, 5−10

enhances the identification of vulnerable groups, and helps minimize errors while improving the interpretability of results in small populations. In cases of scarce human biomonitoring data, *in vitro* to *in vivo* extrapolation (IVIVE) also enables precise health risk predictions at the individual level for small populations. It also provides confidence for the routine use of chemical prioritization, hazard assessment, and regulatory decision making. Recently, Han et al. systematically summarized and discussed IVIVE methods in next-generation HRA and innovatively proposed prospects from two aspects.[6] The first is to expand the scope of IVIVE, such as focusing on the joint risk of parent compounds and their metabolites; the second is to integrate new technologies like systems biology, multiomics, and adverse outcome networks in IVIVE, aiming at a more microscopic, mechanistic, and comprehensive risk assessment.

(II) Incorporating multilevel omics data into small-sample HRA can provide a more comprehensive and accurate understanding of how environmental exposures affect health. It has been well-known that environmental exposures typically influence biological processes at various levels, leading to adverse health outcomes. However, a limitation of multiomics studies is the high cost, with the expense of single-cell transcriptomics ranging from $1500 to $2000 per person, resulting in a typically small-sample size for high-throughput omics assessments. The "Curse of Dimensionality" is a significant modeling challenge in multiomics data. As the dimensionality of the data increases, the sparsity of the data space grows exponentially, causing the occurrence probability of certain feature combinations to become extremely low or even completely unobserved. This sparsity makes it difficult for models to learn meaningful patterns, thereby affecting their generalization ability. Compounding this challenge, multiomics data obtained through different measurement techniques often exhibit distinct characteristics and distributions.[7] Faced with the challenges of small-sample size and high dimensionality, deep learning algorithms can capture complex nonlinear relationships and automatically learn high-quality feature representations from low-level omics data while performing dimensionality reduction. For example, Cao et al. developed a modular framework called GLUE that utilizes a knowledge-based graph to simulate cross-layer regulatory interactions, linking the feature spaces between omics layers. In this graph, the vertices correspond to features from different omics layers and the edges represent regulatory interactions between them.[8] This model was employed to integrate unpaired single-cell multiomics data while simultaneously inferring regulatory interactions, demonstrating exceptional robustness in small-sample scenarios. To mitigate the multicollinearity issue in multiomics data, it is essential to ensure high data quality and diversity, select appropriate model training and regularization strategies, and integrate existing biological knowledge (e.g., gene regulatory networks and protein−protein interactions) to improve model efficiency.

Third, deep learning methods offer tremendous potential to address the problem of small-sample size. Despite living in an era of big data with increasingly available resources such as large-scale biomonitoring databases, wearable devices, bioimaging, and multiomics, we still face the challenges of integrating multiple source information for HRA studies. These challenges are often accompanied by issues such as missing or noisy data, data imbalance, high dimensionality, and overfitting. The corresponding small-sample learning models can be broadly categorized into four approaches.

(I) Data augmentation. This is a technique used to produce new artificial data that closely resemble real data, thereby enhancing the dataset. In this context, AI-generated content (AIGC) leverages large amounts of unlabeled data to learn data distributions and generate samples, facilitating the creation of images, text, audio, and video while enhancing model generalization capabilities under limited labeled data conditions. Common models include generative adversarial networks (GANs), which generate realistic samples through adversarial training between a generator and a discriminator, leveraging the flexibility of latent space to effectively capture data features. Diffusion models generate high-quality data from random noise through a gradual denoising process, allowing for effective utilization of limited data and producing diverse outputs. Variational autoencoders (VAEs) compress data into latent space via an encoder and then decode it to generate new samples, providing a stable generation process that preserves data diversity and structural integrity in small-sample learning scenarios. AIGCs have made significant advances in molecular property prediction[9] and drug design[10] but remain underexplored in datasets related to HRA.

(II) Transfer learning. Machine learning or deep learning models can be trained on large general datasets and fine-tuned with smaller, domain-specific datasets for downstream tasks. One of the key applications of transfer learning in HAR research is using insights from the structure−property relationships of chemicals and shared biological networks underlying diseases to apply knowledge from conventional pollutants to emerging pollutants and from common diseases to rare diseases. Upon application of transfer learning, it is essential to align key factors between the source and target populations, including exposure characteristics (e.g., dose−response relationships), population structures, disease incidence, and disease types. Fine-tuning is crucial in transfer learning and involves strategies like adjusting layer-specific learning rates, freezing early layers while tuning later ones, and using data augmentation or regularization to prevent overfitting. Generative pretraining (GPT), leveraging attention mechanisms and unsupervised pretraining, has become a pioneering model for natural language processing tasks. Inspired by GPT, researchers have applied self-supervised pretraining to large-scale single-cell transcriptomics data to develop foundational models that can be applied to smaller patient data sets, aiding in the identification of candidate therapeutic targets for diseases.[11] In the near future, models similar to ChatGPT may be trained on human biomonitoring

https://doi.org/10.1021/acs.est.4c11832
Environ. Sci. Technol. 2025, 59, 5−10

databases, although further research is needed to validate these applications.

