Artificial Intelligence & Environment

https://doi.org/10.66178/aie-0026-0008

Perspective

Open Access

Advancing AI/ML-driven chemical exposomics to identify biologically relevant environmental exposures

Hemi Luan[1,2*] and Tiangang Luan[1,3]

Citation: Luan H; Luan T. Advancing AI/ML-driven chemical exposomics to identify biologically relevant environmental exposures. *AI Environ.* 2026, 1(2): 77-82.

Received: 19 January 2026

Revised: 11 March 2026

Accepted: 2 April 2026

Published online: 30 April 2026

Abstract

Exposomics aims to systematically measure environmental exposures that an individual experiences across all stages of life, from conception to death, and link them to health and disease states. The ability of high-resolution mass spectrometry (HRMS) to perform non-targeted detection of thousands of chemicals in biological and environmental matrices has made this goal feasible. Artificial intelligence and machine learning (AI/ML) have become a transformative force in parsing this complexity, primarily by enhancing the ability to identify previously unknown chemicals from spectral data. However, this perspective argues that while expanding the catalog of detected chemicals is essential, the ultimate goal of AI/ML in chemical exposomics is shifting towards functional studies of chemicals with the greatest potential to disrupt biological systems and cause disease, referred to as functional chemical exposomics, which is the comprehensive study of biological activities, mechanistic pathways, and phenotypic responses triggered by all environmental exposures throughout a lifetime. We propose a paradigm shift in which AI transitions from a "discovery engine" to a "functional prediction engine," integrating multimodal data (chemical structures, predicted toxicity, and biological response profiles) for exposomic screening. This article outlines the framework of this vision, discusses pioneering research supporting this transition, and explores key challenges and future directions for achieving proactive, health-related functional chemical exposomics.

Keywords: Exposomics, AI/ML, Environmental Exposures, Mass spectrometry

Introduction

Exposomics refers to the totality of environmental exposures in the lifetime of an organism. These exposures possess immense chemical complexity, encompassing thousands of synthetic and natural compounds, their metabolites, and transformation products, presenting an analytical challenge. Non-targeted analysis based on high-resolution mass spectrometry (HRMS) has become the major technique for discovering these exposures within complex environmental and biological systems, consequently generating massive datasets. However, this capability reveals a critical bottleneck: our current ability to generate data in the field of exposomics far exceeds our capacity to interpret it for meaningful health risk assessment. Traditional workflows heavily rely on existing reference standards and researcher expertise in mass spectrometry-based experiments, which often proves inefficient and fails to comprehensively discover potential exposures. Consequently, a substantial portion of the chemical exposome remains in the "dark matter" state—either undetected or with its health risks uncharacterized[1].

Artificial intelligence and machine learning (AI/ML) are beginning to bridge this gap in mass spectrometry data interpretation. A growing number of applications, including deep learning models for peak detection, deconvolution, and compound annotation, are enhancing the throughput and accuracy of identifying known chemicals. Tools such as CFM-ID[2], BUDDY[3], and SIRIUS[4] have emerged. Furthermore, advanced techniques like Graph Neural Networks (GNNs) demonstrate the capability to directly infer unknown structures from molecular data and establish reliable structure-activity relationships[5].

* Correspondence: Hemi Luan (luanhm@gdut.edu.cn)

Full list of author information is available at the end of the article.

https://doi.org/10.66178/aie-0026-0008

Although these advancements have greatly contributed to advances in chemical exposomics, they have primarily focused on expanding the list of chemicals associated with environmental exposures[6]. This has gradually made us realize that there are more chemicals than previously thought, but we are still faced with a lack of systematic strategies to identify which are the most important and health-risking exposures. A necessary paradigm shift is required in AI/ML-driven exposomics, moving from a focus primarily on chemical structure to biological function. The next frontier lies in building intelligent systems that not only detect and quantify more chemical exposures but also rapidly evaluate them, predict their interactions with biological systems, and assess their potential to disrupt health homeostasis. By directly linking exposure features to possible biological activities, AI/ML can bridge the gap between exposomic discoveries and actionable insights in environmental health risk assessment. AI/ML-based approaches will enable the translation of mass spectrometry data into prioritized candidate lists for mechanistic validation and public health interventions.

