scientific reports

Check for updates

OPEN

Computational insights into exploring the potential effects of environmental contaminants on human health

Fuyan Cao[1,4], Xinyue Zhao[2,4], Xueqi Fu[1,2,3] & Yue Jin[1,2,3]✉

With rapid industrialization and urbanization, the increasing prevalence of air and water pollutants poses a significant threat to public health. Traditional research methods, such as epidemiological studies and in vitro/in vivo experiments, provide valuable biological insights but are often costly, time-consuming, and limited in scale. To address this gap, this study develops a machine learning-based approach to predict the carcinogenicity of pollutants. Using the dataset of carcinogenic and non-carcinogenic molecules that we collected, the pretrained KPGT model trained with molecular fingerprints and descriptors achieved an AUC of 0.83, surpassing traditional machine learning models. To validate this model, common pollutants from air and water sources were analyzed. Further clustering classified these pollutants into five distinct groups. Target prediction analysis identified key genes associated with representative pollutant molecules, such as MAPK1, MTOR, and PTPN11. GO and KEGG pathway analyses, along with survival analysis, revealed potential carcinogenic mechanisms and prognostic implications. Our findings contribute to improved pollution risk assessment and evidence-based environmental policy development, ultimately aiding in the mitigation of pollutant-related health risks.

Keywords Environmental pollutant, Machine-learning, Survival analysis, Human health

Environmental pollution, driven primarily by human activities, has exceeded the environment's self-purification capacity, posing serious threats to public health. Pollutants can either degrade into harmless substances or undergo chemical transformations that enhance their toxicity[1]. Globally, pollution is a major contributor to morbidity and mortality, accounting for 24% of the total disease burden and 32% of deaths[2,3]. The health effects of pollutants range from acute conditions, such as respiratory irritation and allergic reactions, to chronic diseases, including cardiovascular disorders and neurodegenerative conditions[4,5]. These effects are mediated through complex mechanisms, including oxidative stress, DNA damage, and inflammation. Numerous studies have established strong links between exposure to certain pollutants such as asbestos, benzene, and heavy metals and cancer development[6–8].

Cancer remains one of the most pressing global health challenges. By 2022, approximately 20 million new cancer cases and 9.7 million cancer-related deaths were projected worldwide, with lung cancer being the most commonly diagnosed and the leading cause of cancer-related mortality[9]. Lung adenocarcinoma (LUAD), the predominant histological subtype, accounts for the majority of lung cancer cases[10–12]. Air pollution has been identified as a significant environmental factor in lung cancer pathogenesis, contributing to 7 million deaths annually and affecting multiple organ systems, including the respiratory, cardiovascular, and nervous systems[13–17]. Chronic exposure to airborne pollutants, such as fine particulate matter (PM2.5), volatile organic compounds, and heavy metals, has been associated with increased lung cancer risk, likely through mechanisms involving DNA damage, chronic inflammation, and epigenetic modifications[18,19].

Water pollution similarly presents a severe public health hazard. Industrial discharge, agricultural runoff, and domestic wastewater introduce a wide array of contaminants into water bodies, including arsenic, chromium, nickel, beryllium, aniline, benzo[a]pyrene, and other polycyclic aromatic hydrocarbons (PAHs)[20]. Prolonged

[1]Key Laboratory for Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun 130012, China. [2]Edmond H. Fischer Signal Transduction Laboratory, School of Life Sciences, Jilin University, Changchun 130012, China. [3]National Engineering Laboratory of AIDS Vaccine, School of Life Sciences, Jilin University, Changchun 130012, Jilin, China. [4]Fuyan Cao and Xinyue Zhao contributed equally to this work. ✉email: yuejin@jlu.edu.cn

exposure to these pollutants, either through direct consumption or bioaccumulation in the food chain, has been linked to an elevated risk of digestive system cancers, particularly those affecting the esophagus and stomach[21–23]. The mechanisms underlying water pollution-induced carcinogenesis are complex and involve oxidative stress, genotoxicity, and endocrine disruption.

Understanding the link between environmental pollutants and cancer is crucial for developing effective prevention and mitigation strategies[24–26]. Traditional research approaches, such as mass spectrometry, chromatography, cellular assays, and animal models, have provided valuable insights into pollutant toxicity[27,28]. However, these methods are often labor-intensive, time-consuming, and ethically challenging, particularly when dealing with highly toxic substances. Additionally, their scalability is limited, making them less suitable for large-scale systematic analysis. The emergence of bioinformatics and computational modeling has provided new opportunities to study pollutant-induced carcinogenesis by integrating multi-omics data and leveraging machine learning techniques[29,30].

In this study, we developed a machine learning-based approach to predict the carcinogenicity of pollutants and further investigated the potential biological mechanisms of these pollutants. We curated and processed thousands of carcinogenic and non-carcinogenic molecules, applying clustering analysis to explore their structural similarities. Molecules with significant structural differences were removed to refine the dataset, ensuring a robust classification framework. Using molecular fingerprints and descriptors, we trained a knowledge-guided pre-trained graph transformer (KPGT) model[31], achieving an AUC of 0.83, which outperformed conventional machine learning models. Additionally, we collected common pollutants from two different sources air and water to validate the model's ability to predict carcinogenicity. Furthermore, we conducted clustering analysis, target prediction, and survival analysis on these pollutants to gain deeper insights into their potential carcinogenic effects. The workflow of our study is illustrated in Fig. 1.

By integrating pollutant datasets, molecular fingerprints, machine learning, GO and KEGG pathway analyses, and survival analysis, this study proposes a comprehensive approach to evaluating the carcinogenic potential of air and water pollutants. Our findings contribute to a more systematic and scalable risk assessment approach, facilitating the identification of high-risk pollutants and their potential impact on human health. Furthermore, this research provides scientific evidence to support evidence-based environmental policies, ultimately aiding in the mitigation of pollution-related health risks and the development of targeted public health interventions.

Materials and methods

Data collection and standardization

Carcinogenic Molecule Collection: We directly retrieved carcinogenic molecules from the PubChem database by accessing the CCRIS (Chemical Carcinogenesis Research Information System) database (https://www.nlm.nih.gov/toxnet/Accessing_CCRIS_Content_from_PubChem.html), a specialized resource for chemical carcinogenesis research. These molecules were carefully collected and downloaded from the CCRIS database[32].

Data Cleaning and Preprocessing: For the collected carcinogenic molecules, we first removed duplicate molecules to ensure data uniqueness. We then used PubChemPy[33]to retrieve typical SMILES strings[34]and InChI strings[35] for the collected molecular names. Molecules that could not be converted to SMILES were discarded. InChI strings were used as the standard for deduplication. Further cleaning steps included removing counter-ions, solvent components, and salts, followed by the addition of hydrogen atoms and standardization. Ultimately, 5,041 unique compounds were obtained.

Non-Carcinogenic Molecule Collection: We retrieved 3,187 non-carcinogenic molecules from the PubChem database. After applying the same cleaning procedure, 3,184 non-carcinogenic compounds were included in the final dataset.

Collection of pollutant molecules in air and water

Environmental pollutants were curated from publicly available databases, with 1,225 air pollutants retrieved from the EPA CompTox Chemistry Dashboard (https://www.epa.gov/environmental-topics/air-topics)[36] and T3DB (http://www.t3db.ca)[37], while 827 water pollutants (reported between 2020 and 2023) were obtained from the Toxic Release Inventory (TRI) database (https://www.epa.gov/toxics-release-inventory-tri-program)[38]. Following the same rigorous data cleaning and preprocessing steps as applied to carcinogenic molecules, the final dataset comprised 218 unique air pollutants and 139 unique water pollutants. These pollutants were used to evaluate the model's ability to predict the behavior of unknown pollutants.

Molecular features of the dataset

The RDKit cheminformatics library[39]was used to parse the SMILES strings of each molecule, converting them into molecular objects for further analysis. Molecular descriptors were then computed using RDKit's Descriptors[40] and MoleculeDescriptors modules, including Molecular Weight (MolWt), Topological Polar Surface Area (TPSA), Number of Hydrogen Bond Donors (NumHDonors), Number of Hydrogen Bond Acceptors (NumHAcceptors), LogP (MolLogP), and Molecular Refractivity (MolMR). To ensure consistency across the dataset, all descriptors were standardized using the StandardScope tool.