(III) Multimodal learning. In HRA research, the exposure data modalities primarily include numerical vectors (e.g., multiomics data and pollutant concentrations), graphs (e.g., molecular graphs and biological knowledge graphs), text (e.g., electronic health records and protein sequences), visual data (e.g., imaging and remote sensing), and audio data (e.g., environmental noise). In tasks involving systems biology, multimodal learning integrates complementary information from various modalities, enhancing performance under small-sample learning and offering a more comprehensive perspective on environmental health issues.[12] Multimodal learning is a versatile concept that can be addressed with various architectures. A common strategy is to perform joint modeling of data from different modalities, either by designing specialized architectures for each modality or by developing a foundational model that maps similar concepts across modalities into a shared latent space. This approach enables the generation of unified internal representations for the same semantics or concepts (e.g., the unification of remote sensing images, sensor exposure data, and exposure event descriptions) and outputs task-specific results as required. Foundation models have been successfully implemented in integrating single-cell omics data[8] and biomedical data.[13] Another approach is to construct knowledge graphs (KGs) between modalities. In systems biology, multilevel features can be represented as a heterogeneous graph, capturing associations among chemicals, genes, proteins, and diseases, with graph neural networks used to uncover unknown relationships among these entities. Multimodal models also offer a new approach for exposome, integrating external environmental exposure factors with internal exposure biomarkers to create unified feature vectors, which provide a comprehensive individual exposure profile.

(IV) Sparsely structured network architectures or algorithms. These are designed on the basis of biological information and can reduce model parameters.[14] Biological information design refers to hidden nodes in neural networks that simulate specific biological processes, such as DCell (cellular processes and functions)[15] and P-NET (genes and pathways).[14] More streamlined architectures and sparse network structures enable higher accuracy in small-sample learning, while quantifying the relative importance of biological processes within the network also enhances interpretability. The deep forest model (DF), a multilayer tree-based forest cascade structure, is suitable for data sets of different sizes, few hyperparameters, and adaptive generation of model complexity. The model complexity of DF can be adaptively determined under sufficient training, making it applicable to data sets of small-size scales.[16] Inspired by the adverse outcome pathway (AOP) framework, we can integrate environmental exposures, molecular initiating events (MIEs), key molecular events (KEs), and adverse outcomes into the network architecture, ordered by increasing complexity.[17] It is worth noting that in specific environmental small-sample studies, these methods are not conducted in isolation but often employ a combination of strategies systematically. For example, Huang et al. integrated multimodal data, including genomic data, cellular signaling, gene expression levels, and clinical records, to construct a biomedical knowledge graph and performed self-supervised pretraining.[18] By combining the concepts of multimodal learning and transfer learning, their model was able to accurately predict drug indications and contraindications across diseases under strict zero-shot conditions, including diseases for which no therapeutic drugs are currently available.

OPPORTUNITIES AND CHALLENGES

As shown in Figure 1, we summarize the opportunities and challenges as follows.

(I) Balancing model accuracy and interpretability. While deep learning has achieved remarkable success across various fields, interpretability remains a recurrent issue. In the context of HRA, it is especially important to translate model weight parameters into meaningful biological information. This not only provides reliable information to healthcare providers but also offers potential biological explanations for disease progression. Common approaches include SHAP (Shapley additive explanations) and attention weight in transformer architectures. However, in small-sample scenarios, data-driven interpretability mining can sometimes produce spurious important features, leading to interpretations that deviate from biological facts from the in-depth environmental toxicological studies.[19] Another approach is to evaluate key environmental exposure and biological features relevant to disease through computational perturbation of features (e.g., deletion, random masking, or omics data inversion) and quantifying the impact of these perturbations on predictive outcomes. To enhance interpretability, perturbing known features (e.g., MIE and KE) instead of random features can be useful, though this requires prior knowledge related to toxicity pathways. While the approach described above applies to assessing the importance of individual features, the advantage of deep learning lies in its ability to capture interactions between features. For instance, we can incorporate environmental exposure, biological processes, and disease progression into different hierarchical levels of a model. By assessing neuron activations among these levels, we can identify nonlinear interactions between environmental exposures and biological effects at various levels. Computational perturbation can also be used to evaluate potential interactions, such as perturbing environmental exposure feature embeddings and observing the impact on biological effect feature embeddings to identify targets of environmental exposures. This approach has been applied to the discovery of gene−gene interactions,[11] though it requires significant computational resources. It is worth noting that perturbing two or more environmental exposure features simultaneously to calculate their combined impact on disease or biological effects could potentially evaluate the possible mixed effects of environmental exposures. Additionally, integrating laboratory data with models through active learning and carefully designed querying strategies can focus on the most uncertain or influential unlabeled samples for