AI/ML facilitating the discovery of chemical exposures

Facing the continuous emergence of chemical exposures and their potential health risks, AI/ML-powered non-targeted analysis is advancing chemical exposomics, overcoming the technical hurdles of mass spectrometry data interpretation and serving as a tool to illuminate the "dark matter" of chemical exposures present in the environment and within organisms (Fig. 1).

AI/ML for mass spectrometry data preprocessing

Mass spectrometry data typically represent compounds as peaks, and the mass-to-charge ratio and morphology of these peaks contain information about the presence and quantity of compounds. AI/ML can help with the challenging task of the extraction of these peaks, such as peakonly[7], CPVA[8], MetaClean[9], and other tools, which have been developed to reduce peaks that do not meet data quality requirements. For mass spectrometry analysis of large-scale samples, quality control of data is also crucial. AI/ML is used for batch effect correction to guarantee that large-scale data are robust and reproducible. These tools include statTarget[10], MetNormalizer[11], SERRF[12], iDIA-QC[13], etc. Notably, a probabilistic generative model based on a fully connected variational autoencoder was used for 3D data for mass spectrometry imaging for unsupervised analysis and peak learning to reveal the hidden relationships of chemical exposures in biological tissues when using mass spectrometry imaging techniques[14]. AI/ML-based approaches have advanced the preprocessing techniques for mass spectrometry data and improved the reliability of data quality.

AI-enhanced spectral interpretation

High-resolution mass spectrometry (HRMS) provides information such as mass-to-charge ratio (MS1), isotopic distribution, and fragment masses (MS2) for identifying compound structures. The traditional approach to identifying compound structures from mass spectra is to compare them against spectral libraries, such as NIST, GNPS, and MassBank. The accuracy of library matching heavily depends on spectral similarity algorithms. Traditional spectral similarity algorithms (e.g., cosine similarity) cannot fully characterize structural similarity. Huber et al. developed a novel spectral similarity algorithm, Spec2Vec, based on machine learning[15]. Spec2Vec treats each peak in an MS/MS spectrum as a word in a language, analyzes the distribution of peak sets to generate abstract spectral embeddings, and finally calculates the cosine similarity between these embeddings. Additionally, Convolutional Neural Networks (CNNs) and, more recently, transformer-based architectures are trained to predict Electron Ionization (EI) or Tandem Mass Spectrometry (MS/MS) spectra from molecular structures[16]. This enables "in-silico library generation" and reverse library searching, significantly improving the annotation rate and confidence for compounds missing from physical libraries. Deep learning methods for deconvolution and peak detection also enhance data quality, separating co-eluting compounds and distinguishing true compound signals from noise signals. With the help of AI/ML-assisted spectral interpretation capabilities, chemical exposomics can reveal more of the "dark matter" hidden in mass spectrometry data.

Structural elucidation of unknowns

For features without matches in experimental databases of standard compounds, researchers initially relied on rules of mass spectral fragmentation across instruments, followed by in-silico fragmentation of candidate chemical structures to generate simulated tandem mass spectra. AI/ML has enabled strategies for directly predicting mass spectra from chemical structures. For example, Ruttkies et al. developed the MetFrag[17] tool for inferring chemical structures from mass spectra. This strategy considers a wider range of potential bond cleavages and ranks candidates based on specific MS2 information (e.g., bond dissociation energy, fragment mass, and fragment abundance). Chen et al. used deep learning, training on parameters such as collision energy, instrument type, and adducts accumulated in libraries, to predict in-silico tandem mass spectra (MS/MS) for known or suspected compounds without chemical reference standards[16]. Kind