Cluster analysis of carcinogenic and pollutant molecules

To investigate structural similarities among carcinogenic molecules, air pollutants, and water pollutants, we performed clustering analysis by first extracting SMILES strings from the original dataset and converting them into molecular objects using RDKit (https://www.rdkit.org/). Each molecule was then encoded as a binary fingerprint vector using Morgan fingerprint vectors (length = 2,048), enabling 3D molecular manipulation from

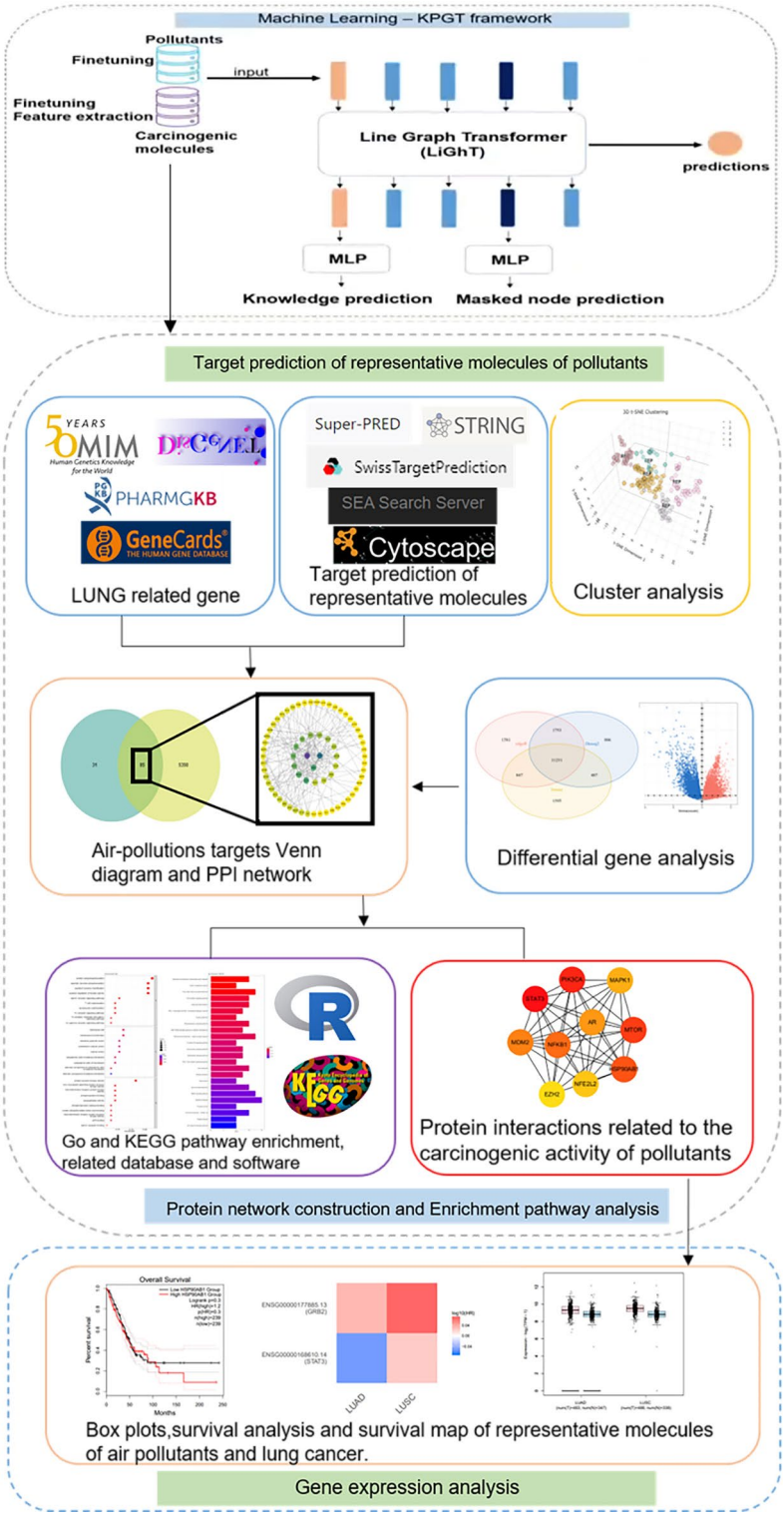

Fig. 1. Overview of the Research Workflow. An illustration of the complete research workflow, detailing each step from data collection to final analysis.

2D representations and facilitating machine learning-based compound descriptor generation, fingerprinting, and structural similarity calculations.

To visualize the high-dimensional fingerprint data of carcinogenic molecules, we applied the t-SNE algorithm, a more efficient clustering technique compared to K-means. Molecules were clustered together if their pairwise distance was less than 1/24 of the maximum pairwise distance in the dataset. Using Matplotlib 3.8.3 (https://matplotlib.org/), we visualized the Clustered Chemical Space Network (CSN), where node radius represents

the molecular distance threshold and edge thickness corresponds to Dice similarity between molecules. The clustering code is available at https://github.com/heyigacu/DistanceClustering[41].

For air and water pollutant molecules, the t-SNE algorithm was similarly applied, reducing the 2,048-dimensional vectors to three dimensions with parameters set as dims = 3, perplexity = 40, and $\theta = 0.0$. Hierarchical clustering was then performed on the t-SNE-transformed data using the ward.D2 method, where Euclidean distances were computed to construct a hierarchical clustering tree, grouping molecules into four clusters. Each molecule was assigned a cluster label accordingly.

For each cluster, we identified representative molecules by calculating the Tanimoto[42]similarity matrix within the cluster. Specifically, we computed pairwise similarities for all molecules within a cluster and determined the average similarity for each molecule. The molecule with the highest average similarity was selected as the representative of that cluster. To ensure uniqueness, only one representative molecule was identified for each cluster. Finally, we visualized the t-SNE reduced three-dimensional data using the plotly package[40], with clusters distinguished by different colors. Representative molecules were labeled as "REP" in the plot for easy identification.

Machine-Learning model

The KPGT molecular pre-training framework[31]integrates molecular fingerprints and descriptors as knowledge-based priors, achieving superior molecular prediction performance across multiple public datasets. This framework incorporates LiGhT (Line Graph Transformer)[43], a high-capacity model specifically designed for accurate molecular graph modeling. By leveraging molecular line graphs, LiGhT captures intricate structural patterns and chemical relationships that conventional Transformer architectures often overlook. To further enhance molecular representation learning, the model employs two positional encoding modules distance encoding and path encoding within the multi-head attention mechanism, improving its ability to model molecular structural information. Built upon a classical Transformer encoder, LiGhT consists of multiple Transformer layers and utilizes a multi-layer perceptron (MLP) to generate both knowledge predictions and masked node predictions. Detailed pre-training strategies and training protocols for KPGT are provided in Supplementary Section S1.

Differential gene analysis

The differential gene expression data were obtained from the UCSC-TCGA database[44] (https://xena.ucsc.edu/, accessed on 17 January 2024), focusing on RNA-seq gene expression datasets for lung adenocarcinoma (LUAD), esophageal carcinoma (ESCA), and gastric adenocarcinoma (STAD). The dataset includes RNA-seq count matrices, sample sizes, and clinical data, ensuring robust and accurate analysis.

To identify differentially expressed genes (DEGs), we applied three widely used statistical methods edgeR, DESeq2, and limma which are well-established for RNA-seq data analysis and known for their reliability in detecting significant gene expression changes. These methods allowed us to systematically identify significantly regulated genes across the studied cancer types, providing crucial insights into pollutant-induced oncogenesis.

Target gene prediction for carcinogenicity assessment

Potential target genes associated with lung adenocarcinoma (LUAD), esophageal carcinoma (ESCA), and gastric adenocarcinoma (STAD) were initially screened using DisGeNET[45] (https://www.disgenet.org/, accessed on 20 January 2024), GeneCards[46] (https://www.genecards.org/, accessed on 20 January 2024), PharmGKB[47] (https://www.pharmgkb.org/, accessed on 20 January 2024), and OMIM[48] (https://www.omim.org/, accessed on 20 January 2024). To refine the target predictions, we further analyzed active components and disease-related target genes using Super-Pred[49] (https://prediction.charite.de/, accessed on 20 January 2024), SEA (Similarity Ensemble Approach)[50] (http://sea.bkslab.org/, accessed on 20 January 2024), and Swiss-Target-Prediction[51] (http://swisstargetprediction.ch/, accessed on 20 January 2024). These databases enabled a comprehensive assessment of potential pollutant-associated molecular targets, facilitating the investigation of their roles in carcinogenesis.

Protein–Protein interaction network

The protein-protein interaction (PPI) network was constructed using the STRING database[52] (http://string-db.org/, accessed on 25 January 2024) to evaluate physical and functional interactions among selected target genes. This network provided insights into the molecular relationships underlying pollutant-associated carcinogenesis.

For further analysis, Cytoscape software[53] (version 3.9.1; https://cytoscape.org/, accessed on 25 January 2024) was used to assess the topological properties of the network and identify key regulatory nodes. Based on network analysis metrics, the top 10 central genes were selected as potential oncogenic targets for further investigation.

Gene ontology and Kyoto encyclopedia of genes and genomes enrichment analysis

To identify key biological processes and pathways associated with pollutant-induced carcinogenesis, we conducted Gene Ontology (GO)[54]and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses[55–57] using R-based bioinformatics tools. GO analysis provided insights into biological processes, molecular functions, and cellular components, while KEGG analysis identified key signaling pathways relevant to pollutant-associated cancer mechanisms. The results were statistically filtered and visualized using bar plots and bubble charts to highlight the most significant pathways.