https://doi.org/10.1021/acs.est.4c11832
Environ. Sci. Technol. 2025, 59, 5−10

validation, reducing annotation costs and enhancing model interpretability.

(II) Integrating multimodal data. Currently, AI applications in HRA often focus on solving single tasks using one type of data. The future direction lies in integrating multimodal data to systematically determine how environmental exposures induce disease through multilevel biological processes. Given the increasing volume of multimodal data, there is an urgent need to develop suitable frameworks that align and fuse structured and unstructured modalities to generate accurate feature representations while preserving each modality's biological information. One challenge in integrating exposome data is that different modalities often have varying resolutions and spatiotemporal scales. For example, chemical measurements in biological samples and gene expression or metabolite data reflect exposure only during specific time windows. In contrast, health outcomes, such as the onset or progression of chronic diseases, may not align with these time frames, as they often develop over longer periods. Another issue is that multimodal data often suffer from missing samples in one or more modalities. Because multimodal data complement each other, known modalities can be used to generate and impute missing features in others, improving the completeness of the data. Furthermore, the use of knowledge graph-based methods for multimodal data integration is still underexplored. KG can provide deeper biological insights for exposome data fusion, but this approach is still in its early stages and requires further development and optimization of graph construction and reasoning algorithms. As a result, the development of high-quality exposome data-sharing platforms has become urgent, as quality data are essential for effective modeling. The collection, cleaning, annotation, and integration of various datasets related to HRA face numerous challenges. While there are several large, high-quality databases, a significant amount of data remains dispersed across the literature, with inconsistencies in data sources, formats, and sample sizes. Therefore, the integration and standardization of data are crucial for creating high-quality open data-sharing platforms, such as ExposomeX and TOXRIC, to facilitate data growth.

(III) Constructing appropriate AI-Based models with stronger generalization ability. Before the selection of an algorithm, it is essential to adopt various strategies to address the challenges of small-sample sizes. Even the best models cannot derive valuable insights from inadequate data. Key strategies include actively collecting and curating diverse sample data (e.g., wearable device data and continuous environmental monitoring data), with efforts to maximize standardization and automation, thoroughly evaluating the distribution of small-sample groups to avoid biases, leveraging biological and environmental knowledge to guide feature engineering, and eliminating noise and outliers while appropriately handling missing data. When the appropriate model algorithm is being chosen, factors beyond accuracy must also be considered, such as model complexity, robustness, generalizability, and interpretability. A more systematic evaluation is required to ensure these aspects are balanced. In the context of transfer learning using large human biomonitoring databases, it is also important to account for potential biases arising from population selection. Furthermore, designing sophisticated models that incorporate existing prior biological knowledge is a promising approach. After model development, it is essential to assess the representativeness of small-sample size in HRA and consider the risks introduced by various biases.

AUTHOR INFORMATION

Corresponding Author

Bin Wang − *Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China; Institute of Reproductive and Child Health, School of Public Health, Peking University, Beijing 100191, P. R. China; Key Laboratory of Reproductive Health, National Health and Family Planning Commission of PR China, Beijing 100191, China; Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China; Laboratory for Earth Surface Processes, College of Urban and Environmental Science, Peking University, Beijing 100871, China;* orcid.org/0000-0002-1164-8430; Email: binwang@pku.edu.cn

Authors

Tianxiang Wu − *Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China; Institute of Reproductive and Child Health, School of Public Health, Peking University, Beijing 100191, P. R. China; Key Laboratory of Reproductive Health, National Health and Family Planning Commission of PR China, Beijing 100191, China;* orcid.org/0000-0001-9456-4082