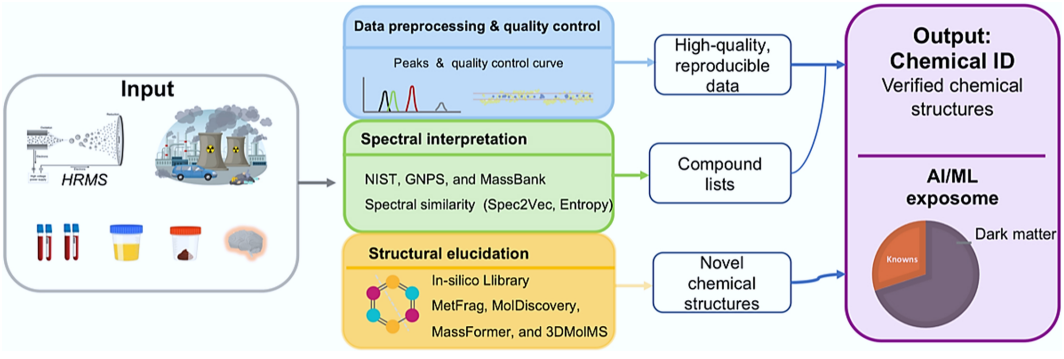

Fig. 1 AI/ML-assisted HRMS helps identify the chemical exposome and integrates with processes such as peak detection, data quality control, mass spectral deconvolution, and compound annotation.

et al. employed a comparable strategy for predicting the chemical structures of lipids from mass spectra[18]. This approach relies on underlying statistical models and explicit training sets containing experimental mass spectra for specific compounds. Other tools in this category include MolDiscovery[19], MassFormer[20], and 3DMolMS[21].

Paradigm shift: from "chemical structure" to "functional activity"

The goal of environmental exposome is to examine the impact of the environment on health from a global and dynamic perspective. Chemicals from multiple environmental exposures exert endocrine disruption effects (bisphenol, phthalates) by binding to various types of receptors (estrogen, aryl hydrocarbon receptor, etc.), inhibiting the activity of metabolic enzymes, disrupting metabolic pathways in the body, inducing oxidative stress, and altering epigenetic markers[22]. Faced with continuously discovered new chemical exposures, the question arises as to which exposures or combinations of exposures are associated with health risks. Transforming exposomics from "chemical structure" to the discovery of "functional activity" in toxicology and risk assessment has become particularly important. It is essential to integrate the chemical identity of exposure groups with multimodal AI/ML frameworks that combine multiple streams of biological and toxicological data (Fig. 2).

Knowledge-driven data

This knowledge includes AI-annotated chemical structures and knowledge information from HRMS, which encompasses public toxicity databases such as ToxCast[23] and PubChem BioAssay, providing experimental activity data; computational toxicology parameters, such as physicochemical properties, absorption, distribution, metabolism, and excretion (ADME) parameters predicted by QSAR, and toxicity endpoints from TOXRIC[24]; transcriptomic, metabolomic, or proteomic profiles from post-exposure studies in vitro or in vivo, which provide biomarkers for biological perturbation characteristics. Based on the constructed knowledge graph and multimodal AI/ML model, these multimodal data can be integrated to create a unified "chemical-biological" spectrum for each feature to form a multimodal correlation artificial intelligence model. A robust, data-driven approach is to model statistical relationships between external exposure groups (HRMS characteristics) and internal responses (other omics layers) in the same cohort. Sparse multivariate models (e.g., penalized regression, typical correlation analysis combined with machine learning) can identify which specific exposome traits are most strongly associated with specific metabolic or gene expression changes that indicate stress pathways, inflammation, or cellular dysfunction. This establishes a direct link between exposure and biological effects.

Causal inference with machine learning

Correlation between chemical exposure and health risk does not imply causation. To move beyond correlation, exposomics can adopt causal inference methods from epidemiology and machine learning. Techniques like instrumental variable analysis, propensity score matching (adapted for high-dimensional data), or causal structure learning can be applied to observational human data to estimate the potential causal effects of exposure mixtures on health outcomes while controlling for confounders. It is also worth noting that such methods require rigorous assessment. Causal inference tools may face serious challenges due to high collinearity and unmeasured confounders. For example, unmeasured confounders are common; factors such as environment, diet, and genetics affect exposure and outcomes, but are rarely fully captured. In addition, large sample sizes are also required for causal structure learning[25]. One method linking exposure-health relationships to intermediate metabolic mechanisms is the "meet-in-the-middle" strategy[26]. This approach can identify environmental factors significantly associated with health status, then map health-related metabolic pathways via non-targeted metabolomics, and finally intersect these exposure-health relationships with the perturbed metabolic network. This reduces false positives in high-dimensional data and helps link environmental triggers to biological consequences. Finally, experimental validation of causal hypotheses derived from AI/ML—derived causal hypotheses, such as targeted in