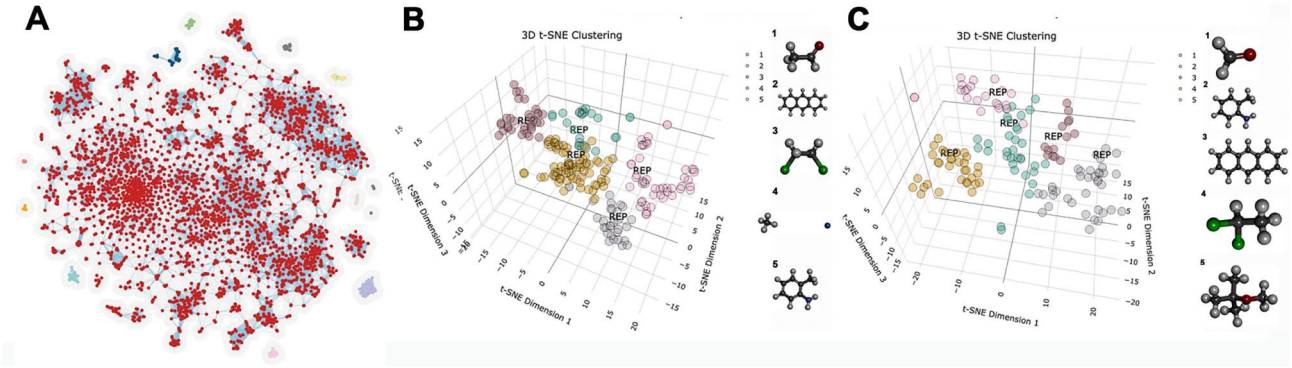

Fig. 2. Clustering Analysis of Molecules. (**A**) Cluster map of carcinogenic molecules. (**B**) Air pollutant clustering results, color-coded with representative molecules. (**C**) Water pollutant clusters and representative molecules.

Category	Number	Representative molecular name	Smiles	Structure
1	26	acetaldehyde	CC = O	
2	32	anthracene	C1 = CC = C2C = C3C = CC = CC3 = CC2 = C1	
3	34	ethylidene chloride	C(CCl)Cl	
4	78	uranium carbide	C.[U]	
5	48	o-toluidine	CC1 = CC = CC = C1N	

Table 1. Cluster analysis results for air pollutant molecules.

Gene expression level analysis

We utilized GEPIA2[58] (http://gepia2.cancer-pku.cn/) to perform survival analysis on the identified hub genes. GEPIA2 leverages data from TCGA and GTEx, offering rapid and customizable analysis functionalities. This platform was employed to assess the expression levels of the top 10 hub genes within the protein-protein interaction network across tumor and normal samples, and to conduct survival analysis, determining the survival contribution of these target genes.

For lung adenocarcinoma (LUAD), esophageal carcinoma (ESCA), and gastric carcinoma (STAD), we selected genes linked to air pollutants for box plots, survival analysis, and mapping. Similarly, for ESCA and STAD, genes associated with water pollutants were analyzed to evaluate their impact on patient survival.

Results and discussion

Analysis of carcinogenic molecule clustering

Carcinogenic molecules were first collected from public databases and subjected to preprocessing, including de-duplication and SMILES standardization, yielding 5,041 unique compounds. To refine the dataset, clustering analysis was conducted using the t-SNE algorithm and Chemical Space Network (CSN) visualization (Fig. 2A). Molecules were grouped into 14 clusters based on structural similarity, with light blue edges indicating relationships above a predefined threshold. Molecules with significant structural differences were excluded, resulting in a curated subset of 3,028 carcinogenic molecules.

These refined molecules served as the positive dataset for training the pre-trained KPGT machine learning model, with non-carcinogenic molecules forming the negative dataset. The clustering step ensured that structurally coherent molecules were retained, enhancing the model's learning efficiency and predictive accuracy. For model validation and generalization, the dataset was randomly partitioned into training and test sets at an 8:2 ratio.

Clustering analysis of air and water pollutants

Air pollutant molecules were clustered into five distinct categories (Fig. 2B; Table 1), containing 26, 32, 34, 78, and 48 molecules, respectively. The representative molecules identified for each cluster were acetaldehyde, anthracene, ethylidene chloride, uranium carbide, and o-toluidine. Similarly, five clusters were identified for water pollutants (Fig. 2C; Table 2), containing 33, 37, 14, 33, and 22 molecules, with the representative molecules being formaldehyde, o-toluidine, anthracene, ethylidene chloride, and methyl tert-butyl ether. These representative molecules were subsequently used for target gene prediction and enrichment analyses, providing a foundation for exploring pollutant-induced carcinogenic mechanisms.

Category	Number	Representative molecular name	Smiles	Structure
1	33	formaldehyde	C = O	
2	37	o-toluidine	CC1 = CC = CC = C1N	
3	14	anthracene	C1 = CC = C2C = C3C = CC = CC3 = CC2 = C1	
4	33	ethylidene chloride	CC(Cl)Cl	
5	22	methyl tert-butyl ether	CC(C)(C)OC	

Table 2. Cluster analysis results for water pollutant molecules.

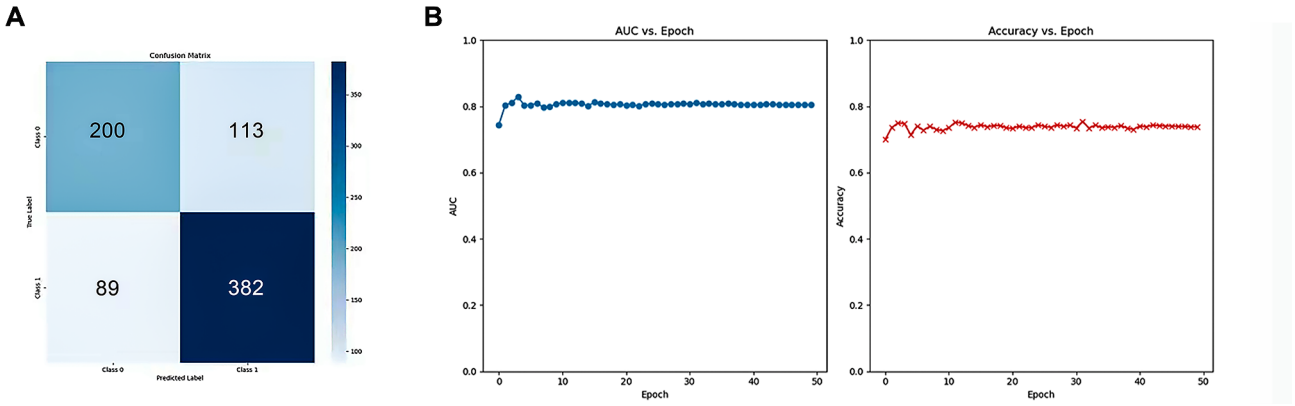

Fig. 3. Machine learning. (**A**) Confusion Matrix. (**B**) After training for 50 epochs, the AUC and Accuracy of KPGT.

Model	AUC	SE	SP
SVM	0.81	0.60	0.76
RF	0.80	0.57	0.76
XGBoost	0.80	0.65	0.76
KPGT	0.83	0.63	0.81

Table 3. Model comparison.

Machine-Learning model performance

To assess the predictive capacity of the KPGT model, we trained the preprocessed positive and negative datasets, leveraging molecular fingerprints and descriptors for enhanced prediction accuracy. Air and water pollutant molecules served as an external validation set, allowing evaluation of the model's ability to predict the carcinogenicity of unseen compounds (Fig. 3A). After 50 epochs of training, the KPGT model achieved an AUC of 0.83 (Fig. 3B), with an accuracy of 0.74, a balanced accuracy (BA) of 0.73, and an F1 score of 0.66. Furthermore, the model was compared with SVM, RF, and XGBoost models with similar functionality (Table 3), demonstrating its effectiveness in predicting the carcinogenicity of pollutant molecules.

Differential gene expression analysis

Differential gene expression analysis using edgeR, DESeq2, and limma on the UCSC-TCGA database revealed significant transcriptional changes associated with pollutant exposure. A total of 11,231 genes were upregulated and 10,109 were downregulated in esophageal cancer (Fig. 4A), while 14,219 upregulated and 9,561 downregulated genes were identified in gastric cancer (Fig. 4B). In lung adenocarcinoma, 9,338 genes exhibited upregulation, whereas 5,349 were downregulated (Fig. 5). These findings established a comprehensive gene expression landscape, forming the foundation for target gene identification and functional enrichment analysis, further elucidating the molecular mechanisms underlying pollutant-induced carcinogenesis.