Lu Zhao − *Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China; Institute of Reproductive and Child Health, School of Public Health, Peking University, Beijing 100191, P. R. China; Key Laboratory of Reproductive Health, National Health and Family Planning Commission of PR China, Beijing 100191, China*

Mengyuan Ren − *Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China; Institute of Reproductive and Child Health, School of Public Health, Peking University, Beijing 100191, P. R. China; Key Laboratory of Reproductive Health, National Health and Family Planning Commission of PR China, Beijing 100191, China;* orcid.org/0000-0003-3819-4181

Song He − *Department of Bioinformatics, Academy of Military Medical Sciences, Beijing 100850, China;* orcid.org/0000-0002-4136-6151

Le Zhang − *School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China*

Mingliang Fang − *Department of Environmental Science and Engineering, Fudan University, Shanghai 200433, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.4c11832

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acs.est.4c11832
Environ. Sci. Technol. 2025, 59, 5−10

Biography

Bin Wang is a tenured Associate Professor and Vice Dean at the Institute of Reproductive and Child Health, Peking University. He also serves as an adjunct professor at the College of Urban and Environmental Sciences, Peking University. His primary research focuses on exposomics and AI-driven environmental health risk assessment. In collaboration with Prof. Mingliang Fang from Fudan University, he co-developed the integrative exposomics platform ExposomeX (www.exposomex.cn), accelerating research into the "Exposure−Biology−Disease" nexus. Prof. Wang has made significant contributions to the field by constructing statistical models to predict levels of common pollutants in the human body across specific regions. He has quantitatively assessed the links between pollution exposure in pregnant women from high-pollution areas and adverse reproductive health outcomes, providing critical evidence on the impacts of environmental pollutants on reproductive health. He is a pioneer in education, offering the course "Exposomics" to undergraduate and graduate students, as well as teaching in the prestigious international master's program in global health and public health, "Environment & Health".

■ ACKNOWLEDGMENTS

The authors thank the working group of environmental exposure and human health of the China Cohort Consortium (http://chinacohort.bjmu.edu.cn/). This study was supported by the National Natural Science Foundation of China (Grant 42477455), and the Strategy Priority Research Program (Category B) of the Chinese Academy of Sciences (XDB0750300).

■ REFERENCES

(1) Fang, M. L.; Hu, L. G.; Chen, D.; Guo, Y. M.; Liu, J. M.; Lan, C. X.; Gong, J. C.; Wang, B. Exposome in human health: Utopia or wonderland? *The Innovation* **2021**, *2* (4), 100172.

(2) Zhao, F. R.; Li, L.; Chen, Y.; Huang, Y. C.; Keerthisinghe, T. P.; Chow, A.; Dong, T.; Jia, S. L.; Xing, S. P.; Warth, B.; Huan, T.; Fang, M. L. Risk-Based Chemical Ranking and Generating a Prioritized Human Exposome Database. *Environ. Health Perspect.* **2021**, *129* (4), 47014.

(3) Wu, L.; Yan, B.; Han, J.; Li, R.; Xiao, J.; He, S.; Bo, X. TOXRIC: a comprehensive database of toxicological data and benchmarks. *Nucleic Acids Res.* **2023**, *51* (D1), D1432−d1445.

(4) Stelzer, I. A.; Ghaemi, M. S.; Han, X.; Ando, K.; Hédou, J. J.; Feyaerts, D.; Peterson, L. S.; Rumer, K. K.; Tsai, E. S.; Ganio, E. A.; Gaudillière, D. K.; Tsai, A. S.; Choisy, B.; Gaigne, L. P.; Verdonk, F.; Jacobsen, D.; Gavasso, S.; Traber, G. M.; Ellenberger, M.; Stanley, N.; Becker, M.; Culos, A.; Fallahzadeh, R.; Wong, R. J.; Darmstadt, G. L.; Druzin, M. L.; Winn, V. D.; Gibbs, R. S.; Ling, X. B.; Sylvester, K.; Carvalho, B.; Snyder, M. P.; Shaw, G. M.; Stevenson, D. K.; Contrepois, K.; Angst, M. S.; Aghaeepour, N.; Gaudillière, B. Integrated trajectories of the maternal metabolome, proteome, and immunome predict labor onset. *Sci. Transl. Med.* **2021**, *13* (592), No. eabd9898.