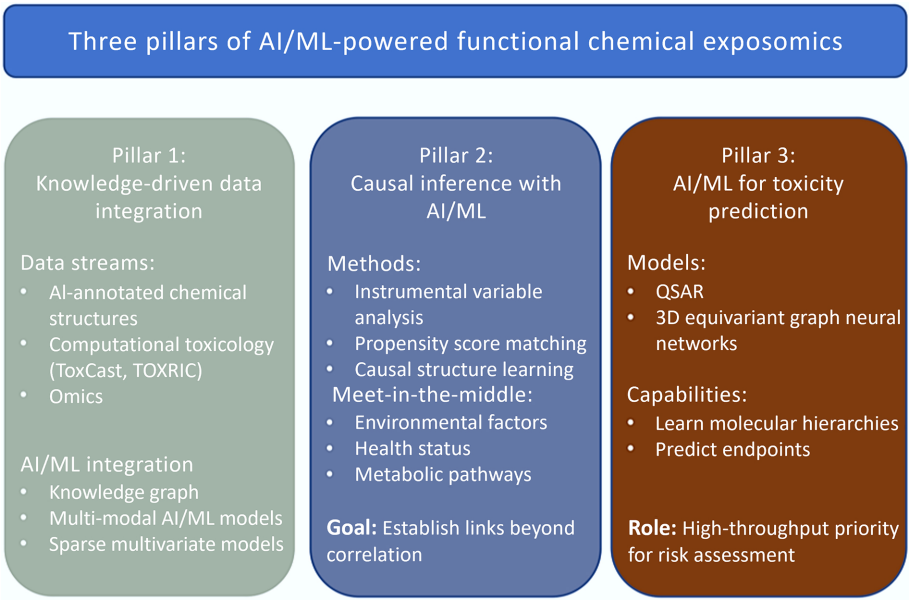

Fig. 2 A functional chemical exposomics approach transitioning from "chemical structure" to discovering "functional activity", combining the chemical identity with multi-modal data under an AI/ML framework.

https://doi.org/10.66178/aie-0026-0008

vitro systems, organoids, or in vivo models—will help form a closed loop between prediction and mechanism.

AI/ML for Toxicity Prediction

Compared to traditional QSAR models, chemo-informatics AI/ML models can directly predict specific toxicological endpoints for chemicals with putative structures. Cremer et al. developed a method using Equivariant Graph Neural Networks based on 3D structure learning to integrate molecular conformation directly into machine-learning-based prediction workflows for toxicity in physical space[5]. Their work demonstrated that this approach yields results comparable to state-of-the-art 2D graph-based models and outperforms SMILES string-based Transformer models across all datasets and tasks. GNNs learn hierarchical representations of molecules, capturing features like functional groups and molecular shape that govern interactions with biological targets. These models can be trained to predict endpoints such as endocrine disruption potential, acute toxicity, or developmental neurotoxicity, serving as primary priority filters[27].

By building a multimodal AI/ML framework, AI/ML-Powered Chemical Exposomics provides not only a chemical inventory but also a priority dashboard. Each chemical receives a comprehensive "biological functional activity risk score," including its predicted toxicity endpoints, hypothetical molecular initiating events, and the strength of its correlation with adverse omics features. As shown in Fig. 3, we further propose a framework that provides a roadmap for implementation and validation. This framework directly provides environmental scientists and regulators with alerts on the most likely health risk threats.

Conclusion and perspective

Functional chemical exposomics is an integrated analytical framework that combines HRMS metrology, AI/ML-driven bioactivity prediction, and multi-omics response analysis, directly linking known or unidentified environmental chemical features to specific biological perturbations and toxicity pathways. It shifts the analytical endpoint from chemical identification to functional prioritization, enabling risk assessment even for unknown or poorly characterized compounds. The implementation of robust, function-oriented exposomics also faces significant challenges. First, high-quality, well-annotated "chemical structure-biological activity" data are still scarce, especially for environmentally relevant metabolites and mixtures[28]. There is an urgent need for interdisciplinary collaborative databases that link the biological activity data of these exposomic chemicals to other biological experiments, as well as changes in temporal and spatial dimensions. Second, models must be built with interpretability in order to obtain regulatory approval and a solid scientific foundation[29]. Third, having good generalization capabilities is essential for models trained in known chemical spaces to perform well on novel, structurally different environmental unknown substances, and training on broader