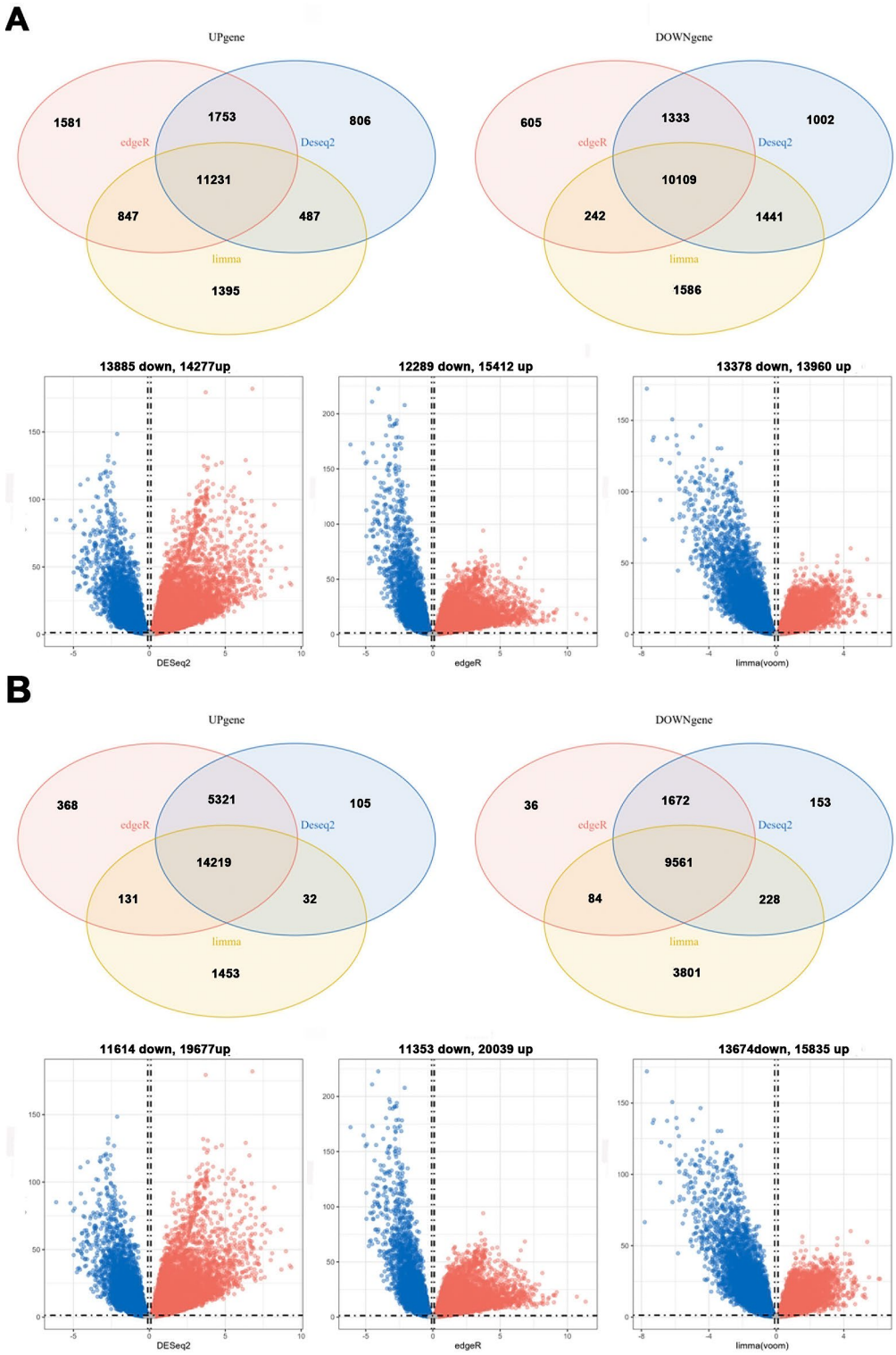

Fig. 4. Differential gene analysis of esophageal cancer and gastric cancer genes in the UCSC-TCGA database was performed using edgeR, Deseq2, and limma methods. Three differential analysis methods for analyzing upregulation and downregulation genes in Venn plots. Use volcano plots to display the upregulation and downregulation genes of three different differential analyses (the blue dots represent downregulated genes, while the red dots represent upregulated genes). (**A**) Water Pollutants-Esophageal Cancer System; (**B**) Water Contaminants-Gastric Cancer System.

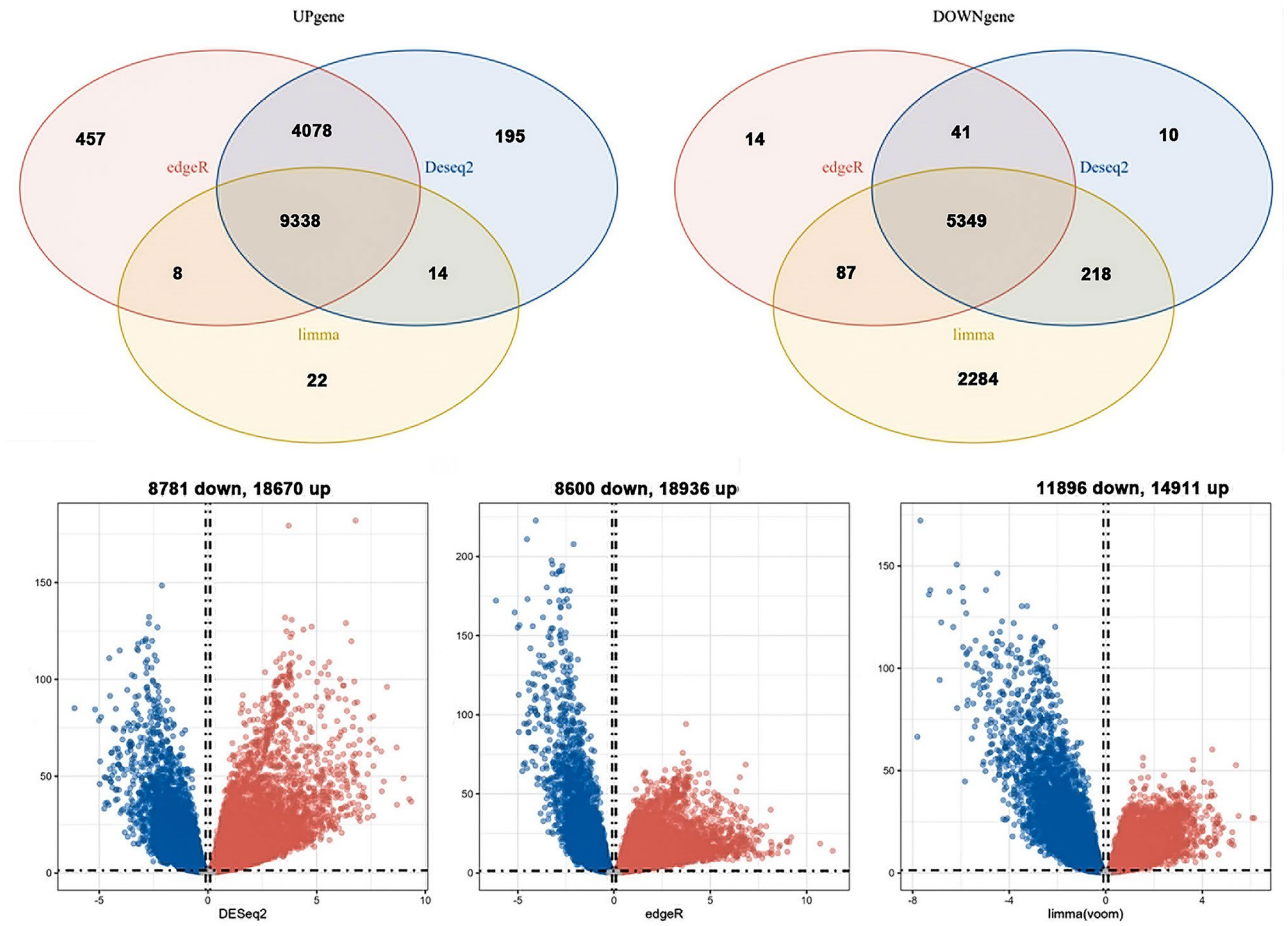

Fig. 5. **Differential gene analysis of lung adenocarcinoma genes in the UCSC-TCGA database was performed using edgeR, Deseq2, and limma methods.** Three differential analysis methods for analyzing upregulation and downregulation genes in Venn plots. Volcano plots displaying upregulated (red dots) and downregulated genes (blue dots) for each differential analysis method.

Prediction of Pollutant-Associated target genes and PPI network construction

To identify potential pollutant-associated target genes, we screened DisGeNET, GeneCards, PharmGKB, and OMIM databases for lung adenocarcinoma, esophageal cancer, and gastric cancer. In DisGeNET, we identified 380, 130, and 254 potential target genes for lung adenocarcinoma, esophageal cancer, and gastric cancer, respectively (score > 0.1). In GeneCards, 3,053, 1,721, and 2,067 target genes were identified (score > 10), while in PharmGKB, 100, 14, and 54 genes were associated with these cancers (score > 100). In OMIM, we identified 14, 5, and 11 genes with relevant associations. After removing duplicates across the four databases, we obtained 6,052 unique target genes for lung adenocarcinoma, 1,770 for esophageal cancer, and 2,119 for gastric cancer. By intersecting these with differentially expressed genes, we refined the dataset to 5,435 genes for lung adenocarcinoma, 1,509 for esophageal cancer, and 1,789 for gastric cancer (Fig. 6).

Identifying pollutant-specific gene targets involved applying Super-Pred, SEA, and Swiss-Target-Prediction to predict molecular interactions based on the active components of acetaldehyde, anthracene, ethylidene chloride, uranium carbide, o-toluidine, formaldehyde, and methyl tert-butyl ether. The number of overlapping genes identified between pollutants and cancer types varied, with lung adenocarcinoma showing 58 genes for acetaldehyde, 74 for anthracene, 89 for ethylidene chloride, 70 for uranium carbide, and 66 for o-toluidine. For esophageal cancer, the respective counts were 33 for formaldehyde, 28 for o-toluidine, 36 for anthracene, 46 for ethylidene chloride, and 49 for methyl tert-butyl ether, while in gastric cancer, 37 genes were linked to formaldehyde, 32 to o-toluidine, 40 to anthracene, 49 to ethylidene chloride, and 43 to methyl tert-butyl ether. The Venn diagram and PPI network (Fig. 7), constructed using STRING and Cytoscape, revealed key molecular interactions, highlighting potential mechanisms of pollutant-induced carcinogenesis.