(5) Argentieri, M. A.; Xiao, S.; Bennett, D.; Winchester, L.; Nevado-Holgado, A. J.; Ghose, U.; Albukhari, A.; Yao, P.; Mazidi, M.; Lv, J.; Millwood, I.; Fry, H.; Rodosthenous, R. S.; Partanen, J.; Zheng, Z.; Kurki, M.; Daly, M. J.; Palotie, A.; Adams, C. J.; Li, L.; Clarke, R.; Amin, N.; Chen, Z.; van Duijn, C. M. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nat. Med.* **2024**, *30* (9), 2450−2460.

(6) Han, P.; Li, X.; Yang, J.; Zhang, Y.; Chen, J. Advancing Toxicity Predictions: A Review on in Vitro to in Vivo Extrapolation in Next-Generation Risk Assessment. *Environ. Health* **2024**, *2* (7), 499−513.

(7) Hu, Y.; Wan, S.; Luo, Y.; Li, Y.; Wu, T.; Deng, W.; Jiang, C.; Jiang, S.; Zhang, Y.; Liu, N.; Yang, Z.; Chen, F.; Li, B.; Qu, K. Benchmarking algorithms for single-cell multi-omics prediction and integration. *Nat. Methods* **2024**, *21*, 2182−2194.

(8) Cao, Z.-J.; Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **2022**, *40* (10), 1458−1466.

(9) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **2022**, *4* (3), 279−287.

(10) Igashov, I.; Stärk, H.; Vignac, C.; Schneuing, A.; Satorras, V. G.; Frossard, P.; Welling, M.; Bronstein, M.; Correia, B. Equivariant 3D-conditional diffusion model for molecular linker design. *Nat. Mach. Intell.* **2024**, *6* (4), 417−427.

(11) Theodoris, C. V.; Xiao, L.; Chopra, A.; Chaffin, M. D.; Al Sayed, Z. R.; Hill, M. C.; Mantineo, H.; Brydon, E. M.; Zeng, Z.; Liu, X. S.; Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature* **2023**, *618* (7965), 616−624.

(12) Liu, W.; Chen, J.; Wang, H.; Fu, Z.; Peijnenburg, W. J. G. M.; Hong, H. Perspectives on Advancing Multimodal Learning in Environmental Science and Engineering Studies. *Environ. Sci. Technol.* **2024**, *58* (38), 16690−16703.

(13) Moor, M.; Banerjee, O.; Abad, Z. S. H.; Krumholz, H. M.; Leskovec, J.; Topol, E. J.; Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature* **2023**, *616* (7956), 259−265.

(14) Elmarakeby, H. A.; Hwang, J.; Arafeh, R.; Crowdis, J.; Gang, S.; Liu, D.; AlDubayan, S. H.; Salari, K.; Kregel, S.; Richter, C.; Arnoff, T. E.; Park, J.; Hahn, W. C.; Van Allen, E. M. Biologically informed deep neural network for prostate cancer discovery. *Nature* **2021**, *598* (7880), 348−352.

(15) Ma, J.; Yu, M. K.; Fong, S.; Ono, K.; Sage, E.; Demchak, B.; Sharan, R.; Ideker, T. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **2018**, *15* (4), 290−298.

(16) Wu, L.; Gao, J.; Zhang, Y.; Sui, B.; Wen, Y.; Wu, Q.; Liu, K.; He, S.; Bo, X. A hybrid deep forest-based method for predicting synergistic drug combinations. *Cell Rep. Methods* **2023**, *3* (2), 100411.

(17) Ciallella, H. L.; Russo, D. P.; Aleksunes, L. M.; Grimm, F. A.; Zhu, H. Revealing Adverse Outcome Pathways from Public High-Throughput Screening Data to Evaluate New Toxicants by a Knowledge-Based Deep Neural Network Approach. *Environ. Sci. Technol.* **2021**, *55* (15), 10875−10887.

(18) Huang, K.; Chandak, P.; Wang, Q.; Havaldar, S.; Vaid, A.; Leskovec, J.; Nadkarni, G. N.; Glicksberg, B. S.; Gehlenborg, N.; Zitnik, M. A foundation model for clinician-centered drug repurposing. *Nat. Med.* **2024**, *30* (12), 3601−3613.

(19) Li, H.; Yin, N.; Yang, R.; Faiola, F. Advancing Environmental Toxicology In Vitro: From Immortalized Cancer Cell Lines to 3D Models Derived from Stem Cells. *Environ. Health* **2024**, *2*, 332−349.

https://doi.org/10.1021/acs.est.4c11832
Environ. Sci. Technol. 2025, 59, 5−10