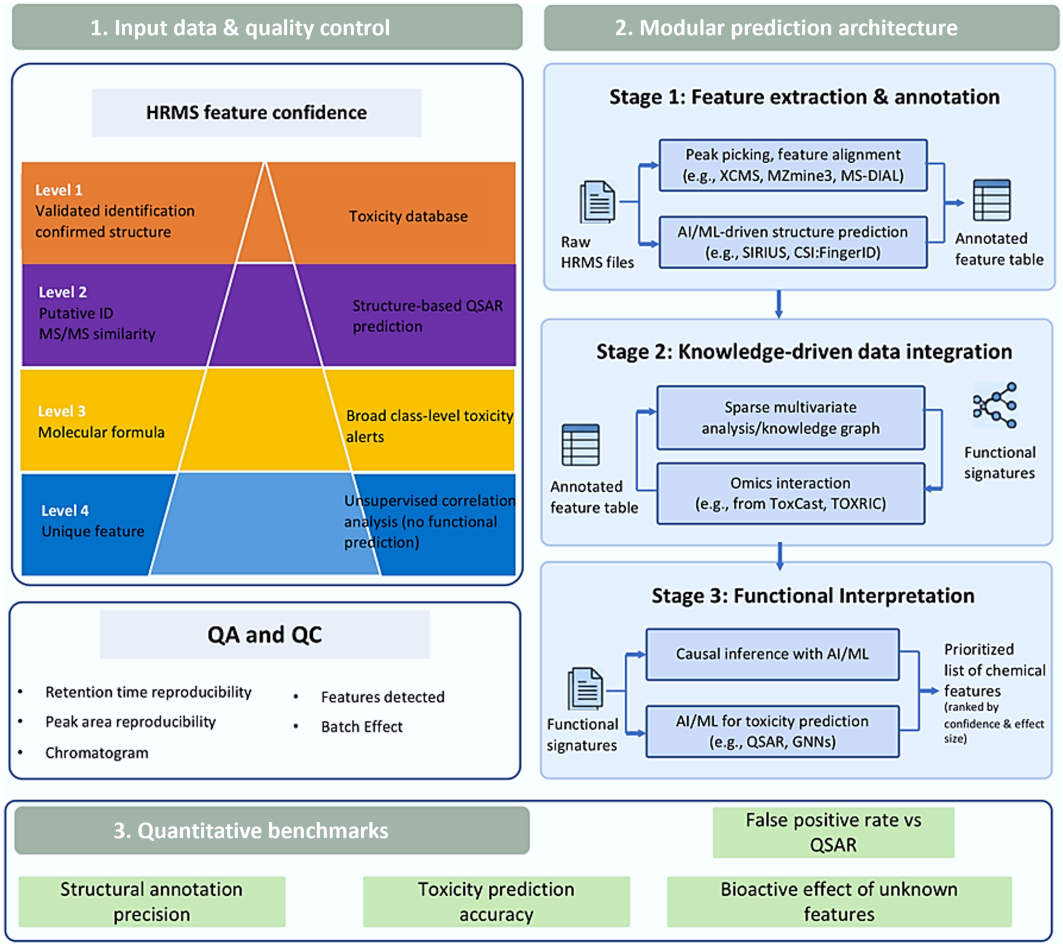

Fig. 3 A proposed framework that provides a roadmap for implementation and validation. (1) Input data & quality control; (2) Modular prediction architecture; (3) Quantitative benchmarks.

biochemical principles rather than just structural features is necessary. Fourth, real-world exposures are complex mixtures, and AI/ML models are being developed to predict additive, synergistic, or antagonistic effects. These models capture relationships between features using techniques such as GNNs, and calculate the combined impact of multiple environmental chemicals on disease or biological outcomes, and they may assess potential mixed effects of complex mixtures[30–34]. The rapid development of large language models can help extract the "chemical-activity" relationships hidden in vast and unstructured toxicology literature, supplementing structured databases[35]. AI/ML-assisted exposomics will enable the flow of HRMS data through AI systems for advanced annotation and structural inference; these chemical identities are then input into integrated toxicology knowledge and multi-omics associations, resulting in a high-priority candidate list sorted by biological activity potential. This list is ultimately used to guide targeted analysis, experimental validation of mechanisms, and informed risk assessment. Achieving this vision requires a higher level of synergy. Chemists, bioinformaticians, computer scientists, toxicologists, and epidemiologists must work closely together to develop, benchmark, and validate these activity-centric next-generation AI/ML tools that transform the science of exposomics from an exposure inventory into a predictive, preventive framework for public health action, ultimately allowing us to focus resources on mitigating environmental chemicals that are truly important to human health.