Identification of key hub genes in pollutant-associated carcinogenesis

To pinpoint key molecular drivers of pollutant-related tumorigenesis, hub genes were identified using CytoHubba with the Maximum Clique Centrality (MCC) algorithm. The top 10 hub genes were selected for each pollutant, providing insights into their potential roles in tumor progression, invasion, and drug resistance (Fig. 8).

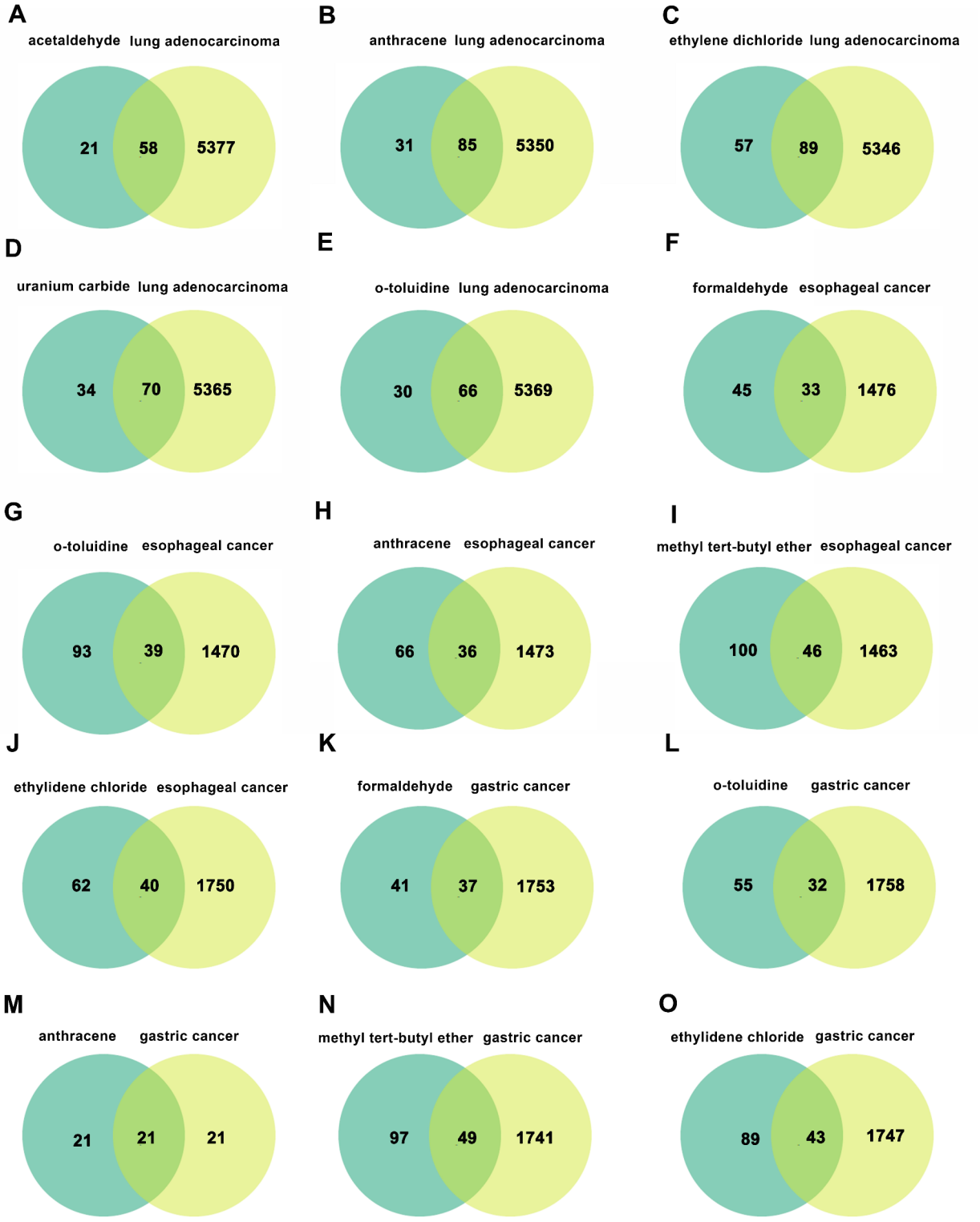

Fig. 6. Venn diagrams showing the cross-genes shared between pollutants and cancer. (**A**) Acetaldehyde and lung adenocarcinoma, (**B**) Anthracene and lung adenocarcinoma, (**C**) Ethylene dichloride and lung adenocarcinoma, (**D**) Uranium carbide and lung adenocarcinoma, (**E**) O-toluidine and lung adenocarcinoma, (**F**) Formaldehyde and esophageal cancer, (**G**) O-toluidine and esophageal cancer, (**H**) Anthracene and esophageal cancer, (**I**) Ethylidene and esophageal cancer, (**J**) Methyl tert-butyl ether and esophageal cancer, (**K**) Formaldehyde and gastric cancer, (**L**) O-toluidine and gastric cancer, (**M**) Anthracene and gastric cancer, (**N**) Ethylidene and gastric cancer, (**O**) Methyl tert-butyl ether and gastric cancer.

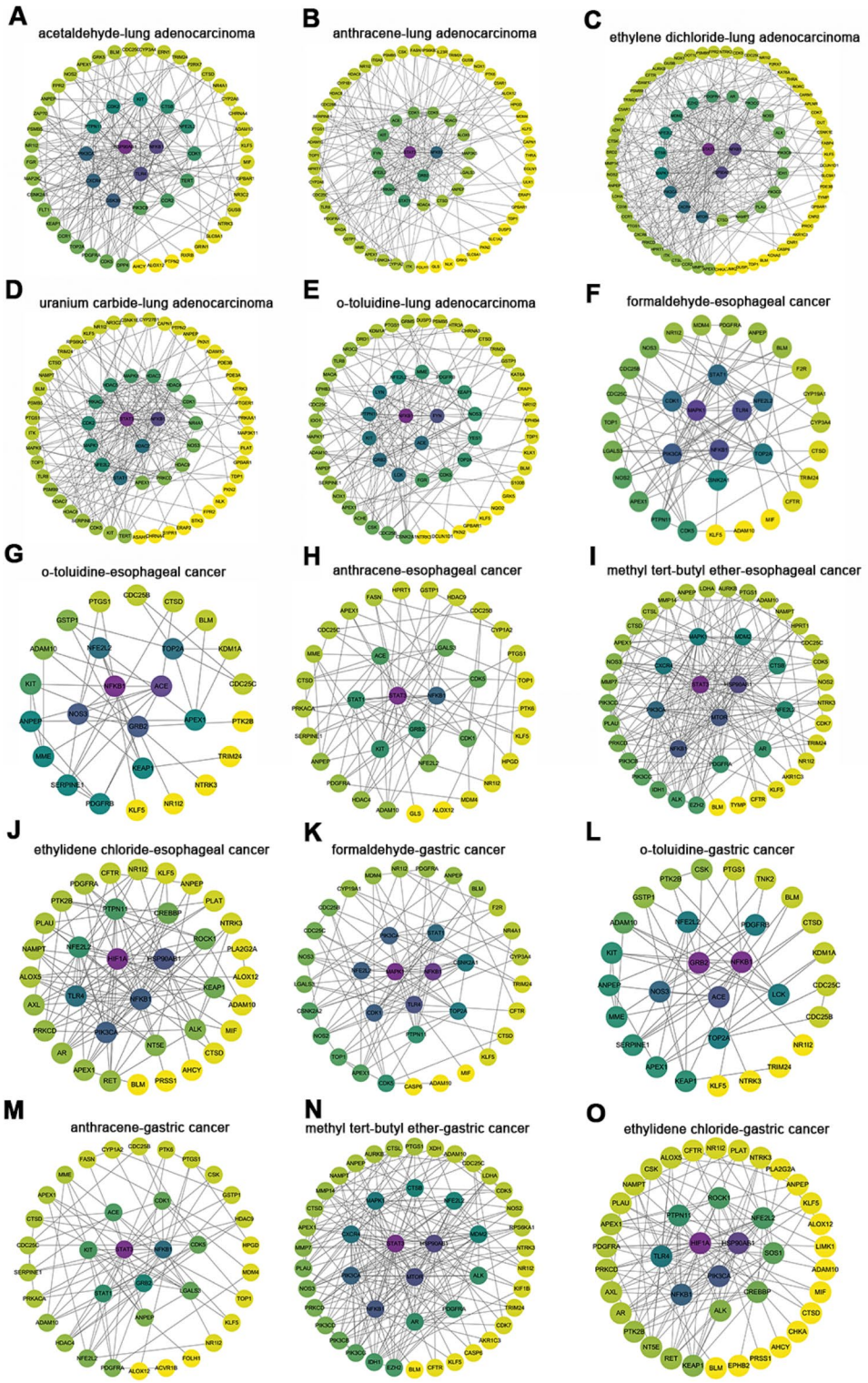

Fig. 7. PPI network analysis. (**A**) Acetaldehyde and lung adenocarcinoma, (**B**) Anthracene and lung adenocarcinoma, (**C**) Ethylene dichloride and lung adenocarcinoma, (**D**) Uranium carbide and lung adenocarcinoma, (**E**) O-toluidine and lung adenocarcinoma, (**F**) Formaldehyde and esophageal cancer, (**G**) O-toluidine and esophageal cancer, (**H**) Anthracene and esophageal cancer, (**I**) Ethylidene and esophageal cancer, (**J**) Methyl tert-butyl ether and esophageal cancer, (**K**) Formaldehyde and gastric cancer, (**L**) O-toluidine and gastric cancer, (**M**) Anthracene and gastric cancer, (**N**) Ethylidene and gastric cancer, (**O**) Methyl tert-butyl ether and gastric cancer.

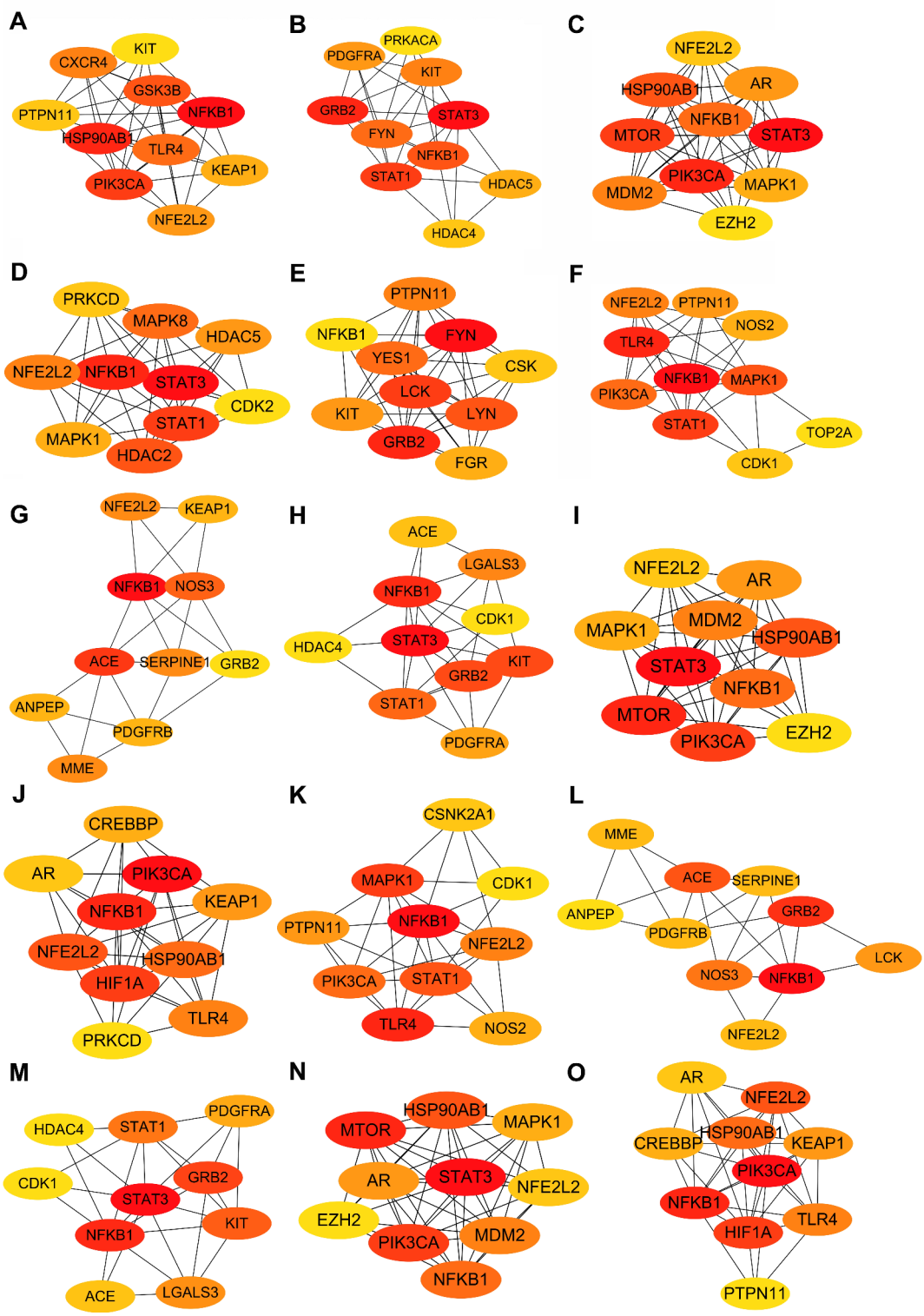

Fig. 8. Protein interactions related to the carcinogenic activity. (**A**) Core targets of acetaldehyde and lung adenocarcinoma, (**B**) Core targets of anthracene and lung adenocarcinoma, (**C**) Core targets of ethylidene chloride and lung adenocarcinoma, (**D**) Core targets of uranium carbide and lung adenocarcinoma, (**E**) Core targets of o-Toluidine and lung adenocarcinoma, (**F**) Core targets of formaldehyde and esophageal cancer, (**G**) Core targets of o-toluidine and esophageal cancer, (**H**) Core targets of anthracene and esophageal cancer, (**I**) Core targets of ethylidene and esophageal cancer, (**J**) Core targets of methyl tert-butyl ether and esophageal cancer, (**K**) Core targets of formaldehyde and gastric cancer, (**L**) Core targets of o-toluidine and gastric cancer, (**M**) Core targets of anthracene and gastric cancer, (**N**) Core targets of ethylidene and gastric cancer, (**O**) Core targets of methyl tert-butyl ether and gastric cancer. These interactions indicate links between central targets associated with lung adenocarcinoma caused by air pollutants and central targets associated with esophageal and gastric cancers caused by water pollutants.

GO and KEGG enrichment analysis

GO and KEGG enrichment analyses were conducted to explore the biological processes and pathways associated with pollutant-induced carcinogenesis, particularly in lung adenocarcinoma and air pollutants, as well as esophageal and gastric cancers with water pollutants. The results, visualized through bubble plots and histograms (Figures S1–S15), demonstrated strong associations between pollutant exposure and critical cellular functions.

GO analysis revealed that environmental pollutants influence key biological processes related to cancer progression. Among air pollutants, anthracene and uranium carbide were linked to the cellular response to peptides, while acetaldehyde, anthracene, and o-toluidine were involved in protein autophosphorylation. Additionally, acetaldehyde, uranium carbide, and o-toluidine were associated with peptidyl-serine phosphorylation, and acetaldehyde and ethylidene chloride were implicated in cell chemotaxis. Several pollutants, including anthracene and uranium carbide, were also involved in oxidative stress response, a critical factor in carcinogenesis. At the cellular component level, acetaldehyde and ethylidene chloride were enriched on the external side of the plasma membrane, while anthracene and o-toluidine were linked to membrane microdomains. Furthermore, anthracene, ethylidene chloride, and o-toluidine were enriched in the cytoplasmic vesicle lumen and secretory granule lumen, suggesting their involvement in intracellular signaling and transport. At the molecular function level, acetaldehyde, anthracene, ethylidene chloride, uranium carbide, and o-toluidine were associated with protein kinase activity, particularly protein serine/threonine kinase activity. Acetaldehyde and o-toluidine were also linked to p53 binding, highlighting their role in tumor suppression and genomic stability regulation. Additionally, anthracene and uranium carbide were enriched in histone deacetylase activity, suggesting their influence on epigenetic modifications, which are critical in cancer development and progression.

These pollutants also demonstrated significant enrichment in key cancer-related pathways. Air pollutants such as acetaldehyde, anthracene, ethylidene chloride, uranium carbide, and o-toluidine were linked to lipid and atherosclerosis-related pathways, including fluid shear stress and atherosclerosis, which are associated with chronic inflammation and tumor progression. They were also involved in pathways related to prostate cancer, the chemokine signaling pathway, and progesterone-mediated oocyte maturation, indicating endocrine-disrupting potential. Furthermore, acetaldehyde, anthracene, ethylidene chloride, and o-toluidine were strongly associated with PD-L1/PD-1 checkpoint regulation in cancer, suggesting that these pollutants may contribute to immune evasion in tumor cells.

For esophageal cancer, GO analysis of water pollutants indicated that formaldehyde and ethylidene chloride were involved in the cellular response to chemical stress and blood coagulation, while formaldehyde and anthracene were linked to the regulation of blood coagulation. At the cellular component level, formaldehyde and ethylidene chloride were enriched in the ficolin-1-rich granule and secretory granule lumen, while ethylidene chloride and methyl tert-butyl ether were associated with the cytoplasmic vesicle lumen and vesicle lumen. At the molecular function level, formaldehyde, o-toluidine, ethylidene chloride, and methyl tert-butyl ether were linked to p53 binding, supporting their role in DNA damage response and tumor suppression. Additionally, formaldehyde and ethylidene chloride were related to protein serine/threonine kinase activity, while o-toluidine and anthracene were associated with DNA-binding transcription factor activity and exopeptidase activity.