Ethical statements

During the preparation of this work, the authors used Grammarly (Date: 2026-03-11) for Grammar Check. The authors reviewed and edited all content produced with the assistance of this tool, verified its accuracy, and take full responsibility for the integrity and originality of the final manuscript. This work represents the authors' own intellectual contribution, and no AI tool is credited as an author.

Author contributions

The authors confirm their contributions to the paper as follows: Hemi Luan: Conceptualization, Writing – Original Draft. Tiangang Luan: Writing – Review & Editing. All authors read and approved the final manuscript. All authors disclosed no relevant relationships.

Data availability

No data were used for the research described in the article.

Funding

The authors would like to acknowledge the financial support from the National Natural Science Foundation of China (Grant Nos. 22376034 and 22536007).

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Biomedical Engineering, School of Biomedical and Pharmaceutical Sciences, Guangdong University of Technology, Guangzhou 510006, Guangdong, China; [2]Guangdong Provincial Laboratory of Chemistry and Fine Chemical Engineering Jieyang Center, Jieyang 515200, China; [3]School of Environmental and Chemical Engineering, Jiangmen Key Laboratory of Synthetic Chemistry and Cleaner Production, Wuyi University, Jiangmen 529020, China

References

[1] Luan H, Sun Z, Zhao Y, Luan T. 2025. Mass spectrometry insight into environmental chemical exposome and metabolome in the microbiota-gut-brain axis. *TrAC Trends in Analytical Chemistry* 192:118361

[2] Wang F, Allen D, Tian SY, Oler E, Gautam V, et al. 2022. CFM-ID 4.0-a web server for accurate MS-based metabolite identification. *Nucleic Acids Research* 50:W165–W174

[3] Xing S, Shen S, Xu B, Li X, Huan T. 2023. BUDDY: molecular formula discovery via bottom-up MS/MS interrogation. *Nature Methods* 20:881–890.

[4] Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, et al. 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* 16:299–302

[5] Cremer J, Sandonas LM, Tkatchenko A, Clevert DA, De Fabritiis G. 2023. Equivariant Graph Neural Networks for Toxicity Prediction. *Chemical Research in Toxicology* 36:1561–1573

[6] Bao Q, Yu N, Jiang Q, Wang X, Jiao Z, Wei S. 2025. Recent Advances in Implementation of Machine Learning for Environmental Nontarget Identification. *Environment & Health* 3:1286–1298

[7] Melnikov AD, Tsentalovich YP, Yanshole VV. 2020. Deep Learning for the Precise Peak Detection in High-Resolution LC-MS Data. *Analytical Chemistry* 92:588–592

[8] Luan H, Jiang X, Ji F, Lan Z, Cai Z, Zhang W. 2020. CPVA: a web-based metabolomic tool for chromatographic peak visualization and annotation. *Bioinformatics* 36:3913–3915

[9] Chetnik K, Petrick L, Pandey G. 2020. MetaClean: a machine learning-based classifier for reduced false positive peak detection in untargeted LC-MS metabolomics data. *Metabolomics* 16:117

[10] Luan H, Ji F, Chen Y, Cai Z. 2018. statTarget: A streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data. *Analytica Chimica Acta* 1036:66–72

[11] Shen X, Gong X, Cai Y, Guo Y, Tu J, et al. 2016. Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* 12:89

[12] Fan S, Kind T, Cajka T, Hazen SL, Tang WHW, et al. 2019. Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Analytical Chemistry* 91:3590–3596

[13] Gao H, Zhu Y, Wang D, Nie Z, Wang H, et al. 2025. iDIA-QC: AI-empowered data-independent acquisition mass spectrometry-based quality control. *Nature Communications* 16:892

[14] Abdelmoula WM, Lopez BG, Randall EC, Kapur T, Sarkaria JN, et al. 2021. Peak learning of mass spectrometry imaging data using artificial neural networks. *Nature Communications* 12:5544