These pollutants were further enriched in key oncogenic pathways. KEGG pathway analysis (Figures S6-S10) revealed that formaldehyde, o-toluidine, anthracene, ethylidene chloride, and methyl tert-butyl ether were significantly associated with pathways related to prostate cancer and microRNAs in cancer, suggesting their involvement in gene regulation and tumor progression. They were also linked to the PD-L1/PD-1 checkpoint pathway, the HIF-1 signaling pathway, and the sphingolipid signaling pathway, implicating them in hypoxia response and lipid metabolism, both of which are crucial in tumor microenvironment regulation. O-toluidine, anthracene, ethylidene chloride, and methyl tert-butyl ether were further associated with central carbon metabolism in cancer, indicating their potential role in metabolic reprogramming, a hallmark of cancer progression.

For gastric cancer, GO analysis revealed that formaldehyde and anthracene were involved in the cellular response to peptides, while o-toluidine and ethylidene chloride were associated with oxidative stress response. Additionally, anthracene and ethylidene chloride were implicated in the response to reactive oxygen species, suggesting their role in oxidative DNA damage and tumorigenesis. Ethylidene chloride and methyl tert-butyl ether were associated with blood coagulation and regulation of body fluid levels, processes that may contribute to cancer metastasis. At the cellular component level, formaldehyde, ethylidene chloride, and methyl tert-butyl ether were linked to the ficolin-1-rich granule, while ethylidene chloride and methyl tert-butyl ether were also enriched in the cytoplasmic vesicle lumen and vacuolar lumen. At the molecular function level, formaldehyde, o-toluidine, ethylidene chloride, and methyl tert-butyl ether were linked to protein serine kinase activity and p53 binding, further supporting their role in tumorigenesis. O-toluidine and methyl tert-butyl ether were also associated with DNA-binding transcription factor binding, RNA polymerase II-specific DNA-binding transcription factor binding, and transmembrane receptor protein kinase activity, indicating their potential regulatory effects on gene expression and signal transduction.

KEGG pathway analysis (Figures S11-S15) showed that formaldehyde, o-toluidine, anthracene, ethylidene chloride, and methyl tert-butyl ether were significantly associated with microRNAs in cancer, prostate cancer, and the HIF-1 signaling pathway. Additionally, these pollutants were linked to PD-L1/PD-1 checkpoint regulation, further suggesting their role in modulating immune responses in the tumor microenvironment. These findings highlight the intricate interactions between environmental pollutants and key oncogenic pathways, underscoring their potential role in cancer initiation and progression.

Survival analysis and clinical relevance of pollutant-associated genes

To assess the clinical significance of pollutant-associated genes, survival analysis was performed using Hazard Ratio (HR) values and Kaplan-Meier (K-M) survival curves. Genes with $p < 0.05$ were considered significantly

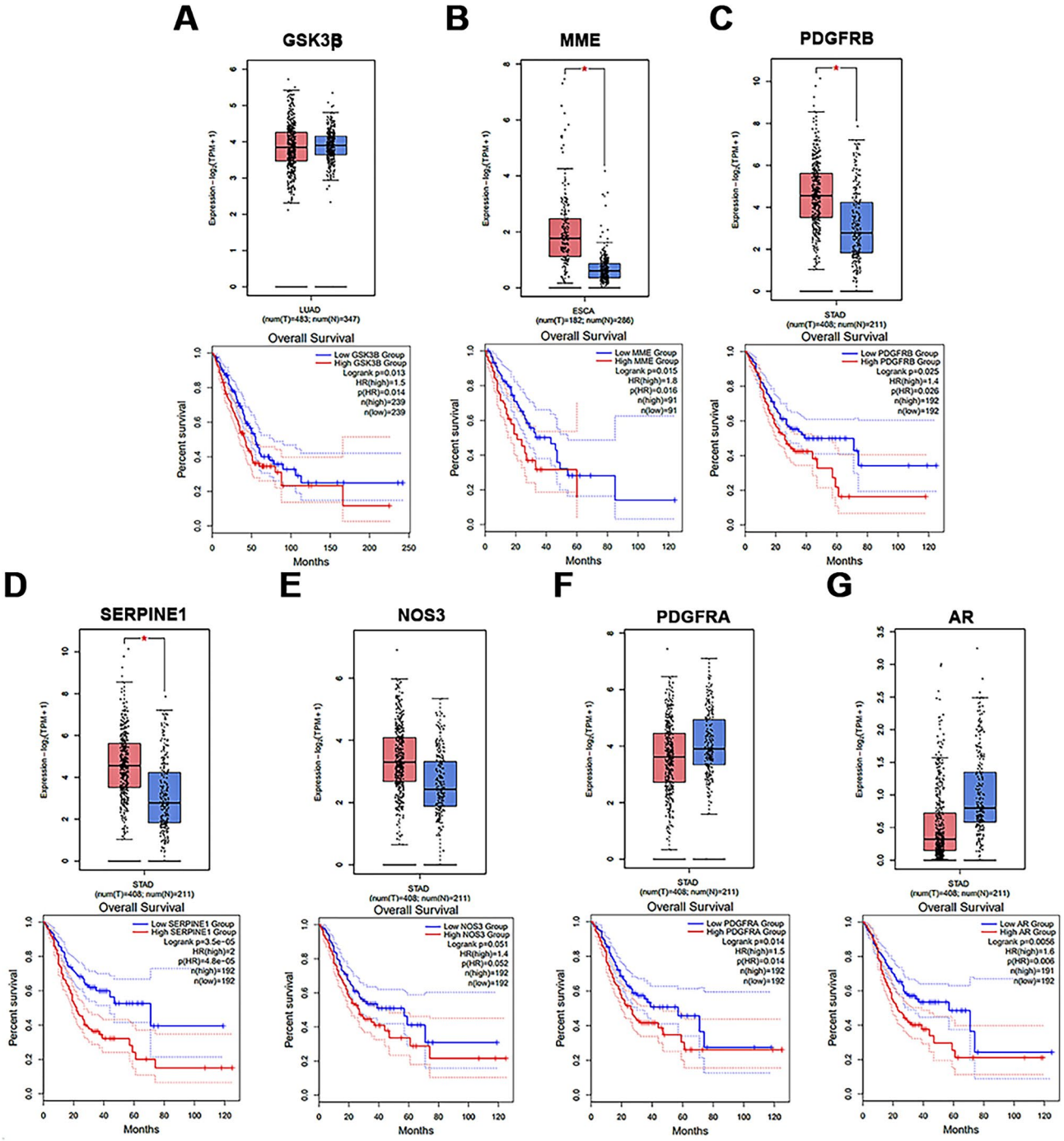

Fig. 9. Pollutants Represent Box Plots and Survival Analysis of Genes. Box plots and survival analysis of genes in molecules strongly associated with poor cancer prognosis: GSK3β, MME, PDGFRB, SERPINE1, NOS3, PDGFRA, and AR.

associated with poor prognosis. Kaplan-Meier (K-M) survival curves were constructed to compare high- and low-expression groups.

As shown in Figures S16 and Fig. 9A, high expression of GSK3β (linked to acetaldehyde) was significantly associated with poor prognosis in lung adenocarcinoma. Similarly, MME (associated with o-toluidine) correlated with poor survival outcomes in esophageal cancer (Fig. 9B). In gastric cancer, PDGFRB and SERPINE1 were identified as potential prognostic markers (Fig. 9C and D). Further analysis revealed strong associations between PDGFRB, SERPINE1, and PDGFRA (o-toluidine), as well as AR (ethylidene chloride and methyl tert-butyl ether) with poor prognosis in gastric cancer (Fig. 9E and G).

Conclusions

This study presents a comprehensive computational framework for assessing the carcinogenic potential of air and water pollutants by integrating pollutant datasets, molecular fingerprints, machine learning, functional enrichment analyses, and survival analysis. The pre-trained KPGT model, trained using molecular fingerprints and descriptors, demonstrated high predictive performance, achieving an AUC of 0.83, outperforming traditional machine learning models.

Our findings underscore the complex biological interactions of pollutant molecules, implicating key oncogenic genes such as MAPK1, MTOR, and PTPN11 in pollutant-associated carcinogenesis. Functional enrichment analyses further identified critical pathways, including oxidative stress response, immune checkpoint regulation (PD-L1/PD-1), and metabolic reprogramming, revealing potential mechanisms through which pollutants contribute to cancer development and progression. Survival analysis highlighted the prognostic significance of pollutant-associated genes, emphasizing their potential as biomarkers for cancer risk assessment.

By offering a systematic and high-throughput approach to evaluating environmental carcinogenicity, this study provides a valuable tool for pollution risk assessment and precision environmental health research. These insights contribute to a deeper understanding of pollutant-driven oncogenesis, supporting the development of evidence-based policies aimed at mitigating pollutant-related health risks.

Data availability

These datasets are from CCRIS database (https://pubchem.ncbi.nlm.nih.gov/bioassay/1259411) , EPA CompTox Chemistry Dashboard (https://echo.epa.gov/trends/loading-tool) , T3DB (http://www.t3db.ca/) , TRI database (https://www.epa.gov/toxics-release-inventory-tri-program) . The datasets and code used in the present study are available from the corresponding authors on reasonable request.

Received: 31 October 2024; Accepted: 26 March 2025
Published online: 06 April 2025

References

1. Mou, Y. et al. Environmental pollutants induce NLRP3 inflammasome activation and pyroptosis: roles and mechanisms in various diseases. *Sci. Total Environ.* 900. (2023).
2. Parvez, S. M. et al. Health consequences of exposure to e-waste: an updated systematic review. *Lancet Planet. Heath.* **5** (12), E905–E920 (2021).
3. Fuller, R. et al. Pollution and health: a progress update (6, Pg e535, 2022). *Lancet Planet. Heath.* **6** (7), E553–E553 (2022).
4. De Bont, J. et al. Ambient air pollution and cardiovascular diseases: an umbrella review of systematic reviews and meta-analyses. *J. Intern. Med.* **291** (6), 779–800 (2022).
5. Kelly, F. J. & Fussell, J. C. Air pollution and public health: emerging hazards and improved Understanding of risk. *Environ. Geochem. Health.* **37** (4), 631–649 (2015).
6. Antwi, S. O. et al. Exposure to environmental chemicals and heavy metals, and risk of pancreatic cancer. *Cancer Causes Control.* **26** (11), 1583–1591 (2015).
7. Sassano, M., Seyyedsalehi, M. S. & Boffetta, P. Occupational benzene exposure and colorectal cancer: A systematic review and meta-analysis. *Environ. Res.* **257**, 119213 (2024).
8. Metintas, M., Ak, G. & Metintas, S. Environmental asbestos exposure and lung cancer. *Lung Cancer*. **194**, 107850 (2024).
9. Li, C. et al. Global burden and trends of lung cancer incidence and mortality. *Chin. Med. J. (Engl).* **136** (13), 1583–1590 (2023).
10. Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca-a Cancer J. Clin.* **71** (3), 209–249 (2021).
11. Cheng, T. Y. D. et al. The international epidemiology of lung cancer: latest trends, disparities, and tumor characteristics. *J. Thorac. Oncol.* **11** (10), 1653–1671 (2016).
12. Testa, U., Castelli, G., Pelosi, E. & Cancers Lung Cancers: Molecular Characterization, Clonal Heterogeneity and Evolution, and Cancer Stem Cells. 10, (8). (2018).
13. Xue, Y. et al. Air pollution: A culprit of lung cancer. *J. Hazard. Mater.* **434**, 128937 (2022).
14. Eckel, S. P., Cockburn, M. & Shu, Y. H. Air pollution affects lung cancer survival. *Thorax* **71** (10), 891–898 (2016).
15. Wang, Y. et al. A novel concern from two sample Mendelian randomization study: the effects of air pollution exposure on the cardiovascular, respiratory, and nervous system. *Ecotoxicol. Environ. Saf.* **284**, 116871 (2024).
16. Fiordelisi, A. et al. The mechanisms of air pollution and particulate matter in cardiovascular diseases. *Heart Fail. Rev.* **22** (3), 337–347 (2017).
17. Costa, L. G. et al. Effects of air pollution on the nervous system and its possible role in neurodevelopmental and neurodegenerative disorders. *Pharmacol. Ther.* **210**, 107523 (2020).
18. Hill, W. et al. Lung adenocarcinoma promotion by air pollutants. *Nature* **616** (7955), 159– (2023).
19. World Health Organization. *WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide* (World Health Organization, 2021).
20. Jiang, X. et al. Comprehensive analysis of the association between human diseases and water pollutants. *Int. J. Environ. Res. Public. Health.* **19**, 24 (2022).
21. Moody, S. et al. Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. *Nat. Genet.* **53** (11), 1553–1563 (2021).
22. Fuller, R., Landrigan, P. J. & Balakrishnan, K. Pollution and health: a progress update (6, Pg e535, 2022). *Lancet Planet. Heath.* **6** (7), E553–E553 (2022).
23. I. Davidson, C., F. Phalen, R. & A. Solomon, P. Airborne particulate matter and human health:a review. *Aerosol Sci. Technology: J. Am. Association Aerosol Res.* **8**, 39 (2005).
24. Sherif, A., Benhammuda, M., Fares, S. & Oroszi, T. L. Environmental Factors and Cardiovascular Diseases. (2017).
25. Kim, K. H., Kabir, E. & Kabir, S. A review on the human health impact of airborne particulate matter. *Environ. Int.* **74**, 136–143 (2015).
26. Xu, X. et al. Environmental pollution and kidney diseases. *Nat. Rev. Nephrol.* **14** (5), 313–324 (2018).
27. Rauert, C. et al. Extraction and Pyrolysis-GC-MS analysis of polyethylene in samples with medium to high lipid content. Journal of Environmental Exposure Assessment 1, (2). (2022).
28. Velimirovic, M. et al. Mass spectrometry as a powerful analytical tool for the characterization of indoor airborne microplastics and nanoplastics. *J. Anal. Spectrom.* **36** (4), 695–705 (2021).

29. Huang, S. Efficient analysis of toxicity and mechanisms of environmental pollutants with network Pharmacology and molecular Docking strategy: acetyl tributyl citrate as an example. *Sci. Total Environ.* **905**, 167904 (2023).
30. Zhang, Y. et al. A network pharmacology-based strategy Deciphers the underlying molecular mechanisms of Qixuehe capsule in the treatment of menstrual disorders. *Chin. Med.* **12**, 23 (2017).
31. Li, H. et al. A knowledge-guided pre-training framework for improving molecular representation learning. *Nat. Commun.* **14** (1), 7568 (2023).
32. Cameron, T. P., Stump, J. M. & Schofield, L. Chemical Carcinogenesis Research Information System (CCRIS) data bank, 1981June 1986 (1988 version).
33. Swain, M. PubChemPy documentation. Release. (2014).
34. Weininger, D. SMILES, a chemical Language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28** (1), 31–36 (1988).
35. O'Boyle, N. M. Towards a universal SMILES representation-A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.* **4**, 1–14 (2012).
36. Listed, N. A. Proceedings of the 14th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences (Part 2). May 24–28, 2010. Montreal, Canada. Sar & Qsar in Environmental Research 22, (1–2). (2011).
37. Wishart, D. et al. T3DB: the toxic exposome database. *Nucleic Acids Res.* **43** (D1), D928–D934 (2015).
38. Landgrebe, M. et al. The tinnitus research initiative (TRI) database: A new approach for delineation of tinnitus subtypes and generation of predictors for treatment outcome. *BMC Med. Inf. Decis. Mak.* 10. (2010).
39. Lovric, M., Molero, J. M. & Kern, R. PySpark and RDKit: Moving towards Big Data in Cheminformatics. Mol. Inf. 38, (6). (2019).
40. Sievert, C. *Interactive web -based data visualization with R, plotly, and shiny* (Chapman 1178 and Hall/CRC, 2020).
41. Hu, C. et al. Molecular insights into chronic atrophic gastritis treatment: Coptis chinensis Franch studied via network pharmacology, molecular dynamics simulation and experimental analysis. *Comput. Biol. Med.* **178**, 108804 (2024).
42. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for 1176 fingerprint -based similarity calculations. *J. Cheminform.* **7**, 1–13 (2015).
43. Yun, S. Graph transformer networks. *Adv. Neural. Inf. Process. Syst.*, 32. (2019).
44. Goldman, M. J. et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38** (6), 675–678 (2020).
45. Piñero, J. et al. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48** (D1), D845–D855 (2020).
46. Stelzer, G. et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Curr. Protocols Bioinf.* **54**, 1301–13033 (2016).
47. Whirl-Carrillo, M. et al. An Evidence-Based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **110** (3), 563–572 (2021).
48. Joanna, A., Bocchini, C. A., Scott, A. F. & Ada, H. McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Res.* 37, (Database), D793-D796. (2009).
49. Nickel, J. et al. SuperPred: update on drug classification and target prediction. *Nucleic Acids Res.* **42** (W1), W26–W31 (2014).
50. Keiser, M. J. et al. Relating protein Pharmacology by ligand chemistry. *Nat. Biotechnol.* **25** (2), 197–206 (2007).
51. Daina, A., Michielin, O. & Zoete, V. Swiss target prediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Res.* **47** (W1), W357–W364 (2019).
52. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47** (D1), D607–D613 (2019).
53. Shannon, P. et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13** (11), 2498–2504 (2003).
54. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25** (1), 25–29 (2000).
55. Kanehisa, M. & Goto, S. Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28** (1), 27–30 (2000).
56. Kanehisa, M. et al. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51** (D1), D587–D592 (2023).
57. Kanehisa, M. Toward Understanding the origin and evolution of cellular organisms. *Protein Sci.* **28** (11), 1947–1951 (2019).
58. Tang, Z. F. et al. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* **47** (W1), W556–W560 (2019).

Acknowledgements

This work was supported by grants from National Key R&D Program of China (2021YFA1500403), Jilin Provincial Scientific and Technological Development Program (20220101276JC, 20210402018GH).

Author contributions

Conceptualization, Q.F. and Y.J.; Methodology, F.C. Formal analysis and Investigation, F.C., X.Z. ; Writing-original draft preparation, F.C. and X.Z.; Writing-review and editing, F.C., X.Z. ; Funding acquisition, Q.F. and Y.J.; Resources, Q.F. and Y.J.; Supervision, Q.F. and Y.J. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-96193-2.

Correspondence and requests for materials should be addressed to Y.J.