[15] Huber F, Ridder L, Verhoeven S, Spaaks JH, Diblen F, et al. 2021. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology* 17:e1008724

[16] Chen B, Li H, Huang R, Tang Y, Li F. 2024. Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules. *Nature Communications* 15:8396

[17] Ruttkies C, Neumann S, Posch S. 2019. Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics* 20:376

[18] Kind T, Liu KH, Lee DY, DeFelice B, Meissen JK, Fiehn O. 2013. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nature Methods* 10:755–758

[19] Cao L, Guler M, Tagirdzhanov A, Lee YY, Gurevich A, Mohimani H. 2021. MolDiscovery: learning mass spectrometry fragmentation of small molecules. *Nature Communications* 12:3718

[20] Young A, Röst H, Wang B. 2024. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence* 6:404–416

[21] Hong YH, Li SJ, Welch CJ, Tichy S, Ye YZ, Tang HX. 2023. 3DMolMS: prediction of tandem mass spectra from 3D molecular conformations. *Bioinformatics* 39:btad354

https://doi.org/10.66178/aie-0026-0008

[22] Hou L, Zhang X, Wang D, Baccarelli A. 2012. Environmental chemical exposures and human epigenetics. *International Journal of Epidemiology* 41:79–105

[23] Kim D, Choi J. 2025. AI-based toxicity prediction models using ToxCast data: Current status and future directions for explainable models. *Toxicology* 517:154230

[24] Wu L, Yan B, Han J, Li R, Xiao J, et al. 2023. TOXRIC: a comprehensive database of toxicological data and benchmarks. *Nucleic Acids Research* 51:D1432–D1445

[25] Schisterman EF, Perkins NJ, Mumford SL, Ahrens KA, Mitchell EM. 2017. Collinearity and Causal Diagrams. *Epidemiology* 28:47–53

[26] Demetriou CA, Esposti DD, Fedinick KP, Russo F, Robinson O, Vineis P. 2020. The exposome and meet-in-the-middle as tools in addressing open questions in air pollution research. *European Journal of Public Health* 30:V559–V559

[27] Zhao X, Zhang S, Zhang T, Cao Y, Liu J. 2025. A small-scale data driven and graph neural network based toxicity prediction method of compounds. *Computational Biology and Chemistry* 117:108393

[28] Vermeulen R, Schymanski EL, Barabasi AL, Miller GW. 2020. The exposome and health: Where chemistry meets biology. *Science* 367:392–396

[29] Singh C, Askari A, Caruana R, Gao J. 2023. Augmenting interpretable models with large language models during training. *Nature Communications* 14:7913

[30] Guo J, Luo YW, Fang C, Jin JS, Xia P, et al. 2024. Advancing the Effect-Directed Identification in Combined Pollution: Using Pathways to Link Effects and Toxicants. *Environmental Science & Technology* 58:18642–18653

[31] Cremer J, Medrano Sandonas L, Tkatchenko A, Clevert DA, De Fabritiis G. 2023. Equivariant graph neural networks for toxicity prediction. *Chemical Research in Toxicology* 36(10):1561–1573

[32] Limbu S, Glasgow E, Block T, Dakshanamurthy S. 2024. A machine-learning-driven pathophysiology-based new approach method for the dose-dependent assessment of hazardous chemical mixtures and experimental validations. *Toxics* 12(7):481

[33] Song Y, Asselman J, De Scharnphelaere KAC, Salbu B, Tollefsen KE. 2018. Deciphering the Combined Effects of Environmental Stressors on Gene Transcription: A Conceptual Approach. *Environmental Science & Technology* 52:5479–5489

[34] Wu TX, Zhao L, Ren MY, He S, Zhang L, et al. 2025. Small-Sample Learning for Next-Generation Human Health Risk Assessment: Harnessing AI, Exposome Data, and Systems Biology. *Environmental Science & Technology* 59:5–10

[35] Lane TR, Vignaux PA, Harris JS, Snyder SH, Urbina F, Ekins S. 2025. Machine Learning and Large Language Models for Modeling Complex Toxicity Pathways and Predicting Steroidogenesis. *Environmental Science & Technology* 59:13844–13856

Copyright: © 2026 by the author(s). Published by NEW HORIZON PRESS LIMITED, TSIM SHA TSUI, HK, China. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.